IMPROVING PATIENT FLOW:
EXAMINING THE APPLICATION OF LEAN MANAGEMENT AND THEORY OF CONSTRAINTS ACROSS DIFFERENT HEALTHCARE SETTINGS

TALAL ALSHARIEF

A thesis submitted in partial fulfilment of the
requirements of Nottingham Trent University
for the degree of Doctor of Philosophy

September 2020

Copyright Statement

ACKNOWLEDGEMENTS

This study would not have been possible without the help and encouragement I obtained from a number of valuable people. I would like to thank everyone who provided me with direct or indirect support. First of all, I would like to express my sincere gratitude to my supervisor Dr Roy Stratton for his assistance, support, and patience in guiding me through this project. His vision and vast experience in Operation and Supply Chain Management has inspired me to carry out this research. Without his continued support and tremendous knowledge, this research would not be possible, and I am very grateful for all he has taught me.

I would also like to express my thanks to Dr Donald Harradine, my second supervisor, for his helpful feedback, constructive advice, and suggestions. His words have always been a source of encouragement to me when I felt frustrated. In addition, I would like to thank all of the NHS staff and other experts who participated in this study for their cooperation and valuable information.

I am deeply grateful to all my family members, especially my brothers and sisters, for their love and support. A special thanks to my mother for always believing in me, encouraging me, and protecting me with her prayers in the most difficult times.

Most of all, I would like to thank my wife Mai, who has stood by me with every step. She was patient and tolerant; she shared the most difficult moments with me and gave endless encouragement and support throughout my academic journey. I cannot thank her enough for having faith in me and being such a supportive wife. My love and gratitude to my son Azzam; I was always busy with my research and may not have been the ideal father during the most recent period, but I hope he will be proud of me and my achievements.

Finally, my heartfelt thanks go to the Ministry of Education in Saudi Arabia for giving me the opportunity to study in the United Kingdom and for offering me financial support. Also, I would like to thank Nottingham Trent University for providing all the support I needed during my studies.

ABSTRACT

The importance of patient flow as a construct is clearly evident today in healthcare institutions and conferences, but, as with manufacturing, there needs to be a mechanism tailored to the instability of needs in this environment. This is reflected in the development of different manufacturing mechanisms, including Ford's physical flow lines, Ohno's Kanban control, and Goldratt's time buffer management (TBM), but little is known about the factors underpinning these distinctions and how they translate to the underlying healthcare environments. This research aimed to gain a deeper academic understanding of how established flow management mechanisms have been developed to meet the distinct conditions within the wider healthcare system, with specific reference to the origins of lean management (LM) and the theory of constraints (TOC).

To accomplish this, a multi-case study approach was conducted, based on four healthcare organisations across three NHS Trusts in the UK. The cases incorporated acute and rehabilitative hospital care, social care, out-patient services, and GP-led community care and community mental health.

Patient flow was improved and managed effectively in those cases where the flow mechanism was applied proactively to ensure that patients progress through each stage of the delivery system to minimise the length of stay (LOS) and target causes of flow disruption. This was evident in the TBM solutions witnessed, including the need to subordinate the flow focus to the planned discharge date (PDD), thus synchronising all activities around this patient, whether in out-patient or in-patient environments.

Healthcare is characterised by a range of variability and uncertainty, and, where there is an opportunity to define the process flow tightly, LM can be readily applied. This includes patient flow, but these findings suggested it is more suited to elective surgery or emergency pathways that are predefined. Where the pathway is unknown, poorly defined, and involves transfer of care, flow mechanism is better suited to TBM. This research contributes to new knowledge by providing a deeper theoretical understanding of different manufacturing flow mechanisms and how these mechanisms can be used selectively in various healthcare environments.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **4**

ABSTRACT **5**

TABLE OF CONTENTS **6**

LIST OF FIGURES **10**

LIST OF TABLES **11**

LIST OF ABBREVIATIONS **12**

CHAPTER ONE: INTRODUCTION **14**

1.1. Research Background 14

1.2. Research Aims and Research Questions 18

1.3. Thesis Outline 18

CHAPTER TWO: LITERATURE REVIEW **20**

2.1. Exploring the Operational Challenges Facing the Health and Social Care System 22

2.1.1. Overview of the UK Health and Social Care System's Structure 22

2.1.2. Challenges Facing Healthcare 28

2.1.3. The Complexity of the Health and Social Care System 31

2.1.4. The Importance of Managing Patient Flow 36

2.2. Exploring the Environmental Conditions Associated with the Origins of Flow Mechanisms and their Adoption in Different Manufacturing Environments 42

2.2.1. Aspects of the Instability in Production Environments 42

2.2.2. The Contribution to Flow Improvement 46

2.2.3. Flow Mechanisms in Different Manufacturing Environments 50

2.2.4. TOC's DBR-TBM VS. LM-Kanban 72

2.2.5. Comparing the Functions/Rules of Kanban and TBM 78

2.2.6. Conclusion 79

2.3. The Influence of LM and TOC Approaches on Flow Management in Healthcare 81

2.3.1. LM and TOC Approaches in Healthcare 81

2.3.2. LM in Healthcare 83

2.3.3. Translation of LM into Healthcare 85

2.3.4. TOC in Healthcare 100

2.3.5. Translation of TOC into Healthcare 101

2.3.6. Current Approaches Applied in the NHS 111

2.3.7. Conclusion 111
2.4. Summary of the Chapter and Justification of Research 113
2.4.1. Summary of the Chapter 113
2.4.2. Focus and Justification of Research 115

CHAPTER THREE: RESEARCH METHODOLOGY 118
3.1. Research Philosophy 118
3.2. Justification for Following the Interpretive Approach 120
3.3. Matching the Purpose of Research with Methods 122
3.4. Rationale for Using a Case Study Approach 125
3.5. The Use of Case Study Research 127
3.5.1. Advantages of Case Study 131
3.5.2. Case Study Limitations 134
3.6. Best Practice for Conducting Case Study Research 136
3.7. Eisenhardt's Process Theory Development from Case Study Research 141
3.7.1. Step 1. Getting Started 141
3.7.2. Step 2. Selecting Cases 142
3.7.3. Step 3. Crafting Instruments 143
3.7.4. Step 4. Entering the Field 144
3.7.5. Step 5. Analysing Data 144
3.7.6. Step 6. Shaping Hypotheses 146
3.7.7. Step 7. Enfolding Literature 146
3.7.8. Step 8. Reaching Closure 147
3.8. Applying Eisenhardt's 8-Step Framework 148
3.9. Validity 166
3.10. Reliability 167

CHAPTER FOUR: INDIVIDUAL CASE ANALYSIS 169
4.1. Case 1 170
4.1.1. Case Background 170
4.1.2. Before the Intervention 170
4.1.3. Nature of the Intervention 172
4.1.4. The Intervention 172
4.1.5. Results 176
4.1.6. Sustained Performance 177
4.2. Case 2 179
4.2.1. Case Background 179
4.2.2. Before the Intervention 179
4.2.3. Nature of the Intervention 179

4.2.4. The Intervention 180
4.2.5. Results 182
4.2.6. Sustained Performance 183

4.3. Case 3 184
4.3.1. Case Background 184
4.3.2. Before the Intervention 184
4.3.3. Nature of the Intervention 185
4.3.4. The Intervention 185
4.3.5. Results 185
4.3.6. Sustained Performance 186

4.4. Case 4 188
4.4.1. Case Background 188
4.4.2. Before the Intervention 188
4.4.3. Nature of the Intervention 189
4.4.4. The Intervention 189
4.4.5. Results 192
4.4.6. Sustained Performance 193

4.5. Summary 194

CHAPTER FIVE: CROSS-CASE ANALYSIS AND DISCUSSION 196

5.1. Exploring the Environmental Conditions Associated with Different Flow Mechanisms and the Translation of these Mechanisms to Healthcare Environments 198
5.1.1. The Value of Flow 198
5.1.2. Mechanism for Managing Flow 203
5.1.3. Focusing on Ongoing Improvement 212
5.1.4. Local Optimisation Measures Inhibit Flow 218

5.2. Exploring the Assumptions Underpinning the Effective Use of Kanban and TBM in Healthcare Settings 225
5.2.1. Sensitivity of Medical Flow Path 225
5.2.2. The Choice of Buffering 226
5.2.3. Instabilities in Demand 227
5.2.4. Detecting Variability and Encouraging Continuous Improvement 227

5.3. Exploring the Relationship between the Current Development of Healthcare Flow Management and the Manufacturing Flow Mechanisms 230

CHAPTER SIX: CONCLUSION 233

6.1. Reviewing the Research Aim and Objectives 233

6.2. Response to Research Questions 235

6.2. Research Contribution 241

6.3. Limitations of Research 243

6.4. Suggestions for Future Research 244

REFERENCES 246

APPENDIX 1...276

APPENDIX 2...278

APPENDIX 3...281

APPENDIX 4...282

LIST OF FIGURES

Figure 1: Outline of the Literature Review Chapter 21
Figure 2: Health and Social Care Structures in England (Source: RCP London, 2018, p.3) 23
Figure 3: Health and Social Care System (Source: Stratton and Knight, 2010, p. 485 (Modified)) 34
Figure 4: Product Process Matrix (Source: Hayes and Wheelwright, 1979, p. 3) 42
Figure 5: The Toyota Production System (Source: Liker, 2004, p. 48) 55
Figure 6: Comparison of MRP and Kanban (Hopp and Spearman, 2001, p. 163) 56
Figure 7: The Rules of OPT (Source: Goldratt and Fox, 1986, p. 179) 60
Figure 8: Traditional DBR (Source: Stratton, 2015, p.32) 62
Figure 9: Simplified DBR (Source: Stratton, 2015, p. 34) 63
Figure 10: TBM Functions (Source: Stratton, 2012, pp. 12-15) 64
Figure 11: The Development of Lean Management (Source: Laursen et al., 2003 cited in Lin and Chiu, 2017, p. 531) 83
Figure 12: Buffer Zones for A&E Patients (Source: Umble and Umble, 2006, p. 1068) 104
Figure 13: Buffer Zones for Acute Hospital Admission Delays (Source: Umble and Umble, 2006, p. 1069) 104
Figure 14: Time Buffer Management: Conceptual View Typically Used in Healthcare (Source: Knight, 2011, p.9) 105
Figure 15: Discharge QFI Jonah Top Delay Reasons by Region Presented to Cross-buffer Meetings (Source: Stratton and Knight, 2010, p. 10) 108
Figure 16: Process before the Intervention 171
Figure 17: Process after the Intervention 174
Figure 18: The Top Reasons for Delay 181
Figure 19: Chapter 5 Outline 197

LIST OF TABLES

Table 1: Kanban Functions and Rules (Source: Ohno, 1988, p. 30) 57
Table 2: Differences and Assumptions Associated with Flow Mechanisms 72
Table 3: Assumptions Underlying Kanban and TBM (Source: Stratton et al., 2014, p. 9) 73
Table 4: Functional Comparison of TBM and Kanban (Source: Knight and Stratton, 2010, p. 27) 78
Table 5: The Translation of LM into Healthcare 99
Table 6: The Translation of TOC into Healthcare 110
Table 7: Comparison of Positivism and Interpretivism (Source: Saunders et al., 2012, p. 136) 120
Table 8: Matching Research Purpose with Methodology (Source: Voss et al., 2002, p. 198) 125
Table 9: Relevant Conditions of Various Research Strategies (Source: Yin, 2003, p. 6) 129
Table 10. Design Tests and Tactics in Case Study Research (Source: Yin, 1994, p. 33) 133
Table 11: Portfolio of Case Study Research Design: differences in underlying elements (Source: Ridder, 2017, p. 12) 139
Table 12: Process of Building Theory from Case Study Research (Source: Eisenhardt, 1989, p. 533) 140
Table 13: List of Case Studies 152
Table 14: Data Collection Tools 153
Table 15: Steps Taken to Ensure the Validity and Reliability of this Research 168
Table 16 Summary of Cases 195
Table 17: Assumptions Underlying Kanban and TBM in Healthcare Settings 229

LIST OF ABBREVIATIONS

A&E	Accident and Emergency
CCR	Capacity Constrained Resource
COVID	Coronavirus Disease
DBR	Drum Buffer Rope
DTOC	Delayed Transfers of Care
ED	Emergency Department
5FS	Five Focusing Steps
GP	General Practice
JIT	Just in Time
LM	Lean Management
LOS	Length of Stay
MDT	Multi-Disciplinary-Team
MTO	Make-to-Order
NHS	National Health Service
OM	Operations Management
OPT	Optimized Production Technology
PDD	Planned Discharge Date
PPM	Product Process Matrix
SDBR	Simplified Drum Buffer Rope
SPC	Statistical Process Control

TBM	Time Buffer Management
TOC	Theory of Constraints
TPS	Toyota Production System
UK	United Kingdom
VMMC	Virginia Mason Medical Centre
VMPS	Virginia Mason Production System
VSM	Value stream mapping
WIP	Work in Progress

CHAPTER ONE: INTRODUCTION

1.1. Research Background

Worldwide, healthcare service providers have been exposed to growing pressures to simultaneously minimise operational costs and improve access to, and quality of, the care they deliver. Changing demographics and the increasing demands facing healthcare delivery systems have created more complex pathways to access healthcare services and necessitated the integration of social care into them, in order to enable the safe discharge of patients (Alderwick et al., 2015). The present demand in the United Kingdom (UK) for health and social care has grown more rapidly than the healthcare service providers' capacity to provide the required care in a timely way (Campbell et al., 2017).

Considering the current budgetary and capacity challenges faced by the National Health Service (NHS) and the fact that over 40 per cent of NHS England's budget is spent on older people, the importance of patient flow between acute services, primary care, and social care in the UK healthcare delivery system can never be overlooked (Karakusevic, 2016; Robineau, 2016). The lack of control of patient flow has been associated with negative impacts on the performance of health and social care systems, with problems in poor patient flow including delayed discharges and long queues, reduced productivity, variability in workload, waste of resources, bottlenecks, and high length of patient stay (Noon et al., 2003; Haredn and Resar, 2004; Walley et al., 2006).

In more recent literature, the delayed discharge has been associated with increased costs of providing healthcare (Rojas- Garcia, 2018), poor patient experiences (Everall et al., 2019), increased exposure to diseases (Barnable et al., 2015; McCloskey et al., 2014; Walker et al., 2011) and increased mortality rates (Rosman et al., 2015). Walker (2011) found delayed discharge of patients can increase delays in providing medically necessary care for other patients, cancellation of scheduled appointments and backlogs in emergency services. Baumann et al. (2007) also found that an increase in LOS correlates with a likely a greater risk of infections and higher costs for treating the infections. Younis et al. (2012) and Chidwick et al. (2017) have associated delayed discharge with bed blockage, which eventually causes delays in transfer within the hospital.

Delayed transfer of patients is indicative of this deeper failure of the system to provide the right care at the right place and at the right time (Humphries, 2017). This failure was further intensified by the emergence of the Coronavirus disease (COVID-19) pandemic. While Søreide et al. (2020) view these periodic challenges as nothing new to wider healthcare systems, they acknowledge the unparalleled impact that the pandemic has brought to patients and service providers. This indicates operational challenges that reflect the primary need to synchronise capacity and demand in enabling patient flow (White et al., 2011).

Significant applications that have a notable impact on the productivity of healthcare delivery systems are best coordinated at a system-wide level. Kämäräinen et al. (2016) found that attempts to maximise productivity at the unit level within organisations often result in problems for other levels or units in the organisation. They further reported that it is often easier to adopt cost minimisation strategies than to improve outcomes, leading to long waiting times for patients, delayed discharges, and an inefficient allocation of resources. In the same vein, Fillingham et al. (2016) believe that the different elements of the healthcare system are directed, supported, controlled, and structured in silos that contradict the imperative of flow. These issues together require healthcare administrators to look at the whole picture of the healthcare delivery system and to concentrate on patient flow, in order to improve safety and service quality, satisfy patients, and reduce costs.

Therefore, it is not surprising that attention on improvement has been directed towards patient flow with a significant interest in potential parallels with the manufacturing sector. According to Olsson and Aronsson (2015), many healthcare services across the globe have implemented process improvement approaches, such as lean management (LM) and the theory of constraints (TOC), which focus on flow as a proxy for system productivity. These approaches are concerned with how continuous improvement can be achieved in production systems (Antunes, 1998) and have been developed to reduce and manage wasteful variation in the delivery system to meet the needs of a variety of manufacturing environments (Stratton et al., 2014). Both approaches have techniques that can address the issue of synchronisation. TOC includes a drum buffer rope (DBR) flow control mechanism, while LM follows the Kanban mechanism (Antunes, 1998; Bhasin and Burcher, 2006).

Where LM and TOC have been applied, there have been mixed results in terms of the success and assumptions underlying the application of such approaches to healthcare (Blackmore and Kaplan, 2016; Radnor et al., 2012; Stratton and Knight, 2010). In the context of the growing use of LM applications in healthcare, the changes implemented have been criticised for being disjointed and not sustained (Radnor et al., 2012; Hallam and Contreras, 2018), for enhancing local rather than wider improvement (Young and McClean, 2008; Brandao de Souza, 2009; Parkhi, 2019), and for having a limited impact on patient flow (Bhasin, 2008; de Vries and Huijsman, 2011; Poksinska et al., 2016). Although there are notable cases of the successful implementation of LM in healthcare, including The Pittsburgh Way (Grunden, 2007), The Royal Bolton Hospital (Fillingham, 2007) and The Virginia Mason Medical Centre (VMMC) (Kenney, 2011), much of this has involved secondary processes involving consumables or pharmacy and test procedures, rather than patient flow, and, although there have been significant improvements in patient flows, these are typically associated with the redesign of medical pathways (Fillingham, 2007; Murphy et al., 2019; Van Calster et al., 2019).

Aside from these exceptional case studies, there is also a lack of adept discussion in the literature about the translation of LM into healthcare (Brandao de Souza, p. 123; Andersen and Røvik, 2015; Rees and Gauld, 2017; Parkhi, 2019), and to what extent healthcare providers can benefit from LM practices (de Vries and Huijsman, 2011; Flynn et al., 2019), especially with regard to the management of patient flow. For example, Brandao de Souza (2009) argues that there has been no comprehensive review of literature on Lean in Healthcare since the management approach has been translated into healthcare. Andersen et al., (2014) found that existing literature on Lean implementation in healthcare is yet to fully address contextual factors and mechanisms that influence the sustainability of Lean efforts. Anderson and Rovik (2015) concluded that Lean thinking is translated into several types as variations occur from management who introduce the concept, consultants who share the information, and employees who experience the actual changes in systems and processes.

Rees and Gauld (2017) also argued that although implementation variations and deployment challenges have made existing evidence about Lean application less conclusive, further research is still needed to extensively and fully address the subject of the Lean application in healthcare. Flynn et al. (2019) concluded that while sustainability of Lean application in healthcare is an important aspect, it remains an understudied area of implementation research. The limitations in translation contribute to insufficient evidence and have an impact on the success or failure of Lean Management approaches.

In contrast, prior researchers have identified a significant improvement in patient flow following the application of TOC (Knight, 2003; Umble and Umble, 2006; Stratton and Knight, 2010; Knight, 2011; Stratton et al., 2014). However, there is limited research evidence concerning the sustainability of these approaches in the long term (Stratton and Knight, 2010; Stratton et al., 2014), and how to represent the optimal benefits that can be achieved from these applications in a variety of healthcare settings (Umble and Umble, 2006; Tabish and Syed, 2015). The different cases of failure and success demonstrate the mixed results arising from the application of these improvement approaches in healthcare.

It is not so clear under which circumstances LM and TOC approaches can be translated to manage patient flow in healthcare and social care organisations (Andersen, 2015). Empirical research that offers an in-depth understanding of the circumstances under which the application of LM and TOC is more effective in managing and improving patient flow remains limited (Brandao de Souza, 2009; Stratton and Knight, 2010; de Vries and Huijsman, 2011; Radnor et al., 2012; Tabish and Syed, 2015; Blackmore and Kaplan, 2016; Poksinska et al., 2016; Rees and Gauld, 2017; Flynn et al., 2019). Therefore, this research project aimed to gain a deeper academic understanding of how existing flow management mechanisms have been developed to meet the distinct conditions present in the wider healthcare system.

This research, therefore, explored the work of Hayes and Wheelwright (1979) regarding the level of instability involved in the operational structure in various manufacturing environments to establish the link between environmental conditions and flow mechanisms. Aa critical review of the origins of flow mechanisms and the aspects of manufacturing environments that enable the effective use of these mechanisms was undertaken, followed by a review of the applications of LM and TOC in healthcare. The research used a multiple case study approach, in line with the best practice advocated by Eisenhardt (1989). Individual and cross-case analyses were included in this research, followed by a discussion of the results. The study then outlined the broader implications of the research and concluded with recommendations for future research.

1.2. Research Aims and Research Questions

This research aimed and gained a deeper academic understanding of how established flow management mechanisms have been developed to meet the distinct conditions within the wider healthcare system, with specific reference to the origins of LM and TOC. The study had the following research objectives:

- To build an understanding of the environmental conditions that led to the development of flow mechanisms associated with LM and TOC, before considering how they have been interpreted in their transfer to health and social care environments;
- To investigate how the LM and TOC flow mechanisms can be applied effectively in the health and social care environments to manage and improve patient flows;
- To provide practical guidance on the selection of flow mechanisms that can help healthcare organisations improve patient flow performance.

To fulfil the research aim, the current study developed three research questions, which are:

1. How can LM and TOC approaches be applied effectively to improve patient flow across health and social care?
2. Why do LM and TOC approaches work better in different health and social care environments?
3. What are the assumptions underpinning the effective use of Kanban and TBM in different healthcare environments?

1.3. Thesis Outline

Chapter One	Introduction	Provides background on the research problem, its aims, and questions and presents a description of the thesis structure.
Chapter Two	Literature Review	Explores the challenges facing healthcare organisations in relation to patient flow management. It also explores the origins of system flow approaches, in particular LM and

		TOC, and their adoption for flow management in manufacturing environments. The review considers the influence of LM and TOC approaches on flow management in healthcare alongside the development of healthcare-specific flow management approaches. The chapter then concludes by identifying gaps in knowledge and by drawing on the limitations in existing attempts to contribute to new knowledge, leading to the development of the research question statement that will guide the thesis toward its overarching goal.
Chapter Three	Research Methodology	Discusses the research design and the approach underlying the thesis. It explains the justification for choosing a multiple case study approach and the protocols and procedures used to ensure confidence in the research findings.
Chapter Four	Individual Case Analysis	Presents a case-by-case analysis of four healthcare organisations, highlighting their original delivery design, the process issues in the original design, the changes recorded, and the results achieved.
Chapter Five	Cross-Case Analysis and Discussion	Discusses the findings from case-study analysis by drawing relationships between constructs and folding these findings back into the literature. The cross-case analysis presented in this chapter facilitates the identification of commonalities and differences in the tasks, events, and processes, which are the units of analyses in the case studies, in relation to the translation of LM and TOC management into healthcare settings.
Chapter Six	Conclusion	Presents the conclusions of the thesis, including a review of the research aims and the answers to the research questions, its contribution to knowledge, the limitations of the study, and the recommendations for future research.

CHAPTER TWO: LITERATURE REVIEW

The focus of this research is the adoption in healthcare of flow-based improvement approaches that originated in manufacturing. Therefore, it is important to understand the nature of these transformations in their original setting before considering how they have been more widely applied across health and social care. This chapter reviews the work of scholars on patient flow within healthcare and the influence of LM and TOC in ensuring better flow. The chapter also positions the study by identifying gaps in knowledge and, drawing on the limitations in existing research, attempts to contribute to new knowledge by addressing these gaps and answering the research questions. According to Creswell (2014), the review of literature allows the researcher to position the study in terms of its relationships and differences with what has been investigated previously. Relating this study to the existing knowledge of patient flow in health and social care allows the researcher to extend those existing studies and generate new knowledge by filling the gaps in the research. The literature review also helps the researcher to develop a conceptual framework and explore the key themes through which the study is undertaken (Creswell, 2014). The chapter, therefore, is divided into four sections, as follows.

Section One begins with an overview of the structure of the UK healthcare system, followed by an explanation of the terminologies used in this research. The challenges faced by healthcare providers and the complexities associated with the health and social care system in the UK are then explored. This highlights the wide variations and instability in primary patient flow activities across this broader system. The section then concludes by emphasising the importance of patient flow management and the need to adopt a system-level perspective.

Section Two explores the instability associated with the operational structure in various manufacturing environments in relation to volume and variety. Drawing on the work of Hayes and Wheelwright (1979) it establishes the link between environmental conditions and system flow approaches. An overview of the contribution of operations management (OM) to flow improvement is provided in this section, followed by an exploration of the origins of system flow approaches and their adoption for flow management in manufacturing environments. The purpose of this section is to examine the principles underpinning effective flow by establishing the originating environmental conditions before considering how these translate to a health and social care settings.

Section Three considers the influence of LM and TOC approaches on flow management in healthcare, alongside the development of healthcare-specific flow management approaches. Section Four then concludes by summarising the chapter and identifying gaps in knowledge and the limitations of the existing literature. This, in turn, leads to the development of the proposed research questions. The outline of this chapter is shown in Figure 1 below.

2.1. Exploring the Operational Challenges Facing the Health and Social Care System
2.1.1. Overview of the UK Healthcare System's Structure
2.1.2. Challenges Facing Healthcare
2.1.3. The Complexity of Health and Social Care System
2.1.4. The Importance of Managing Patient Flow

2.2. Exploring the Environmental Conditions Associated with the Origins of Flow Mechanisms and their Adoption in Different Manufacturing Environments
2.2.1. Aspects of the Instability in Production Environments
2.2.2. The Contribution to Flow Improvement
2.2.3. Flow Mechanisms in Different Manufacturing Environments
2.2.4. TOC's DBR-TBM VS. LM-Kanban
2.2.5. Comparing the Functions/Rules of Kanban and TBM
2.2.6. Conclusion

2.3. The Influence of LM and TOC Approaches on Flow Management in Healthcare
2.3.1. LM and TOC Approaches in Healthcare
2.3.2. LM in Healthcare
2.3.3. Translation of LM into Healthcare
2.3.4. TOC in Healthcare
2.3.5. Translation of TOC into Healthcare
2.3.6. Current Approaches Applied in the NHS
2.3.7. Conclusion

2.4. Summary of the Chapter and Identification of Knowledge Gaps
2.4.1. Summary of the Chapter
2.4.2. Focus and Justification of Research

Figure 1: Outline of the Literature Review Chapter

2.1. Exploring the Operational Challenges Facing the Health and Social Care System

2.1.1. Overview of the UK Health and Social Care System's Structure

First established in 1946 under the National Health Service (NHS) Act, the national system was launched to the public in July 1948. The NHS is structured and organised differently in each of the UK's countries, and similarly to other health systems, healthcare is divided into two broad sections: one concerned with strategy, policy, and management, and another with actual medical/clinical care, which is divided into primary care, secondary care and specialist hospitals (Grosios et al., 2010). The primary includes community care, GPs, Pharmacists and Dentists, whereas secondary involves hospital-based care received through a referral from a general practitioner.

In England, the NHS is the publicly funded healthcare system governed by the Department of Health and Social Care. The NHS provides healthcare to residents of England and other parts of the United Kingdom, with the majority of services being free at the point of use. Certain services, such as emergency care and infectious disease treatment, are provided free of charge to the majority of people, including visitors (DHSC, 2018). NHS England commissions primary care services such as doctors, pharmacists, and dentists. Since 2013, Clinical Commissioning Groups (CCGs) have been responsible for commissioning secondary care services in their local areas. These services include the following:

- most planned hospital care;
- rehabilitative care;
- urgent and emergency care;
- most community health services;
- mental health and learning disability services (NHS England, n.d.)

In England health and social care, decision-making processes can be complex and include a large number of individuals, with various organisations involved in multiple aspects of policy development and delivery. The decision-making process is constantly evolving, with integrated care systems being the most recent initiative to foster the partnership and collaboration between the NHS and social care. Public health, the NHS, and social care all operate under distinct structures and strategies for decision-making, financing, and responsibility (RCP London, 2018). The diagram below depicts a high-level overview of the existing systems and their relationships in England, including healthcare, trust administration, commissioning, funding flow, legislation, and education.

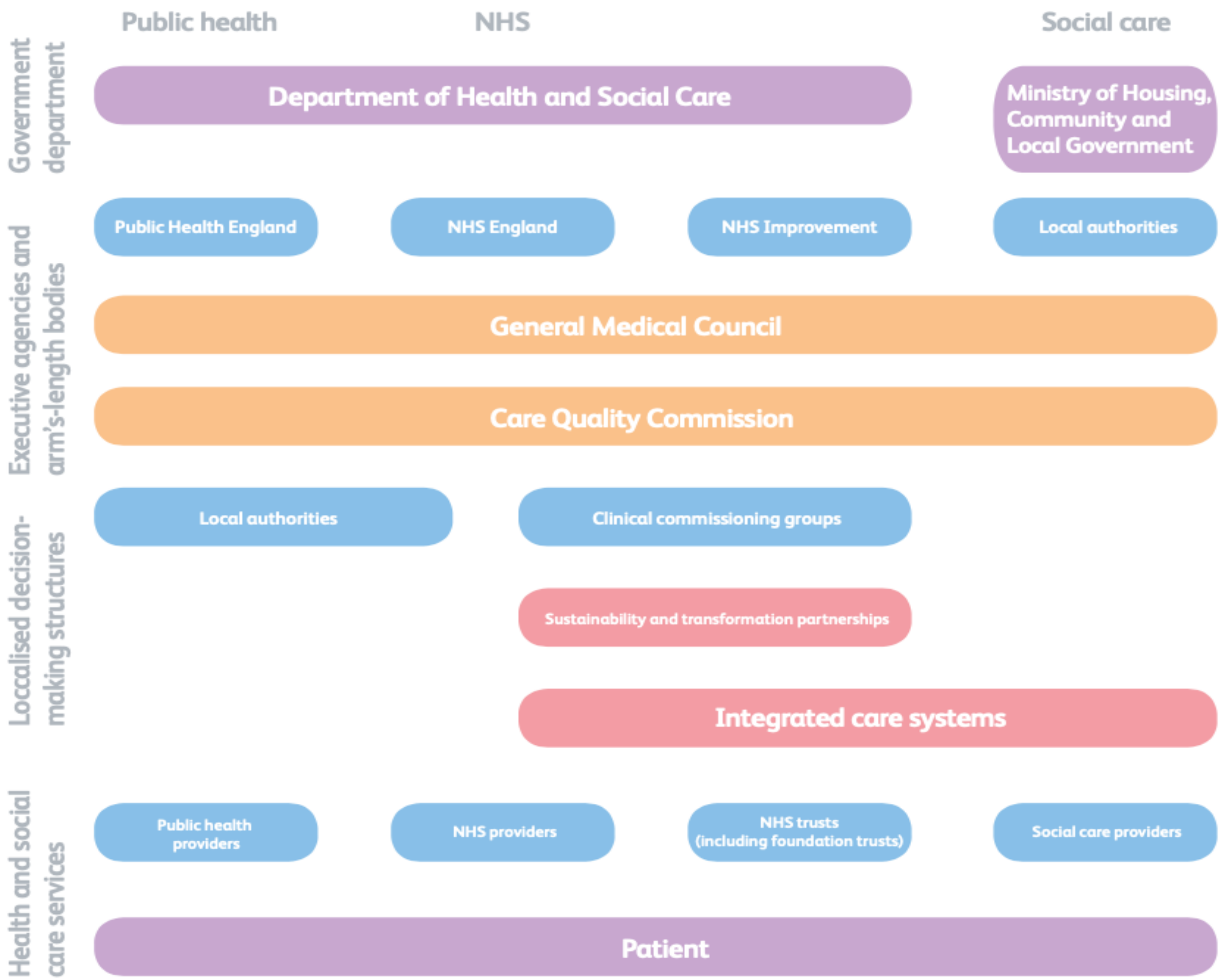

Figure 2: Health and Social Care Structures in England (Source: RCP London, 2018, p.3)

Terminology

As this research focused on how patient flow can be improved across health and social care systems, it is essential to clarify some terminologies before moving to the following sections of this chapter. The terminology used in this research as follows:

- **Healthcare** is a term that refers to a healthcare need that relates to the medication, care, or aftercare of someone who has been diagnosed with a disease, illness, injury, or disability. Individuals can receive healthcare from a variety of medical/clinical services, including general practice (GP), emergency department (ED), community hospitals (CH), residential and nursing care, and acute hospitals (NHS England, n.d).

- **Social care** encompasses assistance with everyday life activities, independence maintenance, social contact, and assisted housing, such as a care home. Individuals who require social care services due to their age, illness, disability, or other circumstances can access them (NHS England, n.d). As indicated in the diagram above, social care is administed locally and is funded differently. This funding distinguishing is resulting in some dysfunctional behaviour, as evidence in some of the case analysis in this research.

- **Integrated care** is a term that refers to a desire to improve the patient experience while increasing the efficiency and value of health delivery systems through improved coordination (Shaw et al.,2011). It is about providing people with the assistance they need, coordinated through local authorities, the NHS, and other stakeholders (NHS England, n.d) and individuals benefit from person-centred care that is coordinated and synchronised across healthcare settings, mental and physical health, and health and social care (UK Government, 2014).

- **Integrated care systems (ICSs)** are modern partnerships between organisations that address health and care needs in a region, coordinating resources and planning to increase public health and decrease inequalities between various groups (NHS England, n.d).

- **Different healthcare environments** refer to different types of healthcare facilities provided to patients. These services could be primary care and hospital services, including inpatient and outpatient care, ophthalmology, and mental health services. There are many different service types that Care Quality Commission (CQC) regulates such healthcare services, community or integrated healthcare, residential social care, community social care and miscellaneous healthcare (CQC. 2015). The flow characteristics in some of these different healthcare environments are subject to a high level of uncertainty and variation in process, demand, and recovery time, whereas variation and uncertainty are low in others.

- **Planned care** is a term that refers to services provided in advance of scheduled health appointments, either in the community or in a hospital setting (NHS England, n.d).

- **Unplanned care** can be described as the healthcare services that patients need that are not expected, unscheduled, or on-demand (Myers et al., 2016).

- **A patient** is a person who has a specific illness or condition and is being treated by a healthcare provider, whether through outpatient or inpatient services (NHS Datadictionary, n.d). Patients' needs vary from one person to another depending on the type of care they require (e.g. primary care, secondary care and specialist hospitals).

- **Patient flow** refers to "the ability of healthcare systems to manage patients effectively and with minimal delays as they move through stages of care" (NHS Improvement, 2017, p.3). Flow in manufacturing refers to reducing inventory in the system and thus reducing lead time, whereas flow in healthcare refers to reducing delay discharge, and hence reducing queues.

- **Measurements used to assess patient flow improvement:**
    - ***Length of Stay (LOS)*** is defined by the NHS as the duration/length of an inpatient episode of care, measured from the date of admission to the date of

discharge and determined by the number of nights spent in the hospital. The average length of stay for patients admitted and dismissed on the same day is less than one day (NHS Scotland. n.d).

- ***Waiting time*** refers to the time a patient waits to receive care and is measured from the time the patient books their first appointment or when the hospital receives their referral letter until the time they receive the care. The government has set the waiting time for patients in A&E to 4 hours from arrival to admission, transfer, or discharge. The treatment target for all other patients is receiving treatment within 18 weeks of referral (NHS England, 2016).
- ***Delayed Transfer of Care (DTOC)*** occurs where a patient is ready to be discharged home or to a supported care facility but still continues to occupy a hospital bed (The King's Fund, 2018).

▪ **Variation** refers to "anything in the system that is not absolutely regular and predictable exhibits variability" (Hopp and Spearman, 2004, p. 145). Apart from the natural variety of patients, there are many causes of variation in health and social care that affect flow. According to the NHS Institute for Innovation and Improvement, the majority of variation is artificial, caused by the way we organise and deliver services. People, information, resources, and processes are all important considerations (NHS England, 2017). These variables differ, as do the types of treatment provided, and they can all influence the flow, cost and results of the process.

▪ The **uncertainty** caused by malfunctions, defects, and disturbances is known as "fluctuations." This refers to any potentially unexpected occurrences of problems, interruptions, or disruptions (Ronen et al., 2012). Uncertainty in healthcare can be associated with demand, process, and recovery times.

▪ **Instability** refers to the variation and uncertainty associated with demand, together with process and recovery times.

▪ **Care/medical/clinical pathways** are a way of laying out a process of best practice that should be followed when treating a patient or client with a specific condition or particular needs (CPA, 2014). Vanhaecht et al. (2007) also described a 'care pathway' or 'pathway' as "a complex intervention for the mutual decision-making and organisation of care processes for a defined group of patients during a well-defined period." This usually entails a specific medical team dealing with particular patients' needs and requirements. If the path of care is predictable, the pathway can be stable and easily managed. However, as patients' needs and conditions become more complicated, the medical path may become complex and unstable.

Terminologies or concepts related to Goldratt's four flow principles (2009) that are used in this study.

- **Value of flow or importance of flow** refers to improving lead-time as a primary consideration in managing manufacturing operations. In healthcare, the value of flow considers ensuring the patient moves to the next stage of care with minimal delays to avoid delayed discharge as the primary objective of managing healthcare operation.

▪ **Flow mechanism** refers to a practical flow mechanism (e.g. Kanban) that ensures when to produce and when not to produce (stope overproduction). In manufacturing, this translates into limiting the release of work in progress (WIP). In healthcare, the flow mechanism can be related to avoiding delayed discharge.

▪ **Focus on Ongoing Improvement** refers to reducing delay further by targeting major causes of delay, be that through reducing inventory (manufacturing) or length of stay (healthcare).

- **Local efficiency measure inhibits flow** refers to the importance of eliminating the focus on local efficiency measure as it negatively impacts on improving flow. In manufacturing, this results in encouraging excess inventory. In health and social care, the protection of budget can result in delayed transfer of care.

2.1.2. Challenges Facing Healthcare

Worldwide, healthcare service providers have been exposed to growing pressure to simultaneously minimise operational costs and improve the accessibility and quality of the care they deliver (White et al., 2011). Today, with the advent of the COVID-19, hospitals across the globe are facing an unparalleled challenge that has a detrimental impact not only on healthcare institutions but also on the economy and the wider welfare society. According to Søreide et al. (2020), while a seasonal syndrome and the periodic cancellation of operations are not new to most healthcare systems, the ongoing pandemic will have unparalleled repercussions for operating centres and for patients with medical conditions.

Recent research has linked delayed discharge to increased healthcare costs (Rojas, Garcia, 2018), negative patient experiences (Everall et al., 2019), increased disease exposure (Barnable et al., 2015; McCloskey et al., 2014; Walker et al., 2011), and increased mortality (Rosman et al., 2015). Walker (2011) revealed that delaying patient discharge can result in increased delays in providing necessary medical care to other patients, appointment cancellations, and backlogs in emergency services. Additionally, Baumann et al. (2007) found that as the LOS increases, there is a greater risk of infection and associated costs of treatment. Younis et al. (2012) and Chidwick et al. (2017) have linked delayed discharge to bed blockage, which ultimately results in delays in transfer within the hospital.

Considering the current budgetary challenges faced by the NHS, the current capacity challenges in the healthcare delivery system, and the fact that over 40 % of NHS England's budget is spent on older people, the importance of patient flow between acute services, primary care, and social care in the NHS and the wider UK healthcare delivery system can never be overlooked (Karakusevic, 2016; Robineau, 2016). Dunn et al., (2016), for example, reported that for the year ending 2015/2016, NHS providers and commissioners reported a deficit of £1.85 billion. The Forward View has previously challenged the NHS to realise savings of £22 billion between 2014/2015 and 2020/21, representing an equivalent 2–3 per cent saving per year (NHS England, 2014). Different parts of the health system have been affected in different ways by the funding challenges, with the acute sector overspending for the previous three years ending 2014 and reporting a 2.6 billion deficit for the year ending 2015/16 (Robertson et al., 2017).

Healthcare expenditure and economic recessions are closely linked, and, as a result, macroeconomic factors, such as insufficient public funds and ageing populations, are becoming challenges for the providers and receivers of healthcare. For example, Hall (2013) noted that total expenditure on healthcare in the USA amounted to nearly 17.9% of the country's GDP in 2010, representing a doubling of costs in a decade, in absolute terms. The NHS in the UK faces the same stresses, leading to huge operational and financial pressures on both health and social care (McKenna, 2016).

Recent data from the Office for National Statistics (ONS, 2020) showed that total health expenditure in the UK in 2018 was reported at £214.4 billion, including both government and non-government expenditure on healthcare, an increase of 5.3 % over the previous year. This represents almost 10.0 % of UK GDP in 2018 compared to 9.8% in 2017 and 6.9% in 1997. According to McKeon et al (2014), the pressure and strain that has been put on the NHS by the ageing and growing population may require an additional 17,000 hospital beds by 2022. With a required increase in the staff from nurses to doctors and equipment, the strain has come when the NHS is already facing a challenge of having to accomplish more with limited resources, emphasising the importance of patient flow.

Expenditure is also constantly rising to keep up with the growing demands of an ageing population (Bergeron and Audet, 2015). This is happening in every country, as people are living longer due to improvements in understanding the causes of diseases and in diagnostic techniques and treatments. The latest data from the Organisation for Economic Co-operation and Development (OECD) reveals that the average life expectancy in OECD countries is now 81 years (OECD, 2019). This has doubled in the last four decades and is continuing to increase. However, people are not only living longer, they are also living longer with chronic diseases (CGI, 2014). Reports highlight the increase in patients' demand for care: in particular, the 40% increase in the number of people aged over 85 between 2003 and 2015 has resulted in increased hospital activities, from accident and emergency (A&E) to referrals to out-patient services, from diagnostic tests to elective admissions (Dunn et al., 2016). These challenges result in a struggle to gain access to care in the right place and at the right time.

Changing demographics and the increasing demands facing healthcare delivery systems have created more complex pathways for accessing healthcare services and have necessitated the integration of social care to enable the safe discharge of patients (Alderwick et al., 2015). While the ageing population continues to grow in the UK, the advances in healthy life expectancy have failed to keep track. The NHS has failed to meet its aim of admitting, discharging, or transferring 95% of patients attending A&E within four hours (The King's Fund, 2016), and, by 2017, over 40% of UK hospitals had cited major problems arising from a limited number of beds unable to match the high demand from patients (Campbell et al., 2017). In December 2019, only 79.8 % of patients attending A&E were admitted, discharged or transferred to another care unit or hospital within four hours compared to the 95 % norm (Stoye, 2020). Due to the challenges posed by the ageing population and the growing burden of disease, integrated care has become a major focus of the UK's health service reforms in recent years (Alderwick et al., 2015). Present demand in health and social care has grown more than the healthcare service providers' capacity to provide the required care in a timely manner. Delays the discharge of older patients once they medically fit and their treatment is complete represent a lack of coordination in different areas of the system, resulting in a waste of approximately 2.7 million hospital day-beds bed days across the entire NHS (National Audit Office, 2016). These delaying older patients' discharge can cost the NHS £820 million per year, the main reason being an increase in patients waiting longer for care packages (National Audit Office, 2016).

Despite receiving an increase in funding for five years from 2019/2020 to 2023/2024, the NHS still continued to experience challenges even before the outbreak of the Covid-19 pandemic (The King's Fund, 2021). Before the outbreak of the pandemic, the NHS was still struggling to manage the growing demand and social care services have experienced the highest strain (West et al, 2020). The growing pressures on the NHS have been felt by patients and employees alike. With staff operating under such immense pressure, patients have experienced longer waiting times for treatment.

The whole health and social care system have been fully stretched by growing complex demand. As the Covid-19 pandemic drove hospitals to prioritise patients in need of critical care, pressures continue to grow and cause delays in some procedures that had already been planned (Trenholm and Wyman, 2020). The NHS, however, still continues to operate under the strain to deliver health and care needs for individuals with or without Covid-19. Dayan (2017) argues that although the NHS has been successfully treating more patients than before, there are still significant issues, for example, the growing need for care, increase in costs and a tight budget have put pressure on managers and employees to manage performance and achieve standards set by the service.

Although improvements have been made, demand has increased at an even faster rate for the service to cope. With approximately 140 000 individuals having been waiting for care for over a year towards the end of 2020, the NHS had already begun experiencing challenges regarding waiting times even before the outbreak of the pandemic (The King's Fund, 2020). A workforce crisis, seen in over 40 000 vacant positions for nursing staff, has meant the NHS needs to effectively and efficiently utilise available resources in providing quality care.

2.1.3. The Complexity of the Health and Social Care System

Apart from the funding issues and the growth of ageing populations, the instability and complexity associated with the health and social care system are perceived to be a significant operational challenge. Although the healthcare sector is considered to be a service sector, the sector cannot be classified as a service or manufacturing organisation (Jarrett, 2006). Sampson and Froehle (2006) presented the unified services theory (UST) to clearly distinguish services from non-service processes and identified the key features of seemingly diverse business enterprises. With services, the authors argue that the customer makes important inputs to the production process, but, in manufacturing, the only primary input of individual customers is the choice and use of the products. UST, therefore, describes a service design process as one that depends on customers' input; for all service processes, customers act as suppliers. In healthcare, for example, customers or patients certainly provide significant inputs into the care process, as they have to be physically present. Such inputs are more uncertain in terms of their needs.

Within the healthcare system, there are different interdisciplinary systems which, as they interact, can create limitations on reaching the goal of the system. Hicks et al., (2015) recognise that systems can have different objectives and factors which constrain the various interactions and flows in the overall system. The dissociation of different elements across the healthcare delivery system indicates the complexity of the system. The increased number of these different elements that influence each other in the system increases the complexity of the flow. According to Kannampallil et al. (2011), the task in healthcare is inherently collaborative. It relies on partnerships between and within groups of different practitioners, from doctors to social workers, and decisions within any unit will be influenced by relations outside the unit, including links with other hospital units, social services, the families of patients, and, occasionally, local law enforcement agencies (ibid). This wider system is subject to high levels of variability and uncertainty due to the involvement of its different components.

Some argue that the health and social care systems in the UK are similar to non-standard manufacturing settings in which there are statistical variations and dependent procedures that characterise flow complexity (Howells and Smeeton, 2015). In manufacturing, instability refers to the variability and uncertainty involved in the production processes and demand. Hopp and Spearman (2004, p. 145) define variability as follows: 'anything in the system that is not absolutely regular and predictable exhibits variability.' The source of variability in manufacturing can take many forms, such as variability in process times, yield rates, delivery times, demand rates, and staffing levels (ibid). Deming (1994) identifies two types of variation in any organisational system: common cause and special cause variation. Variations of the common cause are the natural consequence of the system. The common cause variation in a stable environment can be expected within certain boundaries. By contrast, an unusual incident outside the system, for example, a natural disaster, is a special cause variation (ibid). In healthcare, a random variation in predictable healthcare procedures is common cause variance. The particular cause is an accidental divergence from a reason which is not an inherent part of a specific procedure of healthcare (Bowen and Neuhauser, 2013).

Litvak and Long (2000) identify three types of variation in the healthcare production process: clinical variability related to various diseases, their severity, and therapeutic responses to them; demand variability due to the unpredictability of some forms of patient flow; and variability of professional care due to differences in strategies, expectations, and skill levels. Health and social care organisations deal with unpredictable demands, including a wide range of unexpected human illnesses and injuries. Some patients need immediate emergency care, while others need longer-term treatment.

Health care is not one single organisation; it is a very difficult task to manage and improve healthcare because a hospital consists of four complex groups that function separately from each other (Glouberman and Mintzberg, 2001; cited in Jacobsson, 2012, p.5). These four groups are physicians, nursing staff, managers and the community. The different functions performed by each group highlight the complexity of care (ibid). Since there are many parts and many different interdependencies between the different parts of the system, health and social care is a complex system, which makes it difficult to analyse, develop, manage, and predict the effects of change.

In the UK, healthcare is provided to individuals with different needs through a variety of channels, as shown in Figure 3, such as general practice (GP), emergency department (ED), community hospital (CH), residential and nursing care and acute hospitals. Social care, meanwhile, provides treatment and support to individuals who are in need due to their age, sickness, disability, and other circumstances. This provides assistance for important daily tasks such as food and washing, and involvement in any aspect of life, for example, employment and socialisation. Health and social care are complementary, as some patients are often unable to leave hospital, even if they are medically fit, until the social care package is ready for them.

Figure 3 depicts the activity chain in a health and social care system, reflecting the varied complexity of the patient pathway. There are plenty of distinct patient pathways, and any given day may include a mix of unplanned and emergency admissions, as well as scheduled outpatient and inpatient treatments. The patient journey might involve sequences of events that a patient experiences within a healthcare system from visiting ED to admission, transfer, or discharge.

Each patient has a different need, and he or she usually visits the ED via GP referral, ambulance, or self-referral. Some may be older people with a long-term condition who are unable to see their GP and must be taken to the ED. They may spend hours waiting to be admitted and then remain in the hospital for hours, days, weeks, or months longer than necessary due to the inability to be discharged for non-medical reasons. (e.g. waiting for a social care package or waiting to be transferred to residential care etc.).

The increase of patients' LOS or DTOC (delayed discharge) represent poor patient flow in the healthcare system. The patient flow also can be affected by a lack of managing resources, staff and information. Information flows between wards and operations may be disjointed, and the complexity and convoluted design of NHS systems exacerbate the problem, resulting in reactive operational supervision (Mortali, 2019). Therefore, there is a more significant cause of variation in the length of stay: delay or disruption, perhaps during the patient's journey or at the end of their treatment.

Patient flow is disrupted when patient discharges are excessively delayed, and it may slow or stop entirely if the number of admissions exceeds the number of discharges. Extreme overload often results in patient backlogs that reflect negatively on the hospital, resulting in delays for patients in the ED department. Patient flow and length of stay (LOS) is not simply a function of the time for medical recovery especially where there is transfer of care and the associated delays. Hence, the increased focus on measuring delayed transfers of care (DTOC) to improve the system flow.

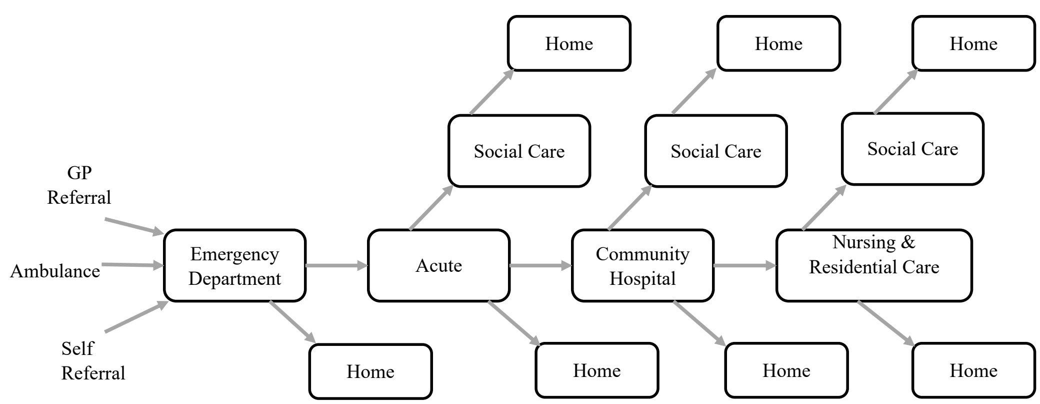

Figure 3: Health and Social Care System (Source: Stratton and Knight, 2010, p. 485 (Modified))

Health and social care organisations have many independent and interdependent elements, which need to respond effectively to the constantly changing demands of patients. The effectiveness of the system as a whole depends on the successful contribution made by many of these and other elements. Hospitals suffer from demand and workload uncertainty, and this complexity in the healthcare system is compounded by the difficulties of integrating the various parts of the organisation. The complexity of this wider healthcare system constitutes an operational challenge that reflects the underlying need and wider objective to coordinate capacity and demand. Although most literature on healthcare service provision has placed considerable emphasis on capacity or demand, it has neglected to explore how the two factors interact to jointly impact patient flow (White et al., 2011).

LaGanga and Lawrence (2007), Qu et al. (2007), Cayirli et al. (2008), Gupta and Denton (2008), Hassin and Mendel (2008) and Klassen and Yoogalingam (2009) have, on another hand, focused on issues of outpatient scheduling. Although Wijewickrama and Takakuwa (2006) and Kopach et al. (2007) explored the impact of outpatient scheduling on patient flow, there is no exploration of how outpatient scheduling and capacity management can jointly affect patient flow. Similarly, Swisher and Jacobson investigated the impact of capacity management on patient flow without evaluating the joint impact of capacity management and outpatient scheduling (2002). Both Berg et al. (2009) and Vermeulen et al. (2009) explored outpatient scheduling and capacity management without focusing on their impact on patient flow while Potisek et al (2007) evaluated patient flow without a consideration of how it might be jointly affected by capacity management and patient scheduling. Only a few researchers have attempted to capture the joint impact of patient scheduling and capacity management on patient flow (Santibañez et al., 2009; White et al., 2011).

The integration of demand management and capacity allocation in a healthcare service setting (Cayirli and Veral 2003; Qu et al. 2007; White et al., 2011) highlights the importance of managing patient flow. The NHS long-term plan now places great emphasis on integrated care to improve patient experiences, service quality, and discharge experiences (Shaw et al., 2011; Lee, 2019), reflecting the importance of optimising patient flow across the healthcare delivery system.

2.1.4. The Importance of Managing Patient Flow

In contemporary health care provision in the UK, managing patient flows has become an integral part of organisational strategy. Considering the growing challenges experienced by global healthcare systems, improving patient flow in hospitals has always been a critical element of achieving efficient and effective provision of quality health care. Following the emergence of the COVID-19 pandemic, there have been interruptions in routine hospital systems around the world. Millions of operations (elective surgery) could have been cancelled or delayed, further highlighting the significance of timely patient discharge.

Patient flow, which ensures that patients receive the care they require, when and where they require it, is one of greatest challenges currently experienced in the health and social care environments (Kreindler, 2017). Improved patient flow has been recognised as of great importance in order to boost the performance of health and social care organisations, as flow performance is a significant feature of organisational performance (Drupsteen et al., 2013). The term 'patient flow' has become one of the most important topics for health managers, academic institutions, and conferences.

Patient flow in hospitals has been of special interest to both practitioners and researchers in healthcare (Armony et al., 2015). Improving patient flow can significantly influence both patient satisfaction and the quality of care. In a similar way, the medical industry has recognised the significance of managing patient flow (JCAHO, 2004). The principal uses of patient flow are: to estimate waiting time and visiting time (Armony et al., 2015); to identify bottlenecks in service provisions (Rojas et al., 2016); to optimise schedules (Thi et al., 2018); and to plan for future reorganisation and resource allocation (Nguyen et al., 2018). By contrast, inefficient patient flow in health and social care settings not only affects the quality of service, crowding/congestion, and timeliness, but also affects patient satisfaction (Aeenparast et al., 2019; Leviner, 2020). However, efficient patient flow through all parts of the healthcare system is rarely achieved in practice, as bottlenecks interrupt the flows and slow down patients' journeys (Pearson, 2008).

According to Pearson (2008), bottlenecks and poor processes result have considerable adverse effects on patient flow. For example, patients need to wait for elective surgeries for longer time periods than are clinically appropriate when hospitals do not have the capability to increase usage or throughput and address inefficiencies in capacity. Hospitals may have to cancel patients' elective surgery at short notice because of poor processes or capacity issues. For example, with the emergence of the COVID-19 pandemic, routine hospital systems around the world have been interrupted, resulting in an estimated 28,404,603 operations (elective surgery) that may have been cancelled or delayed during the 12-week period of disturbance (Collaborative, 2020). This would, of course, ensure the safety of patients and help protect both patients and staff from viral transmission, but it could also lead to declining health, deteriorating quality of life, and premature deaths (Grass et al., 2020), indicating a growing challenge for healthcare providers across the world.

Another important cause of bottlenecks and delays in hospitals' patient flow systems is variation in the level or amount of difference in processes or patient groups (Pearson, 2008). For instance, hospitals treat patients with various injuries or illnesses and try to manage differing numbers of patients from their emergency departments on different days and at different times. These variations are normal and are expected within hospitals. According to Pearson (2008), most variations are avoidable, although they negatively affect patient flows. Hospitals unintentionally develop variations in the ways they manage their processes and systems, for example in the ways they schedule admissions, manage the discharge process, and manage timings for ward rounds. These types of variation could result in considerable differences in the average length of patients' stay across hospitals, even in the same conditions. Consequently, the requirement for hospitals to manage these variations effectively and keep them to an absolute minimum is fundamental to improve patient flows.

The first attempt to improve patient flow was made in 2003 by the Institute for Healthcare Improvement (IHI) as a part of a series of development initiatives (IHI, 2003). The IHI argued that smoothing patient flows could reduce the high variations in occupancy rates and prevent the flood of in-patient visits that result in congestion, poor handovers, and interruptions in care (Hostetter and Klein, 2020). Since then, there is mounting evidence which demonstrates that poorly managed patient flows in emergency departments, intensive care units, and hospitals' other departments result in adverse outcomes, as it is impossible for support staff in laboratories and to keep up and there is insufficient time for physicians and nurses to concentrate on individual patients. For example, the study of McHugh and Ma

(2013) showed that, for each extra patient with pneumonia, heart failure, and myocardial infraction assigned to a nurse, the odds of readmission increased by between 6% and 9%. Poor patient flow has also resulted in increased mortality rates when the ratio of surgical nurses per patient falls, while higher rates of mortality and morbidity occur when patients remain in emergency departments until in-patient beds are freed up (Hostetter and Klein, 2020). For example, the study by Singer et al. (2011) found that mortality rose with increased boarding time, from 2.5% in patients boarding for less than two hours to 4.5% in patients boarding for twelve hours or more. Despite this evidence, overcapacity remains common and persistent in hospitals and other healthcare facilities settings. According to experts in OM this is because healthcare administrators tend to treat the symptoms, rather than causes, of the problems.

For instance, Kreindler (2017) argued that several initiatives focused narrowly on smaller segments of the patient journey thereby failing to affectively address the impairments to flow, whilst in some cases, sometimes improvement initiatives have introduced additional capacity but without processes to link capacity and targeted population. Hewitt and Chreim (2015) also found that healthcare providers generally did not prioritise problems that could be fixed thereby adopting a "fix and forget" approach. When patient care is coordinated and designed with the patient's needs in mind, both staff and physicians would benefit from systems that are streamlined and that can support them in spending more time providing the patient with the required patient care (Bisognano, 2016).

Another noticeable issue in most health practices is a silo mentality, which indicates the absence of a system-based perspective that facilitates the linkage between all parts of the healthcare delivery system (Crawford-Mason, 2002; Knight, 2000; Umble and Umble, 2006). All parts of the system are interrelated and, so, the system needs to be recognised and controlled as a whole, in order to achieve reliability and effectiveness (Langley et al., 2008). For instance, individuals and teams may be drawn to focus on cost-effective decisions in their working, failing to pay attention to the needs of the patient and dismissing the larger perspective of the organisation (Steiger, 2009). According to Steiger (2009), significant interventions that have a notable impact on the productivity of healthcare delivery systems are best coordinated at a system-wide level. Kämäräinen et al. (2016) found that attempts to maximise productivity at a unit level within organisations often result in problems for other levels or units in the organisation. They further reported that it is often easier to adopt cost minimisation strategies than to improve outcomes, leading to long waiting times for patients, delayed discharge, and inefficient allocation of resources.

Additionally, hospitals have inherent complexities and uncertainties in their subsystems. Here, uncertainties emerge primarily because of the randomness in the inter-arrival times of unscheduled and scheduled patients, the non-appearance of the scheduled patients, time-randomness of services at different stages in care, uncertainties in the transitions in the health status of the patients and of patients between different care locations (Bhattacharjee and Ray, 2014). According to Bhattacharjee and Ray (2014), patient flows have three main characteristics, as follows:

- There are different probable care pathways for nearly all the subsystems of a hospital, based on various inter-related factors, for example: the seriousness of the illness, decisions taken by physicians, the development of the patients' health status, and different routing rules and resources limitations.
- There are a number of stages in some instances where patients require combinations of a large number of services.
- There could be different rules of priority for patients to be seen by doctors or to be assessed at radiology departments. Additionally, there are different rules that govern admissions and reservations to allocate beds to different types of patients.

In the UK, the ageing population has made it increasingly important to consider healthcare delivery systems from a wider perspective, recognising that the broader healthcare delivery system is made up of both health and social care systems which should be aligned for effective and efficient healthcare delivery (Stratton and Knight, 2010). The importance of flow across the health and social care system has been emphasised by the introduction of new measures such as DTOC. This measure indicates delayed days as a percentage of all individual occupied bed days at each NHS Trust, thus giving a day-level metric proportionate to the organisation's size (NHS Digital, 2019). A monthly report is collected that shows the total monthly delayed days from the relevant hospitals.

The DTOC report shows some details regarding the delay provider (Social services, NHS, or both), the level of treatment received (acute or non-acute), and the causes of delay. The total delayed days in March 2018 were 154,602, of which 45,457 were attributable to social services (Thorlby et al., 2018). Older people, people with mental health problems, and homeless people are the main groups impacted by DTOC (National Audit Office, 2016). Failure to discharge patients when they are medically fit also leads to high bed occupancy in hospitals, thereby leading to bed blockage. The DTOC is indicative of this deeper failure of the system to provide the right care at the right place and at the right time (Humphries, 2017), reflecting a lack of coordination and interrelation between the different components of the system.

While challenges to patient path flow are clearly evident in healthcare environments, the inefficient use of equipment, employees, and other resources is also apparent (Camgöz-Akdağ et al., 2017). These challenges can create patient flow problems, increasing delays and waiting times for patients and forcing patients to repeatedly execute tasks already performed, while resources are idle and are not utilised efficiently. Studies highlight that resources are rarely the cause of delays; rather the problem lies in patient flow (Haraden and Resar, 2004; Stratton and Knight, 2010; Radnor et al., 2012). This has led healthcare institutions to acknowledge the need to focus on flow and to reduce variability in order to enhance the overall performance of the system. According to Karakusevic (2016), tracking variations in patient flow can, for example, offer service providers an early warning about their system's performance and highlight areas which may require process improvements.

Like other service sector organisations, healthcare institutions have attempted over many years to apply manufacturing-based practices in order to enhance their service performance. For example, the Statistical Process Control (SPC), developed by Walter Shewhart (1939), was widely used for quality control within medical practices (Best and Neuhauser, 2006). Deming's (1982) Plan-Do-Study-Act (PDSA) cycle was also widely adopted as an improvement methodology for solving key problems in the healthcare system (Grol et al., 2007) and for standardising practices through the introduction of medical pathways (van Herck et al., 2004).

The importance of managing patient flow has been acknowledged by many healthcare organisations which have taken a significant interest in system flow approaches from manufacturing. Among these approaches that are commonly used in the healthcare sector are LM and TOC, which focus on flow as a proxy for system productivity (Olsson and Aronsson, 2015). These approaches are concerned with how continuous improvement can be achieved in production systems (Antunes, 1998) and have been developed to reduce and manage wasteful variation in the delivery system, to meet the needs of different manufacturing environments (Stratton et al., 2014). The primary patient flow in healthcare is a broad mix of instability that can be compared to the instability in manufacturing ranging from standardised to customised requirements. Choosing which approach might be suitable to meet the unique condition associated with this wider healthcare system requires an understanding of the environmental conditions that led to the development of LM and TOC. Therefore, the next section critically reviews the origins of these approaches and how they have been applied effectively to manage flow in different manufacturing environments, before considering how they can be applied in a range of healthcare environments.

2.2. Exploring the Environmental Conditions Associated with the Origins of Flow Mechanisms and their Adoption in Different Manufacturing Environments

2.2.1. Aspects of the Instability in Production Environments

In order to understand the environmental conditions associated with different system flow approaches, it is important to clarify the aspects of instability relating to the variation and uncertainty involved in manufacturing operations. The variation of demand and process refers to the degree of non-uniformity of demand or output within the supply chain that may occur from the variation in external demand (e.g. mix and volume changes) or variation in internal processes (e.g. set-up time), whereas uncertainty exhibits the level of unpredictability of the related variation (Stratton, 2008, p. 10). The need to acknowledge the instability in manufacturing operations was stressed by Hayes and Wheelwright (1979), who developed a product-process matrix (PPM) that represents environmental factors relating to volume and variety in different production processes. To explore the aspects of instability inherent in the supply chain of different manufacturing environments, we refer to the model of Hayes and Wheelwright (1979), as shown in Figure 4 below.

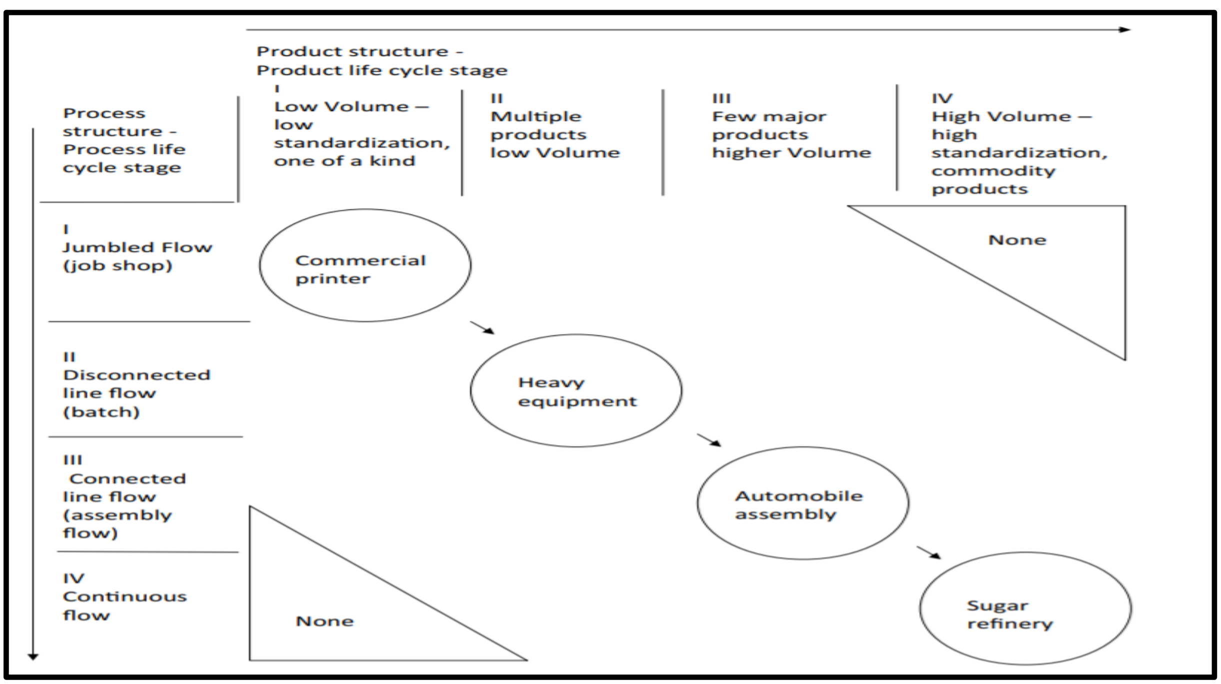

Figure 4: Product Process Matrix (Source: Hayes and Wheelwright, 1979, p. 3)

The PPM represents environmental factors relating to volume and variety in different production processes in order to help understand the strategic choices made by the company. It provides a direct relation between the stage of the product life-cycle (horizontal axis) and stages of the process life-cycle (vertical axis). The distinctions between product structure and process structure are illustrated by using the dimensions of the product life-cycle (production volume and standardisation (low to high)) on the x-axis and those of the process life-cycle (job shop, batch flows, line flows, and constant flow) on the y-axis. The matrix shows the relationship between the volume-variety position of the process and its design characteristics. It can be used for any type of process, whether it is the production of products or services (Slack, 2013, p. 106). PPM gives companies options for the type of processes they use to produce their goods or to deliver their services and also to trade off between the volume and variety of products (Jones, 2008, p. 4).

Demand characteristics are the main driver for choosing a process. At one extreme, a manufacturing company can choose to compete in a market dominated by uncertain demand for many low-volume product variants. At the other extreme, a corporation may choose to compete in a market that is marked by high-volume demand for standardised products. Historically, jobbing and lines have catered to the needs of these two extremes, respectively. Jobbing is linked to customer specifications, such as in the service sector, which involve high levels of instability (variability and uncertainty) and where delivery time and quality are difficult to guarantee, as every job is different.

According to the model developed by Hayes and Wheelwright (1979), processes lie at the end of a continuum of process preference with a high level of customization and a low volume. At this extreme, processes and operations in the sequence are unique, and they create one-of-a-type services or products which are directly created for customers' orders. Under the PPM, firms generally market themselves based on their capabilities instead of based on the particular goods and services. Under the PPM, projects are generally complex, consume large amounts of time, and tend to be large, wherein various interrelated tasks need to be completed, necessitating close coordination. The PPM has been developed by mapping the rows that represent the main stages in the evolution of the process, from fluid to systematic;

the columns in the PPM represent product life-cycles, from the wide product varieties in start-ups to more standardised commodities. On the diagonals, a number of examples are used in this figure to define the intersection of each of these stages. Such trade-off models are now widely used (Hill & Hill, 2012; Slack et al., 2016). The following sections describe the product-process matrix of Hayes and Wheelwright (1979).

2.2.1.1. Job Shop Process

The job shop process is the next process choice in the product process matrix, creating the flexibility that is required to manufacture a variety of goods or services in considerable quantities. The customisations are comparatively high at this point, and there is a relatively low volume for any service or product. In this stage, firms generally use a flexible workforce and equipment to handle different tasks. Additionally, firms generally produce products or services to order, and they do not produce the products or services ahead of time. Furthermore, each order is considered as a single unit, similar to a job. Therefore, the job shop process uses the flexible flow strategy in the product process matrix, organising resources around the process. According to Hayes & Wheelwright (1979), the jobs will mostly have a varying order of processing steps.

2.2.1.2. The Batch Flow Process

The batch flow practice varies in terms of quantity, volume and variety from the job shop procedure. The first difference is related to quantity, whereby the customer groups or manufacturing lots are arranged in greater numbers of quantities, or batches, compared to the job shop processes. At first, the customer group or batch of one product is managed, and then production is transferred to the next one, and, finally, the first service or product is produced again. In terms of volume, the batch flow method differs from the job shop method due to higher volume, as similar or identical products or services can be produced repeatedly. The last difference is related to variety, in that a narrower array of products or services can be produced in the batch flow process. Consequently, the batch flow process uses the assemble-to-order strategy, as opposed to the make-to-order strategy that is used in the job shop process. Additionally, some of the parts for the final product or service could be produced through the batch flow process in advance. According to Hayes & Wheelwright (1979), the batch flow process has moderate volumes and more varieties, which require dedicating more considerable resources to produce each product or service. Additionally, the flow pattern is mixed with no considerable array of operations throughout the facilities. However, according to Hayes and Wheelwright (1979), the batch flow process can result in some more dominant paths compared to the job shop process, and some of the processes' segments have linear flows.

2.2.1.3. The Line Flow Process

The line flow process lies between the continuous and the batch processes, where products or services are standardised, and volumes are higher, allowing firms to organise their resources around the products or services. In the line flow process, firms can move their materials in a linear way, from one operation to the next operation in a fixed sequence, therefore requiring fewer inventories between operations. The firms can operate each activity through the same process over and over with less variability in the production of the products or services. Additionally, in the line flow process, the production orders are not directly associated with customers' orders, unlike the job or batch flow processes. Consequently, organisations can adopt a make-to-stock strategy in their production process, which allows firms to hold inventories for the products to supply as per customers' orders. Therefore, the line flow process is frequently referred to as mass production. However, the line flow process can offer two other possibilities, mass customisation and an assemble-to-order strategy. Product varieties are possible by carefully controlling the addition of standard options to the main products or services. According to Hayes and Wheelwright (1979), production in the line flow process can be either worker-paced or machine-paced.

2.2.1.4. The Continuous Flow Process

The continuous flow process refers to the extreme end of standardised and high-volume production processes, incorporating strict line flows and closely associated process segments. According to Hayes and Wheelwright (1979), materials constantly move through continuous flow process, and the process is generally capital-intensive and runs around the clock to maximise utilisation and avoid costly start-ups and shutdowns.

Hayes and Wheelwright used this model to classify the distinctions that Skinner (1969) originally identified, where he argued the trade-offs and choices needed to reflect the different environments. He also stressed the need to avoid cost and efficiency which encourages local, as opposed to system-wide optimisation.

> Its [manufacturing] management concepts are outdated, focusing on cost and efficiency instead of strategy, and on making piecemeal changes instead of changes that span and link the entire system (Skinner, 1971, p. 62).

Hayes and Wheelwright (1979) and others, such as (Hill 1985), established that not all manufacturing needed to be treated the same way, but rather that the design needed to reflect the level of instability, thus illustrating the reasons for the lack of focus indicated by Skinner's focused factory concept (1974). These developments have emphasised the importance of aligning trade-off choices, but they do not address continual improvement and do not provide a means of continuously managing and reducing variability and uncertainty in the manufacturing delivery system. Subsequently, continual improvement approaches were acknowledged in the West that made use of the systems approach by introducing a mechanism to emphasise flow as a proxy for system performance and to provide a means of continual improvement. As a result, the underlying drivers of trade-office choices were eliminated. Here, two major developments are notable for their widespread use, including in healthcare, but they were originally designed for different levels of instability, and these are the focus of the next section.

2.2.2. The Contribution to Flow Improvement

The adoption of a systems approach to managing operations is commonly associated with managing variability and flow, as opposed to local optimisation based on cost and efficiency. These developments have been sustained because they were accompanied by seminal innovations that offered a practical means of managing variability and flow across the delivery system. This is illustrated by Ford's (1926) physical flow lines, Shewhart's (1931) statistical process control, Ohno's (1988) Kanban control and Goldratt's (1999) time buffer management (TBM). These flow mechanisms provided a means by which management could adopt a systems perspective that guided the organisation in enhancing value.

Walter Shewhart (1939) developed a SPC tool that interprets statistical data as a means of controlling the process and systematically reducing the associated variability. This tool provides management with signals that alert it about when to intervene, as well as how to evaluate the outcome of any improvement activity. The development of SPC resulted in the scientific experimental approach, now commonly referred to as the Shewhart/Deming (1982) cycle of Plan, Do, Study, Act (PDSA), which emphasises the importance of understanding variability and its impact on delivery system performance. This, in turn, was considered a major problem facing management in all aspects of organisations (Deming, 1986). One of Deming's main theoretical contributions was the Chain Reaction Model (1986), which argues that cost reductions are effectively achieved by continuously improving processes. Deming stressed the main action that most companies took during a crisis to minimise costs in order to increase profits. He argued that the emphasis should be on continuous improvement rather than reduction of costs (Deming, 1986).

Each of these developments was concerned with adopting a longer-term systems approach to managing operations, one that is now embraced in operations theory: in cumulative capability (Ferdows and de Meyer, 1990), performance frontiers, and swift and even flow (Schmenner and Swink, 1998). All of these developments have had wide-ranging influence on operations practice, but two, in particular, have contributed to flow management in health and social care, namely LM and TOC.

System flow mechanisms highlighted the importance of detecting and addressing the delays and defects in the process. This, in addition to reducing or controlling overproduction and variability, is the key to improving delivery system performance in manufacturing environments and to meeting customer requirements. Although these mechanisms originated in different manufacturing environments, many scholars have investigated the relationship between these different approaches and acknowledged their significant role in improving flow. For example, the cumulative capability theory of Ferdow and De Meyer (1990) examined sustained improvement in manufacturing related to the development of quality management (QM) and the toyota production system (TPS) management approach. Likewise, swift and even flow theory (Schmenner and Swink, 1998) provides a theoretical foundation for the relationship between LM and TOC-related system flow mechanisms, as they concentrate on flow as a proxy for a system’s productivity.

In his landmark article of 2009, ‘Standing on the Shoulders of Giants’, Goldratt presents a new insight that provides four flow concepts related to the innovations in supply chain management embraced by Henry Ford and Taiichi Ohno. He claims that, while flow mechanisms, including Ford’s flow lines, Ohno’s Kanban system, and Goldratt’s DBR approach, have been implemented to address various degrees of instability, they are all underpinned by four fundamental flow concepts. Goldratt acknowledges that both Ford’s flow line and Ohno’s TPS/ Kanban are systems with the primary objective of achieving a smooth production flow and that the common approach they have adopted can be illustrated by four key concepts (Goldratt, 2009, p. 3), as follows:

First concept*: Improving flow (or equivalently lead-time) is a primary objective of operations*.

Flow in operations entails the speed with which inventories in the operation are moved. Moving inventory and lower speeds create inventory accumulation which takes up space in operations and reduces efficiency (Goldratt, 2009). Both Ford and Ohno realised that increase in work in progress (WIP) causes flow delays, which lead to longer lead times. Ford’s argument was underpinned by the idea that improving the flow of products and materials within an operation and the subsequent reduction of lead time was the key for efficient production (Ford, 1926; Goldratt, 2009). In the same way that Ford sought to increase throughput of the production system, Ohno’s goal in TPS was to improve flow and reduce lead times (Ohno, 1988).

Second concept: *This primary objective should be translated into a practical mechanism that guides the operation when not to produce (prevents overproduction).*

Goldratt (2009) stresses the second concept as a key to Ford’s success, and that of TPS, in achieving smooth production flow. He states that Ford’s flow line and the Kanban system are primarily mechanisms that tell the work centre when production should be stopped. For example, when one work centre stops on the production line, all other centres will stop, as the line stops, and, if all of the other centres continue to run, there is no place to put materials. Likewise, job centres avoid operating in a Kanban system if there are no Kanban cards. Ohno envisioned a mechanism that would determine when not to produce and, from that, designed the Kanban system: a system that directs each work centre when and what to produce, hence preventing overproduction (Goldratt, 2009).

Third concept: *Local efficiencies must be abolished.*

Henry Ford used space to restrict production, whereas Taichi Ohno built a Kanban system to prevent overproduction. If we enforce a system that purposely prevents resources from being overproduced, then the third concept (abolition of local efficiencies) is inevitable (Goldratt, 2009).

Fourth concept*: A focusing process to balance flow must be in place.*

According to Goldratt (2009), the Ford and Toyota mechanisms not only avoided restricting production, but also made use of these circumstances to improve processes that optimised and increased flow speed. Once the integrated mechanisms – space or inventory – generate a path delay or stoppage, the causes of the stoppage can be seen clearly, thus pointing to issues that must be addressed to balance flow. The scale of improvements Ford and Toyota were capable of achieving, in terms of improved speed and lower overall costs over their rivals, attests to the success of their strategies.

These four concepts are widely recognised by many scholars and have received widespread attention, as indicated by their reference in *The Goal* (2004) and in *the theory of constraints handbook* (Cox and Schleier, 2010). In manufacturing, the value (importance) of flow refers to lead-time as a primary consideration in managing operations while in healthcare the value of flow considers ensuring the patient moves to the next stage of care with less disruption to avoid delayed discharge as the primary objective of managing healthcare operation. In healthcare as with manufacturing, the value (importance) of flow should be translated into a flow mechanism to protect the flow from any disruption by controlling variability and uncertainty involved in the patient journey. Having a flow mechanism to control the flow across the system should eliminate the focus on local optimisation measures and enhance the process of focusing on ongoing improvement.

These concepts are useful to interpret the key developments in managing flow in manufacturing and, therefore, will also be used as a structure to interpret the management of flow in healthcare environments. The key developments in managing flow in different manufacturing environments are explored now in the next section.

2.2.3. Flow Mechanisms in Different Manufacturing Environments

2.2.3.1. The Value (importance) of Flow

A good flow means that work flows and progresses according to the requirements of the system in meeting customer needs. When flow is encouraged, the delivery times to the customer can be shorter, whereas poor flow can lead to longer delays for customers' orders, increasing WIP, so that the system becomes out of control. This has led to the development of flow mechanisms to meet the needs of different manufacturing environments. As we have noted, these approaches include Ford's physical flow line (1926), Ohno's (1988) Kanban control, and Goldratt's (1999) TBM, which have all recognised the importance of flow by ensuring that the release of work into the system is based on market needs. Delivery time can, therefore, be reduced and guaranteed to satisfy customers' expectations.

Ford (1926) recognised the importance of improving flow by introducing the flow line concept to meet the needs of the market for high-volume and low-variety products. Following Taylor's practices (1911), the development of Henry Ford's moving assembly line (1926) was a notable evolution in mass production which introduced mass automobile manufacturing, where consumer choice was not essential. Ford's early manufacturing approach had the greatest impact on the development of automotive design and manufacturing. The automotive manufacturing environment had been complex, which led Ford (1926) towards pioneering the moving line of flow by standardising the Model T car. He invested in the flow line in order to deal with the volume of mass production and create a long-term flow of consistent products that were created by workers on dedicated special-purpose equipment (Piore and Sabel, 1984).

> It remained for Henry Ford (1863-1947) to make high-speed mass production of complex mechanical products possible with his famous innovation, the moving assembly line (Hopp and Spearman, 2001, p. 24).

Ford recognised the importance of improving flow by reducing lead time, as customers generally want products as soon as possible with relatively little effort. The establishment of flow line production underlines the need to reduce WIP and remove waste, in order to minimise lead time and improve the overall flow across system operations.

> The thing is to keep everything in motion and take the work to the man and not the man to the work. That is the real principle of our production, and conveyors are only one of many means to an end (Ford 1926, p. 103).

> Having stock or raw materials or finished stock in excess of requirements is a waste (Ford, 1926, p. 99).

As with Ford, the value of flow in improving delivery system performance was stressed by Ohno, the originator of the TPS, which is also recognised as lean production. Ohno was an uncompromising admirer of Ford's thinking and took inspiration from Ford's flow concepts in order to advance them to the next level, to meet the challenge of maintaining a smooth production flow in the context of a diverse product mix (Hopp and Spearman, 2001, p152).

A look at the history of Toyota highlights the special circumstances that led Toyota to come up with a system that was substantially different from the traditional methods of mass production. As Cusumano (1988) explains, the end of the Second World War caused a radical shift in customer demand, from creating trucks for army use to an evolving passenger car market. Ohno (1988) realised that Ford's flow line model was the foundation for improving flow across the system, so he built on Ford's work and further developed it to cope with much lower production volumes.

> Toyota's production organization [. . .] adopted various elements of the Ford system electively and in unbundled forms and hybridized them with their ingenious system and original ideas. It also learnt from experiences with other industries (e.g. textiles). It is thus a myth that the Toyota Production System was a pure invention of genius Japanese automobile practitioners. However, we should not underestimate the entrepreneurial imagination of Toyota's production managers (e.g. Kiichiro Toyoda, Taiichi Ohno, and Eiji Toyoda), who integrated elements of the Ford system in a domestic environment quite different from that of the United States. Thus, the Toyota-style system has been neither purely original nor totally imitative. It is essentially a hybrid (Fujimoto, 1999, p. 50).

With an even lower volume and manufacture of increased varieties in more complex environments, Goldratt (1990) acknowledged the importance of flow and developed the TOC, based on works by Goldratt and Cox (1984) and Goldratt and Fox (1986), to deal specifically with make-to-order (MTO) environments. Goldratt recognised the need to minimise the lead time in complex manufacturing environments and increase the flow of manufacturing operations across the entire system (Schragenheim et al., 2009, p. 20). He recognised the need to identify and exploit the constraint on the system, representing how critical the flow value was from his point of view. If the constraint (time) can be effectively and systematically exploited, market demand can be met.

Goldratt (1984) proposed TOC in manufacturing industries for generic problem solving. It appeared as a piece of production scheduling software called Optimized Production Technology (OPT) for the planning and control of inventory in MTO environments (Olhager, 2013). In his book *The Goal* in 1984, Goldratt launched the TOC philosophy to guide organisations to continuously accomplish their goals (Cox and Goldratt, 1986). TOC is an approach that identifies the constraint and comes up with solutions to resolve or mitigate the constraint (Goldratt and Cox, 1984; Tsou, 2013). The constraint is anything that stops the system from obtaining greater productivity against its target. According to Goldratt (1990), the objective of a business is to create cash now and, in the future, and the word ‘constraint’ is defined as any component that hinders the accomplishment of the company’s ultimate goal. TOC can be described as a management philosophy with the aim of continuous improvement, which leads to enhanced organisational performance (Inman et al., 2009).

Goldratt describes Five Focusing Steps (5FS) as a process of ongoing improvement. The 5FS are: identifying system constraints, choosing how to manage the constraints, subordinating anything to the prior choice: the actions of the whole system must be subordinated to the constraints outlined in the first step, raising the system constraint and, if a constraint was breached in the past step, return to the first step (Goldratt and Cox, 1984; Goldratt and Cox, 1992). According to Goldratt (1990), three fundamental questions about change must be addressed while facing constraints; what to change? what to change to, and how to implement change? TOC is fundamentally about change, and the 5FS deal straight with these three fundamental issues of change that every manager wants to understand. These defined 5FS can be incorporated by change agents in order to strengthen the weakest link or constraint, and it can lead to substantial improvements in the organization’s operations.

All of these distinct approaches focus on flow and reduce the lead time to meet market demand. Acknowledging flow as the primary goal has led to the introduction of different practical mechanisms designed to meet the needs of different manufacturing environments, aiming at controlling variability and uncertainty to protect the flow from any disruption. These different flow mechanisms are discussed in the following section.

2.2.3.2. Mechanisms for Managing Flow

From the manufacturing point of view, improving flow (lead time) is reflected in the ability to meet customers' needs by ensuring the product reaches the customer as and when they need it. To achieve this, there should be a practical mechanism to encourage and manage the flow and to produce precisely what the market needs. Ford, Ohno, and Goldratt translated the value of flow into a practical mechanism, which, in each case, operates under different circumstances. These different mechanisms are discussed below.

2.2.3.2.1. Ford Flow Line

In Ford's flow line, the value of flow is ensured by restricting the space for WIP between the two workstations. When the allocated space is full, workers who continue to feed the space should stop producing, hence stopping overproduction. This is a means of controlling the release of work into the system, ensuring that only what is actually needed on the market is produced and preventing the accumulation of inventories. This creates the foundation for a mechanism that prevents overproduction and encourages flow. This way of thinking made it possible for Ford to significantly reduce production costs and achieve high-speed mass production of complex mechanical products (Dettmer 2001; Hopp and Spearman, 2001).

In terms of Skinner's (1969) strategic focus, Ford focused on speed which he linked theoretically to keeping costs down in order to gain competitive advantages. According to Hopp and Spearman (2001, p. 25), Ford realised that elevated throughput and small inventories would allow him to keep his expenses low enough to maintain an advantage over his competition and to price his product to be accessible to a larger market. However, the moving flow line relied on volume and the use of dedicated routes, eliminating the need to handle changeovers. The mechanism invented by Ford is, therefore, only applicable to the production of a highly standardised car model, such as the Model T.

> Therefore in 1909 I announced one morning, without any previous warning, that in the future we were going to build only one model, that the model was going to 'Model T,' and that the chassis would be exactly the same for all cars, and I remarked: 'Any customer can have a car painted any color that he wants so long as it is black' (Ford 1926, p. 27).

2.2.3.2.2. Ohno's TPS and Kanban System

With growing market demand for a wide range of automobile products, the use of dedicated resources was limited, as machine changeovers were needed and economic batch sizes were determined. As the demand market shifted to a small volume of cars, Ohno realised that dedicated lines were limited and that using limited space as a mechanism to stop production would lead to jams, as not all parts are ready for assembly while the assigned space is already full. Hence, to limit WIP, Ohno (1988) took Ford's thinking further and developed the concept of pull, through the use of a Kanban system to produce a much wider variety of products with much lower and uncertain volumes.

To illustrate this, it is important to provide an overview of the elements associated with the TPS, which Liker (2004, p. 32) referred to as the TPS House. Figure 5 shows the structure of TPS highlighting the two outer pillars of TPS, just-in-time (JIT) and Jidoka, which play a significant role in linking the different elements of the house and maintaining the objectives of quality, safety, morals, cost, and lead time. JIT means generating and delivering completed products JIT for sale and sub-assemblies JIT for assembly into finished goods and buying equipment JIT for its transformation into manufactured parts (Schonberger, 1982). Jidoka, meanwhile, is concerned with avoiding manufacturing that would, in theory, mean 'never letting a defect pass into the next station and freeing people from machine's automation with a human touch' (Liker, 2004, p. 47). Other underlying aspects of the house include the need for stable and standardised procedures, levelling the manufacturing schedule, and visual management, in addition to a philosophy of long-term thinking. Levelling production by levelling customer demand over time can ensure a better use of resources and continued production. All the components depicted in this house need to be robust because the whole system will be affected by any failure (Liker, 2004). In addition to the role of Kanban in managing the operations of TPS, this house is supported by the use of LM tools and other techniques such as 5S, quality check sheets, standardised work chart, SPCs, seven waste, and error proofing, in order to help reduce variation and eliminate waste in production processes.

Within the TPS LM value flow means creating value in a reliable system design that enables the smooth flow of production with minimal breakage, delays, and variations, in order to satisfy customer demand (Womack and Jones, 1996). Ohno's aim was to obtain the required material from the upstream workstations precisely JIT or as required. JIT is concerned with flow by ensuring production responds to the drumbeat of the system or at the level of demand takt time. To enable this, Ohno established the concept of pull, through the use of a Kanban system to control inventory and manage flow across the TPS.

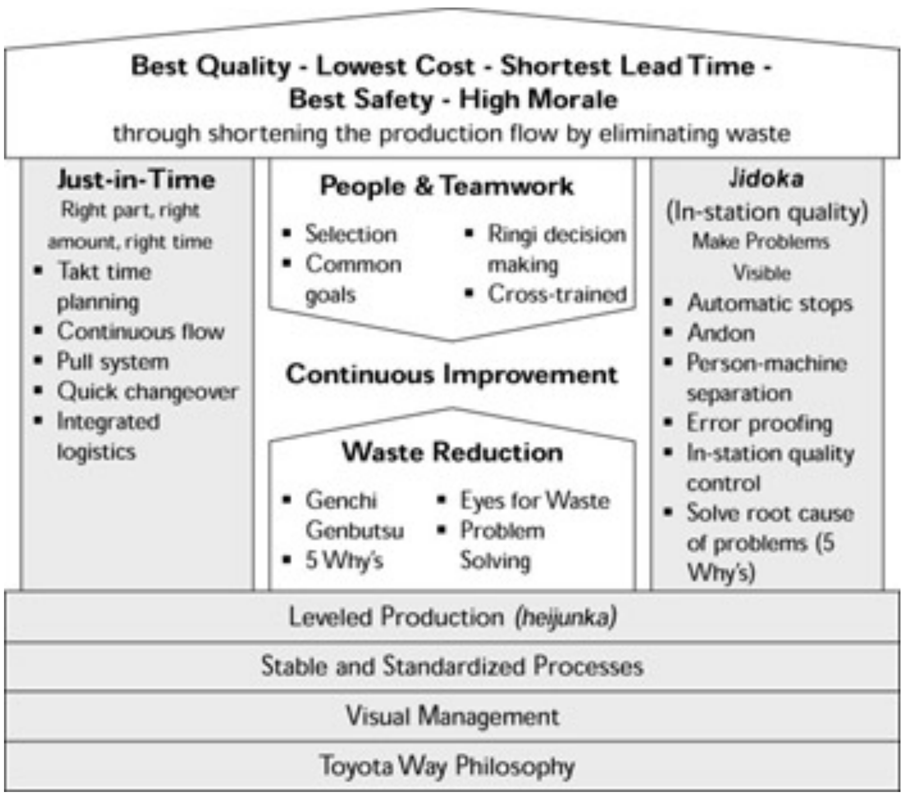

Figure 5: The Toyota Production System (Source: Liker, 2004, p. 48)

The fundamental belief of the five LM principles is the introduction of a pull control mechanism to produce work based on the needs of the system (Womack and Jones, 1996). Kanban is a flow mechanism developed by Ohno (1988) in Toyota that is considered the key to managing variation and flow through the distribution chain. The most famous American advocates of JIT, Hall (1983, p. 39) and Schonberger (1982), describe pull systems as being similar to the Toyota style Kanban system. A pull system is where the release of work is only authorised on the premise of the needs of the system, as opposed to a push system, where work is released without respect to the status of the system (Hopp and Spearman, 2001, p. 340). It is hardly surprising, therefore, that the word pull is often seen as synonymous with

Kanban (Hopp and Spearman, 2001, p. 340), while the term 'push system' is often used as a synonym for scheduling based on material requirements planning (MRP) (Gupta and Boyd 2011; Hill and Hill 2012). Figure 6 shows a comparison of pull (Kanban) and push (MRP) systems.

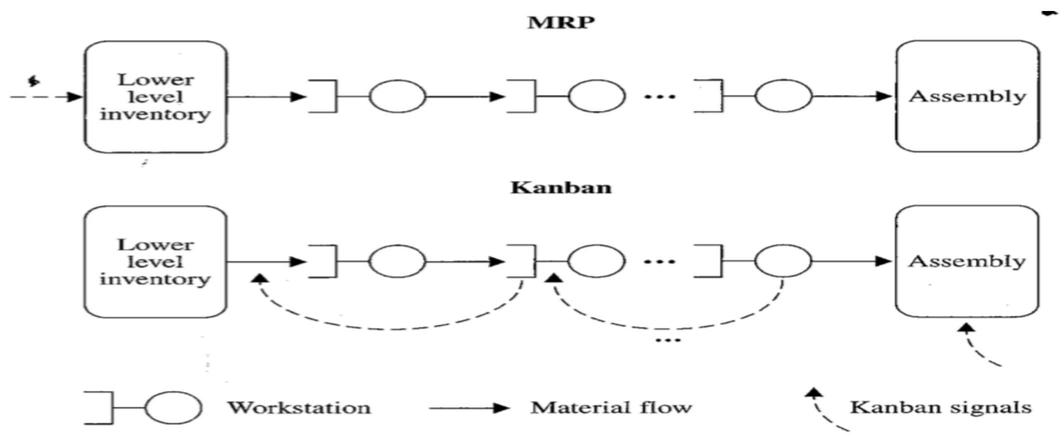

Figure 6: Comparison of MRP and Kanban (Hopp and Spearman, 2001, p. 163)

To deal with lower volumes, Ohno used the resources available to reduce the need to hold inventory and then reducing batch size by reducing setup time.

> When a general-purpose machine... has excess capacity it is an advantage to reduce the lot size as much as possible aside from the separate problem of shortening setup time (Ohno, 1988, p. 56).

The new TPS development model was designed to produce in small batch sizes and with a quick setup. According to Shingo (1989, p. 92), mass manufacturing companies like Ford produce parts in large volumes to prevent changeovers, while the Toyota system takes the opposite approach.

> Our production slogan is small lot sizes and quick setups (Ohno, 1988, p. 6).

This led to buffering with capacity, rather than inventory, which enabled smaller batches and adjusted capacity to meet actual demand. Thus, a capacity levelling schedule can be enabled. This illustrates the law of variability buffering, by effectively choosing capacity over inventory: as Hopp and Spearman (2001, p. 295) indicate, variability in a delivery system will be buffered by some combination of inventory, capacity, and time.

The LM model is influenced by the simplification of material flows by the standard integration of production and supporting services in value streams that generally contribute to physical restructuring from a functional to a cellular organisation (Maskell et al., 2012). With this restructuring in place and to enable the production of smaller volumes with a wide range of products, the Kanban system provides an effective mechanism to manage flow across the whole TPS (Sugimori et al., 1977; Ohno, 1988) by producing the parts required at the right time and in the quantity needed.

Kanban, meaning signal, is aptly named and refers to the Japanese term for the cards used in the Toyota Kanban system to control material flow through the plant (Hopp and Spearman, 2001, p. 162). A lean and just-in-time technique created to control the production, supply, and inventory levels of components (Junior and Filho 2010), Kanban has been considered a useful material flow control tool to manage quantities and control times of production and/or the delivery of products and services when they are required by the customer. With Kanban control, authorisation is granted for the release of work to the shop floor or the transfer of a batch of material between two operations or service centres. According to (Darlington et al., 2014, p. 491), each individual card allows the upstream producing resource to generate and/or transfer a pre-designated amount of a particular part to a pre-designated downstream resource location. Thus, without a Kanban card (signal), nothing can be produced, thereby avoiding overproduction. Ohno considers Kanban to be a central component in the success of TPS in bringing together Kanban's six rules/functions.

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

These six functions/rules associated with Kanban system are presented in Table 1 below:

Table 1: Kanban Functions and Rules (Source: Ohno, 1988, p. 30)

Functions of Kanban	Kanban Rules of Use
F1. Provides pick-up or transmission information	1. Later process picks up the number of items indicated by the Kanban at the earlier process

F2. Provides production information	2. Earlier process produces items in the quantity and sequence indicated by the Kanban
F3. Prevents over production and excessive transport	3. No items are made or transported without a Kanban
F4. Serves as a work order attached to goods	4. Always attach a Kanban to the goods
F5. Prevents defective products by identifying the process that makes the defectives	5. Defective products are not sent on to the subsequent process. The result is 100% defect-free goods.
F6. Reveals existing problems and maintains inventory control	6. Reducing the number of Kanbans increases their sensitivity

With direct reference to Toyota's original Kanban, a review by Junior and Filho (2010) indicated the original Kanban functionalities as follows:

1. Two interaction signals (Kanban system dual card). This dual card utilises manufacturing signals which allow a method to generate a certain quantity of product and transport signals authorising a set quantity of products to be transported downstream (Sipper and Bulfin, 1997).
2. Pulled production. Manufacturing is undertaken depending on the inventory rate or the last station planning.
3. Decentralised control. Monitoring of the manufacturing flow is carried out by personnel at each phase of the manufacturing process, via visual control.
4. WIP Limited. The stock amount for each workstation is controlled, which implies that the buffer capacity is limited, according to the number of signals. (Junior and Filho, 2010).

In the same vein, Stratton and Knight (2010) illustrate the functions / rules of the Kanban system as follows:

- Functions 1, 2 and 4 of Kanban relate to the transfer and processing of information relating to the standard predefined requirements, routing, and transmission of data.

- Function 3 is critical for the LM emphasis on just-in-time production and for maintaining a predefined rate of inventory between each work centre.
- Function 5 guarantees immediate visibility of the cause of defects, thereby ensuring rapid problem detection and resolution.
- Function 6 ensures continuous improvement.

Overall Kanban acts as a flow mechanism in the form of a card to restrict the amount of inventory between work centers or lines, and, hence, all parts across the system only work based on the authorised inventory levels identified by Kanban. Function and Rule 3 of Kanban, as displayed in Table 1, emphasises preventing overproduction and, thus, WIP is controlled effectively and flow can be encouraged. With the Kanban system, the stability of workload and processing time needs to be established in order to effectively control the release of work and enable a continuous improvement process (Dettmer, 2001; Goldratt, 2009).

2.2.3.2.3. Goldratt's DBR and Time Buffer Management

With a much higher degree of environmental instability in the context of MTO environments, the principle underpinning the Kanban method does not fit, as Kanban is inherently affected by relatively low fluctuations in demand and cannot be effective in coping with an absolute magnitude of variability (Hall, 1981; Monden, 1983; Shingo, 1989; Stratton et al., 2008). To enable flow value in this type of environment, Goldratt (1999) developed the DBR technique, as a practical mechanism for limiting time, rather than space and inventory.

The DBR originated from Goldratt's experience with the OPT software production for shop floor scheduling. As the use of the early version of OPT software was limited to repetitive environments, Goldratt realised that it was not necessary for all machines to be utilised 100% of the time – only the constraints needed this. Therefore, the OPT software was redesigned to limit non-constraints to only the work needed to keep the constraints properly fed. However, it was difficult for supervisors to distinguish between non-constraint and constraint resources to meet the necessary timetables when these schedules required less than 100% usage. Later on, Goldratt considered that a lack of understanding of the OPT methodology contributed to its failure (Watson et al., 2007). Goldratt was aware that only bottlenecks had to be scheduled, because other stations have sufficient resources and can maintain pace, and therefore, data accuracy is required only for constraints. Thus, Goldratt and Fox (1986) developed a new version of OPT software to reflect the nine OPT rules as illustrated in Figure 7. The focus of OPT was on finding the bottlenecks in the production process, so that the scheduling activities were focused on these bottlenecks (Schragenheim and Ronen, 1990). This software planned resources depending on the detection and control of the bottleneck constraints. All these rules of the OPT software acknowledge the importance of flow and Rule 1, in particular, highlights the importance of stopping overproduction. Based on Goldratt's experience with OPT software, the DBR technique was developed as a solution for flow management in complex manufacturing environments.

Rule 1: *Balance flow not capacity.*
Rule 2: *The level of utilization of a non-bottleneck is not determined by its own potential but by some other constraint in the system.*
Rule 3: *Utilization and activation of a resource are not synonymous.*
Rule 4: *An hour lost at a bottleneck is an hour lost for the total system.*
Rule 5: *An hour saved at a non-bottleneck is just a mirage.*
Rule 6: *Bottlenecks govern both throughput and inventories.*
Rule 7: *The transfer batch may not, and many times should not be equal to the process batch.*
Rule 8: *Process batches should be variable not fixed.*
Rule 9: *Schedules should be established by looking at all of the constraints simultaneously. Lead times are the result of a schedule and cannot be predetermined.*

Figure 7: The Rules of OPT (Source: Goldratt and Fox, 1986, p. 179)

DBR is a pull control system introduced by Goldratt and Cox (1984) as a practical scheduling mechanism for implementing TOC in manufacturing or service environments. According to Darlington et al. (2014), DBR is a major pull system within TOC; it inculcates the whole throughput of the system. It creates a plan which supports best utilisation of capacity in order to increase the profitability of the company. The TOC claims that that the best place to set and control the pace of the system is at the constraint indicated by the lowest capacity station which controls the throughput rate for the entire production line. This constraint station is known as the 'drum'. Management policy may be a constraint, but, in most situations, the constraint represents market demand (e.g. master production schedule (MPS)) or bottleneck resources (e.g. capacity-constrained resource (CCR)) in terms of the flow rate of the constraint (Schragenheim and Ronen, 1991).

Gupta and Boyd (2008) define the 'drum' as the most extensively used resource, which acts as a capacity constraint in a particular manufacturing system, and, therefore, regulates the whole system's throughput. A rope is a mechanism to control work release based on the system rate (Bicheno, 2004), which connects the drum at the start of the production sequence with the gateway resource (Darlington et al., 2014). Time is viewed as the buffer and exists to protect the released work from variation and uncertainty, as well as to ensure that the drum is never left starved of work (Goldratt and Cox, 1984). It is known as time buffer management (TBM) or buffer management (BM), which provides four management functions (prioritise, expedite, escalate, target). The first three steps can be seen in the DBR system. Step 1 defines a company's drum. Step 2 creates shipping buffers and inner resource constraints if they occur. At Step 3, to keep the buffer steady, the rope is linked between the buffer and material discharge (Cox and Schleier, 2010; Blackstone, 2010, p149).

Several authors, such as Goldratt and Cox (1984), Goldratt (1990), Schragenheim and Ronen (1990), Schragenheim and Ronen (1991), Spencer and Cox (1995), Umble and Srikanth (1997), Schragenheim and Dettmer (2000), Stratton et al. (2008), Stratton (2012), Gupta and Boyd (2008), Darlington et al. (2014), and Stratton (2015), have reported and addressed this method, as described below.

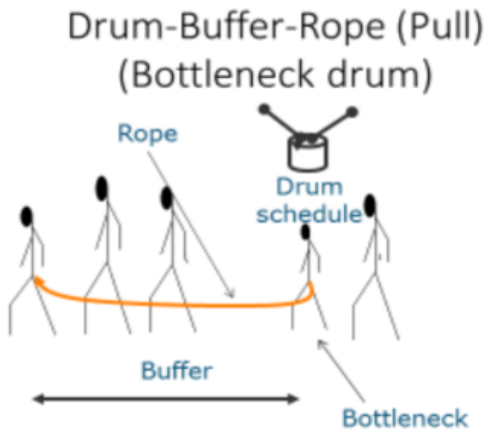

Figure 8: Traditional DBR (Source: Stratton, 2015, p.32)

There are two types of DBR: traditional DBR and simplified DBR. In traditional DBR, the aim is to control WIP flow across a manufacturing line with or near the highest potential of the most constrained resource within the supply chain. Planning of products using CCR as a drum was very traditional in this environment, as illustrated in Figure 8, but it was also recognised that only some products were limited by market demand (Stratton et al., 2008). For optimal flow, work entry is aligned with the current output rate of the CCR, which is generally connected to a drum's rhythm and regulates the rest of the system (Schragenheim and Dettmer, 2000). The rope in this case is tied to the bottleneck that acts as the drum. The rope is a contact tool that links the CCR (bottleneck) to the release material point, ensuring the raw material is not brought more quickly into the processing cycle than the rate at which it can be accommodated by the CCR. To prevent the CCR from 'starving', a time buffer is generated to make sure that the WIP reaches the CCR well before it is expected to be handled (ibid).

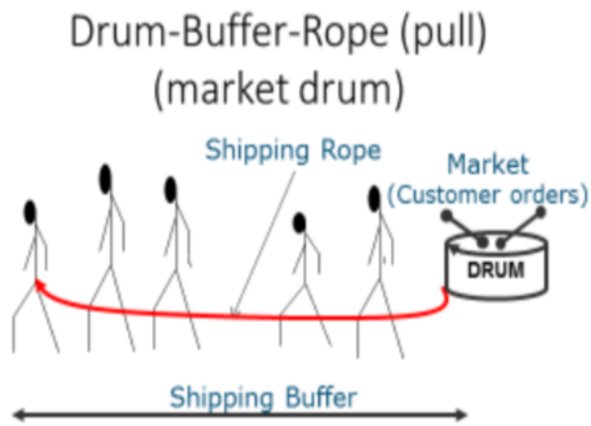

Figure 9: Simplified DBR (Source: Stratton, 2015, p. 34)

SDBR, by contrast, was introduced by Schragenheim (2000) and is based on the same principles as traditional DBR so that it is compatible with TOC and the 5FS. However, the difference is that the drum is always expected to be market demand. With SDBR, even if there is a sudden internal capacity constraint, market demand is the main constraint of the system (Schragenheim and Dettmer, 2000).

The logic of this assumption is that, if we do not sufficiently fulfil the demands of the market, our potential market demand will reduce (Schragenheim and Dettmer, 2000). In this case, as shown in Figure 9, the rope is linked to market demand and there is only one long rope. This is not the case with traditional DBR, as there could be a series of short ropes when there are more bottlenecks in the system. For highly variable environments, using an aggregated buffer (rope) can reduce the impact of variability. According to Hopp and Spearman (2001), variability pooling helps to mitigate total variability by making it less likely that one event dominates output. In the event of a rise in internal capacity, a means of capacity adjustment must be in place to ensure market demand is fulfilled.

The DBR application performs an important role in solving two key problems: the plant's capacity to deliver the planned product flow within a particular timeframe and the impact of continuous deviations on the planned product flow (Pandit and Naik, 2009). DBR offers a mechanism to synchronise product flow and to safeguard the complete system throughput by using time buffers. TBM is involved in delivering management signals to sustain value flow (Stratton et al., 2014) through its four functions, which are illustrated below in more detail.

TBM is a prerequisite for an efficient DBR mechanism. Schragenheim (2010, p. 216) defines TBM as 'a proactive mechanism to handle uncertainty by monitoring information that points to a threatening situation and taking corrective actions accordingly'. TBM sets priorities for processing according to the level of buffer time consumed. It comprises four main functions, and the buffer is divided into three zones (green, yellow, and red), each of which is normally 1/3 of the entire buffer. If the buffer is green, no intervention is required; if the buffer is yellow, a possible problem exists; and if the buffer is red, the problem is urgent and needs to be resolved. Figure 10, as proposed by Stratton (2012), demonstrates the four functions of TBM.

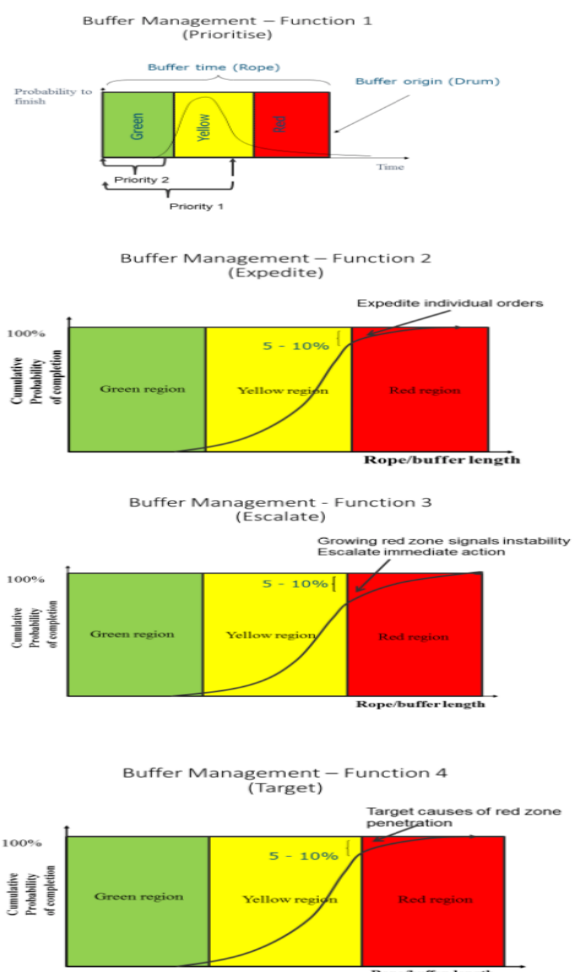

Figure 10: TBM Functions (Source: Stratton, 2012, pp. 12-15)

1. Prioritize: prioritising the flow of work (orders or tasks) based on the penetration or consumption of the buffer. As can be seen in Figure 10, the job order assigns the colour green when the buffer level is higher than 67%. It is therefore fair to assume that the order will eventually be delivered on schedule, since there is still plenty of time left to complete it.
2. Expedite: indicating when the individual orders or tasks are expedited, especially those that are at certain risk, identified by bordering the red area of the buffer. Since the buffer status is between 33% and 67% (yellow zone), orders located in this zone suggest disruptions to the regular flow, and there is a possibility that additional disturbances may prolong these orders. However, progress needs to be made more quickly on these individual orders to achieve the highest priority.
3. Escalate: immediate action is taken when there is a possibility that the system may go out of control. Orders placed in the red zone indicate that the condition of the buffer is less than 33%. No action may be needed if the order is nearly completed. If the order is not at the stage of delivery, action is needed to limit the likelihood of a delay (e.g. a growing number of red zones penetrations), usually by adding capacity or reducing the intake of orders.
4. Target: identify key causes of delay to target improvement. By tracking major causes that lead to delays over time (e.g. causes of red zone penetration), continual improvement of the system can be attained (Goldratt, 1999).

The DBR approach and TBM functions offer a practical mechanism to manage flow in more complex manufacturing environments that involve a wide variety of products with uncertain demand and processes. The concept of the ‘rope’ restricts work release by reflecting the manufacturing lead time and is linked to the drum, which is the system constraint. The ‘drum’ sets the beat of the system and, in most cases, represents market demand. When work is released, TBM facilitates the delivery phases, including prioritising work. In manufacturing with a bottleneck (resource) constraint, a buffer is required to ensure that the constrained resource is never starved of work and that the rope is, therefore, tied to this critical resource, the system drum. In this environment, the inventory buffer is needed ahead of the limited capacity resources, because, if the available buffer capacity is being used, the inventory buffer builds up naturally. Therefore, a mix of buffer capacity can protect inventory from starvation through one aggregated buffer and, thus, reduce the total inventory required in the system.

To sum up, three distinct flow mechanisms, including Ford's flow line, Ohno's Kanban system, and Goldratt's TBM, were explored in terms of how these approaches address instability and enable the flow of production. Each of these distinct flow mechanisms works under different circumstances in manufacturing environments and offers a means to adopt the flow mechanism. The practical mechanism associated with these approaches is the key element to enable the flow across the system. The Ford flow line enables the production of high volumes with a lower variety of products, by physically restricting the space between processing steps, effectively combining process steps with dedicated resources, reducing the risk of workstations being run separately. Ohno's Kanban system was created to be more effective in managing flow with much lower volumes and a broader and more consistent range of products. The TPS was clearly established to achieve Toyota's needs with relatively stable demand and a standard pathway design in a dedicated facility. Under the Ohno Kanban system, inventory is controlled at each workstation, and, only when a product is taken, will the trigger be sent to the preceding work station to produce the amount which has been used. With complex manufacturing environments characterised by a high level of instability in low-volume production and with a wide variety of goods in the form of MTO (Jobbing), TOC's DBR has developed as a robust mechanism to protect system requirements from variability and uncertainty, by restricting time, rather than inventory or space, and by explicitly restricting the release of work based on the 'rope' principle, to avoid early release.

2.2.3.3. Focusing on Ongoing Improvement

Shewhart (1939) opened the door for the process of ongoing improvement by developing the SPC tool as a means of controlling the process and systematically reducing the associated variability. SPC provides the management with signals that alert it about when to intervene, as well as about how to evaluate the outcome of any improvement activity. This has led manufacturing and other sectors to pay attention to the importance of managing to reduce variability, as it significantly harms system performance and is considered the main cause of flow disruption. As the law of variability proposed by Hopp and Spearman (2001, p. 295) states: increasing variability always degrades the performance of a delivery system. So, if variability falls, the efficiency of the delivery system rises, and, thus, a balancing of the flow is achieved.

As mentioned earlier, each of the distinct flow mechanisms works under different circumstances in manufacturing environments and offers a means to establish a culture of continuous improvement. Ford and Ohno were certainly the first to find a systematic way to create a philosophy of continuous improvement. They ensured that, in each organisation and at each level of the organisation, a culture of continuous improvement is built to recognise change that leads to better flow and better use of resources, and which eliminates waste and developments ways to undertake operations with less waste.

> We do not make changes for the sake of making them, but we never fail to make a change once it is demonstrated that the new way is better than the old way (Ford, 1926, p. 53).

> All we are doing is looking at the timeline, from the moment the customer gives us an order to the point when we collect the cash. And we are reducing the timeline by reducing the non- value-adding wastes (Ohno, 1988, p. ix).

The standardisation of work was an essential part of the innovation at Ford and Toyota, in order to accomplish and maintain continuous improvement.

> Where there is no standard, there can be no kaizen (Ohno, 1988).

Goldratt expanded the systemic routes to continuous improvement to involve more complex manufacturing environments. He acknowledged that standardisation of work in a complex system would always be challenged. Where variation in demand or process times increases, standardisation of processes cannot be suitable. In this regard, trying to eliminate variation across the system very rapidly would often be a huge task and very time consuming. So, instead of focusing on all aspects, the focus would be on what needs to be done.

> Focus on everything, and you have not actually focused on anything (Goldratt, 1990, p. 58).

Ford, Ohno, and Goldratt all used various mechanisms, as illustrated earlier, to identify areas that need improvement and to ensure that the flow between demand and supply is balanced. In the assembly line, Ford used space to restrict production. If one workstation fails, all the others will stop, because the line stops, and, if any other station continues to produce, there is no room to put more material. The practical mechanism of using space to limit production obviously leads to clear visibility and direct observation of the assembly line, which helps to identify the reasons for a line stoppage if it happens. Thus, problems to be addressed can be identified, so that the flow can be better balanced.

> If a machine breaks down, a repair squad will be on hand in a few minutes...the machines do not often break down because there is continuous cleaning and repair work (Ford, 1926, p. 103).

Similarly, the use of visible signals to support the continuous improvement process can also be associated with LM/ TPS, since a large part of the TPS approach was based on the idea of Ford's flow line (Ohno, 1988). However, the majority of TPS work centres were not dedicated to a single component, which made it difficult to use direct observation to detect the real problem causing the interruption of flow (Goldratt, 2009). Therefore, the Kanban system was established as a practical mechanism to direct the overall operation across the TPS system, by restricting the release of materials to target variability and thereby encouraging a continual process of improvement.

The process of ongoing improvement by using Kanban in an LM environment can be illustrated through the rocks and river analogy. The inventory level is related to water level, and rocks represent the sources of variability disrupting its flow. Reducing the water level exposes rocks hidden at the bottom of the river and, therefore, the rocks emerging above the water can be removed to improve the flow (Liker, 2004; Lander and Liker, 2007). As a critical aspect of Ohno's TPS, creating the flow of information and materials within a delivery system reduces the water level. It highlights inefficiencies that should draw immediate solutions, as rocks are quickly exposed in production that disrupt the flow (Liker, 2004).

In the Kanban system, job centres stop working if there are no Kanban cards. Ohno used a gradual decrease in the quantity of containers and, consequently, the systematic reduction of components per container. By reducing and controlling inventory, problems in the production process are quickly identified and can therefore be addressed and improved. In reference to Table 1, which represents Kanban's six functions or rules, it is clear that a mechanism of continuous improvement is defined through functions 5 and 6. Function 5 concerns targeting the sources of variability in the production process, by identifying any defective elements in the operation and ensuring that defective products are prevented from passing to the next process. Function 6 exposes existing problems and maintains control of inventory, subsequently resulting in a reduction in inventory over time. Based on the detailed configuration of the Kanban system and the centralised control of mixing and volume, the sources of variation lead to quality problems and delays in processing. The Judoka sends the signal of escalation marked by the 'Andon' signal, which stops the system and specifically identifies the cause of the stoppage (Fujimoto, 1999). This prevents the system from transferring any products into the next process and targets the source of the stoppage. The stoppage of operations might frequently occur because of the system's sensitivity, but Ohno often made buffer capacity available by a policy of rotating shifts that allows the workday to be extended as required (Hopp, 2011).

In more complex production environments, controlling all aspects of the system through direct observation and establishing clear visibility, as with Ford's flow line and TPS Kanban, is hard to achieve due to the instability of the flow paths. Without the need for predefined flow paths, continual improvement in these environments can be achieved. The 5FS of TOC (Goldratt, 1990) reflect the application of DBR, as the rope and TBM work as a flow mechanism to control when work takes place and how the flow time can be improved in accordance with the system requirements. Balancing the flow and not the capacity is the first rule of OPT, which emphasised the focus on ongoing improvement (Goldratt and Fox, 1986). The four functions of TBM act as a flow mechanism that directs the system when to expedite, when to escalate, and, most importantly, when to target variability reduction. As indicated in Figure 10 (Section 2.2.3.2.3), function 4 of TBM concerns improvement by identifying the cause of flow disruption (causes of delays) resulting in the penetration of the red zone. These delay issues can then be resolved and, as a result, the culture of continuous improvement is achieved.

Overall, all these distinct flow mechanisms acknowledged that limiting the accumulation of inventory would promote better flow and ensure efficient operations. Restricting the build-up of inventory within operations can increase the visibility of real challenges that hinder or threaten the flow within operations. The resulting visibility of challenges to flow can then support the elimination of work stoppages, balance flow, and increase the throughput of the production system, more importantly, abolishing local efficiencies. How these different mechanisms have avoided the localised measures to encourage the flow is discussed in the next section.

2.2.3.4. Local Efficiency and Cost Measure Inhibit Flow

A system-based perspective can contribute to the dominance of local cost and efficiency measures. Focus on local efficiency and cost is one of the common issues which inhibits flow improvement. This has been emphasised by the work of a number of scholars, including Shewhart's (1931) SPC, Skinner (1969), and Deming (1986), but this phenomenon has been controlled in manufacturing to a remarkable degree through the development of flow mechanisms. Ford's (1926) physical flow line, Ohno's (1988) Kanban system, and Goldratt's (1999) TBM were distinctive innovations that adopted a system-based approach, concentrating primarily on improving flow, which, in turn, dominated over cost and efficiency measures.

With the Ford flow line, balancing the line and using space as a mechanism to signal when to limit production prevents local efficiency. If one workstation of the flow line fails, the whole flow line stops immediately, and, if any other station continues to produce, there is no room to put more material. Thus, localised measures are effectively controlled. This practical mechanism is a means of adopting a system-based perspective that enabled Ford to achieve the shortest production cycle of the automobile industry at that time.

> Our production cycle is about eighty-one hours from the mine to the finished machine in the freight car (Ford, 1926, p. 118).

The Kanban system also addresses the issue of local efficiency by directing each workstation when the products must be produced and when they must not be produced. So, if there is no Kanban card, it means no production and, therefore, workers only work based on the system requirements and only JIT, when products are needed.

> Kanban is a way to achieve just-in-time; its purpose is just-in-time. Based on this, production workers start work by themselves, and make their own decisions concerning overtime. The Kanban system also makes clear what must be done by managers and supervisors. This unquestionably promotes improvement in both work and equipment (Ohno, 1988, p. 29).

In more varied and uncertain settings, localised measures are most likely unavoidable, because it seems challenging to interlink various parts of the system, particularly if the tasks and resources vary from department to department and sometimes may rely on each other. Dr. Goldratt introduced a series of innovative ideas directed at shifting focus from a measurement of local efficiency to improving organisations' global performance, by concentrating on a few points of system leverage. He stressed the main problems with traditional methods of concentrating on local optimisation of labour as a slogan that, in practice, hinders the achievement of the overall target.

> If a worker does not have anything to do, find him something to do
> (Goldratt, 1990, p. 88).

This contributes to unnecessary manufacturing and inventory accumulation and obscures how important this is by generating the incorrect perception that elevated rates of effectiveness exist (Jones and Dugdale, 1998). Goldratt's philosophy is that there are many activities in almost any system that will add to the efficiency of the system. Instead of focusing on everything, focus on doing what should be done.

Goldratt and Fox (1986) stressed the need to abolish local efficiency to enable flow as the first rule of OPT. This is also reflected in the application of DBR, associated with the TOC approach, by providing a mechanism to ensure that work can be initiated and processed based only on the system constraint. As illustrated earlier, through the rope and TBM, the DBR technique provides a method to determine when the work should be released into the system. Once the work is released, the four functions of TBM provide a robust mechanism to ensure the released work flows without any interruption by subordinating resources to focus on completing this work and meeting the priorities of the system. This way of thinking ensures all resources across the system focus on a common buffer, and that tasks are carried out based only on the system requirements; hence the local efficiency measure is protected.

Overall, avoiding local cost and efficiency measures is a key element in adopting a system-based approach. In Ford's environment, work station efficiency is established at the design phase through line balancing and, hence, using space to limit WIP between workstations made it possible to avoid controlling local efficiency. Within TPS, Kanban's mechanism also addresses the issue of local efficiency by directing each workstation when the products must be produced and when they must stop being produced. So, if there is no Kanban card, it means no production, and, thus, workers perform work only when products are required (JIT). The prevention of local cost and efficiency measures in more complex manufacturing environments has been demonstrated through the use of DBR, which provides a mechanism to ensure that work can be started and processed based only on the system constraint and that everything else, including resources, should be subordinated to this constraint.

From the previous review of flow mechanisms, Table 2 outlines differences across a number of characteristics between these approaches.

Table 2: Differences and Assumptions Associated with Flow Mechanisms

Flow Mechanisms	Supply Volume	Product Variety	Level of Instability (Variation & Uncertainty)	Cycle Time	Batch Size	Process/steps Sequence	Buffering Mechanism			Emphasis	Mechanism Managing the Flow	Ongoing Improvement Process	Abolish Local Efficiency
	High/Low	*Wide-Narrow*	*High/Low*	*Fixed/ Variable*	*Large/Small*	*Defined/Undefined*	*Inventory*	*Capacity*	*Time*				
Ford/Flow Line	High	Narrow very standard product	Low/Stable flow	Fixed	Large No set-ups (Changeovers)	Defined	√			Flow/Lead time	Space	Clear visibility and direct observation	Predictable and feasible when overproduction is stopped
Ohno's TPS/Kanban	Low	Wide but the range is limited	Low/Inherently stable flow	Fixed	Small Set-ups reduction	Defined	√	√		Flow/Lead time	Kanban System	The root cause of disruption "Rocks & Water" Analogy	Predictable and feasible when overproduction is stopped/ Just-in-time production
Goldratt TOC/BM	Low	Wide and unlimited range	High/Complex flow	Variable	Small	Undefined		√	√	Flow/Lead time	BM	Recording and identifying the cause of disruptions	Predictable and achievable when overproduction is stopped /Priority system

These different flow mechanisms have adopted a system-based approach to meet the needs of different manufacturing environments, but LM and TOC, in particular, have contributed to flow management in health and social care. Given the focus of the study on the impact of TOC and LM approaches on improving patient flow, it is important, therefore, to compare the two flow mechanisms (Kanban and TBM) associated with these approaches, to establish the principles underpinning their effective use.

2.2.4. TOC's DBR-TBM VS. LM-Kanban

TOC and LM have acknowledged the need for synchronising production and the creation of a systematic process for continuous improvement (Antunes, 1998). Within TOC, this is achieved through the logic of DBR-TBM (Umble and Srikanth, 1997; Schragenheim and Dettmer, 2000; Stratton et al., 2008), whereas LM production follows Kanban logic (Bhasin and Burcher, 2006; Goldratt, 2008; Stratton and Knight, 2010). Kanban and TBM are the means of managing variation and uncertainty in the manufacturing delivery system. Both mechanisms have similar functions in managing flow, but they operate under different environmental conditions. In order to compare these distinct flow mechanisms, we build on the work of Stratton et al. (2014), who discussed the assumptions associated with Kanban and TBM in manufacturing environments, as displayed in Table 3.

Table 3: Assumptions Underlying Kanban and TBM (Source: Stratton et al., 2014, p. 9)

TPS Kanban assumes:	TOC TBM assumes:
Predefined process steps, time and transfer paths	No predefined processing steps, time and transfer paths
Buffering is based on inventory and held at each processing step	Buffering is based on time and aggregated (pooled)
Process delays (quality problems) are not passed on to the next process	'Delays' are expedited based on rules concerning the consumption of the aggregated buffer.
Level scheduling	Demand may vary, triggering (timely) escalation
Continual improvement is encouraged through reducing inventory to expose problems that are then targeted.	Continual improvement is enabled by targeting the causes of delay (e.g. red zone penetration) then reducing the buffer.

The authors differentiate between these two methods in five aspects concerning the instability associated with their original environments. These five aspects are explored in more detail to illustrate the assumptions underpinning their effective use in manufacturing environments as follows.

2.2.4.1. Sensitivity of Material Flow Path

Both Kanban and TBM address variability to manage and improve flow, but their sensitivity to material flow path is different. With Kanban in the TPS environment, predefined process steps, time, and transfer paths are necessary to balance resources and equipment with the products needed to meet customer demand. Spear and Bowen (1999) reveal the tacit knowledge behind TPS in the four basic rules that drive design operation and enhancement for each product or service.

- Rule 1: The scope, sequence timing, and outcome of all the work shall be highly defined.
- Rule 2: Any communication between the customer and supplier must be direct, and an unambiguous way of sending requests and receiving answers must exist as yes or no.
- Rule 3: The pathway for each product and service should be simple and direct.
- Rule 4: Any improvements in accordance with the scientific method must be made at the lowest possible level in the organisation under the guidance of a teacher.

All rules require that operations, links, and flow paths have built-in tests to automatically signal problems. This seemingly rigid system is made flexible and adaptable to changing circumstances by continuous response to challenges (Spear and Bowen, 1999). From these rules, it becomes apparent that LM relies on the stream of values and a layout allowing the predefined 'simple' and 'direct' pathway for each product. According to Feld (2000), Toyota's workload balancing involves man-time, machine-time, and set-up time analysts in each separate job cell. This analyst compares the time-reflection of product demand (Takt time) with each of these three aspects individually. Success in balancing the workload with takt time would encourage one-piece flow. But balancing workload is a complex task, and the result is that, if LM actually follows it, almost every phase in the manufacturing process can be entirely loaded and functioned as safety nets without buffers (Dettmer, 2001). The system in this state is extremely susceptible to disturbance at any stage. It is all or nothing in continuous flow systems. This is why the focus is on decreasing variability (ibid).

While Kanban is more sensitive to material flow path, TBM has been developed for an unstable flow environment (jobbing) where process times, product mix, and demand are subject to high levels of variation and uncertainty (Chakravorty and Atwater, 1996; Goldratt 2009). Instead of managing the workflow in each cell individually, TBM provides a technique to smooth the workflow through the production system. Goldratt indicates that DBR-TBM offers a much simpler way to smooth the workflow by increasing speed, achieving flexibility, and further improving delivery efficiency, without making extensive attempts to balance each cell's capacity on its own (Dettmer, 2001). TOC argues that variability cannot really be eliminated entirely in unstable environments and that uncertainty is also less controlled, since it is not simple to achieve perfection in the process, to the extent of full accuracy, in such complicated and unstable environments (Dettmer, 2001). Therefore, TBM is less sensitive to predefined process steps, time, and transfer paths.

2.2.4.2. Buffering Choice

Pooled buffers are a useful way to contrast these two flow mechanisms. While Kanban focuses on variability reduction and is highly susceptible to variability detection, TBM focuses on variability management and is less sensitive to sources of variability. Chakravorty and Atwater (1996) indicated that TOC's TBM is ideal for systems with instability and comparatively high downtimes, whereas JIT/Kanban is better for systems with lower downtimes and variability. In the same manner, Pacheco (2018) stressed that TOC addresses variance and uncertainty in operational requirements using strategic buffers (physical, time capacity) while LM approach aims to reduce variability on a continuous basis. TOC uses strategic buffers to address variation and uncertainty in operational requirements, and LM seeks to constantly decrease variability.

Reducing and managing variation is a priority in the design of both LM and TOC approaches, but the key difference is that the TBM mechanism is time-based while the Kanban is inventory-based (Dettmer, 2001; Goldratt 2009). Kanban serves as a signal mechanism, in the form of a card, to restrict the amount of inventory between work centres or lines, to ensure that all parts of the system are produced based on the specified amount of inventory, and, as a result, the capacity should be in place to meet customer demand. With the TPS approach to enable the Kanban system, load levelling and set-up reduction must be implemented (Shingo, 1989). In TBM, as the process and product range are uncertain, the total work release into the system is limited by the rope length. Instead of having inventory buffers to reduce variation at each workstation or individual process, the buffering can be aggregated in the face of the constraints of either the market or resources. The focus on one aggregated buffer allows non-constraint capacity to be freed up to support the requirements of the system and to ensure customer demand is met (exploiting the constraint), and hence variability and uncertainty can be managed effectively.

> Combining sources of variability so that they can share a common buffer reduces the total amount of buffering required to achieve a given level of performance (Hopp, 2003, p. 120).

In TBM, priority is based on the percentage penetration buffer, without any direct relation to steps along the way of processing or routing, therefore managing variability. This is unlike Kanban, where the focus on reducing variability considers workload balance and one-piece flow (Dettmer, 2001). Within TBM, the material is released according to the time buffers and the actual running time of the constraint. Material is not released into the system until a customer order is placed; thus, the WIP in the system relies on the customer's requirements and what the constraint may generate.

2.2.4.3. Fluctuations in Demand

The TPS Kanban system is naturally influenced by relatively low fluctuations in demand and uncertainty and may not be effective in dealing with an absolute magnitude of variation (Hall, 1981; Monden, 1983; Shingo, 1989; Stratton et al., 2008). Goldratt (2009) stresses that one key assumption made by TPS is that of a stable production environment. The TPS Kanban system is naturally influenced by relatively low fluctuations in demand and uncertainties, which are now expressed in LM practices (Womack and Jones, 1996; Stratton et al., 2008). The importance of levelling production to prevent fluctuations is stressed by Shingo (1989, p. 187), as Kanban may not be efficient in dealing with an absolute magnitude of variation. Stratton et al. (2008) stated that Ohno emphasises the adoption of the Kanban system to support flow, but he also stresses that withdrawals should not fluctuate to achieve it.

> The greater the fluctuations in quantity picked up, the more excess capacity is required by the earlier processes... Ideally, levelling should result in zero fluctuations in the final assembly line (Ohno, 1988, pp. 36–37).

Kanban originates in a stable manufacturing environment and is not applicable to other environments characterised by high levels of instability. According to Hall (1981), 'Kanban is intrinsically a system for repetitive manufacturing. It will not work in a shop controlled by job orders.' (cited in Spearman et al., 1990, p. 879). It is not valuable in the context of expensive, rarely ordered products, as at least one product of each type must always be in the inventory (Spearman et al., 1990). As with Hall (1981) and Goldratt (2009), Monden (1983) also shares this assumption. It was assumed by Monden (1983, p. 64) that Kanban is difficult or impossible to use when there are: (1) short-production work orders; (2) substantial set-ups; (3) scrap losses; or (4) large, unpredictable demand fluctuations. The stability in the Toyota environment is essential to enable the effective utilisation of Kanban, hence reducing variation and encouraging flow. This pull system can be used effectively in Toyota due to the stability of demand and load levelling.

In unstable production environments, demand is subject to high degrees of variability and uncertainty, indicating difficulty in level scheduling. TBM is less prescriptive regarding level scheduling and, therefore, TBM functions are capable of prioritising and monitoring the released work and, more importantly, escalating when the system is going out of control (Goldratt 1990; Schragenheim, 2010; Knight 2014).

2.2.4.4. Detecting Process Delays

Both Kanban and TBM offer a mechanism through their functions/rules to detect any quality or delay issues in the production process. As explained earlier in section 2.2.3.3, TBM only escalates and expedites for action when delivery is threatened, accelerating any delays in the process if the red zone of the pooled buffer is breached. Under Kanban, the distribution of centralised buffers between job centres makes it possible to overcome problems immediately and to address them without handing the problem over to the following job centres (Ohno, 1988). Therefore, any quality problems or delays in the process between individual buffers will not be passed on to the next process.

> To get continuous-flow systems to flow for more than a minute or two at a time, every machine and every worker must be completely capable. That is, they must always be in proper condition to run precisely when needed... By design, flow systems have an everything-works-or-nothing-works quality which must be respected and anticipated (Womack and Daniel,1994, p. 60).

2.2.4.5. Encouraging Continuous Improvement

Both LM/Kanban and TOC/TBM are concerned with how continuous improvement can be achieved in production systems (Antunes,1998). According to Moore and Scheinkopf (1998), the parallels between TOC and LM are: the understanding of value from a customer perception; value stream, pull of production and flow; flow control approaches using the principle of market demand pulling; and the quest for constant improvement.

Both Kanban and TBM are committed to supporting management in continuous improvement by systematically reducing variation. As illustrated earlier, continuous improvement is encouraged in Kanban by reducing the inventory to identify problems that can then be addressed (e.g. the rocks and river analogy), whereas in TBM the causes of delay (e.g. red zone penetration) are targeted and then the time buffer can be reduced.

2.2.5. Comparing the Functions/Rules of Kanban and TBM

Both TOC and LM are proxies for productivity and approaches for continuous improvement. As mentioned in the previous section, TBM and Kanban are designed for very different levels of instability. Knight and Stratton (2010) outlined a functional comparison between TBM and Kanban as shown in the Table below.

Table 4: Functional Comparison of TBM and Kanban (Source: Knight and Stratton, 2010, p. 27)

TBM Functions	**Kanban Functions**
F1- Prioritize. Provides relative priority based on planned completion time rather than intermediate processing steps and inventory.	F1 – Pull intermediate inventory F2 – Pre-planned quantity and routing sequence F3 – Prevents over production at each stage F4 – Predefined works order data
F2- Expedite. Proactive time based mechanism of potentially late completion (red zone penetration)	F5 – Quality (variability in the process) signals immediate action.
F3- Escalate. Proactive time based mechanism of growing levels of expediting	

F4- Eradication. Targeting the repeated causes of expediting (red zone penetration) reduces the need for time buffers and improves flow	F6 – Reducing the number of Kanbans (inventory) is used to highlights causes of disruption to flow.

It is clear that both TBM and Kanban have functions and rules that act similarly in principle but that are different in application. The DBR planning process does not aim to bring each activity into line with a timetable, but rather to make the entire system flow stable enough to meet customer demand. The goal is not to protect that task's ability to be on track but to ensure that the requirements of the whole system are met on time. The time buffer is to prevent interruptions in the flow through the system. The DBR provides a significantly higher degree of planning reliability, instead of ensuring protection for each step in the process as in the Kanban pull system. TBM and Kanban are considered to be important elements in protecting the system flow from disruption, and both are designed for very different levels of instability. While Kanban is designed to cope with lower levels of instability, TBM is effective in addressing high levels of instability in the production environment.

2.2.6. Conclusion

Manufacturing environments are characterised by different levels of variation and uncertainty in terms of customer demand, product range, and process sequences. Line environments are distinguished by constant high-volume production and a limited variety of standard products. Some environments have smaller volumes and a wider variety of consistent products, while others are characterised by the production of lower volumes with a greater variety of product range.

Three distinct flow mechanism, namely Ford's flow line, Ohno's Kanban system, and Goldratt's TBM, were explored in terms of how these approaches address instability and enable the flow of production. Each of these distinct flow mechanisms works under different circumstances in manufacturing environments and offers a means to adopt a flow mechanism. The practical mechanism associated with these approaches is the key element in enabling flow across the system. Whereas the encouragement of flow for high volumes and low variety in production have been found to be associated more with Ford's flow line, TPS's Kanban system is found to be more effective in managing flow in producing lower volumes with a consistent range of products. The TPS was clearly designed to meet the needs of Toyota and the relatively stable demand and standard pathway design in a dedicated facility (Hall, 1981; Monden, 1983; Ohno, 1988). In complex manufacturing environments that are characterised by high levels of instability and produce low volumes with a wide range of products in a form of MTO (jobbing), TOC's TBM approach was found to be a robust mechanism to protect the system requirements from variation and uncertainty and to enable flow (Chakravorty and Atwater, 1996; Umble and Srikanth, 1997; Goldratt and Cox, 1984; Schragenheim and Dettmer, 2000).

All these distinct approaches emphasise flow and reduce the lead time to meet market demand. The practical mechanism introduced by each approach plays a significant role in preventing overproduction by effectively managing the release of work based on market demand. The most notable task that these different mechanisms perform and which cannot be overlooked is to address the causes of the flow disruption. Following this systemic way of managing the flow through the entire system, the natural default for local efficiency measures can be eliminated.

Both TOC and LM acknowledge flow as a proxy for productivity with the aim of establishing a continuous improvement process. The Kanban and TBM functional components are built for very various degrees of instability. While Kanban is designed to cope with lower levels of instability (Ohno, 1988; Spearman et al., 1990; Spear and Bowen, 1999), TBM is effective in addressing high levels of instability in the production environment (Chakravorty and Atwater, 1996; Dettmer, 2001; Goldratt 2009; Stratton and Knight, 2010; Pacheco, 2018).

TBM and Kanban can be seen as a significant element to protect the system flow from disruption, as well as to continually address instability. Hence, it is important to take account of the differences between these tools in order for organisations to choose an appropriate approach that can be aligned with their environment.

2.3. The Influence of LM and TOC Approaches on Flow Management in Healthcare

Having reviewed the key systems developments in manufacturing across different levels of instability this next section reviews how these approaches have been adopted in a healthcare setting with a particular focus on patient flow. Firstly, an overview of the interest in manufacturing parallels, particularly LM and TOC, is presented. Secondly, LM and its tools are reviewed, followed by an evaluation of the translation of LM into healthcare. Thirdly, the approach of TOC in healthcare is explored and how this approach has been applied to manage patient flow in healthcare environments. The section then concludes with a summary of the translation of LM and TOC into healthcare.

2.3.1. LM and TOC Approaches in Healthcare

The significant interest in parallels to manufacturing has led many healthcare services across the globe to implement process improvement approaches like LM and TOC that focus on flow as a proxy for system productivity (Olsson and Aronsson, 2015). The last 15 years have seen LM approaches increasingly adapted and implemented in healthcare (Camgöz-Akdağ et al., 2017). Since the beginning of the twenty-first century, LM has been growing in healthcare into a major component of research (D'Andreamatteo et al., 2015).

According to Dekier (2012), the global economic crises which started in the year 2008 further inclined companies to seek opportunities to mitigate the risks of losses imposed by these crises. The main outcomes of these attempts were, as per the study of Dekier (2012), the implementation of methods such as LM and TOC. Radnor et al. (2012) stated that lean-inspired methodologies were gaining momentum in sectors including healthcare and had also been adopted by various hospitals. As Radnor et al. (2012) indicated, LM has emerged when the hospital sector requires improvements in terms of patient care and flow and needs ways to make these improvements in a proactive manner.

By contrast, Mabin and Balderstone (2003) claimed that TOC is regarded as a multifaceted philosophy that came into existence in the early 1980s, inspired by the renowned physicist Goldratt, who addressed the need to bring change to the business world, especially in the US. In addition, a more recent analysis of the application of TOC has been provided by Tabish and Syed (2015), who stated that TOC was being developed as a methodology of improvement for the purpose of creating breakthroughs in performance in complex organisations. Shah and Ward (2007) stated that the term ‘lean’ originated in the production system incorporated by Toyota, in order to deal with relevant market competition.

In contrast, ŞENGÜN (2017) has argued that, as Toyota has completed the transformation of management and production systems within the manufacturing industry, the environment of healthcare sector also demands a new management system. It has been further explained that: the budget of healthcare exceeds its costs; a range of errors are endangering the lives and safety of patients; and there is an increasing bureaucratic inefficiency in health workers and this has led to a search for a new system of management (ibid).

Concentrating on achieving uninterrupted flow through a system by identifying value in each step of a process and derived from the Toyota production system, LM refers to philosophies focused on standardising processes, smoothing out the workflow, and eliminating wasteful steps in the process (Ng et al., 2010). According to Stratton et al. (2014), LM and TOC originated in manufacturing with a focus on reducing and managing variation within delivery systems. In the health and social care context, reducing waiting times between steps and the provision of the next user’s requirements improves the quality and productivity of the delivery system (Ng et al., 2010). Although evidence of success in implementing TOC has come mainly from manufacturing companies (Chou et al., 2012), Cox and Schleier (2010) believe services also offer opportunities to apply TOC and improve delivery. While TOC application has been limited across service industries as a whole, Ronen and Pass (2010) suggest that it is in healthcare where TOC has been mostly developed and where the principles are therefore applicable.

Blackmore and Kaplan (2016) noted that LM ideas have been more widely applied in healthcare than TOC, with significant literature dedicated to supporting the effective use of LM approaches in healthcare. Umble and Umble (2006) and Stratton and Knight (2010), however, argue the application of TOC enables healthcare systems to focus narrowly on managing and improving patient flow, thereby facilitating the integration of health and social care operations.

2.3.2. LM in Healthcare

The Toyota production system, also recognised as lean production, is the outcome of operational excellence centered on techniques and instruments for improving quality and performance. Lean thinking or management is a philosophy that emphasises the elimination of waste or non-value adding elements in processes, so that organisations can provide greater value to their customers (Chan et al., 2014; Dickson et al., 2009; Holden, 2011; Mazzocato et al., 2012). According to Womack and Jones (1996), the lean principle of flow is regarded as the 'progressive achievement of tasks along the value stream so that a product proceeds from design to launch, order to delivery and raw materials into the hands of the customer with no stoppages, scrap or backflows' (p. 306). This means abandoning the traditional batch and queue method of thinking, which seems to be common sense. Robinson et al. (2016) consider the concept of LM as the integration of process improvement tools and operations management to enhance operating and business processes. It is a dynamic process that helps the management to reduce overall waste from the production process.

Following the successful implementation of LM in Japan, the US auto manufacturing sector was the first industry to implement LM outside of Japan, and the concept subsequently spread to other industries, including healthcare, as shown in the following Figure.

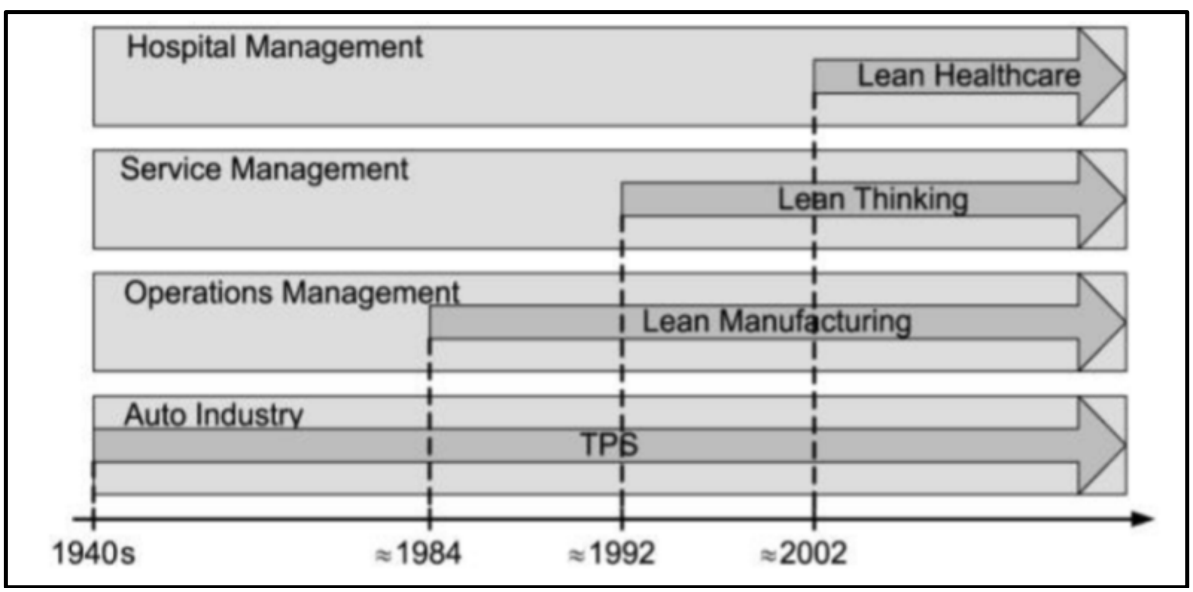

Figure 11: The Development of Lean Management (Source: Laursen et al., 2003 cited in Lin and Chiu, 2017, p. 531)

Since the promotion of LM in the West, and as the work of Womack and Jones (1996) and Porter and Barker (2005) indicates, five LM principles have been endorsed as the general framework for the implementation of LM in organisations. The five key principles of LM are: value specification from the customers' perspective; the identification of the value stream for each product or service; the creation of the value flow; the introduction of pull between all steps where continuous flow is impossible; and management towards perfection (Womack and Jones, 1996). Behind the implementation of its methods and techniques is Toyota's ideology of understanding people's motivations, to be a learning institution, and to focus on long-term thinking (Liker, 2004).

In Toyota's environment, frontline staff and management are accountable for the provision of quality work, and they should be engaged in problem-solving processes through involvement in rapid and constant development workshops, known as Kaizen. The role of leadership in successful lean implementation cannot be overemphasised. Failure by leaders to appreciate and share the importance of change often results in a failure of lean initiatives (Waring and Bishop, 2010; Al-Balushi et al., 2014). Although the support and participation of the leadership and top management are crucial in the implementation of LM, modern prescriptions of LM insist that employees must be incorporated and empowered to assess and improve their own tasks (Dombrowski and Mielke, 2013; Drotz and Poksinska, 2014)

There are a number of tools and methods used in the application of LM (Womack and Jones, 2003; Graban, 2008; Holden, 2011), which include:

- Value stream mapping (VSM) which is a method used to develop diagrams and other methods to describe the existing and desired future steps in processes, including the flow of materials, products, workforce, and information;
- Sessions involving short-cycle continuous improvement (e.g. Kaizen events);
- Standardised work based on evaluation of the presumed 'best way' to perform tasks;
- Tasks performed by multi-skilled teams;
- The 5S method (sort, store, shine, standardise, and sustain) to coordinate and standardise workspaces and solve problems;

- Root cause analysis (5 why);
- Assembly lines and cell-based manufacturing;
- The A3 Report which is a standardised tool to solve problems;
- Prevention of mistakes and failures (poka-yoke);
- Kanban- which is a flow mechanism for inventory management across a decentralised system;
- Andon which is a visual feedback system showing the state of production and warning when support or action is needed.

VSM is one of the most effective tools in the implementation of LM. Although it does not directly enhance flow on its own, VSM enables all industries to appreciate and continue appreciating the applications of LM (Singh and Sharma, 2009). Just as it is widely applied in manufacturing industry, VSM has found applications in the process and service industries (ibid). However, VSM is most effective when applied to systems that follow a linear operation. Once systems become complex, the application of VSM becomes complicated, and it becomes difficult to enhance LM processes in non-standardised and complex systems.

The implementation of LM principles in a delivery system seeks to create a more controlled environment that supports and improves both the impact and management of demand variation (Westwood et al., 2006) and that promotes the creation of value flow throughout every step of the delivery system. Eliminating non-value adding processes from the delivery system to achieve this objective, LM relies on tools and techniques that improve flow and reduce waste, among them VSM, 5S, work standardisation, PDSA cycles, Kanban and Kaizen (Burgess and Radnor 2013).

2.3.3. Translation of LM into Healthcare

This section evaluates the applications of LM in healthcare, based on the four key aspects related flow identified earlier. These four aspects are used because they have informed the effective translation of flow mechanisms in manufacturing environments. This enables a comparison between the implementation of LM in manufacturing and the translation of LM into healthcare.

2.3.3.1. The Value of Flow

As mentioned in Section 2.2.3.1., the value of flow is intended to ensure that market demand is met by ensuring that the product reaches the customer at the right time in the required quantity. Healthcare recognises the importance of managing patient flow by minimising the LOS for patients and ensuring that patients are discharged on time when they are medically fit.

The value of flow was acknowledged among the notable cases of the successful implementation of LM in healthcare, such as the Virginia Mason Medical Centre (VMMC), Royal Bolton Hospital, Flinders Medical Centre, and the Pittsburgh General Hospital (Spear, 2005; Ben-Tovim et al., 2007; Fillingham,2007; Grunden, 2007; O'Connell et al. , 2008; Gubb, 2009; Kenney, 2011; Blackmore et al., 2011; Radnor et al., 2012). It was evident in all case studies that the organisations were experiencing critical challenges that threatened the survival and viability of the healthcare organisations. The desire to ensure quality and safety and reduce patient delays was a consistent driver of the improvement initiatives across all the studies.

In the case of VMMC, as cited by Kenney (2011), the CEO Kaplan acknowledged the urgent need for change and innovation by focusing on quality, rather than on finances, and making the approach patient-centric rather than physician-centric.

> We need to be a great place for doctors, and we need to have physicians in leadership role, but we need to keep in front of us all of the time that it is all about the patient (Gary Kaplan, cited in Kenney, 2011, p. 4).

Value to patients by emphasising patient safety and quality was evident at Royal Bolton Hospital, where the high mortality rate and LOS was longer than average, resulting in a lack of access to diagnostics and clinical decision-making and delayed discharge (Fillingham, 2007; Gubb, 2009). At the Flinders Medical Centre, the ED became congested and was deteriorating, causing frequent cancellations of elective surgery. According to Tomvin et al. (2007), patients visiting the ED were put into different triage categories and this resulted in patients who arrived earlier being attended after patients who arrived later because of the differences in the triage categories. This highlights the need to improve flow to ensure patient safety and quality of delivery are met. Similarly, the Pittsburgh General Hospital recognised the importance of flow and the need for improvement where an infection arising from surgery threatened patient safety and the quality of services (Spear, 2005; Grunden, 2007).

With other LM applications in healthcare, there appeared to be a conflict in specifying value from the customer's perspective, the first step in the LM implementation process, which is perhaps the most significant element of LM. Failure to correctly specify value prior to the implementation of LM techniques can easily drive organisations to provide the incorrect type of service or product when this appears to be the efficient means for the organisation (Womack and Jones, 1996).

Some authors, such as Burgess and Radnor (2013), have pointed out that, as value in healthcare is uncategorised and complex, the interpretations and perspectives of value in that setting vary widely. Uncertainty persists when it comes to identifying customer value, as value can seem the same for both commissioners and patients (Radnor et al., 2012). This is consistent with other LM literature, which has also highlighted the presence of multiple customers in healthcare organisations, revealing a fragmented and complicated scene in which to determine the value (Young and McClean, 2008; Ghobadian et al, 2009).

However, Toussaint and Gerard (2010) suggested that the principles of LM, when applied to healthcare, could be simplified to identify value for the patient, eliminate any form of waste, and minimise the time taken to diagnose and treat the patient, thereby incorporating the patient into the design of care.

2.3.3.2. Mechanisms for Managing Flow

In healthcare settings, the increase in patients' LOS or DTOC indicates an increase in variability that compromises patient flow efficiency. It may be analogous to the state of overproduction associated with manufacturing operations, because DTOC implies that beds are over-used by patients who might not need to stay in the hospital, but may need to move to the next stage of care when they are medically fit. As with manufacturing, the need to reduce variation and uncertainty is essential to enable the flow of patients in healthcare environments. This required a practical mechanism to ensure all elements in the healthcare delivery system are aligned to meet patients' needs on time.

In all the successful cases examples mentioned in the previous section, a practical mechanism to reduce variation and enable flow was driven by the adoption of a structured approach to redesign the medical pathways. This involved the use of LM tools, such as the 5S, visual management, and other key LM techniques and tools, to standardise work. Across all the organisations, the structured approaches aim to add value to the customer and reduce waste. All the attempts to redesign the care pathways resulted in more standardised approaches to healthcare delivery.

The VMMC started the journey in 2001 by developing the Virginia Mason Production System (VMPS), aimed at redesigning their care pathways and creating standardised instrument trays for surgeries and procedures, as well as standardised secondary processes. In developing the VMPS, the management also applied the VSM tools, particularly 5S. The 5S tool provided a visual system that kept the working environment organised and safe (Kenney, 2011). The success of this approach was motivated by the VMPS, which aimed to standardise procedures and streamline routine treatment aspects in order to minimise the amount of waste and free up staff time for patients (Kaplan, 2010; Plsek, 2013). Standardised procedures included the creation of a laparoscopic cell for surgery check-in with standardised laparoscopic case carts for all surgeons, a one-piece pre-operation flow cell, and a standard room-planning procedure for all surgeries, with significant standard work in place and a patient ready to move into the room in 5 minutes (Black and Miller, 2008).

Beside the standardisation of work, other popular LM tools were applied at the VMMC. These tools including defining the value stream and takt time, load levelling, and mistake-proofing. According to Black and Miller (2008, p. 169), one way to organise load levelling in both radiation oncology and infusion centres was through the principle of cascaded scheduling (i.e. staggered start times for patients and staff). This allowed both clinics to give patients longer operating hours, thereby increasing the volume of patients without stressing staff and also optimising the availability of the equipment required (ibid).

For Hyperbaric Medicine at the VMMC, Kanban cards were used to classify and pull the patient along the course of treatment (Black and Miller, 2008). The laminated Kanban cards had letters large enough to be read from a distance and included the patient’s last name and medical and treatment numbers, as well as the date. A big X is also shown on the card when the patient has to see the doctor after a hyperbaric procedure on that day.

> The idea here was that the patient would move through the system at her own pace, without being routed by clinic staff or held up by a lack of their availability. The use of Kanban cards would enable this (Black and Miller, 2008, p. 184).

Kenney (2011) found the underlying principle of pull allowed patients to move through the system based on their needs and not based on the hospital departments' speed. Eliminating the need for patients to find someone to check them in, the Kanban cards created the smooth flow of patients through the processes. A key feature of the pull system that enabled flow at the VMMC was the ability of the system to reduce intervention in the patient journey by hospital staff, granting more control to the patient. The two characteristics created flow, allowing patients to pull the resources they needed when they required them and to independently manage their treatment (Black and Miller, 2008; Kenney, 2011). However, there is limited evidence regarding applying Kanban to manage patient flow across the whole hospital.

In addition, two-bin Kanban systems were put in place in the nursing units of the VMMC to avoid excess inventory, to detect unloaded supply demand, and to adjust stocks. This led to a 30% reduction in the inventory of nursing units and a similar 30% reduction in the central supply region between 2006 and 2007, decreased inventory value by about $500,000 (Black and Miller, 2008). The redesigning of care processes has led to significant improvements as follows:

- A reduction of 85% in patient waiting times for a lab result.
- An increase in productivity by 93%.
- A decrease in inventory costs of $1 million.
- A reduction of 90% in defects.
- An annual saving of $26,880 by reducing surgical instruments from 74% to 58%.

The Royal Bolton Hospital in the UK applied a similar approach to redesigning the process for patients with fractured hips. The hospital started redesigning care in 2004, which resulted in the development of the Bolton Improving Care System (BICS) in 2006, aiming to improve productivity and quality. By creating a specialised trauma centre and standardised protocols with its own physical space that incorporates all the resources required for patients in a single team, the LOS was reduced by 33%, mortality decreased by 36%, and personnel handover was lowered by 42% (Fillingham, 2007).

The Flinders Medical Centre adopted a programme of redesigning care in 2003 aimed at redesigning whole sequences of care.

> The Flinders Medical Centre has been using lean thinking to make a start on designing whole sequences of care-not simply to provide the care that is right, but right first time, for the right patient, at the right place, and at the right time. As we do so, the size of the challenge and the potential benefits of success become clear. Redesigning Care has made a start, and it is only a start, on this important task (Ben-Tovim et al., 2007, p. 15).

The redesign of the Flinders Healthcare delivery system started with the identification and classification of patients into groups with similar overall care processes and needs to each other, but different from those required by other groups, such that they could not be managed together. This involved classifying patients into two categories by creating two streams of patients based on the likelihood of being admitted and likelihood of being discharged; each patient stream was allocated to specific teams of doctors and nurses (Ben Tomvin et al., 2007; O'Connell et al., 2008).

As with all the above hospitals, the Pittsburgh Regional Health Initiative (PRHI) developed Perfecting Patient Care techniques in 2001 to redesign secondary processes and streamline the Ambulatory Surgery Center at the Pittsburgh General Hospital (Spear, 2005). Promoting the flow of patients through the concept of 'pull' was central to the success of ensuring that patients arrived in the operating room on time, in compliance with their time of surgery and with a set of specific priorities. As illustrated by Grunden (2007, p. 120), pull suggests that the patient should not be prepared for surgery by the Ambulatory Care Center ACC (upstream process) before the patient is required by the Operating Room (OR) (downstream process). Once the OR requires the patient, a signal will be sent to the ACC to have a patient ready for surgery. However, prioritization is needed as patients vary, and the ACC must be able to anticipate when pre-op preparation will be done to avoid delay in the procedure.

The use of Kanban as an inventory management tool was implemented on the whole pharmacy at the Pittsburgh Healthcare hospital (Grunden, 2007). By ensuring that the whole pharmacy operated under the Kanban system, 16 hours of technician time were eventually freed up. When there was enough inventory for the specified number of days, the Kanban cards would be placed at a point where they could act as a trigger for inventory, allowing adequate time for stock replenishment. Overall, the use of LM tools has led to reducing infections by 90% within 90 days and reductions of nearly $500,000 in intensive care expenses per year.

A further redesign of care processes using the LM approach is commonly known in the NHS's Productive Ward programme in the UK. This programme aimed to involve operational healthcare employees in improving productivity and quality. Recommendations of productive schemes are noted for their role in transforming healthcare (Lipley, 2009; Dean, 2009; Snow and Harrison, 2009; Callard, 2008; Manning, 2011). As with all the above cases, the use of LM tools, such as kaizen blitz activities, focused on redesigning the work environment to manage capacity, reduce waste, and support work scheduling.

Among the cases providing evidence of success of LM are Lakeview healthcare (Carman et al., 2014), the Consorci Sanitari del Garraf (Netland and Powell, 2016), and Saskatchewan (Jones, 2015), where the application of LM has resulted in notable improvements in systems operations. At Lakeview healthcare, for example, the adoption of LM has been accompanied subsequently by reduced process times, improved efficiency, and the elimination of waste (Carman et al., 2014). Other notable outcomes included a reduction in process cycle times, efficient staff and patient flow, and shortened travel time for patients and employees (Carman et al., 2014; Netland and Powell, 2016). The Luton and Dunstable University Hospital, a medium-sized district general that had challenges in primary care access, recorded a more than 30% decline in patients from the emergency department and improved patient satisfaction, following process intervention methods that provided more suitable environments to address the care needs of patients (NHS, 2017).

A study of Brazilian hospitals showed that LM techniques, such as VSM, 5S, kaizen, workload balancing, standardisation, cell layout, and quality control, minimised waiting times and the number of elective surgery cancellations, as the hospital struggled to perform high-volume surgery. According to Regis et al. (2018), lean healthcare has helped the hospital to minimise waiting time, decrease the number of cancelled surgeries by 28%, and reduce bedtime from five and half hours to two and half hours. As with Brazilian hospitals, Edwards et al. (2012), as cited by Daultani et al. (2015), presented a lean case in a surgical ward where a team turned two out of ten surgical rooms into 'turbo rooms' that conducted elective surgery only for less complicated patients. This contributed to greater stability and load scheduling. They also noted that, given the diverse mix of patients, emergency departments tended to be ideal for LM because care was performed with little to no planning, unlike elective operations (Daultani et al., 2015).

2.3.3.3. Focusing on Ongoing Improvement

The use of visible signals to sustain the continuous improvement process can also be correlated with LM / TPS. As described in Section 2.2.3.3. earlier in this chapter, a significant part of the TPS approach was based on the Ford flow line concept. With the development of the TPS to produce a lower volume and a wide range of products, the direct visibility of the process across the system became difficult. Therefore, it was essential to establish the Kanban system as a practical mechanism to guide the overall operation of the TPS system and to encourage the continuous improvement process.

In all the successful case examples mentioned earlier, ensuring central visibility was a key characteristic, with process mapping used in all organisations to allow employees to visualise the entire patient journey within the delivery system. The redesigning of medical pathways to deal with particular and consistent patient demand made direct observation by staff possible to identify any problems interrupting the flow path. At the Flinders Medical Centre, for example, the creation of production cells in line with the mapped value streams supported waste reduction and enhanced flow improvements. With each cell directing clinical efforts on particular patient groups, it was easier to standardise care to a certain extent and to deliver patient care as patient needs arose, in contrast to prior approaches patient treatment was offered in batches, which resulted in queues.

The focus at the VMMC on continuous improvement was evident through the establishment of patient safety alerts (PSAs) as part of its Jidoka system. According to Black and Miller

(2008), the PSA system at the VMMC allows any employee to stop the process if anything goes wrong. As the institute's chief, Dr Andrew Jacobs, said:

> Incorporating PSAs has enabled us to catch 'near-misses' and make process improvement (Black and Miller, 2008, p. 170).

According to the same authors, the importance of focusing on LM concepts at the cancer institute at the VMMC, rather than tools, was achieved by placing the patient first, eliminating waste (Muda), introducing continuous flow, creating pull production, and introducing ongoing incremental improvement through Kaizen events.

Harrison (2008) outlines how using productive ward tools supports workers in identifying their failures and offers further examples of actions which can be applied to improve performance. Financial savings were also reported from the use of productive community tools, while specific steps in processes also resulted from adopting the tools (Laurent, 2013). Productive ward programmes have been linked to the reduction of hospital acquired sicknesses and staff sickness (Foster et al., 2009; Smith and Rudd, 2010). Improved patient satisfaction and reduced interruptions are also benefits recognised from adopting the productive ward scheme (Lipley, 2009; Smith and Rudd, 2010).

A study by Regis et al. (2018) on the implementation of LM in three Brazilian hospitals to improve chemotherapy, radiotherapy and the surgery schedule indicated that the use of VSM, PDCA, and DMAIC has facilitated progress in the ongoing improvement culture. The authors found that the application of LM tools, such as VSM, can result in reduced lead time and decreased setup of processes, as VSM facilitates the identification of activities which fail to add value to healthcare processes. The identification of non-value adding activities and their subsequent elimination over time can then bring incremental improvement in the healthcare service delivery. The role of continuous process improvement in the success of LM in healthcare is evidenced in the studies of Regis et al. (2018) and Costa et al. (2017), who repeatedly associated successful LM implementation with kaizen events.

Antierens et al. (2018) argued there was no healthcare system that had adopted all the 14 principles of TPS, with only 50 to 80% of organisations showing evidence of administrators' commitment to the expansion of their 'pull' systems to discourage overproduction, the frequent standardisation of tasks as the basis for continuous improvement, and persistent reflection and continuous improvement. Crema and Verbano (2015) found that, for all healthcare organisations that have adopted LM in the USA, those reporting full adoption do not exceed 4%, resulting in the neglect of sustainability issues, such as developing a culture of continuous improvement (Radnor et al., 2012). However, the success of organisations in creating a lean culture is dependent on continuous improvement and people's involvement, as Liker (2004) suggests that the involvement of people in continuous improvement efforts and the elimination of waste through the Toyota Way is a characteristic feature of advancement towards a lean corporate culture.

Of the measures adopted in notable cases of LM success in healthcare, most of the interventions involved a redesign of a particular pathway and using structured approaches. This allowed direct observation to be possible to detect problems easily, as all pathways were predefined to offer care for particular groups of patients. Although two cases demonstrated the use of Kanban cards for a particular department, such as the operating room and hyperbaric procedures, there remains limited evidence of using a Kanban system as a practical mechanism for controlling and balancing flow through the system in LM applications in healthcare.

Despite the successful implementation of Kanban in the manufacturing sector, production and market conditions are not similar for all organisations, production and market condition are not similar for organisations, more still for the service industry and the manufacturing sector (Papalexi et al., 2015). Thus, the limitations of Kanban in promoting process and material flow within delivery systems have been recognised, with Junior and Filho (2010), Sjoberg et al. (2012), and Lin et al. (2013) acknowledging that, in situations where demand is unstable and unpredictable, where operations are non-standardised and require long setup times and include a great variety of items to be processed, Kanban is not adequate to create flow in the process and materials. The limitations in the suitability of Kanban to implement a pull strategy within a delivery system could possibly explain the weak evidence of the successful adoption of the Kanban system within the healthcare sector (Papalexi et al., 2015).

2.3.3.4. Local Optimisation Inhibit Flow

As with manufacturing, avoiding local efficiency is a key to adopting a flow mechanism, as focusing on local efficiency measures could undermine the widespread improvement of the system and could negatively impact the improvement of flow. This can be reflected in the healthcare environment in terms of meeting government measures, such as DTOC or the 18-week treatment target, where the tension between the various parts in the system may arise, each looking at their own performance and causing conflict in achieving the overall goal of the system. As mentioned earlier in section 2.1.3., one of the most important issues facing healthcare practices is a silo mentality which indicates a lack of a centralised linkage between many elements whereby efforts for local optimisation cannot be avoided. Although hospitals use some tools like the Kaizen blitz to measure and improve flow, this might contribute to localised optimisation gains and might not address the system's requirements as a whole (Radnor et al., 2012). Now, let us see how local optimisation measures have been managed through LM applications in healthcare.

As with manufacturing flow line environments, where there is a predefined pathway with staff allocated for expected demand, the opportunity for localising measures is more avoidable, as there is a standard procedure, and everyone should follow it. This is reflected in all the notable cases of the successful implementation of LM, indicating stability in their environments, which made the structured approach to redesigning the medical path more applicable. Although LM applications have shown significant improvements in these hospitals, it should be noted that LM transformation is not an easy task and that it requires a great deal of effort, as there are always challenges. Here are some quotations that express these challenges.

> A lot of people talk to me, say, 'yes, we're going to do lean and that will sort that problem out'. Lean is paradoxical: if you remove waste, you'll find more problems underneath that waste (Joy Furnival, the Associate Director of Transformation at Bolton NHS Foundation Trust, 'PEX Network', 2012, p. 11).

> We'd been doing all this improvement work for three or four years, some of which was on mortality work in specific pathways. So how could it be if we'd done all this improvement work, we were still really not very good? We had to face up to the fact that we'd done some good work, but we hadn't done enough yet (Joy Furnival, 'PEX Network', 2012, p. 6).

> Lean isn't a magic wand that you can do it for three months and suddenly everything's fixed and I guess part of the challenge for Lean or other methods, like Six Sigma, is there can be this perception that if you're doing it suddenly you'll be perfect but, in fact, potentially the opposite is true (Joy Furnival, 'PEX Network', 2012, p. 13).

> Bolton's early progress on its lean journey has been encouraging, but it has not been without its dilemmas and challenges (Fillingham, 2007, p. 240).

> Applying lean to healthcare in Bolton seems to be achieving just that for those who work there (Fillingham, 2007, p. 241).

Despite the success of the TPS LM as a typical example of a system based approach and a number of publications which have explored this approach, the effectiveness of LM in healthcare is still a subject of debate. Andersen et al. (2014), for example, found that the existing literature on Lean implementation in healthcare does not effectively address contextual factors and mechanisms influencing the sustainability of Lean efforts. Lean thinking takes on various forms as a result of management introducing the concept, consultants sharing information, and employees experiencing actual system and process changes (Anderson and Rovik, 2015).

Holden's study of the application of LM principles in the A&E department, for example, revealed that LM application had reduced the number of patients who left without being attended, waiting times, and the total duration the patients stayed in hospital (Holden, 2011). In a similar way, the application of LM at the Astrid Lindgren Children's hospital in Stockholm has led to tremendous improvements in the hospital's operations (Mazzocato et al., 2012). According to the author, notable achievements following the adoption of LM principles included a two-year sustained improvement of 19 to 24% in waiting and lead times, which was credited to lean's capacity to connect interdependent resources, reduce uncertainty, and standardise work.

Lu et al. (2010) have argued that, as most of the LM studies assumed stable production processes free from process variability, not all attempts to improve processes through LM have created the anticipated results. Andersen and Rovik (2015) stress that evidence regarding the impact of the adoption of LM practices is lacking. While Moraros et al. (2016) believe no evidence on the success of LM outcomes has been reported, Mazur et al. (2012) consider the rate of failed LM implementations in healthcare to be as high as 90%. These arguments build on the earlier discussion by Winch and Henderson (2009), who established that, rather than always solve patient management problems and hospital challenges, LM had frequently added to the healthcare challenges. Winch and Henderson (2009) further argued that, rather than serving the interests of patients, LM has often been implemented in healthcare to serve the interests of healthcare providers, undermining the value of professional healthcare to that of a mere service that is measured in terms of cost, time, and inauthentic concepts of quality.

Such failures of LM approaches in healthcare can, however, be attributed to poor adoption of LM practices. As several studies have established, many healthcare organisations have operated at then adoption stage of LM, restricting LM to particular tools and techniques (Kim et al., 2006; Fillingham, 2007; Radnor et al., 2012; Dannapfel et al., 2014; Kaplan et al., 2014; Holden et al., 2015). These findings on the failure of LM to produce notable results in healthcare are consistent with the argument of Steed (2012), who claimed that one of the primary reasons for the failure of LM failure in healthcare is that several organisations have not adopted the LM approach as an integral component of their operations. Radnor et al. (2012) consider the current state of LM adoption in healthcare organisations, at their adoption stage, to be equivalent to the stage at which the automotive industry found itself during the late 1980s and the early 1990s. They claim that budgetary limitations in healthcare provision make service delivery capacity-driven, limiting the potential to manage demand and fully utilise resources that subsequently become available. Thus, using some specific LM tools, such as Kaizen blitz, can lead to localised improvement and unfilled promise.

Failures to adopt the underlying principles of LM and to implement LM fully is evident even among the notable cases cited as successful case studies of LM implementation. Barnas (2011) cites ThedaCare in Wisconsin USA, a leading healthcare organisation that has successfully adopted and implemented LM approaches, as having adopted only the tools and techniques of LM without adopting the entire LM philosophy. As a result, in its initial LM implementation attempts, ThedaCare experienced challenges in sustaining the LM work on an organization-wide basis (Barnas, 2011; Antierens et al., 2018). According to Liker (2004), failure to understand the principles of the Toyota Way, which consists of 14 management principles, suggests that adopting the TPS tools and techniques and applying them to healthcare will not offer healthcare organisations the anticipated results.

Moraros et al. (2016) state that LM interventions have potential, but inconsistent, benefits in process improvement, such as patient flow, improved discharge, reduced patient visits, and improved time-dependent care. The authors further argue that the implementation of LM in healthcare environments has no statistically significant association with health outcomes and the satisfaction of patients. The implementation of LM was negatively associated with worker satisfaction and financial costs (ibid). Fillingham (2007) considers delivering benefits from LM in healthcare to be one of the most challenging phases. For LM to be more effective, there is a need to have clear aims for the interventions at the beginning, for example, whether the interventions are aimed at improving patient satisfaction, productivity, or reducing mortality (Fillingham, 2007).

Aherne and Whelton (2010) believed LM thinking could be applied effectively to a variety of organisations, among them healthcare and service organisations. As LM is centred on processes, and all organisations, irrespective of industry, consist of a set of activities and steps that make a series of processes through which the organisations create value for their customers, LM can be applied to non-manufacturing processes in a similar way that it is applied to manufacturing operations (ibid). Within healthcare, it can be said the focus of LM is on the elimination of waste in all processes and tasks to facilitate as efficient a use of time, materials, and resources as possible (Aherne and Whelton, 2010). Most of the LM applications in healthcare involved secondary processes, such as consumables, pharmacy, and test procedures, rather than patient flow. While major improvements have been made in patient flows, these are typically associated with the redesign of medical pathways (Fillingham, 2007; Van Calster et al., 2019). These applications are consistent with TPS LM improving flow through the redesign of the physical flow path (Hicks et al., 2015; Murphy et al., 2019).

Attaining the total elimination of non-value adding activities and implementing LM is challenging under circumstances of demand uncertainty and process variability. Research on LM applications in the context of primary care and across the entire health supply chain remains limited (D'Andreamatteo et al., 2015). According to Brandao de Souza (2009), no comprehensive review of the literature on Lean in healthcare has been conducted since the management approach was translated into healthcare. Rees and Gauld (2017) also argued that, while implementation variations and deployment challenges have made existing evidence about Lean application less conclusive, additional research is necessary to extensively and comprehensively address the subject of Lean application in healthcare.

The majority of published cases highlighted as examples of successful LM implementation in healthcare have dedicated their attention to the adoption of certain tools, for example, process mapping, as the means of attaining short-term improvements (Brandao de Souza, 2009; Radnor and Walley, 2008; Mazzocato et al., 2010; de Vries and Huijsman, 2011). The application of LM tools in healthcare is therefore still at its early developmental stages, and currently there are limited cases where an organisation-wide strategy has totally integrated LM interventions (Nelson-Peterson and Leppa, 2007; Radnor and Walley, 2008; Mazzocato et al., 2012). Flynn et al. (2019) concluded that while sustainability of Lean in healthcare is an important aspect, it remains an understudied area of implementation research. The limitations in translation contribute to insufficient evidence and have an impact on the success or failure of Lean Management approaches.

The summary of the translation of LM into healthcare is shown in Table 5 below.

Table 5: The Translation of LM into Healthcare

Translation LM in Healthcare	
The value of flow	- Emphasis on ensuring quality and safety, reducing patient delays (LOS) and increasing productivity (Spear, 2005; Fillingham, 2007; Grunden, 2007; Gubb, 2009; Kenney, 2011) - Conflict in identifying value from customers' perspectives, as the value in healthcare is perceived to be uncategorised and complex (Young and McClean, 2008; Ghobadian et al, 2009; Radnor et al., 2012; Burgess and Radnor, 2013).
Mechanism for managing flow	- The redesigning of medical pathways for specific patients' needs enabled visibility across the pathway and direct observation to manage this particular flow path (Spear, 2005; Ben-Tovim et al., 2007; Fillingham, 2007; Kaplan, 2010; Plsek, 2013; Carman et al., 2014; Jones, 2015, Netland and Powell, 2016) - Using Kanban as a mechanism for particular departments (Black and Miller, 2008; Grunden, 2007; Kenney, 2011)

	- Using Kanban as a practical mechanism across the system was not evident
Focus on ongoing improvement	- VMMC used a Patient Safety Alert System to alert management once errors or problems occurred (Black and Miller, 2008) - The use of LM tools can help identify waste and non-value adding practice through the establishment of clearly defined pathways where direct observation can easily detect any problems interrupting the flow (Spear, 2005; Ben-Tovim et al., 2007; Fillingham, 2007; Harrison, 2008; Kaplan, 2010; Costa et al., 2017; Regis et al., 2018)
Inhibit localised measures	- The dedicated pathway allows the standard procedure to be easily followed and eliminates the opportunity for localised measure. - Although some cases highlighted examples of successful LM implementation in healthcare, the majority of published literature regards the results as short-term improvements (Brandao de Souza, 2009; de Vries and Huijsman, 2011; Radnor and Walley, 2008; Mazzocato et al., 2010), reflecting local improvement rather than system-wide improvement (Nelson-Peterson and Leppa, 2007; Radnor and Walley, 2008; Mazzocato et al., 2012)

2.3.4. TOC in Healthcare

In addition to LM, the application of TOC thinking in healthcare has been encouraged through various education improvement bodies in the NHS and worldwide, typically centering on The Goal (1984). This guidance primarily acknowledges the role of bottleneck management and the need to focus planning and control around constraints (limiting factors) in the delivery system. A broader view of the approach is presented in *Pride and Joy* (Knight, 2014), a healthcare-centred version of *The Goal* (Goldratt, 1986). As with LM, some hospitals and NHS foundation trusts in the UK and the Netherlands have adopted TOC as an overall management approach, typically with the support of TOC expertise, as in the case of QFI Consulting, which developed applications used by two of the cases in this research.

As in manufacturing, TOC thinking in healthcare tended to be applied to more complex healthcare environments typified by A&E and more complex pathways in health and social care (Knight, 2003; Umble and Umble, 2006; Knight, 2011; Stratton and Knight, 2010; Knight, 2014; Stratton et al., 2014). For example, Umble and Umble (2006) outline the use of TOC TBM in three distinct hospital implementations. Significant gains were accomplished nearly instantly with each application using this practice in EDs and acute hospital admissions. The research of Stratton and Knight (2010) has also identified that TBM is a notion associated with TOC that further improves and enhances patient flow throughout planned and emergency health and social care.

According to Knight (2011), TOC is a methodology that has been applied by administrators, nurses, and doctors, because it fits with the problem-constraint system of healthcare. However, it has been argued by Tabish and Syed (2015) that TOC provides synchronisation in the current patients' flow, but it is not sufficient for maintaining trustworthy and robust patient-centred prioritisation. This is due to the fact that priorities do not only rely on updating the rate of a patient's recovery but also on the sound understanding of delays or disruptions (Tabish and Syed, 2015). While negating the argument of Tabish and Syed (2015), Amonge (2015) stated that, as per TOC, each system has a number of constraints, and systems are designed to provide a flow at times when patient-related activities are at their peak and these systems can further help to resolve the problems of emergency departments.

In the same manner, Umble and Umble (2006) and Stratton and Knight (2010) argue that, in the healthcare environment, the application of TOC enables healthcare systems to focus narrowly on managing and improving patient flow, thereby facilitating the integration of health and social care operations. Healthcare environments are more likely to provide unprecedented results for patients through high quality and timely care, for which TOC offers an appropriate methodology and framework (Amonge, 2015).

A review about the translation of the application of TOC to improve patient flow in different healthcare environments is now explored in the next section.

2.3.5. Translation of TOC into Healthcare

2.3.5.1. The Value of Flow

As with LM in healthcare, TOC approaches emphasise the importance of patient flow to ensure the market need for care services is met. The first, focusing step stresses the need to identify the constraint that prevents the system from reaching its goal. In the healthcare environment, time is critical for both patients and healthcare providers, and therefore the value of the flow can be reflected by ensuring that patients can receive care when they need it and, more importantly, that patients are discharged from the hospital when they are medically fit.

Young et al. (2004) and Womack et al. (2005) highlighted the key common factor for both healthcare and manufacturing is ensuring that ‘value’ is delivered to the customer and that all sources of ‘waste’ are removed, such as excess inventory, mistakes, waiting, inappropriate processes, and re-admissions. Within TOC applications in healthcare settings, the importance of setting a common target across the system by making patients a central priority of the system goal is an essential step in managing patient flow. According to Stratton and Knight (2010), establishing a planned discharge date (PDD) is the key to managing patient flow, as it places the patient at the centre of focus across the health and social care system by creating a patient-centred clinically led plan.

The term PDD is defined by Alex Knight (2014), who is concerned with the transformation of the TOC approach in healthcare settings, as a realistic estimation of the time patients need to recover and prepare for discharge to their next place of care. According to Stratton and Knight (2010), the PDD is a new term to many healthcare organisations; it is indicative at first but can be updated as dictated by knowledge and conditions. But today, the concept of PDD has become commonly used by many healthcare institutions (Knight, 2014). It can be related to similar terms, such as the expected, estimated, and anticipated date of discharge.

As previously stated with LM applications and in comparison with more stable healthcare settings, the establishment of a PDD is vital in complex or unstable healthcare environments such as acute, social care homes and community hospitals, where the demand and nature of clinical care are uncertain, to minimise the LOS of patients and to ensure a system-wide focus (Knight, 2014). By creating a plan for every patient, it becomes clear for everyone across the system that the required tasks or support activities be subordinated to the patient’s medical needs. As with TOC in manufacturing, establishing the plan needs a practical mechanism to encourage flow which is protected from variation and uncertainty to meet the goal of the system. How this PDD has been managed is now explored in the next section.

2.3.5.2. Mechanism for Managing Flow

As mentioned earlier, the technique of DBR-TBM was developed to manage the flow in unstable manufacturing environments where the production mix, process, and demand are unpredictable. In unstable healthcare environments that are characterised by uncertainty and variation in demand and process, the redesigning of medical pathways is a more difficult task, as they involve dependent events and statistical fluctuations.

> In healthcare the need for tailored pathways and highly variable procedures, as well as recovery times, adds considerably to this complexity (Stratton and Knight, 2010, p. 3).

With regard to the complexities represented in the healthcare environment, in terms of dynamic curing times, dependent incidents, and stochastic treatment plans, Siha (1999) indicates that the DBR approach may be valuable in this type of environment. Many scholars have documented the use of this approach as a practical mechanism for managing patient flow in this kind of setting (Knight, 2003; Umble and Umble, 2006; Stratton and Knight, 2010; Knight, 2011; Stratton et al., 2014; Tabish and Syed, 2015).

For example, the TBM approach has been introduced to reduce waiting times for both A&E and acute hospital admissions for A&E patients at three UK hospitals, including Milton Keynes District, Oxfordshire Horton, and Oxfordshire Radcliffe (Umble and Umble, 2006). Based on the technique of TBM, the processes of A&E and acute hospital admissions have been viewed as multi-project environments, by assigning a plan for each patient and considering each patient as a new project. According to Tabish and Syed (2015), the DBR is not relevant in this type of environment, as the arrival of patients cannot be scheduled, and, therefore, the buffer control mechanisms for project environments are used. From the time the patients arrive until they are discharged, the method of TBM allowed staff to systematically track every patient's progress until they were released from A&E or transferred to the acute hospital (Umble and Umble, 2006).

As with the TBM approach in manufacturing, the buffer is divided into three equal zonings, based on the A&E target of 4 hours. The three dimensions are divided into three equal 1 hour and 20-minute intervals. As illustrated by Umble and Umble (2006, p. 1068) in Figure 12 below, these three zonings are represented respectively as green (0 to 1hr &20m), amber (1hr

& 20m to 2hrs&40m), and red (2hrs&40m to 4hrs). Patients who passed the 4-hour target are placed in the black zone.

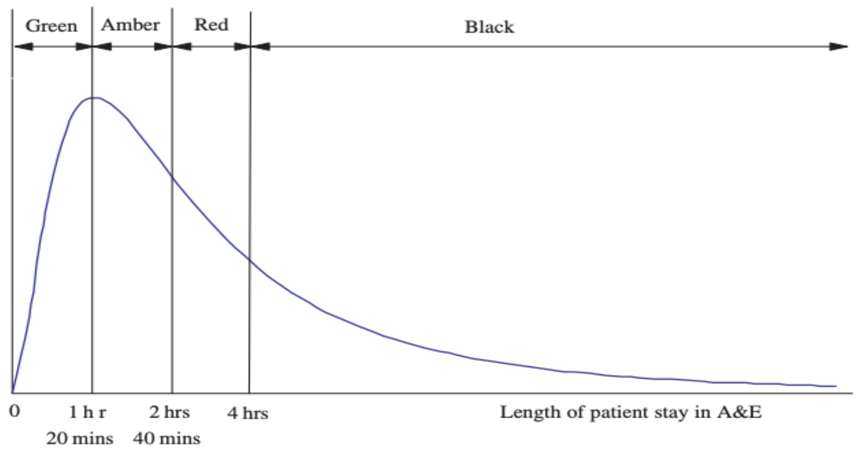

Figure 12: Buffer Zones for A&E Patients (Source: Umble and Umble, 2006, p. 1068)

The TBM is concerned with monitoring the progress of patients effectively once they are positioned in the amber zone, to avoid them heading to the red area. In a situation where there is a potential for a patient to be delayed and passed to the red or black zone, a signal through TBM notifies management to expedite and escalate, if needed, to avoid deterioration in system performance.

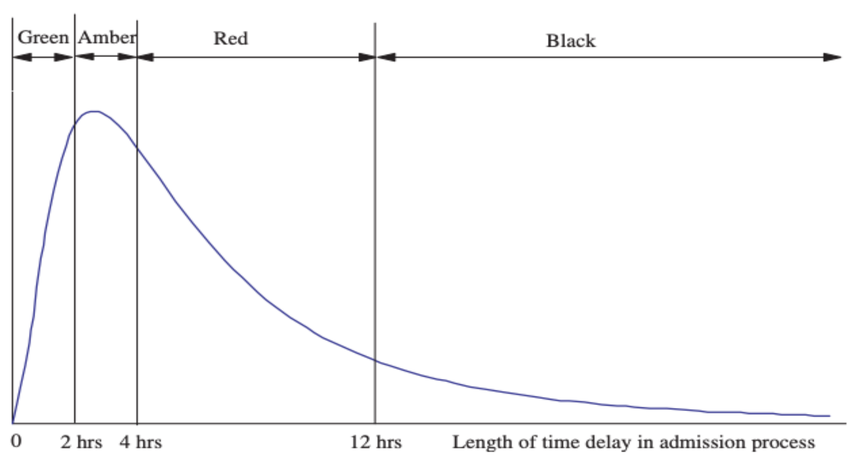

Figure 13: Buffer Zones for Acute Hospital Admission Delays (Source: Umble and Umble, 2006, p. 1069)

The same principle was applied to A&E patients who were admitted to an acute hospital using a different buffer to track patient progress through the process of acute hospital admission. In this case, the buffer time is structured differently, as seen in Figure 13. In this case, as Umble and Umble (2006) stated, because it is preferred for the duration not to extend 4 hours and it is required not to surpass 12 hours, the green, amber and red buffer zones involve unequal time intervals. The green, amber and red regions represent respectively time periods of 0–2 hours, 2–4 hours, and 4–12 hours. Where possible, every patient who enters the red area should be given high priority in the admission process.

Consistent with Umble and Umble (2006) in the evaluation of the TBM approach in A&E and acute hospitals, Stratton and Knight (2010) examined this approach at A&E, but they also explored the effectiveness of a TBM approach in applications to health and social care system where the PDD is uncertain. The authors stressed that TBM emerged as a key mechanism in all the applications of TOC, and this has recently been developed in order to address the needs of patient flow in the health and social care sector. To explore the use of the TBM approach in more varied and uncertain healthcare environments, we refer to Figure 14 to explore the concepts underpinning this approach in these environments. This exploration is based on the work of Stratton and Knight (2010), Knight (2014), and Stratton et al. (2014).

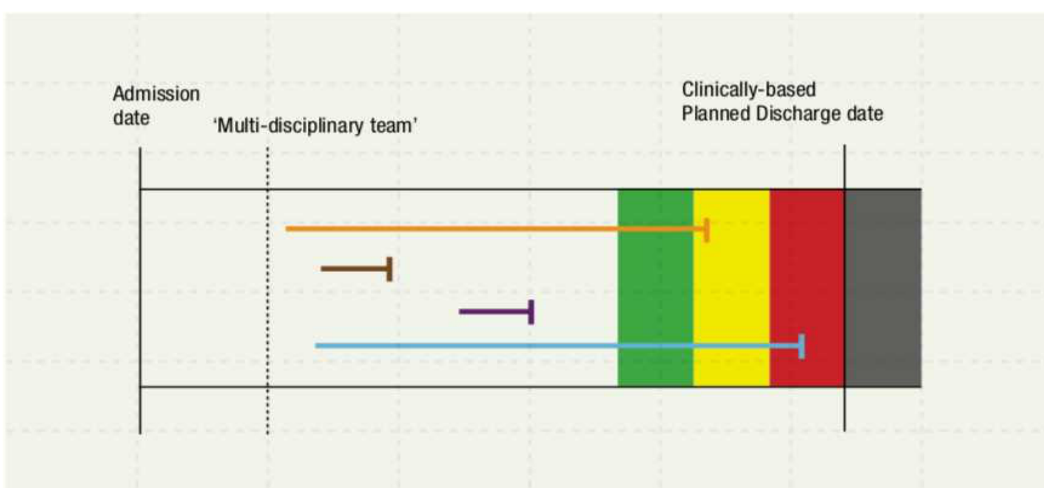

Figure 14: Time Buffer Management: Conceptual View Typically Used in Healthcare (Source: Knight, 2011, p.9)

The figure shows how the PDD for all patients is determined, based on their medical needs. Time goes from left to right. The PDD is to prioritise the tasks which need to be finished by this date to prevent delayed discharge.

This is the means of exploiting the constraint of time. Going beyond this date would result in a delay, typically measured today in the NHS by DTOC. The vertical dotted line represents the current time, signifying a review of this patient by a multi-disciplinary team (MDT). The purpose of the TBM mechanism is to ensure the PDD is met, thereby avoiding any DTOC. To do this, all the activities required for discharge need to be completed by this date, and these activities and their expected durations are represented by horizontal lines in the Figure, where each resource provider is represented by a different colour.

Figure 14 illustrates that the orange activity is in the yellow zone, indicating priority over brown and purple. The blue activity is in the red, which signals the need to expedite. The blue resource (e.g. occupational therapy (OT)) would see this patient on the aggregated worklist in colour order (R, Y, G), prioritizing patients based on their PDD, which is subject to change. A growing number of reds and blacks signals the need to escalate the growing instability of the system. Buffer management meetings are held daily and monthly to address short-term and longer-term improvement activity. TBM manages progress efficiently through four functions, as demonstrated by Stratton and Knight (2010) below:

1. Prioritization. Identify the comparative priority of patient tasks on the basis of the time allocated for planned discharge.
2. Expediting. These particular duties are expedited when they fall into the red area and are discussed specifically in regular buffer management meetings.
3. Escalating. If there is a rise over the standard boundaries of red zone duties, this is an escalation signal for exceptional management intervention, as the system is showing signs of immediate instability.
4. Targeting enhancement. Data on the causes of delays in the red area is continuously collected. These are regularly analysed and discussed at functional buffer management meetings through Pareto analysis, for direct enhancement activities.

2.3.5.3. Focus on Ongoing Improvement

With a predefined medical pathway, direct observation may be used to detect any disturbance that may occur in this pathway. Yet in a complex flow, where multiple resources are working through different pathways, it is difficult to identify the causes of the flow disruption due to the lack of visibility of the whole system. However, the TBM approach, through its four functions, offers a robust mechanism that helps to establish a process of ongoing improvement and improve the flow, as illustrated in the above section.

For example, by applying TBM in the processes of A&E and acute hospital admission across three UK hospitals, waiting time was substantially reduced in both processes, without additional capacity (Umble and Umble, 2006). As stated by one of the A&E managers:

> Buffer management has provided real time information for the first time. This has given us the impetus to change rather than just talking about it (A&E Manager, Umble and Umble, 2006, p. 1074).

The four functions of TBM contributed to the effective use of the system as they supported the management to identify the causes of delays in the red zone, so that improvement could be made. As expressed by an A&E clinical director:

> (Buffer management) has helped those outside the A&E understand that patient waits are not caused by the department but rather constraints within the whole healthcare system. The (positive) effect on A&E staff morale cannot be overestimated (An A&E clinical director, Knight, 2003, as cited in Umble and Umble, 2006, p. 1074).

The introduction of TBM systems for A&E and Discharge across four hospitals in the UK has demonstrated a substantial and dramatic improvement, reducing patients' LOS by more than 20% and dramatically improving A&E performance (Stratton and Knight, 2010). The implementation of the TBM approach was managed by QFI advisory services and applied in the form of a piece of software called QFI Jonah Software. This software allowed any causes of delay to be recorded and presented in a Pareto format analysis, as displayed in Figure 15. These causes can be then addressed via routine buffer management meetings for direct enhancement activities.

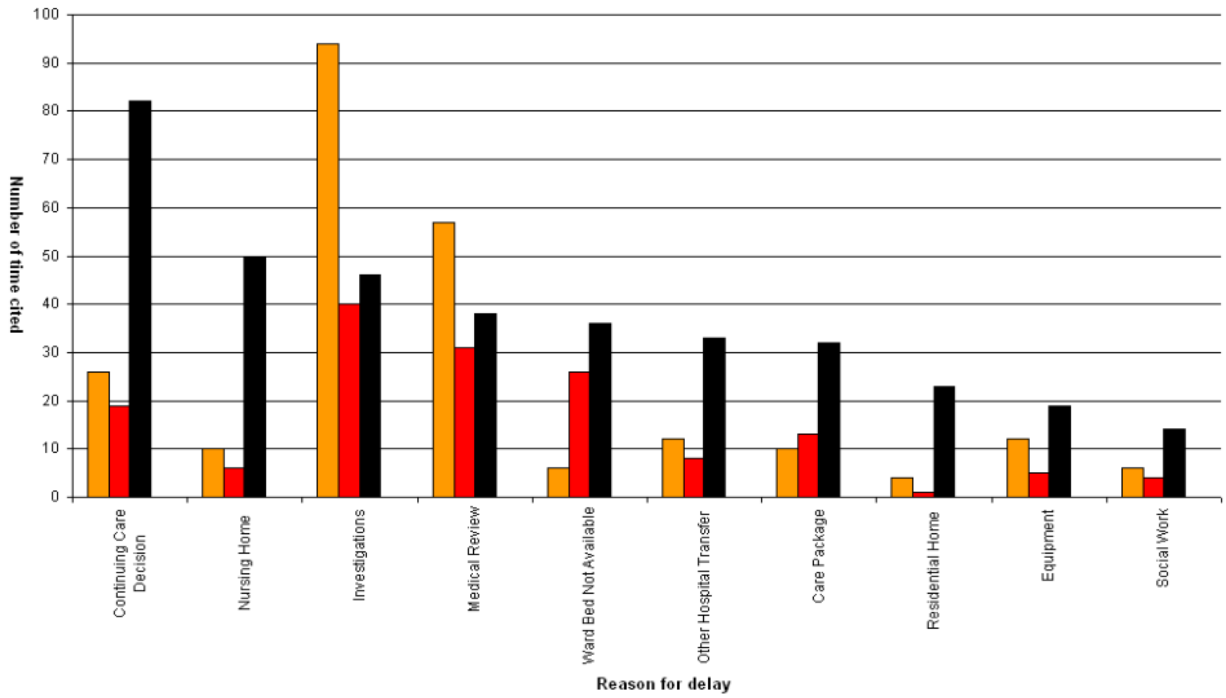

Figure 15: Discharge QFI Jonah Top Delay Reasons by Region Presented to Cross-buffer Meetings (Source: Stratton and Knight, 2010, p. 10)

Although the application of TOC is less widespread, it has enabled a hospital in the UK, within a few months of implementation, to rise from the lower 10 to the higher 10 among 500 hospitals rated according to emergency service time (Knight, 2011). Gupta and Kline (2008) have also reported an improvement in lowering appointments for psychiatric assessment from 43% to 20%, as a part of the application of a TOC continuous improvement process in the Chemical Dependency department in the community centre for mental health. The use of TOC in a large public hospital has shown significantly shortened patient waiting times, improved patient satisfaction, increased performance, smoother workflow, and enhanced morale and retention for the staff, while retaining patient integrity and safety and remaining within specified cost criteria (Mabin et al., 2018).

A recent review of the application of TOC in healthcare was carried out by Tabish and Syed (2015), referring to a study by Lubitsh et al. (2005), which examined the influence of TOC on three NHS trust units: Eyes, ENT and Neurosurgery. In general, the focus was on reducing system waiting lists and improving patient flow performance. This study has shown that the use of TOC had a positive effect on waiting lists and patient throughput in Eyes and ENT, where the medical procedure is standard. However, in neurosurgery, where the current circumstances are more complex and life-threatening emergencies are prevalent, the TOC did not have a positive effect (Tabish and Syed, 2015).

2.3.5.4. Local Optimisation Inhibits Flow

In a complex healthcare system, a lack of strategic alignment between the relevant parts of the system can lead to an increase in localising control practices. As with manufacturing, the application of a TBM approach provides healthcare management with a practical mechanism to avoid focusing on local efficiency measures. This approach enables hospital management to track the progress of patients' journeys through the system. By creating a common goal that is clear to all those involved in day-to-day care activities, and by subordinating everything else to the needs and priority of the system, the risk of localisation practices can be avoided.

A study by Stratton et al. (2014) on community therapy services revealed how the TBM strategy moved the emphasis from decentralised practice, where each therapy was responsible for managing their own queue, to a situation where it was managed by a centralised queue. Taking advantage of the TBM functions to manage one aggregated buffer list made strategic alignment possible among occupational therapists and local service providers, thereby allowing all patients in each location to be given dynamic priority to meet their needs. TBM strategies like daily huddles for a quick patient flow evaluation and measures to preserve providers' time can also maintain successful utilisation of the provider (Cox et al., 2014). According to the authors, the application of TBM challenges traditional measures of organisational efficiency and effectiveness, where all employees try to work at their best, minimising their time for every job, as though they do not have an impact on other practices. Thus, the TBM approach systematically ensures that everybody works together in favour of the goal of the system, eliminating disharmony and leading to a continual improvement culture.

Although the PDD concept is a fundamental part of any scheduling and control system (Stratton and Knight, 2010) and the TBM approach has been applied to manage the PDD by administrators, nurses, and doctors, because it fits with the problem-constraint system of healthcare (Knight (2011), it has been argued by Tabish and Syed (2015) that, although TOC provides synchronisation in the current patients' flow around an initial PDD, it is not sufficient to maintain trustworthy and robust patient-centred prioritisation. According to the authors, priorities should be focused not only on reviewing the current patient recovery rates, but also on understanding disturbances or delays. Hectic goals trigger system-wide chaos and lead to clinical and management personnel moving from disaster to disaster. Even if patient flows are synchronised, there can still be chaos in a priority system.

Overall, TOC has been used in a variety of healthcare settings where positive results have been demonstrated. The TBM approach appears to be very encouraging for many of the major healthcare processes (Umble and Umble, 2006; Stratton and Knight, 2010). However, there is limited research evidence on how sustainable these applications have been (Stratton and Knight, 2010) and how to represent the optimal benefits that can be achieved by using TOC in the healthcare system (Tabish and Syed, 2015). Research on the application of TBM principles in a variety of healthcare settings, with or without the DBR technique, is also limited (Umble and Umble, 2006). Table 6 summaries the translation of TOC into healthcare.

Table 6: The Translation of TOC into Healthcare

Translation TOC in Healthcare	
The value of flow	Focusing on meeting patient needs by setting a PDD for each patient and making the patient a central priority of the system goal (Stratton and Knight, 2010; Knight, 2014)
Mechanism for managing flow	Using the TBM approach and its functions to manage patient flow and ensure the PDD is met (Knight, 2003; Umble and Umble, 2006; Stratton and Knight, 2010; Knight, 2011; Stratton et al., 2014)
Focus on ongoing improvement	The four functions of TBM allowed management to identify the causes of delays (e.g. red zone penetration), and these causes can be then targeted (Umble and Umble, 2006; Stratton and Knight, 2010; Knight, 2011)
Inhibit localised measures	The functions of the TBM enabled resources and tasks to be subordinated to the PDDs and the system's priority (Umble and Umble, 2006; Stratton and Knight, 2010; Knight, 2011)

2.3.6. Current Approaches Applied in the NHS

2.3.6.1. Red2Green

Other approaches have been adopted in healthcare, for example, Red and Green Bed Days. This approach is a visual management system that supports the identification of wasted time in a patient's journey and has been applied in both acute and community settings to minimise internal and external delays and deliver improved patient flow (NHS England, 2019). Red days are when patients receive little or no value-adding care, while green days are when patients receive value-adding care, advancing their path towards discharge.

The red and green days process uses ward improvement boards to link flow, safety, and reliability through visual demonstration (NHS Improvement, 2017). Using ward level metrics, the process identifies constraints for wards to convert red days to green days and proactively manages the constraints at a board round. Where improvement solutions are not immediately available, this results in an escalation process where processes are invoked to seek to actively manage the constraints. Where teams have failed to resolve constraints, the top five constraints are reviewed by senior operational managers and included in local improvement plans (NHS Improvement, 2017).

2.3.6.2. SAFER Patient Flow Bundle

The SAFER patient flow bundle aims to promote engagement and continuous improvement by combining five elements of best practice: senior review; expected discharge date and clinical criteria for discharge for all patients; flow of patients at the earliest opportunity; early discharge; and senior review for patients for management and discharge decisions (NHS Improvement, 2017). The SAFER patient flow bundle works more effectively when implemented with the Red2Green days approach and when all five elements are implemented to obtain cumulative benefits (NHS Improvement, 2017).

2.3.7. Conclusion

This section explored the translation of LM and TOC into healthcare settings. The existing literature highlights how these approaches have dealt with the increase in variability that compromises patient flow efficiency. With the use of LM tools to redesign medical pathways, significant improvements in terms of reducing the LOS of patients, increasing productivity, enhancing quality and safety, and improving the overall performance of healthcare organisation were documented. However, some research points to a number of failures in LM applications in relation to system-wide improvement, patient satisfaction, performance sustainability, and the restriction of LM to only some of its elements, rather than the application of the LM concept across the whole system.

The literature on the applications of TOC highlights that the TBM approach has been shown to be a robust mechanism to manage instability that threatens patient flow performance. As with TOC in manufacturing, using the TBM approach and its four functions demonstrated a significant improvement in patient flow across the system, establishing a continual improvement process. Although this approach has shown promising results and supporting its users in healthcare, it is still early to judge to what extent the results are sustainable.

2.4. Summary of the Chapter and Justification of Research

2.4.1. Summary of the Chapter

Healthcare is an extensive and complex system with significant challenges in terms of increasing DTOC, patients with complex needs, an ageing population, and demographic changes which require improved efficiency and effectiveness. Health and social care systems are highly interrelated, unique, and characterised by uncertain demand and complex pathways, as the needs of patients differ. Issues of an inherent silo culture and instability in healthcare services have demonstrated the need to manage patient flow across the system to meet the needs of patients and avoid localised control measures (Knight, 2000; Crawford-Mason, 2002; Umble and Umble, 2006; Kämäräinen et al., 2016). Healthcare organisations have recognised that focusing on flow and reducing variability is key to improving the healthcare delivery system, which led these organisations to take greater interest in system based approaches from manufacturing, such as LM and TOC (Olsson and Aronsson, 2015).

LM and TOC are distinct system-based approaches that originated to address issues of variability to enable flow and enhance the overall system performance (Antunes, 1998). Both approaches have techniques that can address the issue of instability and enable synchronisation across the whole system. TOC has a DBR-TBM flow control mechanism (Umble and Srikanth, 1997; Schragenheim and Dettmer, 2001; Stratton et al., 2008), while LM follows the Kanban mechanism (Antunes, 1998; Bhasin and Burcher, 2006; Goldratt, 2009; Stratton and Knight, 2010), but they both work under different circumstances in manufacturing environments.

An early review of the conditions associated with LM and TOC approaches found that the LM Kanban approach to TPS is commonly associated with predefined process steps, time, and transfer paths, with relatively low fluctuations in demand and uncertainty (Hall, 1981; Monden, 1983; Ohno, 1988; Spearman et al., 1990; Spear and Bowen, 1999). While the LM approach is more akin to line environments, the TOC TBM approach has been developed for an unstable flow environment (jobbing) where process times, product mix, and demand are subject to high levels of variation and uncertainty (Chakravorty and Atwater, 1996; Dettmer, 2001; Goldratt 2009; Stratton and Knight, 2010; Pacheco, 2018).

The literature on the translation of LM and TOC in healthcare identified that the majority of LM applications involved the adoption of a structured approach to redesigning medical pathways (Spear, 2005; Ben-Tovim et al., 2007; Fillingham,2007; Grunden, 2007; O'Connell et al., 2008; Gubb, 2009; Kenney, 2011; Blackmore et al., 2011; Radnor et al., 2012; Van Calster et al., 2019; Murphy et al., 2019), while the TBM approach associated with TOC was applied to more complex healthcare environments with high levels of instability in terms of patient needs and demand (Knight, 2003; Umble and Umble, 2006; Stratton and Knight, 2010; Knight, 2011; Stratton et al., 2014).

In reviewing the results of these approaches in improving patient flow, there have been mixed outcomes in terms of their success and the assumptions underlying the application of these approaches to healthcare (Blackmore and Kaplan, 2016: Radnor et al., 2012; Stratton and Knight, 2010). While principles of LM have been widely applied and improvements have been documented in certain prominent case studies such as the Virginia Mason Medical Centre, Royal Bolton Hospital, Flinders Medical Centre, and the Pittsburgh General Hospital (Spear, 2005; Ben-Tovim et al., 2007; Fillingham, 2007; Grunden, 2007; O'Connell et al. , 2008; Gubb, 2009; Kenney, 2011; Blackmore et al., 2011; Radnor et al., 2012), the majority of literature considers LM applications in healthcare to be disjointed and not sustained (Radnor et al., 2012; Hallam and Contreras, 2018), promoting local rather than widespread improvement (Young and McClean, 2008; Brandao de Souza, 2009; Parkhi, 2019) and having a limited impact on patient flow (Young and McClean, 2008; Bhasin, 2008; de Vries and Huijsman, 2011; Radnor et al., 2012; Poksinska et al., 2016). Some studies claim that most LM implementations in healthcare have not achieved the desired results, due to poor adoption of LM practices and the restriction of LM to the use of particular tools and techniques (Kim et al., 2006; Fillingham, 2007; Radnor et al., 2012; Steed, 2012; Dannapfel et al., 2014; Kaplan et al., 2014; Holden et al., 2015; Antierens et al., 2018).

Despite less extensive of the applications of TOC in healthcare, there is evidence to indicate that the TOC approach appears to be very promising in managing and improving patient flow across the system (Knight, 2003; Umble and Umble, 2006; Gupta and Kline, 2008; Stratton and Knight, 2010; Knight, 2011; Stratton et al., 2014; Cox et al., 2014; Tabish and Syed, 2015; Mabin et al., 2018). Nevertheless, there is insufficient evidence of the sustainability of improvement and the benefits of adopting this approach to a variety of healthcare settings (Umble and Umble, 2006; Stratton et al., 2014; Tabish and Syed, 2015).

2.4.2. Focus and Justification of Research

Both manufacturing and healthcare environments are dealing with different levels of instability, variability, and uncertainty. While the drive for more customisation, either in manufacturing or healthcare, demonstrates the complexity of the pathways, as each patient has different needs or each customer requires different products, the drive to standardise a specific pathway with dedicated resources indicates that flow operations in these environments are less complex and more stable. The need to acknowledge instability in manufacturing resulted in different flow mechanisms, each of which originated to meet the requirements of different environments.

The review of the origins of the system-based approaches highlighted that the LM approach tends to control all aspects of the system, by establishing clearly defined flow paths and using the Kanban system as a flow mechanism to manage the flow and target variability across the system (Hall, 1981; Monden, 1983; Ohno, 1988; Spear and Bowen, 1999; Feld, 2000). The establishment of clearly defined pathways was also evident in the literature on the applications of LM in healthcare, where patients' needs are less complex (Spear, 2005; Ben-Tovim et al., 2007; Fillingham, 2007; Grunden, 2007; O'Connell et al., 2008; Gubb, 2009; Kenney, 2011; Blackmore et al., 2011; Radnor et al., 2012; Van Calster et al., 2019; Murphy et al., 2019).

By contrast, TOC focuses primarily more on the critical points of the system by setting new rules, using the TBM approach as a practical mechanism to manage variability and uncertainty and ensure the flow is improved and sustained (Goldratt and Cox, 1984; Umble and Srikanth, 1997; Schragenheim and Dettmer, 2000). While the TBM approach originated as a practical mechanism for managing instability and improving flow in unstable manufacturing environments that are more akin to jobbing (Chakravorty and Atwater, 1996; Dettmer, 2001; Goldratt 2009; Stratton and Knight, 2010; Pacheco, 2018), there seems to be a pattern in the translation of this approach to managing patient flow in complex healthcare environments (Knight, 2003; Umble and Umble, 2006; Stratton and Knight, 2010; Knight, 2011; Stratton et al., 2014). Through the use of an aggregated buffer (rope) linked to the drum (market demand) and zoning associated with the TBM approach, variability and uncertainty can be handled proactively in this type of environment (Goldratt and Cox, 1984; Umble and Srikanth, 1997; Schragenheim and Dettmer, 2000; Schragenheim, 2010).

Within healthcare, there is a tendency to adopt one or other approaches, irrespective of the healthcare environment with little consideration of how these manufacturing approaches might be theoretically developed to meet the needs of different health and social care environments. Although many of these interventions have improved local performance, this does not always positively impact the overall system nor is it sustained over time. Therefore, where LM and TOC have been applied, there have been mixed results regarding the success and assumptions underlying the application of such approaches to healthcare (Stratton and Knight, 2010; Radnor et al., 2012; Blackmore and Kaplan, 2016).

No research has attempted to explore the circumstances associated with the success of different approaches to production system flow and how these approaches can be used selectively and effectively in various healthcare settings, with particular reference to the primary patient flow with a wide mix of instability. Empirical research that offers an in-depth understanding of the circumstances under which the application of LM and TOC is more effective in managing and improving patient flow remains limited (Brandao de Souza, 2009; Stratton and Knight, 2010; de Vries and Huijsman, 2011; Radnor et al., 2012; Andersen, 2015; Tabish and Syed, 2015; Blackmore and Kaplan, 2016; Poksinska et al., 2016; Rees and Gauld, 2017; Parkhi, 2019). Therefore, this research aims to gain a deeper academic understanding of how established flow management approaches have been developed to meet the distinct conditions within the wider healthcare system, with specific reference to the origins of LM and TOC. Three questions provide focus for this largely exploratory case study research:

1. How can LM and TOC approaches be applied effectively to improve patient flow across health and social care?

2. Why do LM and TOC approaches work better in different health and social care environments?

3. What are the assumptions underpinning the effective use of Kanban and TBM in different healthcare environments?

CHAPTER THREE: RESEARCH METHODOLOGY

This chapter outlines the justification for applying particular methods and techniques for defining, collecting and analysing knowledge to answer the research questions, thus enabling readers to critically assess the validity and reliability of the study as a whole. The literature review in Chapter 2 identified three research questions as follows:

1. How can LM and TOC approaches be applied effectively to improve patient flow across health and social care?
2. Why do LM and TOC approaches work better in different health and social care environments?
3. What are the assumptions underpinning the effective use of Kanban and TBM in different healthcare environments?

The researcher believes that these questions are inductive in nature, making the case-study enquiry method a suitable approach for answering them, as a part of an interpretivist framework.

Informed by the research questions which have guided the design of the research, this chapter aims to outline that research design and the approach underlying the thesis. It begins with an overview of research philosophies, before focusing on the philosophy that is most relevant to the objectives of this study. This overview establishes the rationale for following the interpretive approach and for using the case study strategy. The chapter then discusses the use of the case study research method in detail and explores the research design framework. The research design for this study follows Eisenhardt's (1989) eight-step framework. The justification for the use of this framework and how it is implemented by the researcher are discussed further in this chapter, including the research tools, data collection, and analysis. Lastly, the chapter discusses the value of this study in terms of its validity and reliability.

3.1. Research Philosophy

This section presents a brief discussion of research philosophies. According to Easterby-Smith et al. (2015), and Jogulu and Pansiri (2011), understanding different research philosophies helps to select an appropriate research design that meets the specific objectives of the research. Positivism and interpretivism are two philosophical traditions, each one taking a different approach to developing knowledge. Cameron and Molina-Azorin (2011) and Easterby-Smith et al. (2015) defined the distinction between these two models as the difference between an ontological assumption and epistemological assumptions. An ontological assumption refers to the way reality is viewed by the researcher; an epistemological assumption refers to the way knowledge is produced.

In the case of the positivist philosophy, knowledge should be collected through measurable and observable facts: 'key properties should be measured through objective methods, rather than being inferred subjectively through sensation, reflection or intuition' (Easterby-Smith et al., 1991, p.22). Positivists believe that truth is stable and can be objectively observed and described (Levin, 1988). Positivism is followed by scientific researchers who work with empirical data and who study quantifiable areas of research (Carson et al., 2001). In a positivist approach, the researcher is considered to be completely independent and objective (Easterby-Smith et al., 1991). The positivists believe that a researcher in social science is isolated from the phenomena being studied and that the research must be free of value. It has been argued that the idea of separating the researchers from their research cannot be applied in practice, as it is challenging to avoid interference from principles or interests while witnessing something (Somekh and Lewin, 2005).

Under the interpretive philosophy, knowledge is obtained by understanding 'the world of human experience' (Cohen and Manion, 1994, p. 36). The interpretivist approach follows Max Weber's school of thought, who argued that the concern of human sciences lay in understanding processes rather than explaining facts (Hughes, 2003). This approach is based on the idea of understanding social phenomena by obtaining the viewpoints of individuals who are involved in them. Scientists should 'understand, explain, and demystify social reality through the eyes of different participants' (Cohen et al., 2007, p.19). Here, meaning can be different from one person to another ', but the truth is a consensus formed by co-constructors' (Crotty, 1998, p.9; Pring, 2000, p. 251). The investigator must look at the various aspects in a specific number of situations, rather than the entire of society, since each circumstance may be various from another. This is clearly appropriate to the object of study for the present research project, where the complex reality of managing patient flow in healthcare contexts is dynamic and constructed subjectively by the social actors involved, rather than being a fixed objective entity. Failure to generalise is one of the limitations of this approach: there is a tendency for interpretivism to investigate a small sample which is not adequate to deduct generalisation about the entire society (Hammersley, 2012). Other scholars have, however, argued that by focusing in depth on a small number of cases, there is potential to gather detailed understanding of a specific event and that the researcher can analyse multiple perspectives on a single subject (Macdonald et al., 2002). Table 7 shows the different assumptions associated with these two distinct philosophies.

Table 7: Comparison of Positivism and Interpretivism (Source: Saunders et al., 2012, p. 136)

Assumption	Positivism	Interpretivism
Ontology	External, objective and independent of social actors	Socially constructed, subjective, may change, multiple
Epistemology	Only observable phenomena can provide valid evidence. Emphasis on causality and law as generalisations, limiting phenomena to the simplest aspects	Subjective meanings and social phenomena. Focus on situation details, a fact behind these details, subjective meanings driving acts.
Axiology	Research is performed in a value-free way, the researcher is independent of the data and has an impartial role	Research is value-based, the researcher is part of the study, cannot be isolated and will be subjective
Commonly used forms of data collection	Highly structured, large samples, measurement, quantitative, but can use qualitative methods	Small samples, in-depth investigations, qualitative

3.2. Justification for Following the Interpretive Approach

As shown in Table 7 in the previous section, each research philosophy has certain assumptions, techniques, methods, limitations, and different ways to assess the quality of the resulting study. Therefore, researchers must consider the essence of their preferred philosophy, select the philosophy in which they work, and record their philosophical preference in writing (De Vos et al., 2011). As is evident in Chapter 2, there are conflicting results, and no single reality concerning the success and assumptions underlying the application of LM and TOC approaches to healthcare. For that reason, this study follows an interpretive approach, aiming to interpret the translation of these approaches to meet the needs of different healthcare environments through the experiences and perceptions of the participants. Understanding the influence of LM and TOC in managing patient flow requires the perspectives of those who experienced the interventions to improve the process. Based on the nature of the proposed research questions, the theoretical samples in this study were carefully selected and not chosen at random. This study is, therefore, best suited to a qualitative interpretative approach, which can provide a rich understanding of the core questions by narrowing the distance between the researcher and the key informants involved in a particular area of concern. In turn, this creates theoretical and practical understanding and generates new, alternative theories and concepts (Bygrave, 1989). Through an interpretive approach, the researcher can interpret social reality through the subjective perspectives of the participants embedded in the context in which reality is situated. This approach enables the researchers to use these perspectives to construct their understanding and interpretation of the data collected.

> Interpretivist researchers view the world through a series of individuals' eyes……. People have their own interpretations of reality, and interpretivists choose methods that encompass this worldview (McQueen, 2002, p. 55).

In interpretivism, the investigator can participate in the study, so that the researcher fully maintains the connection between himself and the subject of investigation, whether that be respondents or research objects (Giddings and Grant, 2007). Considering that the research is founded upon the values of the researcher, Mertens (1998) argues that there should be no independence between researchers and their investigation. From an interpretivist perspective, there is potential for all those involved in the investigation to offer unique contributions, the researcher included (Mackenzie and Knipe, 2006). To understand fully the reality surrounding a phenomenon under investigation, the researcher should consider detail in-depth (Easterby-Smith, Thorpe and Jackson, 1991). It is in this vein that Blumer (1986) argued the need for researchers to go beyond social actors' interpretation of the phenomenon and analyse individual interactions. Additionally, there is a need for interpretivist researchers to understand facts from the experiences of participants instead of trying to understand them objectively and independently of the context. It is the interaction of the researcher with the participants that allows the researcher to identify individual constructs (Guba and Lincon, 1994, p. 111). However, interpretivism has been subjected to widespread criticism for its lack of objectivity (Burrell and Morgan, 2017). The presentation of truth in a subjective way, and the creation of different meanings by each person, give rise to the possibility of participants failing to reach full agreement with the researcher in the way that truth is interpreted (Rolfe, 2006 p. 305). Mack (2010), however, has argued that the analysis of data from an interpretivist perspective represents an objective approach, given the interpretivist's sincere attention to the participants' assumptions in understanding situations rather than their reliance on their previous views. Consequently, although recognising the researcher's involvement in this investigation, interpretation was based on the participants' standpoint, rather than that of the researcher. The researcher does not have any initial hypothesis that has the potential to influence the views of respondents but rather engages with the study only to encourage respondents to offer their own perceptions and assumptions.

Both inductive and deductive steps are integrated into the building and refinement of theory. The relative importance of deduction and induction in the building of the theory is determined by choice of research methods, the extent of the structure adopted in the study, the data which the study generates, and the interpretation generated from the investigation (Yin, 2014). In operations management, there has been a consensus that interpretive methods are suitable for the generation or extension of theory, while rationalist approaches will be more suited for testing or verifying theories (Voss et al., 2002). Following the interpretative approach and using qualitative studies enhance the researcher's depth of understanding, which comes from the presence of several stakeholder viewpoints and thereby presents a justification for the performance of the healthcare systems (Ozawa and Pongpirul, 2013).

3.3. Matching the Purpose of Research with Methods

Operations Management (OM) has developed over the past 30 years as an academic field, as can be seen in the growing number of scholarly journals and publications in the area (Walker et al., 2015). While attitudes to OM studies and related strategies remain to have their origins squarely in positivist school (Pannirselvam et al., 1999), there has been growing attention in, and exposure to, qualitative research in OM, including some thematic issues in journals, has been noted (Kiridena and Fitzgerald, 2006). Qualitative research has an important role to play in strengthening the empirical foundation of OM. According to Soltani et al. (2014), qualitative analysis is vital to improving the theoretical basis of OM, for two key reasons: first, 'OM is applied in its perspective' (Singhal et al., 2008, p.345); and second, 'the study of OM is a social science' (Boyer and Swink, 2008, p.339). The advancement of OM research is increasingly recognised as an overarching criterion for high-quality work involving the researcher's active participation in field-based research (Meredith, 1998: Boyer and Swink, 2008; Bluhm et al., 2011). Westbrook (1995) suggested that OM academics should engage in a productive conflict between theory and experience and that the creation of new theories should be established through the analysis of the real situation. The development of theory remains the most fruitful area of research in OM (Westbrook, 1995; Pannirselvam et al., 1999). A variety of efforts have been made to establish and propose OM theories and practices. These include, for example, the trade-off principle (Skinner, 1969), the process-product matrix (Hayes and Wheelwright, 1979), the TOC (Goldratt and Cox, 1984), cumulative theory (Ferdows and DeMeyer, 1990), and the theory of performance frontiers (Schmenner and Swink, 1998).

Often theories emerge from a new concept or analogy, leading to the creation of a theoretical framework that aims to clarify the subject more effectively. Bacharach (1989) defined a theory as 'a statement of relations among concepts within a set of boundary assumptions and constraints. It is no more than a linguistic device used to organise a complex empirical world' (p. 496). The theory is also defined by Gioia and Pitre (1990) as 'a coherent description, explanation, and representation of observed or experienced phenomena' (p. 587).

According to Wacker (1998), three specific critiques of the theory are of concern to academics: theory does not need to be applied (Shubik, 1987); it does not contribute substantial improvement in the real world (Lindblom, 1987); and theory does not exist due to the absence of a definition of measurement (Churchman, 1961). Wacker (1998) argued that these concerns are based on the concept of theory and, more significantly, on the parameters used to establish 'good' theory. The author suggested three reasons why practitioners and researchers consider theory essential: (a) it provides an analytical structure; (b) it provides an efficient method for developing the field; and (c) it provides clear explanations for the pragmatic universe. Wacker (1998) gives the example of how a 'theory of internationally competitive manufacturing' would provide a framework for evaluating the international competitiveness of manufacturing industries, allowing researchers to set out the exact criteria for classifying companies in degrees of global competitiveness. Indeed, as some academics claim, ‘there is nothing as practical as a good theory’ (Lewin, 1945; Van de Ven, 1989).

> Good theory is practical precisely because it advances knowledge in a scientific discipline, guides research toward crucial questions, and enlightens the profession of management (Van de Ven, 1989, p. 486).

Within the OM field, the case study approach has played a prominent role and has been commonly used as a methodological approach to build theory, when compared against other qualitative approaches, for example, phenomenology and ethnography. Meredith (1998) argued that ‘the natural emphasis of the case study approach on understanding is clearly most directly focused on theory building’ (p. 445). When case studies are used in hybrid forms, this can give the case study approach the potential to obtain even better results: for instance, using them in a grounded theory approach, combining a longitudinal case with action research, or combining multiple reflective cases with a longitudinal case (Kiridena, 2005; Leonard-Barton, 1990; Rytter et al., 2005).

When combined together, the fundamental attributes of case study research make the method a notable competitor in OM theory-building research as opposed to more rational, abstract, and restricted approaches that are detached from the subjects of the study, for example, quantitative modelling, simulations, and questionnaires. While the former is more suitable for developing theory, the latter is better suited to the validation of theory or to testing theory. Wacker (1998) differentiates between analytical conceptual methods and case study methods: ‘...the key difference… is that the empirical case study method uses data to form the theory, and the analytical, conceptual method uses deduction to form theories’ (p. 375). However, case studies, according to Voss et al., (2002), may be used for various research purposes, as illustrated in Table 8 below.

Table 8: Matching Research Purpose with Methodology (Source: Voss et al., 2002, p. 198)

Purpose	Research question	Research structure
Exploration Uncover areas for research and theory development	Is there something interesting enough to justify research?	In-depth case studies. Unfocused, longitudinal field study
Theory building Identify/describe key variables Identify linkages between variables Identify ``why'' these relationships exist	What are the key variables? What are the patterns or linkages between variables? Why should these relationships exist?	Few focused case studies In-depth field studies Multi-site case studies Best-in-class case studies
Theory testing Test the theories developed in the previous stages Predict future outcomes	Are the theories we have generated able to survive the test of empirical data? Did we get the behaviour that was predicted by the theory, or did we observe another unanticipated behaviour?	Experiment Quasi-experiment Multiple case studies Large-scale sample of the population
Theory extension/refinement To better structure the theories in light of the observed results	How generalisable is the theory? Where does the theory apply?	Experiment Quasi-experiment Case studies Large-scale sample of the population

3.4. Rationale for Using a Case Study Approach

The need for case studies emerges from the desire to understand complex social phenomena in the real world (Yin, 2003). Based on the nature of the proposed research questions and the desire to perform an empirical investigation aimed at explaining a contemporary management issue, in the form of improving patient flow in complex healthcare settings, which necessitates a range of sources of evidence, including interviews, observations, and documentary sources, the case study was chosen as the research strategy for this study (Yin, 2003), this is discussed further in pages 136 and 147. Because case study research methods allow researchers to maintain the general and meaningful characteristics of real-life events, such as their organisational structures and processes, case studies have emerged as the preferred strategy for answering how and why patient flow can be effectively improved, where the researcher has little control over events and where the research has an emphasis on the phenomenon of delayed discharge within a health and social care setting (Yin, 2003; Mills et al., 2017).

In case studies, as opposed to experiments in an artificially created context in which the researcher has control over many variables, the case is typically something that already occurs as the source of research, and the researcher has no control over these real events (Yin, 2003). Although case studies may use quantitative data, the key distinction from other methodological approaches is that case studies tend to investigate events in their context rather than on a context-independent basis (e.g. Pettigrew, 1973). The case study is referred to as a triangulated research approach (Eisenhardt, 1989; Stake, 1998; Yin, 1994), as it permits the use of several data collection techniques and evidence sources to investigate a particular phenomenon.

By being applied through a variety of methodological approaches, case study research has developed substantially in the last four decades, growing in reputation as a suitable methodology for the investigation of complex issues in the real-world environment (Mills et al., 2017). The emergence of grounded theory methodology, merging quantitative data analysis methods with qualitative field study methods, stimulated the use of case study research designs in several disciplines, leading to an inductive methodology that analysed data using detailed systematic procedures (Glaser and Strauss, 1967; Johansson, 2003; Anthony and Jack, 2009; Merriam, 2009).

In operations management (OM), and for the development of new theory, case study research has consistently proved to be one of the most powerful research methods (Voss et al., 2002). Field-based approaches to answering research questions are particularly important at a time when OM is going through substantial and significant change, in terms of both technology and management (Lewis, 1998). As the case study method exposes the researcher to the real-life problem, there is opportunity for the development of new and innovative insights, while developing theory in a way that offers the highest degree of validity for the users of the research (Voss et al., 2002). The study is exploratory in nature, helping the researcher address why questions, and therefore providing a deeper understanding of the particular cases under review (Meredith, 1998).

Multiple case study methods produce cumulative knowledge which can be best presented through theory (Kiridena, and Fitzgerald, 2006). The theory then provides explanatory statements regarding the phenomenon, event, or behaviour under investigation. In turn, the emerging theory is presented in models or frameworks that support generalisations (Little, 1992). The explanatory power provided by theory gives the theory its capacity to support predictions, unlike other forms of scientific enquiry theory (Kiridena, and Fitzgerald, 2006). Issues of identification, explanation, and understanding within the case study phenomena – that is the what, how, and why of the study constructs – require detailed exploration (Meredith, 1998). Hudson and Ozanne (1988) see the goal of explanation in qualitative enquiry as the facilitation of predictions, while Bacharach (1989) argues that prediction and explanation are different constructs, with explanation intended to establish meaning and prediction being instrumental in testing forecasted outcomes, thereby validating the theory. Within OM, qualitative research has been seen as diverse, often incorporating a variety of data collection and analysis methods. In addition to the multiple sources and types of data used and reporting methods followed in qualitative studies, several scholars in the field have widely acknowledged that the qualitative research discipline shares a number of distinctive design elements, including its methodological tradition, which has a multi-method focus and a preference for multiple data sources. The approach is also characterised by its naturalistic mode of enquiry, its primarily inductive and interpretive role in interpreting the nature of socially constructed events, and the researcher's active role and participation in the study process (Neuman, 2003). Therefore, irrespective of the discipline used or the epistemological position adopted by the researcher, the methodology is suitable for different forms of research for addressing a variety of research questions.

3.5. The Use of Case Study Research

The case study is a research technique that focuses on understanding situations within a particular environment. Hartley (2004) defined case study research as a detailed investigation of phenomena within their context, with the goal of providing an analysis of the processes and context which clarify the theoretical issues being studied. The case study approach is defined by Yin (1994) as an empirical investigation into a social phenomenon within the sense of real life. Creswell (1998) introduced the concept of a case as a set of limitations. Other writers agree with these views, acknowledging that the case study has an exploratory potential, is grounded in nature, and constitutes an intense, in-depth, natural phenomenon-based investigation (Meredith, 1998; Bergen and While, 2000; Luck et al., 2006). Through case studies, issues in organisations can be better understood by investigating the context leading to these issues and the central issue.

Case study research designs have been applied in different disciplines, especially in business, the social sciences, and health, to answer different research questions (Mills et al., 2017). Yin further advanced the development of case study research designs and strengthened the methodological approach of case study research by blending qualitative methods with applied experimental inquiries (Johansson, 2003). As Yin (2003) emphasised, in his realist approach to qualitative case study research he developed a structured process to guide case study research through formal propositions or theories which guide investigations and are tested as part of the research outcome (Mills et al., 2017). Although Yin's approach to case study design remained inductive and qualitative, it was deterministic in nature and placed emphasis on developing theory, on testing and investigating cause and effect, and on understanding the truth (Brown, 2008; Yin, 2014). While some methodologies have been linked with certain philosophical positions that influence the research process, Rosenberg and Yates (2007) noted that case study research is not linked to a fixed epistemological, ontological, or methodological position. Case studies can, therefore, follow a realist and positivist philosophical orientation, assuming there is only one single reality independent of the individual, or a relativist and interpretivist perspective, which assumes there are multiple realities and meanings, which are all influenced and co-created by the researcher (Lincoln et al., 2011; Yin, 2014; Mills et al., 2017).

Merriam (1988) indicates that case studies can be based on historical, psychological, ethnological, and sociological orientations. According to the author, ethnographic case studies are used to examine the observable and learned behaviour, customs, and lifestyles of a culture-sharing community. Stake (1995) and Harling (2012) categorise case studies as intrinsic, instrumental, or collective. The first two are forms of single case study design, while the last one uses multiple cases. Yin (2003) distinguished between a traditional case study design, consisting of one single case, and multiple or more case studies. He classified four designs as follows: Type 1 (holistic) single-case design; Type 2 (embedded) single-case design; Type 3 (holistic) multiple-case design; and Type 4 (embedded) multiple-case design. These case study designs were then classified as either descriptive, explanatory, or exploratory designs. According to Yin (1994), an exploratory case study seeks to investigate any phenomenon that serves as a point of interest to the researcher. Exploratory designs aim to identify research questions in a subsequent analysis or to assess the viability of research methods involving fieldwork and information gathering prior to defining a research question (Hancock & Algozzine, 2011). Tellis (1997) claims that exploratory cases are often regarded as a prelude to social science research. The authors define explanatory designs as those that pursue relationships between cause and effect, while descriptive designs as those that provide a full description of the phenomenon within its natural context (Tellis, 1997; Hancock & Algozzine, 2011).

Comparing case studies with other research strategies, Yin (2003) considers that case studies are most relevant to the exploratory phase of the study, that the survey and histories are suitable for the descriptive process, and that experiments are the only way to answer explanatory or causal questions. Just as there are these three types of case studies (Yin, 1981), Yin (2003) notes that there can also be exploratory experiments, descriptive experiments, and explanatory experiments. He differentiates between each strategy based on the following three conditions:

A. The form of research question posed.
B. The extent of the researcher's control over real behavioural events.
C. The degree of emphasis on contemporary rather than historical events.

These three conditions can be related to five major research strategies in the social sciences: surveys, experiments, histories, archival analysis, and case studies. The comparison between the case study and the other social science research strategies, in relation to the three conditions listed above, is shown in Table 9.

Table 9: Relevant Conditions of Various Research Strategies (Source: Yin, 2003, p. 6)

Strategy	**Form of research question**	**Requires control over behavioural events?**	**Focuses on contemporary events?**
Survey research	Who, what, where, how many, how much?	No	Yes
Experiment	How, why?	Yes	Yes

Historical research	How, why?	No	No
Archival analysis	Who, what, where, how many, how much?	No	Yes/No
Case study	How, why?	No	Yes

As highlighted in the background and the review of literature, patient flow within the healthcare settings is a broad and complex phenomenon which requires collective evidence to address the real experiences of the environment within their context. The background also highlighted the lack of an adept discussion on how and why the use of LM and TOC approaches in healthcare settings have been associated with improvements in patient flow. The lack of adequate discussion highlights the need for an in-depth evaluation of the circumstances under which these approaches are effective at improving and sustaining patient flow. It is clear that experiments and historical research both address the same type of research questions, but historical research does not deal with current or contemporary events, and experiments need to be able to influence behavioural events. Considering the aim of the study was to understand the application of LM and TOC in their context rather than their application on a context-independent basis, case studies, unlike experiments and several other quantitative approaches became a distinguished methodological approach capable of addressing the research aim and questions (Pettigrew, 1973).

The survey research strategy, which focuses on deduction and prediction of causal relationships would be unsuitable for this investigation at this stage considering the nature of the research questions. Use of a survey falls short of providing the rich data required for exploring and understanding emerging relationships linked to the phenomena under investigation such as managing patient flow in different healthcare environments.

The research questions were focused on gaining an in-depth understanding of circumstances associated with the application of Lean and TOC in effectively improving and sustaining patient flow and thus were associated with the 'how, why and what' type of questions.

additionally, there was no degree of control required for the events under investigation. The research questions focus on current events, seeking to understand how LM and TOC work in present healthcare environments. From the above, experiments are therefore ruled out as a possible strategy. The first research question was a "how" type of question (How can LM and TOC approaches be applied effectively to improve patient flow across health and social care?) and the second was a "why" type of research question (why do LM and TOC approaches work better in different health and social care environments?). From these questions, it is clear that a deeper investigation required could not be achieved through a survey strategy which relies on a sample to gather comparatively smaller amounts of data. Considering Table 9, the most appropriate strategy for addressing these questions appears to be the case study.

3.5.1. Advantages of Case Study

The case study acts as a research strategy for planning how research questions can be addressed. The richness of the detailed interpretation of reality they allow is a key advantage of undertaking case studies in any dynamic setting such as the patient flow in the healthcare environment.

> A case study is a history of a past or current phenomenon, drawn from multiple sources of evidence. It can include data from direct observation and systematic interviewing, as well as from public and private archives. In fact, any fact relevant to the stream of events describing the phenomenon is a potential datum in a case study, since context is important (Leonard-Barton, 1990, p. 249).

In contrast to quantitative approaches, George and Bennett (2005) list four advantages of case studies as follows:

1. Conceptual validity

Conceptual validity refers to defining and evaluating the metrics that best reflect the theoretical concepts to be measured by a researcher. Many of the variables of interest in the social sciences, such as political culture, democracy, power, are difficult to measure. Hence, a 'contextualised comparison' is needed, which 'self-consciously seeks to address the issue of equivalence by searching for analytically equivalent phenomena—even if expressed in substantively different terms—across different contexts' (Locke and Kathleen, 1998, p.11). This involves a more detailed analysis of contextual variables, such as variables and sources of variability that influenced the length of patient stay within the health and social care delivery system, which is common in case studies, but extremely difficult to conduct in statistical research.

2. Deriving new hypotheses

Case studies provide a significant gain in the heuristic detection of new variables and theories, such as how and why patient flow has improved over time in the case study organisations under examination, by analysing divergent or outlier cases and in fieldwork, including archival analysis and interviews with respondents, experts from the field, and historians. Case studies can examine dynamic events qualitatively and take multiple variables into account, specifically because they do not involve a large number of cases or a small number of variables. Quantitative experiments neglect methods to produce new theories inductively. Quantitative methods may distinguish specific cases that could lead to new hypotheses, but these methods themselves lack clear ways to identify those new hypotheses. Without further investigation, including open-ended interviews, inductive methods cannot be identified to classify missing variables.

3. Exploring causal mechanisms

Case studies investigate in-depth the operation of causal mechanisms in individual cases. These causal mechanisms highlight sources of variation and uncertainty during the patient journey in a particular healthcare organisation and how these are addressed through different manufacturing approaches such as LM and TOC. They analyse a large number of intervening variables in a single case and observe some unanticipated aspects of the operation of a specific causal mechanism inductively, or they help to classify the conditions in a case which activate the causal mechanism. In contrast, there are no such causalities in the correlations of quantitative studies. However, it is not necessarily accurate that quantitative analysis does not include causality. Rather, quantitative analysis fails to take account contextual factors other than those codified within the calculated variables; several additional variables are overlooked, which may also be relevant in the given context.

4. Modelling and assessing complex causal relations

Case studies can handle complex causal relations, such as complex interacting impacts, equifinality, and path dependency (Ragin, 1987). Case studies can enable equifinality by creating narrower, more contingent generalisations. Despite this, those who support statistical methods place value on general theories, even though it means they are vaguer and more open to counterexamples.

In addition to the above, there are other advantages that strengthen the use of the case study as a research strategy. Using case studies can enhance the validity of the study through the use of different tactics. Yin (1994, p. 33) argues that, in order to be valid, each research study should comply with and 'pass' certain design tests, with respect to the different levels of the validity of the study. He refers to the following four design tests:

1. Construct validity: establishing correct operational measures for the concepts being studied.
2. Internal validity: establishing a causal relationship, whereby certain conditions are shown to lead to other conditions, as distinguished from spurious relationships.
3. External validity: establishing the domain to which a study's findings can be generalised.
4. Reliability: demonstrating that the operations of a study – such as the data collection procedures can be repeated with the same results.

These design tests and their associated tactics are explained in Table 10 below.

Table 10. Design Tests and Tactics in Case Study Research (Source: Yin, 1994, p. 33)

Tests	Case study tactic	Phase of research in which tactic occurs
Construct validity	Use of multiple sources of evidence	Data collection
	Establish a chain of evidence	Data collection
	Have key informants review draft case study report	Composition
Internal validity	Do pattern matching	Data analysis
	Do explanation building	Data analysis
	Do time-services analysis	Data analysis

External validity	Use replication logic in multiple case studies	Research design
Reliability	Use case study protocol Develop case study database	Data collection Data collection

3.5.2. Case Study Limitations

Despite the many advantages offered by the case study approach, there are also disadvantages. Some criticisms of the case study approach are that it suffers from a lack of rigour and an excess of bias and that it is limited by the researchers' sensitivity and integrity. For example, Guba and Lincoln (1981) stressed their concerns about case studies research, specifically case evaluation as an 'unusual problems of ethics. An unethical case writer could so select from among available data that virtually anything he wished could be illustrated' (p. 378). The disadvantages of the case study approach, in addition to the view presented by Guba and Lincoln (1981), are emphasised by Becker (1986) who claims that researchers may experience 'feeling' about the topic and that the results derived, therefore, suffer from a lack of reliability.

Other weaknesses or limitations include concerns about validity, reliability, and generalisability, which are noted by some authors. Berger (1983) suggests, for example, that it is hard to measure external validity because different conditions are inherently dependent on the phenomenon and context. External validity is not the only weakness in the validity of a case study approach; the internal validity of data may also be affected by the bias of the researcher (Bromley, 1986). In favour of qualitative case studies, however, Shields (2007) argued: 'The strength of qualitative approaches is that they account for and include difference--ideologically, epistemologically, methodologically--and most importantly, humanly. They do not attempt to eliminate what cannot be discounted. They do not attempt to simplify what cannot be simplified. Thus, it is precisely because a case study includes paradoxes and acknowledges that there are no simple answers, that it can and should qualify as the gold standard' (p. 12).

Although case studies have been recorded in various ways and categories with different levels of data collection of analysis methods (Eisenhardt, 1989; Yin, 1994), Kiridena and Fitzgerald (2006) established that academic publications in OM have failed to give sufficient representation to this research approach. Casell et al. (2006) and Gephart (2004) consider the lack of specificity and detail surrounding the philosophical position adopted by the study, the methods adopted for the analysis of data, and the procedures used in the interpretation of findings to be key issues that should be addressed in studies employing the case study research method. The scientific aspects of research, which include a lack of generalizability, are the key concerns for a study relying on this approach (Silverman, 2001). However, the choice of a research strategy is widely recognised as a trade-off: the strengths of one approach outweigh the disadvantages of another approach, and vice versa (Cavaye, 1996). Consequently, when a researcher chooses the case study as a suitable research methodology, the strengths of case-study research are considered as significant, and the methodologically based limitations of the research are accepted (ibid).

The most important feature of the case study approach is that it offers a minimal basis for conventional 'scientific generalisation' (Remenyi et al., 1998; Yin, 1994). As with all experimental studies, the findings of the study can be generalised to conceptual propositions, or theory, and can influence other similar environments, such as complex health and social care settings in other western countries. According to Erickson (1986), given that the general lies in particular, what we learn in a particular situation can be transferred to similar circumstances. Further, Eisner (1991) argues that, in a case study, the colourful explanation of a particular example can build a powerful image: 'a vivid portrait of excellent teaching, for example, can become a prototype that can be used in the education of teachers or for the appraisal of teaching' (p. 199). Readers could learn vicariously from this experience through the detailed explanation of the researcher (Stake, 2005). The purpose of case studies, therefore, cannot be to conclude that one study reflects the results for a worldwide population, but rather to understand and express patterns and connections of the theoretical value.

Overall, the researcher has made every effort to overcome the limitations of the case study method. The researcher is taking a multiple case study approach, the evidence from several case studies frequently being seen as more convincing and more robust than that from a single case. Reliance on evidence from multiple cases also increases the study's external validity (Voss et al., 2002). Multiple cases validate the results by replicating the matching pattern so that assurance in the strength of the theory is increased. Multiple cases facilitate replication and amplification among individual cases and are influential in developing theory since case study research should have methodological rigour in order to be useful in creating theory (Eisenhardt, 1991).

In addition, the researcher is following the guidelines for the validity and reliability of the case study design suggested by Yin (1994), and as indicated earlier. The triangulation technique further overcomes the limitations of the case study method. Triangulation is described by Stake (1995) as using procedures to ensure accuracy and alternative interpretations. Triangulation can be accomplished by using various data sources in case studies (Yin, 1994). The core principle of triangulation is that qualitative and quantitative approaches are meant to be treated as complementary, rather than as rival camps (Jick, 1979). Therefore, in this study, the researcher applies triangulation by using various data sources, such as semi-structured interviews and observations, and, in particular, by using quantitative methods in the form of documentary evidence which highlights improvements by comparing patient flow performance before and after the intervention to support the results of the qualitative research. These specific strategies for validation and reliability are further clarified in the research design phase, in line with the Eisenhardt 8-step process presented in this chapter.

3.6. Best Practice for Conducting Case Study Research

In light of this research, it is important to have a rigorous design for carrying out case study research so that the findings are reliable and accepted. As discussed in chapter 2, there is no existing theory and clear evidence that provides a feasible answer to how TOC and LM approaches can be applied effectively to meet the needs of different healthcare environments. Thus, the literature review identified three inductive research questions for this research. These research questions are critical for health and social care organisations. Although a few recent research studies have attempted to tackle the same problems, this has been insufficient to explore the real problem, as they have never attempted to explore the circumstances associated with the success of LM and TOC approaches in manufacturing and how these approaches can be used selectively and effectively in different healthcare settings. In other words, the majority of existing research does not address these proposed questions at all. As the proposed research questions are inductive in nature, indicating that there is a potential for the emergence of new theories or concepts from this study, this section reviews briefly some of the best research practices for developing theory from case study research, thereby supporting the choice of case study design adopted in this research.

Case studies have been used by many researchers in various fields of research to test or develop theories (Anderson, 1983; Pinfield, 1986; Eisenhardt & Graebner, 2007; Harris & Sutton, 1986; Gersick, 1988; Mintzberg, 1979). Conducting case studies to build theory is a research strategy that employs one or more cases to establish theoretical constructs, hypotheses, and/or midrange theory through case-based analytical evidence (Eisenhardt, 1989). There are distinct approaches to case study research for building theory. Glaser and Strauss (1967), for example, developed the grounded theory approach to inductive case research. Grounded theory proposes that theory emerges as the researcher gathers data about a phenomenon. The approach of grounded theory is focused on the ongoing cycle of empirical data gathering and data analysis to create concepts through a coding process that enables theory to be generated. To promote the theory, the researcher joins the field without a prior hypothesis, but with a substantial background in the literature.

Yin (1994) provides a simple protocol guide for a case study with an emphasis on field procedures and case study questions, and with reference to the final write up. His case studies begin with the construction of particular theoretical hypotheses, extracted from a conceptual framework. The grounded theory approach is, therefore, very different from Yin's (1994) replication/validation approach, in which theory is 'developed' and then tested in case environments at the outset. Yin stresses the need for researchers to review related literature and to provide theoretical hypotheses on the case under analysis before beginning any data collection. This distinguishes it from approaches such as grounded theory and ethnography.

Eisenhardt (1989) offers an alternative approach to the development of theory from case study research. The paper by Eisenhardt (1989) is a common framework representing eight distinct steps for building theory from case studies and is one of the most quoted papers in approaches to methodology (Ravenswood, 2011). In her empirical articles, Eisenhardt, for instance, gives a wider range of case study research models, by combining the building and elaboration of theory (Bingham and Eisenhardt, 2011). Eisenhardt (1989) proposes a combination of grounded theory with Yin's approach. For instance, her approach is close to an inductive approach which is consistent with some elements of grounded theory, such as the opportunistic collection of data and the absence of prior hypotheses. However, there are aspects in which she follows a more structured approach, such as the use of a case protocol.

In contrast to Yin's approach, both Eisenhardt and grounded theory approaches are not based on the validation of current theoretical structure but on creating a conceptual model by continually analysing and evaluating empirical evidence and building on these evidence theoretical concepts/constructs. Given the absence of an existing theoretical framework for managing patients flow in different healthcare settings, the case study design approach by Yin is not considered by the researcher to be suitable for the purposes of this research. With regard to theory building, the 8-step framework and grounded theory approach are both inductive, but each follows a different direction in developing theory from case study research. For example, the number of cases in a grounded theory study cannot be stated at the outset of the research, as it can in the 8-step approach. Eisenhardt recommends selecting cases before joining the field, i.e. early in the research design. This is not consistent with the grounded theory approach, where case choice is based on previous tests, i.e. after joining the field. And this is not the case in this research, as the theoretical sampling was identified before the researcher entered the field.

Ridder (2017) outlines parallels and distinctions between different case study models and analyses their respective theoretical contributions, as seen in Table 11. With respect to contributing to the theory, the design of case studies has various objectives:

1. In the 'No Theory First' design model, Eisenhardt suggests that a broad and tentative research question begins with some provisional variables. The research question, as well as the variables, can be altered during the study. This design prevents proposals concerning relationships. In this model, theoretical sampling seeks to select a case, or cases, that is suitable for the identification of new or extended preliminary constructs and the establishment of new relationships.

2. In ‘Gaps and Holes’, Yin advises the research questions to be firmly connected to current theories, concentrating on queries about 'how and why.' The contemporary theory involves study gaps which, once recognised in the existing theory, contribute to presumed relationships which are the foundation for empirical information matching frameworks and proposals. Theoretical sampling focuses heavily on the purpose of the case study. Extreme and unusual cases have different aims from common or revelatory cases.
3. In Stake’s ‘Social Construction of Reality’, there is no research question at the start, but the curiosity of the case is a facilitator for comprehending the research problem. As in ‘Gaps and Holes’, the sampling is purposeful but for a variety of purposes. Either the case itself is of interest or the case is an opportunity to understand a theoretical issue.
4. Burawoy’s ‘Anomalies’ approach suggests that the theoretical question is motivated by the question of why the current theory cannot explain a given anomaly. What kinds of gaps, silences, or inner inconsistencies reflect the inadequacy of the existing theory? There is a distinct difference between the ‘anomalies’ model and the theoretical sampling in the three models above. Here, the sampling strategy seeks to select a case that demonstrates the theory’s failure.

Table 11: Portfolio of Case Study Research Design: differences in underlying elements (Source: Ridder, 2017, p. 12)

	Case study research designs			
	No theory first	Gaps and holes	Social construction of reality	Anomalies
Representative scholars	Eisenhardt	Yin	Stake	Burawoy
The case	Research question; A priori constructs, variables; No assumed relationships	Research question; Existing theory; Proposition; Framework	Curiosity in the case; Understanding of research issues	Curiosity; Existing theory; Anomalies; Internal contradictions; Gaps, silences
The data	Theoretical sampling; Qualitative data as the primary choice	Purposeful sampling; Qualitative data as the primary choice	Purposive sampling; Thick descriptions; Holistic comprehension	Theoretical sampling; Dialogue of observer and participants; Participant

In all of these designs, qualitative data are preferred, while quantitative data are viewed as a possible means of strengthening the cases. In the 'social construction of reality', however, a powerful focus is placed on comprehensive descriptions and holistic knowledge of the case. This is compared with a more constructed and variable-oriented collection of data in the 'no theory first' and 'gaps and holes' models. Contrary to that, the 'anomaly' tactic is the only strategy which receives data from the dialogue between the observer participant and the observer.

Having illustrated the differences between different case study research designs, this researcher considers Eisenhardt's methodological framework the most suitable design approach to meet the purpose of this research. Researchers in many disciplines have used this framework, and her famous paper 'Building case study theory' has been cited in academic work in a broad range of fields.

Table 12: Process of Building Theory from Case Study Research (Source: Eisenhardt, 1989, p. 533)

Step	Activity	Reason
1 Getting started	Definition of research question	Focuses efforts
	Possibly a priori constructs	Provides better grounding of construct measures
	Neither theory nor hypotheses	Retains theoretical flexibility
2 Selecting cases	Specified population	Constrains extraneous variation and sharpens external validity
	Theoretical, not random, sampling	Focuses efforts on theoretically useful cases – i.e. those that replicate or extend theory by filling conceptual categories
3 Crafting instruments and protocols	Multiple data collection methods	Strengthens grounding of theory by triangulation of evidence
	Qualitative and quantitative data combined	Synergistic view of evidence
	Multiple investigators	Fosters divergent perspectives and strengthens grounding
4 Entering the field	Overlap data collection and analysis, including field notes	Speeds analyzes and reveals helpful adjustments to data collection
	Flexible and opportunistic data collection methods	Allows investigators to take advantage of emergent themes and unique case features
5 Analyzing data	Within-case analysis	Gains familiarity with data and preliminary theory generation
	Cross-case pattern search using divergent techniques	Forces investigators to look beyond initial impressions and see evidence through multiple lenses
6 Shaping hypotheses	Iterative tabulation of evidence for each construct	Sharpens construct definition, validity, and measurability
	Replication, not sampling, logic across cases	Confirms, extends, and sharpens theory
	Search evidence for "why" behind relationships	Builds internal validity
7 Enfolding literature	Comparison with conflicting literature	Builds internal validity, raises theoretical level, and sharpens construct definitions
	Comparison with similar literature	Sharpens generalizability, improves construct definition, and raises theoretical level
8 Reaching closure	Theoretical saturation when possible	Ends process when marginal improvement becomes small

Eisenhardt's eight-step road map, as illustrated in Table 12, is the best way to develop theory from case study research. Case study research is used to study a particular phenomenon using an in-depth study that is narrow in scope. This type of research is used in the early phases of research, where there are no previous hypotheses or work to be used as a guide. Eisenhardt's eight-step road map for developing theories from case studies is suitable for this study as it provides a systematic framework for conducting research. Eisenhardt describes the position of researchers at each step and explains what they should do at each step and why. The approach helps in detailed step-by-step research, taking every minute detail into account. This approach supports studies of both very complex and dynamic phenomena, such as managing patient flow in complex health and social care settings. It is also a very rigorous, but creative, process used for research of an intensive nature. Eisenhardt's eight-step method is discussed in the next section.

3.7. Eisenhardt's Process Theory Development from Case Study Research

3.7.1. Step 1. Getting Started

Although Eisenhardt (1989) recommends that the research should start as close as possible to the ideal, with no preconceived theories or hypotheses to test, she suggests that researchers should determine some variables and the research question to address before the research starts. Eisenhardt (1989) argues that defining the research question initially is important for building theory from case studies. Defining the research question has the same rationale as in research which tests a hypothesis – to give the research a focus and to determine the type of organisations to be targeted and the kind of data to be collected. The research questions provide a focus for the research, without which the researcher would easily become overburdened by the volume of qualitative data gathered during the research process (Yin, 1994).

Eisenhardt further highlights the importance of a prior specification of the constructs which shape the initial design of theory-building research, by allowing researchers to accurately measure and investigate these constructs. While identifying and defining the research question and possible constructs are essential, Eisenhardt (1989) advises researchers to recognise both of these as tentative, since research questions may change during the process and alter the role of certain constructs in the resultant theory. Although it is impossible to truly accomplish, beginning theory-building research as close as possible to the ideal of no existing pre-conceived theory and no hypotheses is important in order to limit the bias involved in predetermined theoretical views and propositions which may have a bearing on findings. Wherever possible; however, as researchers formulate a research problem and variables considering the existing literature, they should guard against considering specific relationships between variables and theories when the research process begins.

A prior specification of constructs is one of the three essential issues in getting started. The researcher can follow two distinct approaches to using current theoretical constructs as a guide to the development of the theory (Anderson and Aydin, 2005). In one instance, the researcher would work within a clear conceptual framework which comprises selected concepts and highlights the relationship among such concepts grouped in a way that provided a simultaneous overview of the relationships between the major concepts (Miles and Huberman, 1994; Pare, 2002). In the other instance, the researcher would not depend on the constructs presented by prior theory, but, rather, the purpose of the investigation focuses on developing relevant theory, constructs, and hypotheses (Pare, 2002). This is in line with Yin's (2014) argument that, even for studies that are exploratory in nature, a conceptual framework is key in defining the constructs under investigation. The constructs identified were, however, only tentative, considering the aim of the investigation was to build theory (Eisenhardt, 1989).

3.7.2. Step 2. Selecting Cases

The selection of cases in case study research is an essential element of theory building (Eisenhardt, 1989; Yin, 2014). While the concept of population is crucial in case study research, the sampling of cases is rare, as cases are selected for theoretical rather than statistical reasons (Glaser and Strauss, 1967; Eisenhardt, 1989). The selection of cases, according to Eisenhardt (1989), can be to extend emergent theory, to replicate previous cases, or to fill theoretical categories. While Eisenhardt (1989) recognises that cases can be randomly selected, it is suggested that this type of selection is not necessary, as researchers should focus on selecting cases that present extreme situations and in which the process they are interested in is transparently evident. As Pettigrew (1990) points out, considering the small number of cases that can typically be studied, it makes sense to select extreme situations and polar forms, in which the mechanism of interest is 'transparently observable'. While Meredith (1998) indicates that multiple numbers of cases as low as 2 to 6 are generally acceptable where there is some understanding of the phenomenon, Eisenhardt (1989) recommends case numbers in the range of 4 to 10 according to the findings. While the sample size of qualitative research appears to be small, it is justified by the depth of case-based research, which is central to this type of investigation (Sandelowski, 1996). The number of replications remains a matter of judgment and discretion on the part of the researcher, with the certainty the researcher wants to obtain from the multiple-case findings being a determining factor (Yin, 2014).

3.7.3. Step 3. Crafting Instruments

The combination of multiple data collection instruments is typical for theory-building researchers (Eisenhardt, 1989). Case study research uses various data collection instruments, for example, observations, archival sources, and interviews, which support data triangulation and the development of stronger hypotheses and constructs (Eisenhardt, 1989). The use of a variety of data collection instruments is a significant tactic in enhancing the reliability of case-study research (Yin, 2003).

Eisenhardt (1989) emphasises the importance of combining quantitative and qualitative data collection instruments in a single case study. Data triangulation supports the use of qualitative and quantitative data. According to Eisenhardt (1989), this supports an understanding of the theory underlying the relationships predicted by quantitative data but eliminates an overreliance on untrue, but strong, impressions arising from qualitative data. According to Yin (2014), the rationale behind the combination of several data collection methods typically adopted by theory-building researchers is similar to that for hypothesis-testing research. Several data collection methods support the possibility of triangulation, making the validation of the constructs stronger (Eisenhardt, 1989). It has been recommended by some researchers that, wherever possible, both qualitative and quantitative data should be collected and used in any single study (Pare, 2002).

Wider coverage and scope are obtained by using different methods and sources in collecting different types of data. This results in a deeper insight into the phenomenon being investigated, which would otherwise not have been attained (Yin, 2014).

The importance of integrating various data collection instruments in developing theory is also emphasised by Mintzberg (1979), who suggests that, although the foundation of theories is built on systematic data, it is the subjective data that support their construction. Theory building, therefore, requires the rich description that is offered by subjective data; while quantitative data uncovers existing relationships, these relationships can only be explained through qualitative data ((Eisenhardt, 1989, p. 587).

3.7.4. Step 4. Entering the Field

Eisenhardt (1989) recommends that, once they enter the field, researchers should overlap data collection and analysis, so that they can alter the data collection process when new data collection opportunities emerge or because of a new line of thinking which arises during the research process. The frequent overlap of data analysis with data collection thus becomes a key characteristic of theory-building case-study research, since researchers can gain new theoretical perspectives that can better ground the theory by altering data collection during analysis (Eisenhardt, 1989, p.539). The joint collection, coding, and analysis of data which presents this overlap is also called for by Glaser and Strauss (1967). Eisenhardt (1989) suggests the overlap can be achieved using field notes. The freedom to adjust the data collection process can include adding cases to probe certain themes emerging from the analysis and to improve the resultant theory.

3.7.5. Step 5. Analysing Data

The analysis of data is central to the development of theory from case studies. This process is considered the most complicated and least structured in the research design (Eisenhardt, 1989; Miles & Huberman, 1994). Qualitative studies create vast amounts of data which cannot easily be mechanically manipulated, analysed, and reduced (Yin, 1994). However, the relative challenge of analysing qualitative data does not invalidate the data or the final conclusions (Cavaye, 1996). The established rules of logic can be extended to verbal data to make sense of the evidence and to analyse the data formally (ibid). Case studies using both within and cross-case analysis are more successful in developing theoretical constructs and formal hypotheses than research which uses either within or cross-case analysis (Barratt et al., 2011). Eisenhardt divides the data analysis into two steps:

3.7.5.1. Within case analysis

Data analysis begins at the level of within-case analysis once data collection is completed, giving the researcher the opportunity to develop in-depth familiarity with each case and supporting cross-case analysis (Eisenhardt, 1989). This generally includes a comprehensive write-up for each case. Sandelowski (1996) argues that 'looking at and through each case in a qualitative project is the basis' of theoretical interpretations and generalisations (p.525). Examination of particular cases allows the researcher to consider those elements of experience which do not occur as individual 'units of meaning', but as part of a pattern created by the convergence of meanings within individual accounts. Eisenhardt (1989) suggests the need for researchers to closely familiarise themselves with each case as an independent unit and to draw unique patterns from each case before a general pattern is drawn across all cases.

3.7.5.2. Cross-case analysis

Cross-case analysis, following within-case analysis of data, may take various forms, such as drawing variances and parallels between cases in certain categories; dividing data by source of data; and evaluating pairs of similar cases for similarities and differences. Eisenhardt (1989) points out that there is a danger that investigators may draw premature, and even false, conclusions as a result of 'information-processing biases', such as jumping to conclusions based on insufficient evidence (Kahneman & Tversky, 1973) or being motivated by more elite participants (Miles & Huberman, 1984) and by the vividness of particular data (Nisbett & Ross, 1980). In order to counteract these trends, Eisenhardt (1989) suggests two techniques: (a) to select dimensions or categories, and then to check for within-group similarities paired with inter-group differences; and (b) to choose pairs of cases, and then to identify similarities and differences between the pairs. Mapping the relationships between variables supports cross-case analysis while exploring the common themes across all cases (Miles and Huberman, 1994). Comparison across cases depends on linking variability to consistency in the process and results (Pettigrew, 1985), using the principle of literal and theoretical replication (Lee, 1989; Yin, 1994). Cross-case analysis tactics, according to Eisenhardt (1989), enhance the likelihood of researchers capturing novel findings in the data.

3.7.6. Step 6. Shaping Hypotheses

While similarities identified in the analysis of cross-case data strengthen the findings, contradictions lead to further examination of the data, making the theory-building process iterative (Eisenhardt, 1989). A match between the theory and data leads to an empirically valid theory (Eisenhardt, 1989). According to Miles and Huberman (1994), the shaping of hypotheses in theory-building research entails the measurement of constructs and the verification of relationships. A detailed comparison of the data gathered with the constructs established will lead to a case study hypothesis, which only emerges after the analysis of data and not a priori to the investigation (Eisenhardt, 1989). The application of a construct to each case and the emergence of evidence supporting the construct in each case leads to acceptance of the hypothesis and gives the study internal validity (Eisenhardt, 1989). The next step in this extremely iterative process is regularly comparing the new theory with the data from each project to determine how well or negatively it matches the results. The fundamental idea in the shaping of the hypothesis is a constant comparison of data with theory, to move towards a theory that thoroughly reflects the patterns emerging from the data (Eisenhardt, 1989).

Since theory building cannot rely on statistical tests to validate the findings, the shaping of hypotheses in theory building involves a more judgment-based approach (Huberman and Miles, 2002). The strength and uniformity of relationships within and across cases are used with supporting qualitative data to justify why the emergent theory holds, while the presentation of evidence allows readers to measure, by their own standard, the rigour adopted in the research process and to ascertain the validity and reliability of the investigation.

3.7.7. Step 7. Enfolding Literature

According to Huberman and Miles (2002), a comparison of emerging concepts or hypotheses with existing literature is a key element in theory building. After shaping the hypothesis, the researcher starts to uncover the relationship between the existing literature and the emergent concept or theory (Eisenhardt, 1989). During the enfolding literature stage, the researcher evaluates the similarities and variations between the existing literature and the constructs being investigated. While agreement between the constructs and the existing literature confirms the emergent theory, conflicting positions in the literature result in further analysis and creative thinking. Consideration of a broad range of literature is key at this stage to ensure the researcher develops a theory 'with stronger internal validity, wider generalizability, and a higher conceptual level' (Eisenhardt, 1989, p. 544). Huberman and Miles (2002) also argue that trying to establish emerging theory through the existing literature increases 'the internal validity, generalizability and theoretical level of theory building from case study research' (p. 26).

According to Eisenhardt (1989), a comparison of the existing literature with emergent hypotheses, theory, or concepts is a critical aspect of the theory-building process. This comparison entails identification of similarities and reasons for differences. In this study, for each of the emerging themes, the researcher highlighted the extent to which the themes were supported by existing literature and established justifications for deviations from the existing literature, where contradictions were noted between the literature and the emergent concepts. The analysis also included new ideas and perspectives that were emerging. The linking of the existing literature with emergent theoretical propositions establishes the generalisability of the theory and improves the internal validity of case study research (Eisenhardt, 1989).

3.7.8. Step 8. Reaching Closure

When the iteration between theory and data does not lead to any conflicting views, but it only incrementally strengthens the emerging theory, the researcher can close the research process when it is exhausted with the theoretical saturation of the case studies (Eisenhardt, 1989). The qualitative samples must be large enough to gain sufficient data for the phenomena of interest and for the research questions to be addressed sufficiently. For example, in phenomenological research, Morse (1994) suggested at least six interviews and Creswell (1998) recommended between 5 and 25 interviews. Morse (1994) proposed 30 to 50 interviews for grounded theory, while Creswell (1998) recommended just 20 to 30 interviews. With respect to the number of cases, between 4 and 10 cases are recommended by Eisenhardt, while 2-6 cases are proposed by Meredith (1998), and an absolute limit of 15 cases is suggested by Mile and Huberman (1994) on the basis of practicality.

Although researchers may use such guidelines to predict the expected number of participants, Glaser and Strauss (1967) propose the saturation principle to obtain what constitutes a sufficient sample size in qualitative studies. Theoretical saturation occurs when incremental learning has reached its minimum, and the researcher starts observing only phenomena that have already been uncovered (Glaser and Strauss, 1967). At this stage, the decision to stop adding cases then becomes a significant factor in reaching closure.

3.8. Applying Eisenhardt's 8-Step Framework

1. ***Getting started***

The first step is (a) to define research questions and (b) to identify constructs where possible.

A review of the literature on applications of LM and TOC in the healthcare setting and their success and failure in improving patient flows offered some guidance on how research questions were formed. The comprehensive literature on the origins of LM and TOC provided a preliminary understanding of the circumstances associated with their success in manufacturing. From these reviews and early case analysis, the tentative hypothesis was constructed in the researcher's mind. The researcher, however, tried to stay positive and avoid predicting any potential relationships between different constructs in advance. The research questions, the tentative hypothesis, and the constructs were therefore checked continually, while the case data were analysed and tested for fitness.

Following this step, the researcher developed three research questions that offered a clearly defined focus for the investigation and that were also central in determining the choice of research methods and the type of data to be collected during the research process. Following the initial case analysis, the associated constructs were developed. Establishing a tentative hypothesis helped in developing and identifying the constructs (e.g. environmental instability and buffering mechanisms) and associated measures such as LOS and DTOC. From these constructs, interaction with the literature would highlight how different strategies and approaches attempt to reduce and manage variability within different circumstances.

As this study investigates the translation of LM and TOC approaches into healthcare environments, the investigation of these two constructs plays a crucial role in digging deeper into how both approaches manage instability in these settings and encourage the patient flow (e.g. demand uncertainty and variation, delayed discharge, intention of local optimisation). Because this study looked at the circumstances associated with the influence of LM and TOC on patient flow performance, the DTOC or LOS is the key measure used to highlight the improvement.

The constructs and measures developed in this research are described below:

- Instability (Variation & Uncertainty):
  The level of variation and uncertainty associated with demand, together with process and recovery times.

- Buffering mechanism (Time & Capacity):
  Managing instability and any threat to patient flow can be achieved through proactive flow mechanisms, which allow capacity to be exceeded without additional beds, resources, and costs.

- Patient flow improvement:
  Measure of the improvement of patient flow in terms of LOS/waiting time or DTOC.

2. ***Selecting the cases***

After the research questions had been identified, the researcher started to look for potential cases that could be found to contribute to achieving the purpose of this research. At the beginning of January 2017, the researcher identified three potential cases, as these cases had applied the principles of LM and TOC to improve patient flow in the healthcare delivery system. At this stage, two cases were selected to evaluate the adoption of the TOC approach in managing and improving patient flow. One case was selected to investigate the use of LM tools in improving the service delivery system. While conducting the initial case study, the researcher developed networking with some participants, leading him to identify three more cases that have implemented innovative interventions which have contributed to improved patient flow. However, two cases were not accessed due to policy conditions and new regulations within the Trust management. Therefore, four case studies were eventually selected across three NHS trusts, as summarised in Table 13 on the basis of their theoretical usefulness. The cases incorporate acute and rehabilitative hospital care, social care, out-patient services, and GP-led community care and community mental health. These cases contributed theoretically to achieving the aim of this research, based on the following criteria:

- The availability and accessibility to these cases were considered in terms of obtaining ethical approval to conduct the study and geographical location. This process was challenging and involved a long process to obtain ethical approval from the relevant NHS Foundation Trusts for all the cases involved in this research.
- The cases had applied the principles of LM and TOC to transform patient flow and had shown significant improvement. As shown in Table 13, the first two cases were chosen to specifically illustrate the influence of implementing the TBM associated with the TOC approach on patient flow. The third case in the table was the only available and accessible case that implemented LM tools to enhance its delivery system and services. However, many previous studies have examined the application of LM in healthcare. A systematic peer review of successful cases of adopting LM in healthcare was, therefore, undertaken in this study. The final case introduced an innovative local intervention that has contributed to improved patient flow and was chosen in line with the logic of theoretical sampling in the light of earlier case findings.
- The key informants participating in this study had a broad experience of dealing with the translation of LM and TOC approaches in healthcare settings. Those key informants had an in-depth knowledge of how such an approach applied to manage patient flow in their environments.
- The adequacy of evidence sources for the objectives of this research was taken into consideration in selecting these cases. While interviews with key informants were sufficient for the goals of this research, alternative sources of evidence, such as observation and archiving records, were a key factor to be taken into account in selecting these cases to collect the data needed for this research. However, there were some difficulties in obtaining some documentary evidence in case 2 and case 3 due to the confidentiality of data.

The selected four cases represent different healthcare delivery systems that deal with issues in terms of instability (variability and uncertainty) that impact the flow performance, resulting in an increase of LOS/DTOC. Each case has implemented a flow mechanism to manage these issues and improve patient flow. Therefore, all of these cases are comparable as they use the same or similar constructs in terms of flow characteristics, including variability, uncertainty, and the mechanism of flow and the use of similar measurements such as LOS, DTOC, and waiting time.

Choosing these cases is justified from a hypothetical perspective because they are similar to many other health and social care environments characterised by complexity, variation, and uncertainty. As shown in Table 13, the selected cases represent different characteristics of healthcare environments regarding the level of instability (variability and uncertainty) associated with demand, process, and recovery (treatment) times. Case 1, for example, represents a planned healthcare environment to provide care for children in communities and acute hospitals across the county. The care is provided to children who suffer from specific and complex needs (e.g. autistic spectrum disorders, hearing impairment, learning and physical difficulties). The nature of treatment for each patient is varied depending on the needs of the patient. Patients should be referred to the service provider through their GP and then booked for assessment and then treatment once they meet the service criteria. Although the service provider is aware of the list of patients schedueled for appointments, the demand for this service is uncertain across the county. Therefore, this healthcare environment can be classified as planned care with a high level of instability.

Case 2 can also be considered a healthcare environment involving a high level of instability as the nature of treatment, process, and demand is unstable. This healthcare organization is a part of 12 community hospitals offering urgent care, rehabilitation, and mental health services to patients across the county. Usually, services are provided to elderly patients with complex needs who are referred immediately to the community hospital once they meet the required criteria. In this type of healthcare environments, demand, process and recovery times are unpredictable and uncertain. Therefore, this case can be classified as an unplanned health care environment with a high level of instability.

While both Cases 1 and 2 deal with a high level of instability, Case 3 represents one example of healthcare community services across the county that provide care and support to patients at their homes or living in residential care homes. This environment is more stable than the previous two environments in terms of demand, process and recovery time. The community's staff deal with patients whose needs are uncomplicated, and usually, the demand for the service is predictable.

Finally, Case 4 represents an acute hospital that serves one of the neediest local authority regions in England and Wales. The hospital is located in the town that ranked as the highest record in the county of individuals suffering from a long-term illness, with three in five people aged 65 and over having a long-term illness. The service is provided through the acute hospital to meet different patients' needs in some planned and unplanned care units. Demand, process and recovery time are varied depending on the type of required care and season conditions. Therefore, this healthcare environment can be classified as planned/unplanned care with a high level of instability.

All cases were chosen by theoretical sampling and were not picked statistically or randomly. Such cases helped the researcher to obtain an in-depth understanding of the circumstances of the translation of LM and TOC approaches to manage patient flow in the context of health and social care.

Table 13: List of Case Studies

Evaluated Cases	Nature of Healthcare Environment	Nature of the Transformation Approach	Instability Level/Variation and uncertainty
Case 1	Planned Care-Community outpatient and Acute	TOC/TBM	High
Case 2	Unplanned Care-Community/rehabilitation hospital	TOC/TBM	High
Case 3	Planned Care-Community Outpatient	LM/5S	Low
Case 4	planned/unplanned care-Acute	Local intervention/early discharge scheme	High

3. *Crafting instruments*

Case study research offers a variety of methods of data collection, including interviews, documentary evidence, observation, and archived records. Interviews, observation, and document analysis were used consistently across most cases to ensure consistency in data collection. Yin (2014) suggests that research instruments have a high importance in the research and that they should support a suitable and consistent level of data capture when used by several investigators across multiple cases. The research instruments were designed to offer the potential to obtain convergence and confidence in the research findings. In this research, triangulation has been used to make sure the data from the interviews were accurate and reliable. Harnessing different sources in this research provided more rigour in the collected data. The researcher considered documented evidence to support triangulation and validate the findings of the study. A summary of the findings was validated by at least one relevant executive in each case study to confirm the accuracy of the collected data.

Since the study involved only one investigator, there was a reduced risk that protocols and instruments could be misaligned. Details of the collection methods and participants are shown in Table 14.

Table 14: Data Collection Tools

Case	Participants	Data Collection Instruments	Duration & numbers of the interview
Case 1	Associate Director of Quality Improvement Quality Improvement lead Quality Improvement Facilitators Service Admin Therapist Individuals of Consultancy Agency	Semi-structured interviews	7 Participants Between 40 and 90 minutes
Case 2	Chief Operating Officer Associate Director of Strategy Transformation Lead Integrated Community Manager Inpatient Matron Individuals of Consultancy Agency	Semi-structured interviews, observation and documentation	8 Participants Between 40 and 90 minutes
Case 3	Associate Director of Quality Improvement Co-ordinator/Administrator Quality Improvement lead Quality Improvement Facilitators	Semi-structured interviews	5 Participants Between 40 and 90 minutes
Case 4	Local authority team Discharge Team discharge Matron Senior Managers Nurse Team Housing Operations & Safeguarding Manager Housing Needs Manager	Semi-structured interviews, observation and documentation	10 Participants Between 40 and 90 minutes

Semi-structured interviews

Semi-structured interviews were implemented as the primary data collection technique for this research. In-depth semi-structured qualitative interviews investigate participants' experiences and the meanings attributed to their experiences (Tong et al., 2007). Usually taking the form of a one-to-one interview, asking open-ended questions allows participants to discuss with the researcher issues relevant to the research question, as the researcher can re-word, clarify, or re-order questions and probe respondents to provide a deeper understanding of the phenomena under investigation (Yin, 2013; Creswell, 2014). Wright et al. (2004) note the usefulness of semi-structured interviews in the healthcare sector when they claim that such interviews make it possible to identify variables which can be adjusted to improve healthcare.

Semi-structured interviews were chosen because they are the best approach for collecting information on a large-scale or for conducting exploratory research, as the researcher does not have enough prior information about the investigated area and cannot create a list of possible pre-codes. However, analysing the interview data from open questions can be challenging (Mathers, Fox, and Hunn, 2002). The interview questions are designed to give an overview of the original design and performance of the system, process issues in the original design, changes recorded, and results achieved.

These questions are standard and unified for all selected cases. Appendix 1 shows the interview protocol intended for this research. A total of 32 interviews were conducted, and interviews were recorded using the Digital Voice Recorder Olympus DS-3500. This digital device is flexible and easy to use, but, most importantly, it has the ability to encrypt a voice recording and thus ensure those voice recordings are safe from unauthorised people. The length of each interview varied between 40 and 90 minutes. Interviews were conducted with the key informants taking into account their managerial or operational involvement. These key individuals included executive directors, senior managers, operation managers, therapists, nurses, quality improvement personnel, and external consultants. All the interviews were transcribed and stored on a password-protected computer to reduce the risk of unauthorised access. All the interviews took place in a quiet room at the participants' location. The researcher observed all ethical issues related to this research. Issues such as the informed consent of the participants, confidentiality, the right to withdraw without giving reasons, and anonymity were clearly explained and adhered to by the participants.

Although semi-structured interviews are very useful, they have some disadvantages like all other research methods. Among the most frequent criticism of qualitative interviews is that the approach is too subjective, because both the collection and analysis of data are performed by the researcher alone and because the qualitative strategy includes personal contact with the people and institutions under investigation (Patton, 1990). Another criticism is that the researcher cannot be certain the informants have given the correct information. They might say what they think the interviewer wants to hear. The researcher can overcome this barrier by looking for more convincing evidence (Newton, 2010). The overdependence on data whose reliability is determined by the responses of the interviewees is a key limitation of interviews as a research instrument (Yin, 2014). To minimize bias by the researcher or the participants, the researcher asked open questions and did not reveal any specific hypotheses, to avoid directing the participants' answers to a particular assumption. Also, observation and access to documentary evidence in some cases enhanced the validity of information obtained from the interviews. After completing the data analysis, the researcher conducted further interviews with two senior managers, a leading TOC specialist in healthcare and a director of NHS improvement, to examine the findings of the research and ensure its accuracy. Thus, the researcher has made sure that the study's findings are as objective and reliable as possible and were not affected by his own interpretations.

Documentary evidence and analysis

Documentary analysis involves undertaking a systematic procedure to evaluate and review documents (Bowen, 2009). Documents that can be used for systematic evaluation take various forms and may include: agendas; minutes of meetings; brochures; manuals; diaries; background papers; programme proposals, institutional reports; and letters and memoranda among other sources (Bowen, 2009; Yin, 2013; Creswell, 2014). Bowen (2009) summarises the role of documentary evidence in case studies and qualitative research as being to provide background and context for the phenomena being studied, to facilitate the formulation of additional questions to be addressed, to supplement data, to offer a means of tracking change and development, and to verify findings. Researchers, however, should carefully consider how and whether the documentary evidence will serve specific research purposes and acknowledge that documentary sources cannot be used as substitutes for other types of data. For example, researchers cannot learn, through records alone, about the day-to-day operations of organisations (Atkinson and Coffey, 2004; Bowen, 2009).

While document analysis can serve the purpose of verifying findings and validating evidence from other sources (Yin, 2014), researchers can also, through the examination of interim and final reports, use documentary evidence to track organisational or process changes and development and to gain a clear understanding of how a programme or an organisation performed over a specific period of time (Bowen, 2009). Document analysis supports the researcher to 'uncover meaning, develop understanding, and discover insights relevant to the research problem' (Merriam, 1988, p118). The rationale for adopting document analysis in this study is in its methodological role for data triangulation and in its capacity to provide understandings of how and why patient flow has improved over time in the case study organisations under examination. Quantitative data in this study highlighted the comparison of patient flow performance before and after the intervention, involved the evaluation of the LOS (e.g. Run chart), and the main reasons for delayed discharge or delayed transfer. Qualitative data was used to establish the reasons for failure to discharge patients in a timely manner, the challenges encountered in service delivery, and the outcomes of interventions to improve the process.

Observation

Observation was adopted in this research in order to provide further support in gathering data. The study used the observation method to achieve a better interpretation of the data gathered and to document what had been observed. Observation is one of the most effective techniques to help researchers to interact directly with the research community and their informants (Spradley, 2016) Through observation, the researcher can identify the information that has not been revealed by the informants and, therefore, the researcher can be aware of any misrepresentations or uncertainty (Marshall and Rossman, 1995). Observation is a powerful data collection technique that provides the opportunity to improve and update the research questions and hypotheses (DeWalt, 2002). The major disadvantage of observation is researcher bias (Patton, 1986; Woodside and Wilson, 2003). Therefore, the researcher has combined multiple data collection techniques to overcome these limitations.

The instruments used in this research are explained in more detail, case by case, as follows.

Case 1

Data were gathered through semi-structured interview questions with staff members of different teams who were involved in the transformation of the TOC project. A total of 6 interviews were conducted between July 2018 and September 2018, followed by a further interview with the Associate Director of Quality Improvement to cross-check the findings and to investigate his opinion regarding the application of LM and TOC in healthcare. Other sources of data collection, such as documentary evidence that highlights the improvements through a comparison of performance before and after the intervention, was limited in this case, as the Trust provided the information from their computers, which are password protected and require authorisation to access such evidence.

Case 2

Data were gathered through semi-structured interview questions, observations and documentary evidence. Interviews and observations were arranged and were undertaken at the Trust location from October 2018 to April 2019. A series of semi-structured interviews were conducted with approximately eight staff members at different levels of the organisation who had experience of a TOC approach, to reveal the influence of the implementation of TOC throughout the Hospital Trust.

The researcher was able to observe the application of TBM to manage the patient flow through the hospital in March 2019. The event involved three staff members, including the Transformation Manager, Inpatient Matron, and Matron Assistant. Those staff have lengthy experience of the application of TBM associated with the TOC approach. An overview of the application of TBM was introduced by the staff. The TBM was used in this hospital through computer software called 'Jonah'. The functions of TBM were explained through the computer screen, and the researcher took notes to record what he observed. This observation was very important, as this case had implemented the TOC solution for more than 11 years, which helped the researcher to understand specifically how TBM managed patient flow effectively in a complex healthcare environment. Terms like PDD and DTOC were discussed with the staff, and the researcher was able to observe how the system dealt with these terms. Also, the researcher observed how the data was generated from the Jonah software and how it helped the staff to identify the causes of delays and ensure the flow of patients was kept safe from any interruption. Details of this observation are used to triangulate the evidence, that is, to compare the results of one data source with those of another data collection instruments, such as interviews and documents obtained (Kawulich, 2012).

Documentary evidence in the form of a run chart published by the Trust was the third source of data. This data showed the improvements recorded through a comparison of performance before and after the intervention. Documentary evidence in the form of Pareto data provided by the Jonah software was obtained, showing important indicators, such as top reasons for delays and the improvement in LOS over time.

Case 3

This case is similar to case one, as the data was only obtained through the use of semi-structured interviews, with a total of 5 members of staff who were involved in the productive toolkit project (LM tools). As this was the only case available to evaluate the application of LM, the researcher, therefore, conducted secondary research on the implementation of LM in healthcare, in particular through peer review of the successful examples of the application of LM in healthcare.

Case 4

The data collection for this case was conducted between June 2019 and October 2019. A total of 10 interviews were conducted with different levels of staff involved in the new discharge project to improve patient flow performance across the hospital. The interview questions followed the same criteria applied in other cases: semi-structured interviews, for example, aimed at exploring the phenomenon under investigation (e.g. before the intervention, the nature of the intervention, after the intervention, results, sustained performance) which were then compared with the other theoretical sampling that applied

TOC and LM solutions. This enhanced the understanding of the impact of different solutions on managing patient flow in various health and social care environments, as well as of the extent to which such solutions can effectively manage and reduce instability (variation and uncertainty).

Document analysis was available to evaluate this intervention, as it was open to the public domain; however, as per the commitment to the ethical procedure of the Trust, the name of the hospital has been kept anonymised. This helped the researcher to strengthen and validate the data obtained from the interviews, as well as to obtain more data on the benefits of this approach to the hospital.

The researcher was able to do some observations by attending several discharge meetings that involved different staff across the hospital. Filed notes were used to write down what the researcher discovered from listening and seeing how the hospital deals with patients who delayed discharge and what mechanism is used to enhance patient flow. Such notes were used later to recall what was experienced in this hospital environment and then triangulated with other recorded data sets. These observations were important because they helped the researcher to see how this hospital managed the patient flow, which could then be compared to other cases evaluated in this research. Some of these meetings were on a daily basis at the E&A department, aiming to review all patients who might not need to stay in hospital and who might need support from the local authority team to enable the discharge process. One of the important meetings observed by the researcher was the 'Hub meeting', which involved relevant people across various organisations, including representatives from social care, CCGs, the hospital discharge team, community care, and the local authority team. The meeting aimed to discuss patients experiencing delayed discharge by reviewing their record on the computer screen. The meeting leader displayed those patients one by one. This screen showed some data relevant to those patients (e.g. PDD, DTOC, ward number, location, notes, etc.). The attendees discussed the cases of these patients, and each party made its statements and determined its responsibility to participate in solving the problem. At the end of the meeting, an escalation action plan was written on MS Word and sent to the relevant people by e-mail. It was observed that the hospital implemented different software programmes to enhance patient flow performance and these programmes were different to the one observed in Case 2. Therefore, this observation was really essential to better interpret and triangulate the data gathered, as well as to better understand how flow can be managed with different techniques and the limitations of such techniques.

4. *Entering the field*

Following ethical research protocols within the graduate school and then the NHS sites, the data collection for each study site was performed and completed through several phases. The consent forms before the start of the interview were signed by the respondents, and they were informed of the purpose of the study and other ethical issues, such as confidentiality, anonymity, and the right to withdraw without giving reasons. They were also informed that the interview would be audio-recorded after their consent. Before the end of the interview, participants were asked to fill out a form to confirm whether or not their name could be used in the study. After the initial analysis of the data was completed, it was sent back to the participants for their review. Each set of case findings was then sent to at least one of the main informants for cross-checking and for permission to use the results in this study.

While the early review of Case 1 showed extensive data on the implementation of TOC and how it helped improve the patient flow and minimise the waiting times of patents, the researcher was open to new opportunities to expand his understanding of the phenomenon being studied. Ongoing data collection and analysis allowed the researcher to make some adjustments during the data gathering process (Eisenhardt, 1989). For example, the initial case analysis of Case 1 showed that there were some challenges in interviewing people who were currently involved and working within the therapy service (Case 1) as their schedule was always busy. Data were, therefore, collected from participants who took part in the transformation of the TOC intervention in 2015 to improve service delivery of Case 1. Thus, the findings of this case were based on evidence from participants who were no longer involved with the daily activities of the current therapy service. It also appeared that the evidence regarding the system's performance before and after the intervention was inaccessible. This, therefore, led the researcher to identify and include a new case (Case 2) as a suitable alternative to better interpret how the application of the TOC managed patient flow. Case 2 had adopted a long-standing TOC solution, and other data collection sources, such as observation and the archiving of records, could be easily obtained.

In fact, including rich cases that allowed the researcher to use multiple data collection methods and evidence sources offered the researcher the ability to gain insight into the phenomenon from different perspectives.

5. Analysing data

Eisenhardt (1989) suggests the need for researchers to closely familiarise themselves with each case as an independent unit to draw unique patterns from each case before a general pattern is drawn across all cases. As mentioned in the first step of Eisenhardt's 8-Step Framework (section 3.8), Eisenhardt suggests while identifying and defining the research questions, it is essential to develop possible constructs as it allows the researchers to measure and investigate these constructs accurately.

The researcher follows two steps to analyse the data: within-case analysis and cross-case analysis. Through this process, the researcher can explore the relationship between constructs and variables. Unlike Eisenhardt, Yin (2003) suggested five components of research design for case studies as follows:

1. a study's questions;
2. its propositions, if any;
3. its unit(s) of analysis;
4. the logic linking the data to the propositions; and
5. the criteria for interpreting the findings (Yin, 2003, p.27).

This research follows Eisenhardt's 8-Step Framework, which is different from Yin's research design, and therefore, the unit of analysis is referred to as constructs and measures used in this study. The analysis of case study data involved two steps as follows:

(a) Within case analysis

This step involved the development of a case study database and of a logical chain of evidence. The development of a database for each case was undertaken through the collection of reflective remarks from the observed interview notes, the codification and extraction of data from the interview transcripts, the grouping of extracted data to identify patterns of data, and descriptive statistical analysis of statistical data. This involved analysing the effectiveness of the flow mechanism on reducing or managing variability and uncertainty during the patient journey in these organisations. To evaluate how such flow mechanism has been applied effectively to manage patient flow, the coding scheme for this study was divided into the following categories:

- Motivational conditions for intervention
- Nature and techniques of intervention
- Criteria for intervention success

These categories are a sample from the coding scheme established in this study, which is clarified in Appendix 2.

Following this step, a case protocol was developed to writeup and organise each case supported with a logical chain of evidence. This first step was to determine what prompted the need for change. An in-depth analysis of the intervention's motivating conditions revealed challenges for the healthcare organisation as well as problems with the original system design. The analysis then described the flow mechanisms used to address these issues and challenges. The extent to which flow mechanisms resolve problems and enhance patient flow was analysed. This entailed demonstrating the effectiveness of a flow improvement intervention in actively managing the instability and successfully minimising the LOS/delayed discharge. The quantitative and qualitative evidence expressing or demonstrating the system's performance before and after the intervention was then analysed to reflect the results and sustainability of the flow improvement intervention. Examples of reflective quotations from transcripts are shown in Appendix 3.

Each interview was transcribed and written in an MS Word document and stored on a password-protected computer to reduce the risk of unauthorised access. Each transcript file was given a unique and coded name and then placed in a case-specific folder (e.g. Code P1/Folder C1). Participants from all the cases and others, including senior managers and consultants, were numbered P1 to P32. Each transcript of the interview was reviewed while listening to the audio file. This procedure was repeated over and over again, to reduce the risk of missing any crucial details. It was, therefore, necessary to regularly listen to the interview record and read the transcript, in order to understand and interpret precisely what the participant said. For all interview results, any variations or similarities were continuously identified through comparison with the previous one.

Each transcript was written and characterised based on the design of the semi-structured interview, which reflected the logic flow to get an in-depth understanding of the phenomenon being studied. For example, 'issues before the intervention' was created as a heading and followed by subheadings related to the original design of the evaluated system before the intervention. Each key point was highlighted and emboldened in the transcript. Key points such as delayed discharge, waiting times, length of stay, DTOC, uncertainty, variation, and complexity of pathways were highlighted in a yellow colour. Interesting quotes were highlighted in a red colour. An example of an anonymised transcript can be found in Appendix 4.

Once all transcripts associated with a particular case were codified, a separate report was developed and used to write up each case individually in a structured and unified layout in the form of story-telling. More specifically, each case was written up based on the following criteria:

- Case background
- Before the intervention
- Nature of the intervention
- The intervention
- Results
- Sustained performance

MS word was chosen as the data analysis software for the study, given its universal accessibility, its ease of use, its speed, and its overall convenience, compared to other software, such as MaxQDA and Nvivo, which have become commonplace in qualitative research. The use of such software can simplify the counting of codes if the researcher, for example, decides to do so. It can also facilitate the development of complex stratified codes, organised across nodes in various layers. It is, therefore, necessary for the investigator to use the software to understand and interpret the meaning of a statement and not only to bring up data and display the statement itself. Richards and Richards (1994) argue the use of simple software coding by a form of 'code-and-retrieve' or indexing (p. 168) is problematic; these codes can be generated and defined easily, while the 'more vulnerable and tentative ideas emerging from the data are harder to incorporate into ordered categories' (1994, p. 168), so that the indexation process overtakes the more significant analytical coding. It's also easy to interpret details in a way that suggests that you don't have an overall perspective on what's happening (Elliott, 2018). In this study, the researcher found that NVivo 11 was useful in structuring and organising the data, but it was thought that NVivo would not take over the researcher's analysis process (McLafferty, and Farley, 2006). In this way, the researcher believes that data analysis relies heavily on the researcher's way of thinking in making use of the data to meet the research objectives. The individual case analysis is represented in Chapter 4.

(b) Cross-case analysis

In this step, similarities in different cases and differences in similar cases were explored, in order to establish patterns to help the process of cross-case analysis. The relationship between constructs and variables was defined through this process. This step is, therefore, important for internal validity to be established. Chapter 5 of this thesis addresses this step in more detail.

6. Shaping hypotheses

This step aims to compare the new theory systematically with the data from each case to determine how well or improperly it matches the results. Following the process of cross-case analysis, a causal relationship starts to emerge between developing constructs, together with evidence in each case. The underlying principle behind hypothesis shaping is to continuously compare the data with the theory so that the more evidence supports the new theory, the more valid the theory will be. This was achieved by refining the definition of the constructs and their relationship with each other by using evidence. This step helps to establish the validity of the constructs.

The process of shaping the hypothesis started in the early phases of the data analysis and was only subsequently linked to current literature. In the early case analysis process, a replication of the same concepts in the context of managing a complex patient flow formed the hypothesis and refined the constructs in this research. This contributed to the creation of a new theoretical model, which explores the means of managing different levels of instability in the health and social care systems, to improve and maintain patient flow.

7. ***Enfolding literature***

After shaping the hypothesis, the researcher started to uncover the relationship between the existing literature and the emergent concept or theory (Eisenhardt, 1989). In this step, the researcher reviewed literature that might support or conflict with the research findings or the emergent theory. Comparing findings with similar or conflicting literature raises the theoretical level and strengthens the validity of the study. Finding similar literature strengthens and increases the external validity of the study and extends its generalisability while finding conflicting literature builds internal validity and sharpens the definition of constructs.

Following a comparison with existing literature in the context of the application of LM, the findings were found to be only contextually different. They point to similar results, but not in the same direction. For example, some previous studies indicate the failure and limitations of the use of LM in healthcare environments, in contrast to what this research has found. Although previous studies have acknowledged the influence of LM and TOC approaches in healthcare, this research has explored the conditions associated with these approaches in order for them to be adapted effectively to meet the needs of different healthcare environments.

Overall, similar and conflicting literature was incorporated into the literature review to support the validity of this research and to help refine concepts and emerging theory in the context of conflicting literature. This step is commonly explored in a cross-case analysis discussion chapter.

8. ***Reaching closure***

When the iteration between theory and data does not lead to contradictory opinions but rather increments the new theory, and when the study is technically filled with case studies, the researcher can close the research processes (Eisenhardt,1989).

Obviously, the researcher was informed of the theoretical saturation through continuous comparisons and correlations between data and theory through analysis both within and across cases. Through a consistent contrast between evidence and theory, as stated in the early part of step six on both the basis of within and cross-case analysis, the researcher established theoretical saturation. After data collection and the analysis of all the cases, the researcher was able to use the findings to develop the new theoretical model. It was therefore concluded that the research questions had been answered sufficiently and that there was no need for further cases to be added.

This investigation involved four case study projects, some of which were planned in advance, while others were added as the research progressed. Therefore, the number of cases in this study was consistent with the recommendations and met the saturation standards suggested by Eisenhardt (1989), Meredith (1998), and Miles and Huberman (1994).

3.9.Validity

Although the notion of a research study's validity can be viewed from both the quantitative and qualitative perspective, the meaning of validity can be different in the context of quantitative research than it is in qualitative research. According to Fisher (2010), validity refers to the extent to which a study addresses the objectives of the research and to the truthfulness of the findings generated by the research methodology (Creswell, 2003). Golafshani (2003) and Seale (2004) argue, however, that in qualitative research the concept of validity is not as significant as in quantitative research, claiming that the validity of qualitative research methods is ambiguous since the concept is founded on the research processes and intentions. Despite the arguments of Golafshani (2003) and Seale (2004), it can be argued that the validity of qualitative research remains vital for studies that seek a robust way of addressing their aims and objectives. This view is expressed by Flick (2009), who argues that, in the context of complex circumstances, research credibility can be considered an acceptable substitute for the concept of validity. Regardless of the arguments about the significance of validity in qualitative and quantitative research, Fisher (2010) suggests that the concept of validity should be considered from two perspectives, that is, in terms of external and internal validity. The extent to which the research supports the findings and conclusions of the study is considered internal validity, while the extent to which the findings from the study can be generalised is considered external validity (Seale, 2004). With the perspective of the arguments highlighted by different scholars, this study addressed validity through its analysis of study data within and across cases. The adoption of a robust research design, in line with the 8-step practice advocated by Eisenhardt (1989), was supported by a comprehensive and critical review of the literature on patient flow and the TOC and LM, with a view to improving the credibility of the study. The triangulation of data sources and methods in this research design also sought to complement and cross-examine the findings of the study from different viewpoints, thereby enhancing its validity.

3.10. Reliability

Whereas validity measures the degree to which the aims and objectives of the study are addressed by the research, Handley (2005) suggests that the related notion of reliability seeks to measure the degree to which there is consistency in the findings over time and the degree to which similar findings can be obtained when the research instruments are administered under the same conditions. Joppe (2000) defined reliability as the degree to which the study findings are an accurate representation of the total population, are consistent over time, and can be reproduced if the same methodology is administered. Considering reliability in the context of qualitative research, Seale (2010) posits that just like the notion of validity, reliability is sometimes dismissed in qualitative studies, with scholars favouring other approaches. Gibbs (2007), however, considers reliability in the context of consistency and claims that, when viewed from the perspective of consistency, qualitative studies can still demonstrate some degree of reliability. The triangulation of results and data in this study and the cross-examination of findings within and across case studies were undertaken to ensure the reliability of the research. The transcription of interview data to analyse standardised data was a further step in promoting the reliability of this study. The standardised transcription of data collected from respondents to the interviews represents data reliability from a qualitative viewpoint (Gibbs, 2007), while, based on the overall research design, the entire study can be considered reliable from a quantitative standpoint. According to Spencer et al. (2013), the researcher can demonstrate a study's reliability by adopting dependable and recognised approaches to data collection and analysis.

Based on the research design adopted, the steps taken to ensure the validity and reliability of this research are summarised below in Table 15.

Table 15: Steps Taken to Ensure the Validity and Reliability of this Research

Criteria	**Steps followed**
Internal Validity	- Cross case analysis - Cross-checking the case analysis report by key informants - Sufficient quotes within individual case analysis - Enfolding findings with conflicting literature
Construct Validity	- Using a variety of methods for collecting data - Cross-checking the case analysis report by key informants
External Validity	Enfolding findings with similar literature
Reliability	Adoption of robust research design approach: - Using Eisenhardt's 8 step model - Using case study protocol and database

Having outlined the approach, and the research design underlying the study, the following chapter describes the individual case analysis of the chosen four cases in this research.

CHAPTER FOUR: INDIVIDUAL CASE ANALYSIS

The following chapter outlines the outcomes of all cases examined in this research. These cases provide a detailed analysis of the nature of LM and TOC or other interventions to improve patient flow based on the perspectives of staff at a different level of the organisation and through the use of documentary evidence and observation. As discussed earlier in section 3.8., four case studies were selected across three NHS Trusts based on their theoretical value. All cases have introduced interventions that have contributed to patient flow improvement. These cases represent different levels of instability involved in various healthcare environments, including acute and rehabilitative hospital care, social care, outpatient services, GP-led community care, and community mental health.

The outcomes of each evaluated case study are categorised into six parts. The first part discusses the background of the cases under investigation and the challenges they face, in terms of types of provided care services, issues highlighted the need for improvement (e.g. increasing patients waiting time), and the levels of instability in the medical path along the patient's journey. Part two discusses the organisation's original system design and performance before the intervention, clarifying the process and the issues with its original design and execution that induced inadequate patient flow. Part three refers to the nature of intervention applied in the Trust and the basis of the selected intervention (e.g. consultancy leading the change). Part four relates to the intervention with respect to the focus, alignment to specific approach and the mechanism used to manage the flow. Part five discusses the results obtained from the intervention, for example, LOS reductions and delayed discharge. The last part considers the degree to which the interventions delivered the expected outcomes and whether the intervention resulted in sustained improvements.

4.1. Case 1

4.1.1. Case Background

Case 1 offers services to children between 0 and 18 years in communities and acute hospitals across a single large English county. The therapy is provided to children that meet the service criteria to be eligible for assessment, diagnosis, and support, and this is provided from different areas depending on the needs of the children. The community teams are usually the first contact point, and access to further support is determined by the needs identified. Therapists working across the county are responsible for assisting children with more specific and complex needs, which include autistic spectrum disorders, dysphagia, learning difficulties, hearing impairment, stammering, and physical difficulties.

Since waiting for treatment has been an issue associated with the NHS, the government has set a treatment target for all patients to receive treatment no later than 18 weeks after referral. According to the Foundation Trust (FT) target for this service, the patient should receive an initial assessment within 13 weeks of referral and be treated within 18 weeks. Achieving these targets was challenging, and patient waiting time increased, generating huge waiting lists and making it difficult for therapists to manage. This is evident from the interview responses suggesting some therapists considered that nothing was wrong with the system, apart from the patients' complaints about the waiting list.

This case can be classified as planned care, where a service provider requires patients to book appointments to visit therapists. The nature of the treatment and recovery times in this environment is varied and uncertain, as each patient's needs differ. This type of service faces a high degree of variation and uncertainty in demand, as well as instability in the overall load on different resources throughout the system.

4.1.2. Before the Intervention

Under the original system operated in the service provider, the therapists would consider initial assessment appointments and treatment bookings based on the medical needs of the patient. Patients' waiting times were managed according to the severity of their speech and learning difficulties. Therapists had their own independent patient waiting lists, managed according to their schedule. This did not offer greater visibility, as the therapists booked patient appointments in their personal individual diaries.

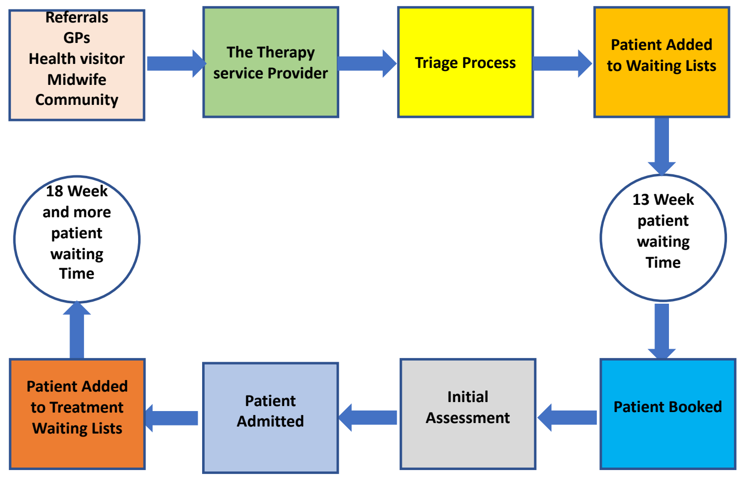

Figure 16: Process before the Intervention

Figure 16 shows the primary process in the service provider system before the intervention, the course by which patients typically move, and the points in the system where the patients are made to wait much of the time, before moving onto the following process or procedure. The diagram indicates that patients are referred from different health centres to the service provider. The therapist receives the referral and checks if the patient meets the service criteria to be eligible for assessment within a week's time. Once the patient meets the criteria for therapy, the patient will then be added to the waiting lists and would wait around 13 weeks for an initial assessment, according to the FT target at that time.

The therapist will decide which patients to choose from the waiting lists according to the severity of their speech and learning difficulties and their waiting time. The patient will then be invited to an initial assessment appointment within 13 weeks, in accordance with the FT target. Based on the therapist's assessment, the patient will often be admitted for out-patient treatment and added to the waiting list. Once the patient has been added to the waiting list, the patient will wait for a treatment appointment. An 18-week government target applied to this waiting time.

With demand increasing, the service provider had growing waiting lists and was no longer able to meet these targets. According to participant P1, there were around '171 different waiting lists contain[ing] more than 2000 patients'. Another participant mentioned that 'one team of therapists with 7 members was responsible for managing a caseload of 450 children' (P4).

After the initial assessment, therapists classified patients as low, medium, and high priority to receive treatment based on their severity of need. So, priority was based on severity and time. If the patients were classified as low priority, it meant they had a delay of around six months (26 weeks), while patients who had a delay of between six and twelve months were classified as medium priority. Patients who ranked as high priority had a delay of twelve months or above.

The analysis found that patients could not get an accurate estimation of the expected waiting time, and Participant P4 believed that giving a treatment plan after an initial assessment appointment was difficult because of the variation in waiting lists.

> It is very challenging to estimate how long the patient stays in the waiting list for the next treatment because there were a thousand children on these waiting lists from all over the county and they were all mixed (P4).

4.1.3. Nature of the Intervention

The process improvement initiative was led by QFI Consulting; a consultancy firm which specialises in solving healthcare problems through the TOC approach. QFI conducts a cause and effect analysis of the wider organisational delivery system to understand the whole system. QFI are specialists in developing generic TOC manufacturing solutions and tailoring these to healthcare and social care environments. In many NHS organisations, it has provided programmes in the form of an aggregated planning and control system, supported by its distinctive Jonah software and achieving significant and sustainable improvements in patient flow within weeks.

4.1.4. The Intervention

In April 2015, the Trust decided to contract QFI to offer a solution in order to improve and enhance patient flow in the service. The QFI consulting team worked with the Trust's quality improvement team to provide the service staff, including clinicians, nurses, frontline staff, and admin staff, with training sessions involving the TOC concepts and their application. The change was delivered in two stages. This was explained by participant P1, as follows:

> So, phase one was about the baselining, the initial education, and understanding the opportunity. Phase two was those four elements around measurement, the education, the packages of care, and the process improvement to make things more visible (P1).

To identify the causes of the increase in waiting lists and to identify the areas for improvement, QFI conducted a cause and effect analysis of the wider organisational delivery system, in keeping with their usual practice. They produced a simple Excel-based system into which the Trust exported information from System One:

> [This showed data] where the appointments were being used, where they were being cancelled, where were we wasting the appointments and things like that (P2).

QFI and the quality improvement team used the data to identify the following problems with the current system:

- The absence of a structured form of service provision was associated with increased waiting times and long waiting lists.
- The average time spent by patients waiting to receive therapy was between 6 and 8 months; sometimes, it was a year or more, with waiting lists of over 2000 patients.
- The data gathered indicates an ineffective booking system, as therapists had their own independent patient waiting lists, managed according to their personal schedule. This inhibited visibility, as therapists booked patient appointments in their own individual diaries.
- Some therapists were working part-time, which presented a challenge for capacity.
- The existing policy prevented patients from booking appointments outside the area where they lived, even in cases where therapists worked across various health centres.
- An evaluation of the causes of delay showed that delay factors included the sickness of staff members, authorised annual leave, and study leave. However, delay factors also included some time that was just not accounted for, with staff focusing on report writing and completing paperwork.

During the time QFI participated in the process improvement, they created a single waiting list for initial assessment, based on a first-in/ first-out approach. Once the referral was received, the patient was added to this waiting list, and the parent or guardian was sent a letter or text message from the Trust within 24 hours. Once they received the letter, the parents/guardians had to phone and request an appointment within two weeks; otherwise, the patient would be officially discharged from the appointments system. If the parents responded within two weeks, then the child would be booked for an initial assessment and would be seen within two weeks. QFI also suggested that therapists offer a short treatment session during the initial assessment appointment and that they book the whole package of care immediately after the initial assessment appointment if the patient needed to be seen for treatment. QFI also recommended allowing patients access to services and treatment from different districts within the county as part of this process.

Figure 17 illustrates the service provider's process after the intervention. This new process enables therapists to pull patients from the new initial assessment waiting list at a level consistent with market demand on a first-in/ first-out basis. Once the patient has entered the list, the booking is made two weeks before the date of therapy. An increase in therapist capacity for a short period (around 3 months) enabled the old waiting lists to be dealt with to meet the two weeks target.

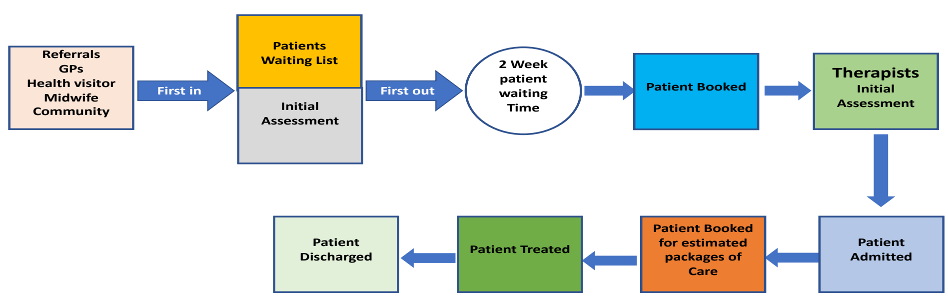

Figure 17: Process after the Intervention

The overall performance of the system was monitored and reviewed through a weekly conference phone call between managers and team leaders. It was evident from interviews that a weekly conference call between managers and team leaders was necessary to review the performance of the system. This process is part of TOC/TBM, where patients and tasks are expedited to deal with any issues influencing the progress of patient appointments. This helped in identifying areas that required improvement and, where necessary, escalated recovery actions before the delivery system was destabilised. According to Participant P1:

> From each meeting will come actions the different individuals were going to take, which was going to drive down the waiting time. It was a visual representation of the waiting times and capacity. (P1)

Within an 18-month period, QFI provided the Trust with training and education about TOC and the proposed adoption of TOC-based software they had developed. The Trust's System One provider, however, would not allow the software to be integrated, as System One is a nationwide system used in most NHS services. Therefore, the Trust preferred to use the same system and to see if this system could be upgraded to include new features used in the Jonah software. As the evidence shows:

> There was reluctance from the IT department to embrace another type of software on top of System One……We tried to work with the Phoenix Partnership (TPP), who are the System One providers. That proved very challenging because they wanted one system, the whole of the country… So we ended up there. A lot of the measurement was offline, so it was Excel spreadsheet type things, which wasn't ideal. (P1)

Since the suggested Jonah software could not be integrated with the Trust's System One, QFI created a spreadsheet that helps staff produce data about delays and monitor system performance based on the Jonah technique. As Participant P1 highlighted:

> We were producing all the monitoring graphs around delays: to what extent are people stacked up, to what extent are things delayed, are we in the green zone, the amber zone, the red zone, the black zone…….maybe it would have been better with Jonah because it is simple to us. (P1)

The use of the spreadsheet allowed staff to measure and monitor the time of the overall patient journey once the patient was admitted. As Participant P3 commented:

> From receiving a referral to [being] registered on the system, to triage, to book an appointment, to attending that first assessment: each of those steps were measured, but we monitored what was waiting [time at] each of those stages. And we set up buffers and understood whether they were in kind of red, amber, green in terms of where they should've been. (P3)

4.1.5. Results

Following the implementation of the TOC management approach within the NHS Trust, there was a reduction in waiting lists and in the length of time patients spent waiting to receive therapy. By adding additional capacity for three months, the Trust managed to streamline the new process and the waiting period for the patient to obtain an initial assessment from the therapist dropped from 13 weeks to 2 weeks. According to Participant P1:

> What we did was we took one of the part-time staff from the service administrators… and she did some work for us for three months in some of her days off to reconfigure some of the staff, so that we could move to streamline that process. (P1)

By moving from managing multiple waiting lists to manage a centralised single waiting list, the service saw improvement in patients' waiting time within six months of implementation.

While there appeared to be a lack of capacity, a closer look at the system revealed that there was considerable unused capacity. As Participant P3 pointed out:

> What they said is it was a lack of capacity, but when we actually looked at it, it was not lack of capacity. It was more around the organisation of it, so what we did was we made everybody use an appointment system on System One that was already there. (P3)

Following the intervention and recommendations made by QFI, the Trust managed to get rid of the individual therapists' diary system and managed patients' bookings through an electronic diary. According to Participant P5:

> Making appointments visible, drove up the amount of activity. So, the staff did more appointments when their appointments were visible. (P5)

Having a centrally managed electronic diary, where appointments were booked, improved visibility and flexibility, allowed the Trust to ensure capacity was used and highlighted where there was additional free capacity that could be used. This was not the case before the intervention, as highlighted by the interview evidence that, while some therapists were doing a great deal of work within their capacity, 'other therapists were doing half as much of the appointments.' (P5)

The interventions created a considerable saving of time eliminating the batching of documents for scanning, filing, and paperwork. Rather than referrals being batched the next day and processed in batches (which staff considered the most effective and efficient use of their time), QFI recommended processing all new files immediately upon receipt, in order to respond to patients' needs in the most timely manner. According to Participant P1:

> The way we were batching changed to not batching like before and this was sustained. There was that individual perspective of what suits me, that, as an individual, I will work quicker if I've batched these things, but there was an impact on the next person. So, that improved, and that has been sustained. So, we're saying serve the flow and sacrifice the time. (P1)

The reconfiguration of System One to allow and support the input of measured therapeutic outcomes made data entry much more accessible and efficient and improved the time available for patients to be seen and, hence, managed patient flow. The interview evidence shows reductions in patients' waiting times for an initial assessment and for treatment, as well as a reduction in the patient waiting list.

4.1.6. Sustained Performance

While the intervention was met with strong resistance and did not continue smoothly after the initial 18-month period when QFI was involved directly, the process did bring changes to service delivery. There was a shift in culture, whereby transparency around appointment bookings and treatment was 'a big and sustained thing'. Participant P1 explained that 'it's now normal to offer assessment appointments out of the area, and it depends on the area: the more affluent area, the easier it tends to be for people to travel out of the area'. In addition to the transparency, there is now more flexibility in the system to adapt to variation in demand across the healthcare system and greater visibility of who is doing what.

Although there was an initial reduction in patient waiting times following QFI's intervention in this service, the current situation exhibits waiting times within the eight weeks target, rather than the initial two weeks achieved under QFI. The findings are, however, based on evidence from participants who are no longer involved with the daily activities of the current service provider. This limits the opportunity to evaluate the improvement generated through the QFI intervention effectively, raising a need to interview people who are currently involved with the system. Evidence regarding the system's performance before and after the intervention was inaccessible.

4.2. Case 2

4.2.1. Case Background

Case 2 is a part of 12 community hospitals offering urgent care, rehabilitation, and mental health services to a population of approximately one million across the county. The county's Trust is one of the largest community healthcare providers in England, with 2,400 staff based in over 60 locations across the whole county. The Trust provides patient care across the county in many sites, including the 12 community hospitals, day hospital services, a learning disabilities centre, out-patient facilities, intermediate care beds, and 25 health centres. This case can be classified as unplanned care, as a patient can be referred immediately to the community hospital for rehabilitation, once the patient meets the required criteria. In this kind of environment, therefore, demand is unpredictable. The varying and uncertain nature of patients and their treatments prove unstable, and this has also created overloading on the various types of resource.

4.2.2. Before the Intervention

In 2007, a managing director recognised high patient flow rates as a critical problem for the Trust and acknowledged a need to decrease the length of in-patient stays (LOS) and the occurrence of Delayed Transfers of Care (DTOC). At that time, the average LOS was 65 days, reflecting the high percentage of older people in trust care, many with mental health problems, who may have exceptionally extended hospital stays.

4.2.3. Nature of the Intervention

The intervention was undertaken by QFI Consulting, a consultancy company providing a service improvement solution based on the TOC methodology and its application, especially in service industries, including the legal sector, financial services, and, most particularly in health and social care. The intervention delivered an aggregated planning and control system, supported by QFI's distinctive Jonah software. This resulted in significant and sustainable improvements in patient flow within weeks.

4.2.4. The Intervention

Before implementing the project, the consulting firm provided an introduction to the TOC and the QFI Jonah approach to staff and managers, so that they could understand how the method works and how it can deliver a sustainable solution to improve patient flow performance. This introduction included a visit to one of QFI's client hospitals to see how QFI-Discharge-Jonah works in practice.

A project manager was assigned by the Trust to supervise and coordinate implementation with the QFI team. The first task of the team was to determine the underpinning constraint, or primary cause, of delay influencing the entire system. This involved analysing the entire process (the current situation before the intervention) by considering how they were synchronising operations to achieve the specified objectives for patients.

Initial analysis found that the problem was not about having enough capacity, but rather the management of the available resources at the right time and right place. According to Participant Q2, the hospital had many therapists available, but they were located in the wrong place. It also turned out from the analysis that there was an absence synchronising activities to meet the defined goals of patients, as each therapist or department tended to work in isolation from others, reflecting the difficulties in managing the discharge process effectively and in a timely manner.

QFI recommended that the community hospital target the improvement of patient flow through systems that synchronise and sequence clinical and administrative tasks. This was to ensure that each patient's care needs were met at the right time, making operations more efficient and reducing their length of stay. Thus, the QFI team introduced a new system called 'QFI-Discharge Jonah', designed to improve patient flow by synchronising and sequencing medical and managerial tasks to meet the care needs of each patient at the right time. Central to this system is establishing a planned date of discharge for each patient within 24 hours of their arrival at the community hospital, with each discharge task synchronised in order to achieve the PDD. The Trust focused on planning the patient discharge date from the moment the patient arrived. According to participant P8,

> One really important thing that has been embedded in the process is that you start planning someone's discharge from the moment they arrive. (P8)

The PDD was set by clinicians on the basis of the medical needs of the patients, as a realistic estimate of the time, they needed to recover and get ready for discharge to their next place of care. Each patient receiving a PDD was then recorded in an active database (QFI-Discharge-Jonah), to which all clinicians, managers, and referrers had access.

This development involved changes to existing policies and procedures, as the focus of staff now shifted to operating from a standard list, managed according to a patient-centred priority system through TBM functions. The QFI team also recommended that all team members use the Jonah system's database every morning to hold daily meetings (huddles) to review the status of each patient to prioritise and assign tasks if necessary.

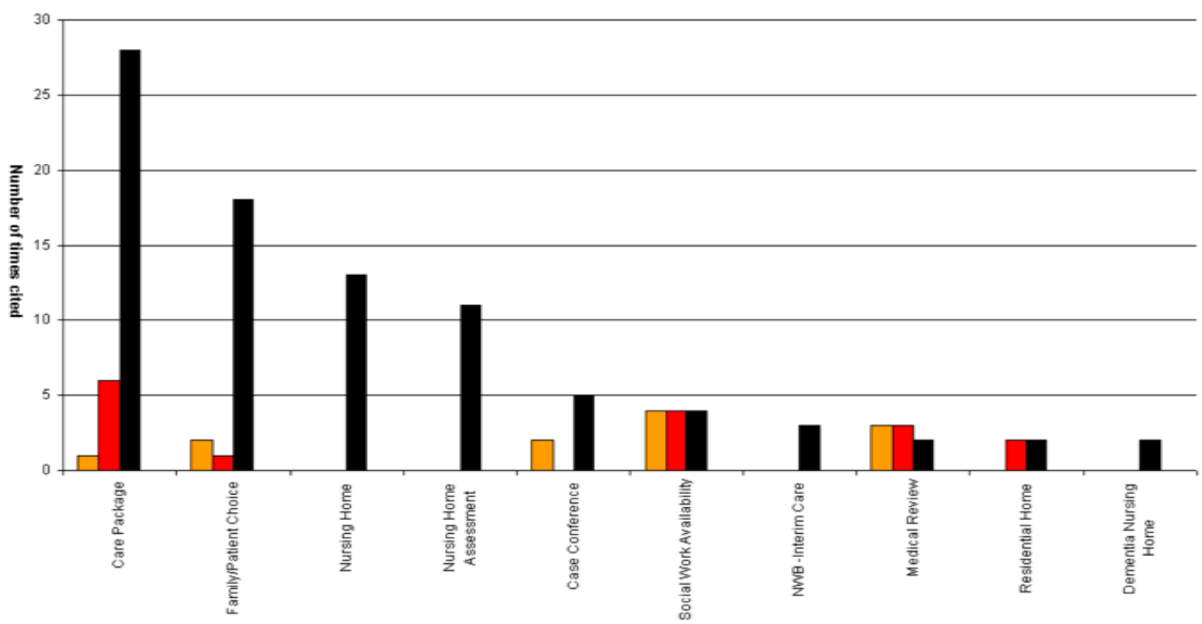

Figure 18: The Top Reasons for Delay

Consistent with the functions of TBM, a multi-disciplinary team (MDT), including ward directors, senior managers, and social workers, then investigates the most delayed patients and causes of delay at the weekly top delay meetings and address these issues through peer-led debate. The MDT reviewed patients who were in the red zone or were already experiencing a delayed transfer through the Jonah discharge database. As illustrated in Figure 18, waiting for care packages was the typical major reason for delay. This Pareto data, provided by the Jonah software, provides focused data to support these meetings.

The Jonah Discharge approach was fully introduced as a pilot at two community hospitals in the summer of 2008 and then rolled out across the remaining ten community hospitals in the Trust.

4.2.5. Results

This TOC approach was introduced at 12 community hospitals, although some have since been closed. The system had been in operation for nine years, and the initial average LOS was drastically reduced by 50% within two months of implementation, from the existing 65 days to 32 days. The reduction in the average LOS continued decreasing, consequently falling by a further 70%, to the current LOS of 20 days.

> Eight or nine years ago, the average length of stay would have been 60 plus days, and I'm delighted to say that now we report exceptions on the basis of any length of stay over 20 days. (P6)

By 2011, all of the Trust's community hospitals had reduced their bed numbers from 525 beds to 370 beds. Today, and through the TOC intervention, the community hospitals have also managed to successfully reduce bed numbers to 148 beds, freeing space on their wards to enhance patient privacy and using less capacity to meet patient needs, as the same resources can be used to deliver better care.

> Certainly, implementing Jonah and the theory of constraints made a major contribution towards making those changes. (P6)

As centralised visibility was enhanced through QFI Jonah, it was possible for the Trust to identify the areas and aspects that required change within the process. Enhancing visibility centrally was consistent with the TOC approach that analyses the process, the organisation, and the industry to identify the elements that require change and to establish the approach to bring about the change that is required to create breakthrough improvements in performance.

The Trust results indicate that, prior to the implementation of this Jonah system, they were not aware of any barriers to flow within the system.

> So, before Jonah, I was not aware of there being any central visibility and understanding whether flow is efficient or whether there are any barriers. (P8)

The participant further said,

> One thing that Jonah did was trying to promote clarity about what we are trying to achieve and how we are trying to do it. (P8)

As a robust patient-centred priority system, the implementation of the Jonah Discharge system helped the staff and management of the Trust to monitor and improve flow by identifying the underlying constraint and then identifying which resources or combination of tasks most often disrupt the patient journey.

> Where you are so extended with different areas, you can't keep an eye everywhere. But Jonah does allow you to do that because you can see what's happening anywhere. (P7)

> The key purpose was to monitor and improve flow…trying to help people achieve their goals as quickly as possible to get through the system as quickly as possible. (P9)

4.2.6. Sustained Performance

After the adoption of QFI's Jonah by community hospitals, all 12 of the Trust's hospitals reported continuous improvements in patient flow and in the efficiency of patient care pathways. The results have been sustained and continue to improve.

The interview evidence, however, suggested that despite the improved and sustained performance in health and social care, the community hospitals now face increased challenges in dealing with complex patient needs, such as mental health issues, which require additional capacity.

4.3. Case 3

4.3.1. Case Background

This healthcare organisation manages care packages for the elderly, frail, or homebound adult patients with a risk of hospital admission. This case involved nurses, occupational therapists, support workers, physiotherapists, and community matrons, who focus on assisting people aged 18 and above to live independently, thereby accelerating their discharge from hospital. Working closely with health and social care workers and doctors to provide patients with the support and care needed, the organisational team provides care services for adults in their homes or living in residential care homes. Working towards the prevention of unnecessary admission and readmission to hospital, community assessors are in contact with patients from local hospitals' medical assessment units and emergency departments in order to refer and direct them to community services. The community services team can then provide services to support recovery, to assist patients in managing their own conditions, and to stabilise patients' symptoms through specialist nurses, occupational therapists, community staff nurses, social workers, and mental health workers, among other specialist care providers. In addition to the home service aimed at supporting patients after their discharge from hospital, the community services also cover clinics within the local area.

This type of healthcare environment is different from that of other health service providers, as services are provided to patients in their homes. It is a planned care environment that deals with stable demand and with patients whose needs are uncomplicated. The patient-facing time involved may vary depending on the condition of the patient, but the overall load of different resources in the organisation is stable.

4.3.2. Before the Intervention

The existing system prior to the improvement intervention was not standardised, with no formal systems guiding some of the processes. The system before the intervention was also characterised by high wastage and lower levels of productivity. The stock room, for example, was characterised by an out-of-date inventory, with employees ordering more stock simply because they did not have access to the resources required. The patient-facing time was low, with non-value adding activities taking up most of the time for nursing and staff.

4.3.3. Nature of the Intervention

The need for change and process improvement was driven by a national agenda to improve service delivery and improve productivity within the workplace. The interventions sought to improve the service to patients, whilst also improving the work environment for staff.

4.3.4. The Intervention

The intervention sought to engage the team and to improve the overall performance of the Trust. The interviews considered the reorganisation of the working environment through the application of crucial success principles relating to structurally reorganise processes and ensure all resources are at the right place at the right time and creating a clean, safe and orderly environment to work in. The whole process involved employees receiving training in nine modules of LM tools, although the modules were designed and presented in the language of the NHS. The first three modules were undertaken under the supervision of the steering group, made up of the quality improvement team and facilitators. With a rollout plan of 18 months, most teams received training in six modules with facilitation and were set to complete the remaining three modules independently.

Productive community toolkits offered a very significant component to create an organised, orderly, clean, and safe workplace, as well as procedures to maintain it that way. This component sought to provide better efficiency and to increase productivity by eliminating waste, standardising processes, and allowing the work to flow. This is reflected in the view of one of the consultants that ‘They're using lean to look at how the environment [can] be better’ (P10).

4.3.5. Results

Although the Productive Community Toolkits project sought to improve overall system performance and engage the team, the intervention also sought to address the working environment, creating a favourable workplace environment that maximised productivity and efficiency. Interview evidence highlighted the following situation prior to the intervention:

> Stock room organisation was terrible, and people would reorder because they couldn’t find something because it was buried in the bottom of other things and things would go out of date. (P9)

Although the system improvement interventions did not directly consider improvements from the flow perspective, the adoption of the Kanban system for stock control, the elimination of waste, capacity management, and scheduling also demonstrated evidence that the adoption of LM tools can maximise productivity and efficiency.

Despite evidence of system improvements from the majority of participants, it was also found that improvements were not widely acknowledged by all system users. A lack of knowledge about LM concepts and its associated tools, such as applying the 5S technique, emerged during the interviews. Some staff saw the Productive Community Toolkit project as being about 'tidying the store cupboard' and, therefore, having no impact on improving patient flow. This was despite the views of other participants, who considered there was a significant improvement in terms of caseload management, the elimination of waste, the standardisation of processes, stock control, and working environment. This, in turn, increased patient-facing treatment time in some communities across the county. By eliminating services and activities that did not add value to the customer, service delivery improvements were noted, which translated to better workflow. After measuring patient-facing time, a key measure of the actual value received by the patient, and observing that the typical customer-facing time within a community nursing environment was approximately 35%, improvement targets for patient-facing time were set at 55%. The increase in the targeted patient-facing time, following the service delivery improvement interventions, indicated the potential for the service to improve its delivery value. Increasing patient-facing time and reducing the time spent looking for equipment and stock and organising the workplace represents value added for patients, and this value-added time can translate to better service delivery for patients. Interview evidence further indicates increased patient-facing time, after distinguishing between activities that were necessary elements of service delivery and those that were not and had to be eliminated.

4.3.6. Sustained Performance

The intervention managed to drive quality improvement and efficiency by applying LM tools through the reduction of waste and non-value-adding services. Adult community services have demonstrated their ability to deliver healthcare services in efficient and effective ways in some communities. The method by which operations have been managed has ensured that the services adopted are influenced by patient experience and choice, and that healthcare outcomes were targeted.

While a national agenda drove this intervention, there was initial resistance within some groups who considered the improvements to be forced upon them without their perspectives being consulted. Although some groups have shown interest in finishing the remainder of the modules on their own, other teams have not welcomed further improvements. It was difficult, therefore, to discuss outcomes, to draw conclusions, and to establish confidence in the findings of this particular evaluation, as the number of participants was very limited. This was in addition to the inability to obtain quantitative evidence to show performance after the implementation of the Productive Community Toolkit project.

4.4. Case 4

4.4.1. Case Background

This hospital serves one of the neediest local authority regions in England and Wales, with a population of 105, 000; nearly 18% of the population consists of people aged 65 and over. The population's life expectancy and health are lower and poorer than the national average. Available statistics indicate the town has the highest record in the county of individuals suffering from a long-term illness, with three in five people aged 65 and over having a long-term illness. The statistics are considerably above the national and regional average.

> It is most probably the third [most] deprived area in the UK in terms of lots of things, such as housing and social deprivation. (P15)

This acute hospital offers services to meet the needs of different patients and includes some planned and unplanned care units. Demand variation and uncertainty can be very high in this hospital, especially in the winter. The instability of the total load on different types of resources is also higher due to the instability of demand over time. The needs and treatment of patients vary and cannot be easily predicted.

4.4.2. Before the Intervention

Prior to the intervention, the hospital was suffering from bed-blocking caused by having patients who do not need to stay in the hospital but are waiting for other support, such as housing support, before they can be safely discharged. These problems led to increased bed blocking and the difficulty of discharging patients from the hospital, which impacted other patients' waiting time to get a bed and obtain healthcare.

The analysis of the interviews demonstrates that there was a lack of communication between the various agencies involved in the hospital, as each side operates independently, with each party attempting not to exceed its budget. Each party believes it is not responsible for the cost of the service to the patient.

As the problems worsened and a solution could not be identified, different parties (stakeholders) decided to meet and discuss the issue, in order to find an immediate solution to bed blocking and enable a smoother discharge process, thereby meeting patients' satisfaction. These parties included the city council, clinical commissioning groups (CCGs), social care, community care, word coordinators, and medics.

4.4.3. Nature of the Intervention

This intervention was led by the local authority in order to reduce discharge delays, by supporting the discharge of medically fit patients from hospital in circumstances where their home or a range of other wider social issues would make it unsuitable for them to return to their home.

4.4.4. The Intervention

In September 2014, the local authority launched the early discharge system to enhance the hospital discharge process by ensuring patients have the required housing assistance when they are discharged. Prior to this, there was a so-called 'perfect week', in the words of the interviewees, where all relevant agencies worked together to see how they could contribute to improving patient flow and getting patients who are medically fit out of the hospital much faster. During this perfect week, the local authority introduced its proposal by offering this new programme. Other relevant stakeholders, meanwhile, evaluated the scheme and agreed to roll it out as a pilot to see the results.

> It was rolled out on a little pilot; I think it was for about six months to see how it works. (P13)

The project was then expanded for another six months until it was fully launched by the local authority with financial assistance from the appropriate CCGs.

The intervention attempted to promote the discharge from hospital of medically fit patients who were in need of accommodation facilities so that they could help to stay secure and safe in their home. The new system recognised a range of ways to achieve this, among them: providing assistance to secure alternative housing; supporting speedy adaptations within patients' homes after discharge; supporting patients with complex needs; and providing temporary housing using local authority housing with funding from the CCG, until appropriate support or accommodation has been organised. Nurses and social workers who organise care and support for patients identify vulnerable patients in hospital before referring the patients to the local authority to ensure patient support following their hospital discharge.

The objectives of the new system included:

1. Improving joint working
2. Tackling bed-blocking
3. Accelerating the transition from hospital to home and improving health outcomes for patients
4. Preventing preventable homelessness
5. Supporting tenants to remain adequately housed
6. Reducing and preventing avoidable and extended hospital or residential care admissions, consequently saving the NHS time and money
7. Accelerating patient discharges from both the emergency department and wards at the hospital and from residential care around the county.

The local authority manages the system through their office based at the hospital. A team of two people from the local authority is assigned to work in the hospital from Monday to Friday, from 8am to 5pm. This team's role is to provide support by rehousing medically fit patients who cannot be discharged because their home is unsafe for them to return to. On a daily basis, the team visits the ED to inspect new patients who might need accommodation assistance, and the team also takes part in the daily meeting at the ED around 9:10 am to review and discuss the patients who do not need to remain in the hospital and may need housing support.

Although the team visits the ED and wards throughout the hospital, the team also provides support when they receive a call or e-mail from any of the hospital staff. For example, when ambulance crews pick patients up from their home, they sometimes inspect the house, and the ambulance crews then inform the hospital staff that the patient's home is not suitable (e.g. hanging wires, water dripping through the roof, or signs of mouse droppings) in case of discharge from the hospital, requiring intervention from the local authority team to make arrangements to enable the patient to leave the hospital.

The intervention sought to provide support to any patient coming in the front door. The team can plan in advance to provide support when they are notified early enough.

> So, they basically try to evaluate any patient coming in the front door, either by ambulance or whatever, if they need something required from the local authority team, they will record and work on it, and they will know in advance. (P20)

The team also participates two or three times a week in a meeting with relevant agencies, discussing the patients experiencing a delayed discharge, and they can offer help if the patients belong to their categories. This meeting involves many members, including social care, community teams, CCGs, the local authority team, and the discharge team.

In addition, the team plays a significant role in helping the hospital to free up beds and to avoid the delayed discharge of patients. As mentioned by participant P15, the team is also sometimes able to recognise patients visiting the ED, which often reflects that these patients are receiving support from the local authority. They may be addicted to drugs or alcohol, which has led to their presence in the ED, and they do not need to stay in the hospital for a long time.

> So that's people that present to ED on a regular occasion. And certainly not our frail elderly, but our say patients, younger patients, they've got drug and alcohol abuse that are in their counsel properties. So, some of these patients have got a history of violence or things like that so that they can tell us up front. (P15)

Those patients who regularly present at the ED can be categorised or grouped as non-health or non-social issues. The local authority team has the expertise and ability to determine patients who come under the 'non-health' category and, thus, prevent hospital beds from being occupied by non-health clients, enabling them to provide care for patients who are in actual need of care.

> They recognise this group of patients that we classified under the non-health category for reasons like this … and they've got the skills and the knowledge, and they know

> the right pathways to support us, because, remember, you know, we're clinicians and nurses, and they're experts in their own field. (P26)

This intervention by the local authority is not only the only contribution to improving the hospital discharge process. There is also coordination across the hospital between different teams. There is a daily board round, between nine and half-past nine in each ward, where all MDT members discuss their patients on the whiteboard to expedite the discharge process.

In addition to the support provided by the local authority team, the Emergency Department Avoidance Support Services (E-DASS) team offers support to the ED, by transferring patients to their homes in cases where the doctor agrees that they do not need to stay in hospital, but the doctor remains concerned about their condition. So, the role of this team is to transfer these patients to their home while checking on them two to three times a day for 72 hours. This, of course, helps to free up some beds to accept new patients.

4.4.5. Results

4.4.5.1. Reduced Hospital Bed Days and Associated Costs

Following the process improvement interventions at the hospital, improvements were recorded, among them, a reduction in the length of the hospital stay, fewer delayed hospital discharges, and reduced readmissions.

Findings from the document analysis and interviews indicated that the early hospital discharge scheme had improved hospital discharge efficiency and reduced the burden on social services and hospital staff. Evidence of efficiency in patient flow was noted, as the study established that the time taken to discharge and rehouse patients and clients within the local district was consistently lower than the time taken in other districts. The document analysis also uncovered evidence suggesting cost savings had been achieved due to a reduced number of excess days spent by patients in hospital. Since the inception of the pilot intervention scheme, it was established that the interventions to improve patient flow under the current scheme had created savings, through reduced bed days, of an estimated £1,142,550 for the entire NHS system.

4.4.5.2. Reduced readmission probability

The interventions and support provided through the early hospital discharge scheme have been vital in reducing the probability of the discharged patient being readmitted into hospital. By ensuring the discharged patients' ability to live independently within their homes and by facilitating patients being easily integrated back into the community, the rates of readmission at the hospital have declined since the implementation of the early discharge scheme. Challenges, such as patients struggling to manage their lives after their hospital discharge, have contributed to the failure in the discharged patients' health, a factor that has been blamed for subsequent readmissions. The discharge of patients without a suitable care package already in place results in the patients' condition declining and requiring their readmission. This scheme has won the NICE Joint Learning Award, whereby the scheme has been able to reduce discharge delays, by ensuring that patients have the required housing support when they leave the hospital. This may be in the form of home modifications or relocations that better meet their needs when they leave the hospital.

4.4.6. Sustained Performance

The patient flow and process improvement interventions at this hospital have created overall efficiency within the entire system. Evidence of overall efficiency improvements and return on investment is seen in the savings generated by the discharge scheme and the overall costs of running the scheme. With the cost of running the early hospital discharge scheme at £340,000 per year and anticipated yearly savings to the system of £1,371,060, the intervention has demonstrated sustained improvements in managing patient flow in the healthcare setting. From July 2015 to April 2016, a total saving of 5,078 bed days was realised, as the local authority interventions through the scheme reduced the delays in patient discharge and improved patient flow within the hospital.

Although this programme has clearly contributed to the positive impact on the flow of patients in the hospital, there are still some obstacles to the sustainability of this programme, especially financial barriers. Local authorities are sometimes unable to provide services to patients, as these services should be provided through healthcare or social care organisations.

In this case, there is an ongoing debate between healthcare and social services providers concerning who will pay for these services, as these patients do not meet their criteria.

> There's a situation where actually patients need to go into a care home, but neither social care nor health pay because the patients don't meet any of their criteria. (P30)

4.5. Summary

This section has presented a case-by-case analysis of four healthcare organisations in terms of the original delivery design, the process issues in the original design, the changes recorded, and the results achieved. Each individual case analysis will be compared with other cases to discuss how the adaptation of manufacturing flow mechanisms has been translated in different healthcare environments. This cross-case analysis is the focus of the next chapter considering the instability reflected in both manufacturing and healthcare environments. An overview of the four cases and their key features presented in this chapter is provided in Table 16 below.

Table 16 Summary of Cases

	Case 1	Case 2	Case 3	Case 4
Level of Instability	High	High	Low	High
Type of healthcare service/ Delivery system	Planned care	Unplanned care	Planned care	Planned/Unplanned care
Patient need	complex needs	It is dependent on the patient's condition. Typically, care is given to elderly patients with complex needs.	Specific and predictable	Mix, however the hospital generally deals with people (aged 65 and higher) suffering from a long-term illness.
Nature of Intervention/Flow Mechanism	TOC/TBM	TOC/TBM	LM/Productive community toolkits	Local authority / early discharge scheme
Timeline of Intervention/Improvement	From 2015 to 2017	From 2009 until now	2009 /ongoing improvement project	Since 2014 to the present
Results	Reduced waiting times for initial assessment from 13 weeks to 2 weeks.	Average LOS reduced from 65 days to 20 days. The number of beds reduced from 525 beds to 148 beds.	Increase staff productivity and patient facing time in some communities.	Reduced delayed hospital discharge and readmission. An estimated saving of £1,142,550 through reduced bed days
Improvement and Sustainability	The intervention show improvement in terms of reducing waiting times for patients, but the evidence regarding the system performance before and after the intervention was inaccessible. Thus, it hard to judge the sustainability of the intervention	Sustainable patient flow improvement for more than 9 years.	Local optimisation	Although the flow mechanism contributed to reducing delayed hospital discharge, local optimisation was found to conflict as there always tension associated with budgetary control between health and social care organisations.

CHAPTER FIVE: CROSS-CASE ANALYSIS AND DISCUSSION

Following the individual case analysis, this chapter seeks to discuss the findings in terms of the relationships between the case studies and the relevant theoretical constructs, folding these findings back into the literature (Eisenhardt, 1989). The overall goal is to advance knowledge on the circumstances associated with the adoption of flow mechanisms in different manufacturing environments and how the adoption of a flow mechanism can be translated to effectively meet the requirements of the health and social care system. Within the cross-case analysis, the relationship between the manufacturing and healthcare contexts is then investigated, exploring which flow mechanism is most appropriate to resolve the instability reflected in the health and social care environments, as well as how and why they can be applied effectively to manage patient flow.

The discussion is structured around the following research questions:

1. How can LM and TOC approaches be applied effectively to improve patient flow across health and social care?
2. Why do LM and TOC approaches work better in different health and social care environments?
3. What are the assumptions underpinning the effective use of Kanban and TBM in different healthcare environments?

The three research questions will be fully addressed in the following chapter. This chapter is therefore structured into the following sections:

Section 5.1. presents a cross-case analysis to consider the research questions one and two. The discussion aims to draw parallels and differences between cases in relation to flow improvement interventions. It also establishes relationships and comparisons between the research findings identified from individual case data analysis and related literature. The section considers the environmental conditions associated with different manufacturing flow mechanisms and the translation of these mechanisms into healthcare. In this context, the section is divided into four subsections based on Goldratt's (2009) flow principles to examine how various flow mechanisms developed in manufacturing can be applied to healthcare environments. Utilising these flow principles through cross-case analysis contributes to explore how these mechanisms can be used selectively and effectively to manage patient flow associated with various levels of variability and uncertainty.

Section 5.1.1. discusses the value of flow in manufacturing and how this value is reflected in healthcare environments. Section 5.1.2. examines how the different flow mechanisms have been used to manage flow in manufacturing environments and how these mechanisms can be translated to different healthcare environments. Section 5.1.3. explores the establishment of a continuous flow management improvement process, and, finally, section 5.1.4. considers the importance of avoiding local efficiency and optimization measures to enhance the adoption of a proactive flow mechanism. These subsections help to explore the link between manufacturing and healthcare regarding the instability of environments, exploring how such flow mechanisms can be used selectively and effectively to manage patient flow.

Section 5.2. considers research question three and through drawing comparisons between the flow mechanisms associated with LM and TOC approaches, namely Kanban and TBM, to explore the assumptions underpinning their practical use in healthcare settings.

Finally, section 5.3. explores the relationship between the current development of healthcare flow management and the manufacturing flow mechanisms. The outline of this chapter is illustrated in Figure 19.

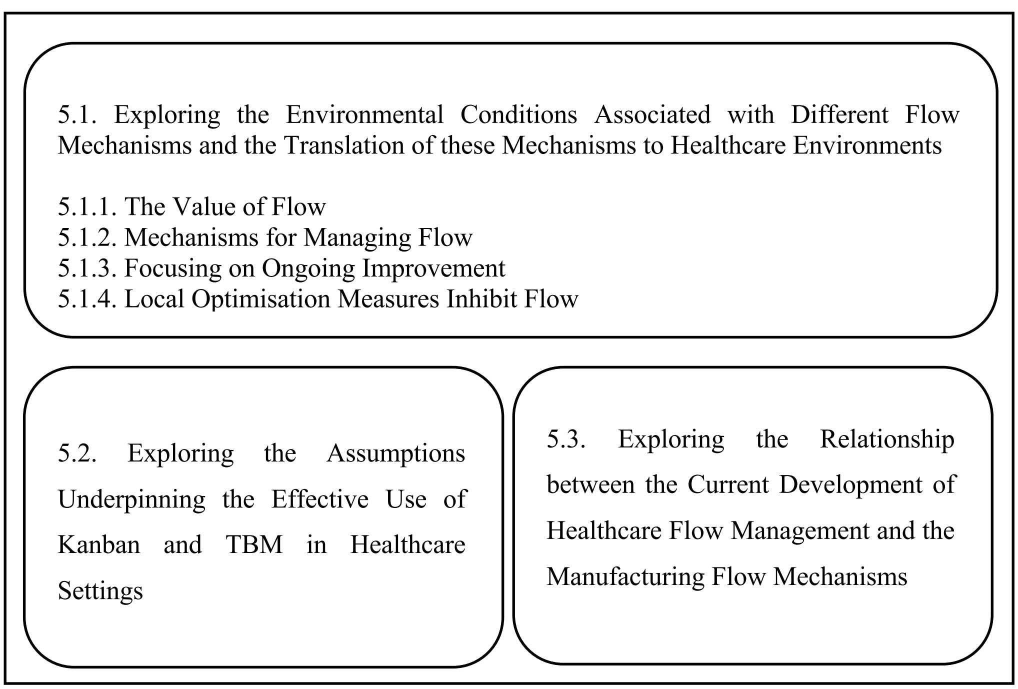

Figure 19: Chapter 5 Outline

5.1. Exploring the Environmental Conditions Associated with Different Flow Mechanisms and the Translation of these Mechanisms to Healthcare Environments

The adoption of a flow-based mechanism is key to effectively managing variation and uncertainty to the delivery system. The level of instability in production environments is varied and depends greatly on what the customers' needs are. This can determine how complicated the process would be, and there is a need for systematic management to ensure customer demand is met, regardless of the level of variation and uncertainty reflected in volume and variety of products. The need to meet customer needs has resulted in different flow mechanisms, each of which is designed to meet the needs of different manufacturing environments. The section uses the same four flow concepts highlighted in section 2.2.2 to reflect the adoption of a flow mechanism in different manufacturing settings and how the adoption of a flow mechanism can be related and translated to healthcare environments through cross-case analysis. These four concepts are:

1. *Improving flow (or equivalently lead-time) as a primary objective of operations*;
2. *Translating the primary objective into a practical mechanism that guides the operation when not to produce (prevents overproduction);*
3. *Abolishing local efficiencies and;*
4. *A focusing process to balance flow.*

5.1.1. The Value of Flow

In manufacturing, flow delays lead to longer lead times and can lead to higher costs as WIP continues to increase and becomes out of control, resulting in inventory accumulation. If the system is overloaded with WIP, it causes quality problems and losses, which have an impact on the system's performance and throughput. This means issues became difficult to manage and poor operational flow exists, which reflects the importance of enabling operations to flow smoothly.

As discussed in Section 2.2.3.1, the importance of flow was established in manufacturing by recognising the requirement to put customer needs first, by creating an order due date. This would determine the point at which the order is to be released into the system, so that the amount of WIP can be better managed. This has led the manufacturing sector to be conscious of the value of flow, ensuring that the work release into the system is based on actual customer demand and, hence, that the lead time can be reduced. From a manufacturing point of view, the value of optimising flow is reflected by meeting customers' needs and ensuring that the product reaches the customer as and when it is needed. How can this value be related to healthcare?

In healthcare, the accumulation of WIP (overproduction) takes on a different form, manifested in the DTOC. Having too many delayed discharges of patients means that long-delayed transfer of care arises, indicating difficultly for staff in focusing on progressing those patients, as it is not clear who has the priority to be progressed first. The assumption of filling beds with almost any patient demand means that the hospital ensures that all beds are fully used. This is a natural reaction if there are empty beds: why not just fill them with patients and fully utilise them? But, in reality, this leads to having more patients staying in beds unnecessarily, and, as soon as the beds are full, the hospital cannot release any more beds, and it becomes hard to manage. Fully utilising beds means that hospital staff always act reactively and are always chasing to get patients out of beds.

> If you take account of beds without improving the flow, you would just make the backlog bigger and the flow worse because you haven't stopped the multitasking and the missed synchronization. (P28)

In both manufacturing and healthcare environments, longer lead times mean poor customer (patient) service, which can influence the progress of other delivery orders (not discharging patients will mean others not being treated). It is therefore clear that all aspects of poor flow indicate an increase in variability and uncertainty, demonstrating how improving flow is an important objective for both the manufacturing and healthcare environments.

Improving the flow of patients is a significant factor for healthcare administrations in all the cases observed in this study, acknowledging that the accumulation of WIP in delaying patients' discharge (beds overused) or increasing patient waiting lists has contributed to poor patient flow.

In Case 1, for example, there were complaints about the growing waiting lists that breached the NHS targets of 13 weeks for assessment and 18 weeks the start of treatment. As illustrated by Participant P1, there were around:

> 171 different waiting lists contained more than 2000 patients. (P1)

Other aspects of poor patient flow were evident prior to the TOC intervention in Case1 and Case 2, where the analysis indicated that there was a gap between local service providers and their doctors, as they tended to work in isolation and to manage their own patient lists separately. In Case 1, the lack of visibility across the system resulted in queues and an increase in waiting lists. According to participant P4, 'One team of therapists with 7 members was responsible for managing a caseload of 450 children' (P4). Poor patient flow and the lack of visibility to management across the system were also evident in Case 2 before the TOC solution, where staff were not aware of any barriers to flow within the system. This contributed to the blocking of beds and an increase in the average LOS for elderly people.

> So, before Jonah, I was not aware of there being any central visibility and understanding whether flow is efficient or whether there are any barriers. (P8)

> Eight or nine years ago, the average length of stay would have been 60 plus days. (P6)

In Case 3, prior to the intervention of LM, there were no formal processes and the processes lacked standardisation, resulting in poor productivity and patient-facing time, while increasing waste as non-value adding activities consumed the nursing staff's time. As highlighted by interview evidence:

> Stock room organisation was terrible, and people would reorder because they couldn't find something because it was buried in the bottom of other things and things would go out of date. (P9)

The notable hospitals leading the application of LM, as highlighted in Section 2.3.3.1., also acknowledged the value of flow where the need to minimise LOS and improve quality and safety was a consistent driver of improvement initiatives in all cases. However, some previous studies regarding the applications of LM in public and healthcare sectors have highlighted the difficulty of identifying value from the customer perspective, due to the presence of different stakeholders involved in these services (Young and McClean, 2008; Ghobadian et al., 2009; Radnor et al., 2012).

Case 4 represents an acute hospital that offers services to meet the needs of different patients including planned and unplanned care units. The interview analysis noted that the high rate of poor patient flow was a result of the blocking of beds caused by patients who did not need to stay in the hospital, but who were waiting for other support packages, such as social care or housing support.

In most cases, the leading cause of poor patient flow was a lack of patient flow management in the form of a system-wide perspective that links different parts of the system to meet patient needs. This underlines the importance of setting a common goal across the system, by making patients a central priority of the system goal. This is considered to be an essential step in managing patient flow. Putting customer needs as the priority of the system is an essential step in manufacturing, which is also relevant to healthcare in terms of ensuring that 'value' is delivered to the customer and that all sources of 'waste' must be removed, such as excess inventory, mistakes, waiting, inappropriate processes, and re-admissions (Young et al., 2004, Womack et al., 2005).

As discussed earlier in section 2.2.3.2.3., setting the focus at the critical points of the system using TBM is a key to managing the flow in complex manufacturing environments. In more unstable healthcare organisations, including acute setting, social care homes, and community hospitals, establishing a PDD is the key to managing patient flow, as it places the patient at the centre of focus across the health and social care system (Stratton and Knight, 2010). The concept of PDD has become commonly used by many healthcare institutions today (Knight, 2014) and can be related to similar terms, such as the expected, estimated, and anticipated date of discharge. Based on an interview during this research with Alex Knight, who is concerned with the transformation of TOC into healthcare settings, the concept of PDD is said to have originated in 1992 during the transformation of the TOC approach at Oxford University Hospital; the PDD is concerned with achieving 'a patient-centred clinically led plan.' The PDD concept is a fundamental part of any kind of scheduling and control system (Stratton and Knight, 2010). In healthcare, as in manufacturing, time is essential, and setting a PDD is critical to exploiting the time constraint in the healthcare environment.

The importance of setting the PDD was evident in three cases, while the PDD in Case 3 is different, because this healthcare organisation offers care services to patients at their homes and, as soon as the staff see the patients, they will immediately be discharged from the system. There is no such PDD in this environment, but the value of improving flow here is manifested in ensuring services are provided to patients at their home, based on the scheduled time. So, if the staff failed to meet this scheduled time, poor flow and an increase of WIP would be generated.

Setting up a PDD for the patient on their first arrival aims to shorten the LOS and gave the system an indication of the importance of focusing on meeting the patient’s needs before this date (Stratton and Knight, 2010; Majeed et al., 2012). The importance of the PDD is clearly demonstrated in Case 2 by the comment made by the Participant P8:

> One really important thing that has been embedded in the process is that you start planning someone’s discharge from the moment they arrive. (P8)

The PDD is set by clinicians based on patients’ medical conditions and may therefore change (Stratton and Knight, 2010; Stratton et al., 2014). As mentioned by some participants in Case 2, ‘We only ever change the PDD if the patient's not medically fit to go’. The PDD is set as a realistic estimation of the time patients need to recover and prepare for discharge to their next place of care. Within Case 4, the term EDD (the expected date of discharge) was evident and was typically set within 48 hours of admission to the ward, based on the medical needs of patients. The PDD is critical in helping to minimise the LOS of patients by making the system focus on meeting patients’ PDD, and this requires that all support activities be subordinated to the patients’ medical needs. Given the instability of the health and social care system, particularly the uncertainty of demand and the complexity of patients’ needs, it may be a challenge for hospitals to manage these PDDs and avoid some form of DTOC without synchronising demand with capacity. This situation can also be seen, whether in a stable or unstable manufacturing environment, where a promised delivery date is given to a specific order for the customer. Therefore, a practical flow mechanism is needed to align all support activities across the system to ensure that the order is delivered at the agreed time (the PDD is met).

Overall, all healthcare organisations observed in this research acknowledge the importance of improving flow to enhance their delivery system performance. The importance of encouraging patient flow was evident from all aspects of poor patient flow, such as increased patient waiting times or length of stay, high rates of waste involved in daily activities, and the lack of visibility across the system. This has led these organisations to translate manufacturing management thinking into their settings, aiming to reduce DTOC and meet their patients' needs and expectations. Acknowledging flow as the primary goal in both environments requires a practical mechanism to ensure that variation and uncertainty are controlled and that the order is delivered at the agreed time (PDD is met). The next section, therefore, discusses how such instability (variation and uncertainty) has been managed in different manufacturing environments and then highlights a cross-case discussion on the translation of such a mechanism to manage patient flow in different healthcare environments.

5.1.2. Mechanism for Managing Flow

Returning to section 2.2.3.1., we can see that all flow mechanisms recognised the value of flow improvement through the emphasis on reducing the WIP and targeting variability, in order to reduce lead times and increase overall flow throughout system operations. However, there are significant differences in the instability associated with their originating environments. The nature of production (volume) or the nature of customer demand (variety) varies from one environment to another. In manufacturing, there can be two extremes in process industries: repetitive/stable environments (high volume and low variety of products) or non-repetitive/unstable environments (low volume and high variety of products). As in the manufacture of the goods or services, demand and procedures can be stable and predictable in healthcare, while other situations can be uncertain and varied. It is, therefore, important to discuss how WIP (overproduction) was controlled in different manufacturing environments, before considering how DTOC in healthcare has been managed through the cases evaluated in this research.

In both environments, the market is typically the drumbeat of the system, but the nature of market demand and needs varies depending on what the environment can deliver. In stable manufacturing environments, the establishment of dedicated pathways is the key to enabling flow. This was evident, as discussed in section 2.2.3.2., in Ford's flow line which relied on the use of dedicated routes, without the need for changeovers to produce high volume and standard products. The limited WIP space between two workstations means that workers who continue to feed the space should stop producing when the allocated space is full, thus preventing overproduction. Creating clearly defined flow paths was also evident in the Toyota environment, as much of the TPS approach was based on the concept of Ford's flow line (Ohno, 1988). With the shift to producing a much wider range of products and much lower and uncertain volumes, machine changeovers and small-batch sizes become necessary. This makes it difficult to use dedicated lines to control WIP, as the majority of TPS work centres were not dedicated to a single component. The Kanban system was, therefore, invented as a flow mechanism for controlling the release of work and directing the overall operation across the TPS system (Ohno, 1988). Thus, without a Kanban card, nothing can be produced, and overproduction is stopped.

As discussed above, Ford and Ohno both acknowledge the need to limit WIP by choking the release of work into the system (preventing overproduction) based on system requirements. In healthcare, instead of producing goods to customers, the same premise can be translated by ensuring that patients receive care when they need it and that they are discharged from the hospital when they are medically fit. The 'patient' flowing from admission to discharge in the healthcare system is a care opportunity for another patient. It is, therefore, necessary to ensure that patients can flow with as little disturbance as possible through the care process. As with manufacturing, one way to manage the flow and keep bed capacity is to focus on meeting system requirements which are patient needs, instead of focusing on filling beds. Therefore, to choke the release of work into the system, the hospital needs fewer beds (limiting the number of beds); rather than focus on filling the empty bed, they should focus on discharging patients to free up beds even though there are some spare beds. The emphasis is to make sure patients leave the hospital when they are ready to leave; otherwise, a state of overproduction occurs, as beds are all being used unnecessarily. To avoid that, establishing a PDD for each patient, and having a practical mechanism to guide the system to ensure the PDD is met, is a means of stopping overproduction (avoiding patients staying in beds unnecessarily).

In many healthcare environments, the market is commonly the equivalent of the drum in manufacturing, but the PDD is more uncertain and, as with manufacturing, it depends on the healthcare environment. The question arises here as to whether Ford's and Ohno's thinking applies to healthcare environments where the PDD is more uncertain.

Referring back to section 2.2.3.2.1., it is clear that both Ford and Toyota/LM approaches tend to control all aspects of the system, by establishing clearly defined flow paths and using a practical flow mechanism to manage flow and target variability across the system. This indicates that, where processes and customer requirements are permanently standard and unchanged, the flow of the environment is more stable and can be easily managed and monitored.

Within healthcare settings, creating dedicated flow paths for a specific care unit, such as the elective surgery unit, is more akin to a flow line environment, where the surgery processing time and demand are predictable and, thereby, capacity is adjusted based on the schedule of demand. As the previous study shows, better stability and load levelling have resulted from the use of LM tools to enhance the performance of elective surgery for patients with less complicated needs (Edwards et al., 2012; Daultani et al., 2015). This is also consistent with the study of Black and Miller (2008), where some LM tools were implemented to organise load levelling and cascaded scheduling in both radiation oncology and infusion centres at VMMC. In these situations, the department knows exactly how long the patients are going to be in the system before they are discharged. Hence, the release of beds in the system can be controlled. Any changes in the procedures which occur (e.g. process time, being short of capacity, increasing demand) may have an impact on the stability of the system. However, any interruption to the flow line can be easily detected and can be addressed.

Applying Kanban in this type of setting can be seen as a signalling mechanism: when the bed in the surgery room is vacant, a signal can be sent to the staff indicating that the bed is ready to be occupied by a new patient (Grunden, 2007). The use of Kanban cards to allow patients to move through the system without being redirected by medical staff was evident in Hyperbaric Medicine at the VMMC (Black and Miller, 2008; Kenney, 2011). However, the evidence regarding the use of Kanban to manage patient flow across a whole hospital remains limited.

Case 3 is the only case adopting an LM-based productive series to improve service delivery that involved nurses, OTs, support workers, and community matrons supporting people to live independently and accelerating their discharge from hospital. The use of 5S helped to create the platform for efficient production planning and to direct the Trust's focus to structurally reorganise the operational process to increase patient-facing time and productivity. The LM approach in healthcare is commonly associated with the creation of a dedicated care pathway that is more like a balanced line in the manufacturing environment. Much of this has involved secondary processes, such as consumables, pharmacy, and test procedures, rather than patient flow. Although there have been significant improvements in patient flows, these are typically associated with the redesign of medical pathways (Spear, 2005; Ben-Tovim et al., 2007; Fillingham,2007; Grunden, 2007; O'Connell et al. , 2008; Gubb, 2009; Kaplan, 2010; Kenney, 2011; Blackmore et al., 2011; Radnor et al., 2012; Plsek, 2013; Hicks et al., 2015).

Yet, in more complex pathways, where demand and the nature of care are unpredictable, capacity levelling planning is limited because of the variation between capacity and demand. This would lead to more complex pathways, and redesigning the flow path or flow line would, therefore, be challenging to implement in this environment. In this type of environment, the drive for customisation requirements is higher, resulting in more complex pathways, making it challenging for the Kanban system to manage the flow due to its sensitivity to demand fluctuation and to the sources of variability (Hall, 1981; Monden, 1983; Shingo,1989; Stratton et al., 2008). While both Ford's and Ohno's ideas for improving operations were based on restricting overproduction by managing the release of inventory, Goldratt (2009) proposes an approach that restricts over-production in time rather than in inventory. While a time-based approach to regulate production within a system restricts the overall amount of work within that system, it also is suitable for unstable environments, which cannot buffer suitably with inventory, as under the Kanban system.

To limit the WIP in this type of environment, a DBR-TBM approach can be implemented, as explained in Section 2.2.3.2.3. In most manufacturing environments, as with Kanban in TPS, the market demand can reflect the constraint that sets the drumbeat of the system. In this case, the rope can be tied to the drum, which represents customer demand. Rather than using a set of short ropes, as with Kanban and TPS, there is only one long rope in the DBR, and only one buffer is used to protect against many sources of variability and to ensure that the drum (market demand) is never starved of orders on time. Is the DBR solution applicable in managing patient flow in complex healthcare environments?

Healthcare is a complex system characterised by a range of variability and uncertainty, and the opportunity to define the process flow is likely to be more limited in this environment. Establishing a planned delivery date for customers' orders or patients' needs is the key to enabling flow in both manufacturing and healthcare environments. Meeting this plan means variation and uncertainty must be controlled, to enable flow and protect it from any disruption. In other words, a synchronisation process between capacity and demand should be in place to ensure that the PDD for each patient is met.

In manufacturing, the release of work into the system begins once the order is received from the market. The rope is defined by the market (backwards) in terms of when the work is released into the system. But when it comes to healthcare, the patient suddenly arrives at the hospital, either for an emergency or another procedure, and the hospital starts planning when the patient should be discharged. The hospital process is, therefore, working forwards, as opposed to backwards. As with the DBR solution in manufacturing, in healthcare it is market demand which most likely remains the constraint, and the rope represents the time offset against the PDD set up in the system. Once the PDD is established, the time remaining for the patient to be discharged is termed buffer time, and it is divided into three equal zones to support the management in taking action if required. The use of the DBR-TBM solution was evident in two different healthcare environments observed in this research.

Case 1 represents an appointment-based service (planned care), but this type of environment involves a highly uncertain demand given different patients' complex needs. This NHS case provides speech and language therapy services in communities and for acute needs across the county, for children between 0 and 18 years of age with more specific and complex needs. It was evident that there were complaints about the growing waiting lists that breached the NHS targets of 13 weeks for assessment and 18 weeks for the start of treatment.

> It is very challenging to estimate how long the patient stays in the waiting list for the next treatment because there were a thousand children on these waiting lists from all over the county and they were all mixed. (P4)

This indicates that the level of instability is higher due to the uncertainty of demand and the nature of treatment. It was evident there was a gap between local service providers and their clinicians, as they preferred to work in isolation and to manage their own patient lists separately. This lack of visibility across the system resulted in queues and increased waiting lists. These issues can be seen in manufacturing where the increase of variation and uncertainty in the delivery system can lead to the accumulation of WIP and, hence, longer lead times cannot be avoided. This is due to a lack of central focus for the system that links different parts of the system to the requirements of the system as a whole.

To ensure that the requirements of the delivery system in Case 1 are met, the consultants established a new process that prioritised patient appointments in one aggregated list on a first-in/ first-out approach, with a two-week initial assessment target. At the initial assessment appointment they also planned the whole package of care, typically involving multiple therapy appointments and effectively resulting in a PDD. In this environment, the drum represents market demand and sets the pace of the system, as demand is unpredictable. The rope defines the time that staff are going to work, once the PDD or initial assessment appointment is established to ensure that patients are able to receive the care and leave the system, and thus to prevent late discharge or a delay in providing care to the patient. The buffer is the time of two weeks which commences on the patient's entry into the aggregated list. With the setting of the PDD, the time to the PDD acts in this case as a buffer.

As with DBR applications in manufacturing, the aggregated buffer option was established to ensure the system requirements were addressed by reducing patients' waiting time and adjusting the available capacity to deal with one aggregated list, instead of multiple waiting lists managed by each therapist separately. This pooled buffer allows the sources of variability to be combined into a single buffer, effectively reducing the buffering required (Hopp, 2003) and ensuring the therapist focuses on managing one aggregated buffer, instead of each therapist dealing with their own buffer of patients in isolation. Thus, moving to a central patient queue, by adopting an aggregated buffer as advocated by TBM, allowed more effective control of priorities with the centralised allocation of therapists, as opposed to a locally distributed resource. As illustrated by Participant P5:

> Making appointments visible, drove up the amount of activity. So, the staff did more appointments when their appointments were visible. (P5)

This was not the case before the intervention, as the interview evidence showed. Although some therapists were doing a great deal of work within their capacity, 'Other therapists were doing half as many of the appointments' (P5). This allowed the waiting time for the initial assessment to be reduced from 13 weeks to 2 weeks.

> Before the intervention of QFI, the waiting time for an initial assessment was 13 weeks. So, when the QFI came in, we did all this work. It went down to two weeks. (P2)

This key feature allows therapists to focus on managing one queue by adopting a common priority system. The scheduling of the treatment plan at the first consultation (initial assessment appointment) is also consistent with flow management by setting a PDD, thereby ensuring that the number of patients in the system is kept low.

It was evident that additional staff needed to be provided to support the process improvement and avoid the system becoming unstable. As stated by Participant P1:

> What we did was we took one of the part-time staff from the service administrators... and she did some work for us for three months in some of her days off to reconfigure some of the staff so that we could move to streamline that process. (P1)

In Case 2, in order to manage instability and ensure the PDD for each patient was met, the TBM was applied in the form of 'Jonah' software. The goal of the TBM approach is to ensure better management in delivering timely care to patients and the improvement of the flow across the hospital. This approach has four functions (prioritise, expedite, escalate and improve), and it has a significant role in supporting management in linking different components of the system towards reaching the overall goal of the system, which is ensuring the PDD for each patient is met. In the opinion of the Participant P9,

> The key purpose was to monitor and improve flow…trying to help people achieve the goals as quickly as possible to get through the system as quickly as possible. (P9)

The role of TBM is to provide management signals to ensure delivery system activities are subordinated to the exploitation of the system constraint. Participant P8 stresses this critical role of the TBM approach in directing resources to achieve the goal of the system:

> One thing that Jonah did was trying to promote clarity about what we are trying to achieve and how we are trying to do it. (P8)

In these types of environments, the equivalent Kanban as a flow mechanism can be limited, due to the high degree of variability and uncertainty. Kanban in TPS requires a well-designed pathway and clearly defined process steps. With Kanban, market uncertainty is often tightly controlled and adjusted centrally, as everybody has to operate according to the system requirements. The demand and nature of treatment in Case 1 and Case 2 is more uncertain, indicating the complexity of pathways involved in their healthcare delivery system. Applying the flow line model by redesigning the medical pathways can be more difficult in these environments, as capacity needs to be equal with actual demand. Kanban in TPS provided a mechanism to support the reduction of variability and to limit the release of material into the system and, hence, encourage flow across the system. Applying Kanban and its rules in these healthcare environments (Case 1 and Case 2) is limited, as it is challenging to achieve levelling scheduling capacity, due to the uncertainty of demand associated with the complexity of patients' needs. Therefore, a Kanban equivalent seems difficult to establish in healthcare, as load levelling and setup reduction must put in place to enable the Kanban system, as with TPS approach (Shingo, 1989).

Thus, TBM is the means of providing management signals to ensure delivery system activities are subordinate to the exploitation of the system constraint. In healthcare, as in manufacturing, this is 'time', and setting a PDD is critical to exploiting the time constraint in the healthcare environment. The establishment of a PDD was key to enabling the other support activities to be prioritised around patient needs and thereby proactively managing the threat of DTOC, a particular issue for the elderly who are transferred between healthcare and social care providers. The fourth function of TBM is to identify strategic opportunities for improvement. The use of aggregated buffers is more accommodating to variability and uncertainty, unlike Kanban, which is more sensitive to flow disruptions and therefore achieves higher performance in a well-defined delivery system.

Case 4 also presents a complex healthcare system, and data from the case analysis show that the town has the highest rate of long-term illness in the region, with three out of five people aged 65 and over suffering from a long-term illness: '…most probably the third [most] deprived area in the UK in terms of lots of things such as housing and social deprivation' (Participant, P15). The inability to manage instability in this environment resulted in bed blocking, as the beds were occupied by patients who needed social care and housing support, rather than staying in hospital. To address this issue and improve patient flow, the solution was driven by the local authority support for the discharge of medically fit patients from hospital in circumstances where their home, or a range of other wider social issues, would make it unsuitable for them to return to their home. The local authority manages the system through their office based in the hospital, including a team of two people to enhance the hospital discharge process by ensuring patients have the required housing assistance when they are discharged. The mechanism acts through this team's daily visits to the ED and wards across the hospital, or by receiving calls from relevant people in the hospital concerning support provided to any patients in need of rehousing when they are medically fit. It was also evident that the intervention was intended to provide support to any patient approaching the front door, and the team can plan ahead of time to provide support when alerted early enough.

> So, they basically try to evaluate any patient coming in the front door either by ambulance or whatever, if they need something required from the local authority team, they will record and work on it, and they will know in advance. (P20)

This mechanism was introduced by the local authority team, in addition to improving the hospital discharge process, preventing hospital beds from being occupied by non-health clients (e.g. people addicted to drugs or alcohol abuse), and helping the hospital to provide care to patients in real need of care.

> They recognise this group of patients that we classified under the non-health category for reasons like this … and they've got the skills and the knowledge and they know the right pathways to support us because remember, you know, we're clinicians and nurses and they're experts in their own field. (P26)

This intervention by the local authority showed a significant improvement in patient flow, by reducing delayed hospital discharges, readmissions, and the burden on social services and hospital staff. It was also established that, by reducing bed days, this intervention had created and estimated saving of £1,142,550 for the entire NHS system.

5.1.3. Focusing on Ongoing Improvement

In manufacturing, choking the release of work into the system through a practical mechanism gives a much better focus and help to identify the points where WIP is accumulating, so that the sources of variability can be easily identified and targeted and thereby balancing flow to ensure there are no disruptions to it. This is achieved in manufacturing through the use of different flow mechanisms to explore the variability and to identify the causes of flow disruption, which make it possible to target these causes an enhance the ongoing improvement process.

In more stable environments, flow paths and processes are predefined providing clear visibility for the flow, and, thus, any problems causing an interruption to the flow path can be easily identified. This clear visibility helps to identify and remove sources of variability which disrupt the flow and supports a process of continuous improvement. With Ford's flow line, as discussed in section 2.2.3.3., any interruption of flow was easily detected because, if one work centre in the flow line stopped, the entire flow line stopped directly.

Focusing on ongoing improvement is also realised in LM/TPS environment through the use of a Kanban system (e.g. the rock and river analogy). When the inventory (water level) is reduced to expose rocks (sources of variability) above the water, then these rocks can be removed to enhance flow (Umble and Srikanth, 1997; Liker, 2004; Goldratt, 2009). Therefore, when the inventory is controlled and reduced, problems in the production process are quickly identified and can be addressed.

As Ohno (1988) emphasizes in section 2.2.3.2.2., Kanban was a central component of TPS, and the six rules of Kanban is the means by which a system-based approach is adopted. Kanban's rules, in particular Rules 5 and 6, embrace the focus on an ongoing improvement mechanism. Through the distribution of centralised buffers between job centres, Rules 5 and 6 of the Kanban system are responsible for overcoming problems immediately and addressing them without handing the problem over to the next job (Ohno, 1989). While Rule 5 focuses on addressing source product variability by identifying the defective process, Rule

6 defines the current issue and ensures inventory control, resulting in a reduction in inventory over time. Any issues of quality and delay in processes, the second pillar of TPS (Jidoka), would ensure the defect was prevented from progressing to the next job and that the source of the stoppage was targeted (Fujimoto, 1999; Liker, 2004).

In unstable manufacturing environments, the complexity of pathways appears to be very high as each job is different. This complexity implies that creating clearly defined paths and controlling many buffers is difficult to achieve. Therefore, using direct observation or a Kanban system is limited in this type of environment. However, as with stable manufacturing environments, the mechanism of continuous improvement can be applied in unstable environments. This was achieved through the use of the TBM approach, as outlined in section 2.2.3.2.3. By following the 4 functions of TBM, the system can be guided to prioritise or expedite tasks, escalate them, if necessary, and, most importantly, identify key causes of delay to target and reduce variability.

These different manufacturing mechanisms are a means of reducing variability and uncertainty and, hence, of improving the flow and ensuring the system requirements are addressed. The critical question is, can these different mechanisms apply a continual improvement process to manage patient flow in healthcare environments? And if so, how can they be implemented effectively?

In manufacturing, chocking the release of work can help to identify any issues that might cause interruptions or delay to the flow. As with Ford and TPS, having a process to balance flow avoids accumulating WIP and ensures flow is protected from any disruption. In healthcare, this equates to bed capacity and the natural tendency to fill beds and only discharge when there is a shortage (Stratton and Knight, 2010). Hence, it is evident that cutting beds can be viewed as a mechanism to force the flow, as in enforced problem solving (the ship and rocks analogy). However, TBM was used effectively in Case 2 to reduce the need for bed capacity more proactively, so subsequently leading to empty beds being cut.

> When I first came here in 1999, we think we've got 560 beds across the 12 community hospitals and would reduce significantly. (P11)
>
> We’ve got 125 beds open at the moment for Assessment and Rehab. (P14)

> So, taking beds down doesn't solve anything unless you improve the flow. Once you've improved the flow, you can take the beds out. (P28)

With the applications of LM in healthcare environments where there is a dedicated process, such as elective surgery, direct observation can easily be used to detect any problem that disrupts the flow, as everything is visible. The system is more likely to be stable in this environment and does not involve a high degree of variation and uncertainty, and, thus, visible signals can support the process of continuous improvement. This was evident in the notable cases of successful LM implementation in healthcare and also in some other cases, as indicated in section 2.3.3.3., where the use of LM techniques, such as VSM, 5S and kaizen blitz activities, have positively supported the culture of continual improvement (Fillingham, 2007; Black and Miller, 2008; Kenney, 2011; Radnor et al., 2012; Costa et al., 2017; Regis et al., 2018).

In addition to the use of these tools to enhance performance, the Jidoka concept was implemented at VMMC by introducing the system of patient safety alerts (PSAs) to detect defects and support the continual improvement process (Black and Miller, 2008). Although there was evidence of the use of Kanban as a pull system in a dedicated pathway (Black and Miller, 2008; Grunden, 2007; Kenney, 2011), it appears that evidence of using Kanban as a flow mechanism to manage the primary flow of patients across the system, as well as to maintain a continuous improvement process, remains challenging to find. Therefore, it seems difficult to establish a direct equivalent of the TPS Kanban in healthcare.

In general, the approach of TBM in healthcare has shown a substantial improvement in patient flow, as it offers a robust mechanism that helps to establish a process of continuous improvement. In healthcare, time appears to be an alternative and significant measure. Looking at where most of the delays in the transfer of care occur, there is an indication of a disruption to the flow, as it could be a step that takes much longer to finish or where there is a wait for other resources.

> We’ve got somebody in a bed, and they shouldn’t be in bed for a variety of reasons… It was actually our stuff not referring to social care in a timely manner to enable the care packages to be arranged. (P11)

Delays in the transferring of care are certainly an indicator of flow disruption (Humphries, 2017) and should, therefore, be considered as part of the flow balancing process. The fundamental principles associated with the TBM method have explicitly carried out this task, as observed in Case 1 and Case 2 of this study. When the patient flow is threatened by a non-medical cause of delay, a robust TBM mechanism helps the hospital management to identify the resources or tasks that are often disrupting the flow.

In both Case 1 and Case 2, as with manufacturing applications (Goldratt, 1999; Goldratt, 2009; Moore and Scheinkopf, 1998; Schragenheim and Dettmer, 2000; Stratton et al., 2008; Umble and Srikanth, 1997) of DBR, the buffer was simply divided into three equal time zones (green, yellow, and red) as also indicated in previous healthcare studies (Stratton and Knight, 2010; Umble and Umble, 2006). The buffers provide data which facilitate a focus on the continuous improvement of the practice.

In Case 1, the first-in/ first-out approach is key to ensuring patients are prioritised based on time. The two-week buffer was set up in three equal time zones (green, yellow, and red) to also ensure that the reds, which are at risk of breaking the target, are being expedited. An increasing number of reds would normally signal timely management escalation with the reasons for delay being targeted for improvement. This concept underlying TBM's priority functions was evident through the use of a rudimentary spreadsheet-based buffer management system, due to policy restrictions on the introduction of TBM software. The purpose of this basic spreadsheet was to help in producing all the monitoring graphs relating to delays, which enable staff to measure and monitor the time of the overall patient journey, once the patient is admitted. As Participant P3 explained:

> From receiving a referral to registered on the system, to triage, to book an appointment, to attending that first assessment, each of those steps were measured, but we monitored what was waiting [time at] each of those stages. And we set up buffers and understood whether they were in, kind of, red, amber, green in terms of where they should've been. (P3)

Although the use of spreadsheet was helpful in improving the delivery system, it was evident from the interviewees that it would have been more effective with Jonah software, because it is simple to use. As Participant P1 highlighted:

> We were producing all the monitoring graphs around delays, to what extent are people

> stacked up, to what extent are things delayed, are we in the green zone, the amber zone, the red zone, the black zone…….maybe it would have been better with Jonah because it is simple to use. (P1)

It was also evident that a weekly conference call held by managers and team leaders was necessary to review the performance of the system, where patients and tasks are expedited to address any issues affecting the progress of patient appointments.

> From each meeting will come actions the different individuals were going to take, which was going to drive down the waiting time. It was a visual representation of the waiting times and capacity. (P1)

In Case 2, the emphasis on improving flow and the process of continuous improvement was supported by the TOC application and the adoption of software (Jonah) to support the application of TBM. TBM acts in the form of Jonah software that generates real-time data which can be reviewed by relevant people at daily huddles and weekly meetings to review the progress of PDDs. They can accelerate actions if necessary, to keep the delivery system stable. Consistent with the functions of TBM, the weekly 'top delay' meeting (Stratton and Knight, 2010; Umble and Umble, 2006) focuses on reviewing patients who are in the red zone or already represent a delayed transfer. These meetings are multi-disciplinary and are attended by general managers and social workers. Waiting for care packages, for example, is the typical main reason for delay, as illustrated in Figure 15 of Chapter 4. The Pareto data provided by the Jonah software provides focused data to support these meetings, which helps to address any issues that affect the progress of patient flow and to escalate recovery, if necessary, before the delivery system is destabilised. Thus, managing variability and balancing the flow within the delivery system is achieved. Identifying and overcoming the principal cause of delay across health and social care provides a systematic process that sustains the continuous improvement culture.

> Where you are so extended with different areas, you can't keep an eye everywhere. But Jonah does allow you to do that because you can see what's happening anywhere. (P7)

This is a long-standing TOC application, including the adoption of software to support patient flow and the application of the TBM. This TOC approach was introduced in 12 community hospitals, although some have since been closed. The system had been in operation for more than 9 years, and the initial average LOS was reduced from over 60 days to an average of around 20 days today.

> Eight or nine years ago, the average length of stay would have been 60 plus days and I'm delighted to say that now we report exceptions on the basis of any length of stay over 20 days. (P6)

It is interesting to note that the Trust retained Jonah, even with the need to duplicate data inputting, as the centralised System One, now rolled out across the NHS, does not provide the data in a form suited to managing patient flow.

It should also be noted that, although Case 1 showed a significant improvement in the flow and reduction of patient waiting time, findings in this case are based on evidence from participants who are no longer involved with the daily activities of the current service provider. This limits the opportunity to effectively evaluate the improvement generated by the QFI intervention, as it was difficult to interview people who are currently working in this service provider, as their schedule was always busy. The documentary evidence regarding system performance before and after the intervention was also inaccessible.

In Case 4, following the process improvement interventions at the hospital, some improvements were recorded, including a reduction in the length of hospital stays, reductions in delayed hospital discharges, and reduced readmissions. However, it was also a very useful opportunity for the researcher to observe how the hospital discharge team and other relevant people manage the EDDs for patients and how the hospital maintains patient flow improvement, regardless of the local authority intervention. It was noted that a meeting called a 'Hub Meeting' usually takes place twice a week. It involves key people from related NHS organisations, including representatives from social care, CCGs, the hospital discharge team, community care team, and the local authority team. This meeting aims to discuss EDDs for patients who have already had a delayed discharge by reviewing their records on the computer screen. At the end of the meeting, an escalating action plan is written in MS Word and sent to the relevant people by e-mail. Although the hospital uses modern computer software to generate data to support these meetings and enhance patient flow, it was evident that the results of these meetings functioned only reactively to address DTOC issues, unlike the buffer management meetings observed in both Case 1 and Case 2, where the data generated from using TBM helped the management to act proactively to avoid any form of

DTOC. Recording resources that interrupt or delay a patient's journey when the patient passes through the green, yellow, red, and black buffer zones helps the management to target the areas that need to be improved (Umble and Umble, 2006). This helps the hospital management to act efficiently and proactively. Therefore, the establishment of a mechanism to focus on ongoing improvement to target the causes of flow disruption and ensure patients' EDDs are met was not evident in Case 4.

It is interesting to note that the interviews with one of senior managers indicate that the TBM approach was introduced in Case 4 by the same consultancy company, QFI, that introduced this approach in both Case 1 and Case 2. According to the senior manager, the basis for using the TBM approach was established, including training the staff and leaders about its concepts, but due to the economic crash in 2008, which impacted on the hospital budget at the same time that the Trust was starting infrastructure to build a massive new hospital, some of executive managers and other key leaders had left the organisation and, therefore, the intervention of TBM was not sustained. Stability among executive mangers and leadership is very important to support the transformation and gain positive results.

> I think a number of things that go on in your organisation... there is a... leadership and focus on things. And it does change over time... I mean, there's a whole theory about organisations that they have, how they learn how to change and have a lead, which leads to sustained change. (P31)

Compared with Case 2, the sustainability of leadership played a significant role in supporting the transformation of the TBM approach. As previously touched upon with Case 2, a long-standing TOC application was still in place, with a significant and sustainable improvement in patient flow.

5.1.4. Local Optimisation Measures Inhibit Flow

In manufacturing, the assumption that every part of the system should work as much as necessary could contribute to a measure of local efficiency. If the work is not started and processed based on system-wide requirements, it means some parts of the system might be busy doing other work, which is not going to serve the system's needs and, thus, local efficiency cannot be avoided. The various practical mechanisms used in the manufacturing process, therefore, concentrate on controlling the release of work into the system, so that work starts only when the system is ready to operate. Those mechanisms demonstrate that the work begins and is carried out according to market demand (drum), and they encourage the link between various parts of the system to support each other, in order to comply with the system's overall requirements. Thus, if there is a practical mechanism that controls operations and resources, determining when to start work ('produce') and when not to work ('stop overproduction'), it means preventing resources from doing any work or only doing work based on the system's requirements. As a result, local efficiencies can be avoided.

As previously mentioned in the above sections, Ford's flow line allowed space to be used as a mechanism to signal when the production process should be stopped. This led to local efficiency measures being prevented, as the whole flow line was forced to stop immediately if the allocated space was full. The Kanban system also addresses the issue of local efficiency, as it directs when each workstation must produce products and when it must stop producing. If there is no Kanban card, it means no production and, therefore, workers only produce when products are needed (just-in-time). These two mechanisms show how local efficiency measures can be controlled in stable manufacturing environments. But in unstable manufacturing environments, how can localisation control measures be avoided, where the pathways are more complicated and less dedicated?

In complex environments, synchronising many independent resources or parts is very difficult, as the tasks and activities vary from department to department across the system. This inevitably leads to a significant reliance on localised management practices, and, as a result, local efficiency measures become inevitable, due to the lack of a centralised mechanism for planning and control. In these environments, instead of controlling all aspects of the system, as with LM environments, TOC only focuses on the critical points of the system and, by using the TBM approach, all parts of the system can be synchronised to meet the system's overall requirements.

By referring back to section 2.2.3.4., we realise that the 5FS of TOC as reflected in the application of DBR – in particular, step 3 (subordinate) – provide a mechanism to ensure that work can be initiated and processed only based on system constraints and not based on local or non-constraint parts of the system. This means that localisation activities are controlled, and, as a result, some proactive capacity becomes available, and the flow through the system is significantly enhanced. Goldratt and Fox (1986) stressed the need to avoid local efficiency measures: as expressed by the first rule of OPT, this is balancing flow, not balancing capacity. This was achieved by applying the DBR solution, where the rope and TBM provide a method for determining when work must begin and how to maintain a continuous improvement process, by addressing the causes of variability. Preventing overproduction and targeting the causes of variability, thereby balancing flow and not capacity, protects against local efficiency measures.

Healthcare is also a complex system, and the opportunity for localised management practices is likely to be unavoidable, due to the difficulty involved in maintaining interdependencies between many components of the system. Time in this environment is critical, as it is in the manufacturing sector, and it needs to be exploited. Localised efficiency measures can be manifested by each department or division across the whole system only being concerned with their own performance, targets, or budgets, thinking that they have done their duty and that no blame should be directed towards them. This can lead to localised optimisation and is not going to reflect improvement across the whole system, due to the lack of central visibility which can otherwise direct all departments to work in line with the system's requirements. Localised control measures were evident in case studies observed in this research.

In Case 1, for example, due to the lack of a centralised control mechanism for the whole system, it was evident that each therapist was dealing with their own buffer of patients in isolation. This local control led to an increase in patients' waiting time, rather than meeting the system requirements, which emphasised providing care to patients according to the target set by the government, whereby the patient should receive an initial assessment within 13 weeks of referral and be treated within 18 weeks. As Participant P4 put it:

> One team of therapists with 7 members was responsible for managing a caseload of 450 children… it is very challenging to estimate how long the patient stays on the waiting list for the next treatment. (P4)

As with manufacturing, to avoid local control, a centralised planning and control mechanism needs to be in place. Hence, by establishing an aggregation pooling and using TBM's priority function, local efficiency measures can be combatted and a better focus on meeting system requirements can be achieved. This was evident in Case 1, where the creation of an aggregated buffer allowed more effective control of priorities, with the centralised allocation of therapists, as opposed to a locally distributed resource. This embraces the law of variability pooling advocated by Hopp (2003, p. 120): 'Combining sources of variability so that they can share a common buffer reduces the total amount of buffering required to achieve a given level of performance'.

The application of TOC TBM provides a reliable mechanism to monitor the system's performance and to identify the causes of buffer penetrations. Once the work is released into the system, the requirements of the system are visible to everyone. One such requirement is meeting the PDD for each patient and, thus, a visible signal through TBM can direct resources when they need to prioritise and expedite tasks. As a result, staff only perform work according to system requirements (ensuring the PDD is met) and avoid local optimisation measures.

> The important thing is what we say is we say to everyone, follow the PDD. So, it's just one list across the whole hospitals. Every resource has one list, which is the PDD orientated list. So, there is no local optimisation. There is just follow the list, there's actually designed in not optimised local optimization. (P28)

This mechanism was witnessed more clearly and more practically in Case 2, as mentioned in section 5.1.3., where the hospital has applied TBM in the form of Jonah software for more than nine years. As with Case 1, the local control measure was evident in Case 2, where beds were occupied not based on system requirements, resulting in an increase of LOS to 65 days and a breach of the availability of hospital beds. This reflects a lack of system focus and the need for the planning and control system to synchronise and link different components of the system to meet its overall needs so that patients' requirements can be achieved on time. The Jonah software underpinning the concept of TBM addressed this issue: by emphasising the focus on PDDs, it managed the variability and uncertainty effectively that otherwise threatened the delivery system performance.

> So, before Jonah I was not aware of there being any central visibility and understanding whether flow is efficient or whether there are any barriers. (P8)

The mechanism of TBM forces management to focus on exactly what the system needs by subordinating all healthcare support activities to the medical needs of the patients that are reflected through their PDDs. In this way, staff and multi-disciplinary teams, for example, were responsible for carrying out important activities to subordinate the focus of flow to the PDD. This was compatible with the assumption of TOC, as it shifted focus from a local efficiency measurement to the performance of the whole system, buffering the system against statistical fluctuations created by unexpected issues (Cox and Schleier, 2010).

In the UK, the importance of flow across the health and social care system has been emphasised by introducing measures such as DTOC, aiming to address issues hindering flow. Although this measure was introduced as an attempt to improve flow, it might contribute to an attitude of blame, as each department looks at their own performance, enhancing local optimisation and causing a negative impact on achieving the goal of the whole system. This was stressed through the further interviews conducted with one of senior managers to test the research outcomes:

> One of the things that actually sustain those gains was actually to bring in this idea of the right to reside, as opposed to, and getting away from the concept of DTOC because the transfer of care tends to blame another organisation. (P31)

The need to address this local optimisation practice has recently been acknowledged in the merger of the NHS and SS (Social Services) at cabinet level in the UK. However, this issue, in particular, focusing on cost-saving and local optimisation was apparent in Case 4, where the budgetary debate was ongoing between health and social care providers. Although the local authority intervention was established informally, as it always under repeated review, it has gained awards and clearly contributed to a positive impact on patient flow improvement across the hospital. These repeated reviews were not concerned about the value of the service provided, but about the tension associated with budgetary control concerning which budget centre should pay for the service. Localised cost control was highlighted by many participants during the interviews and was clearly stated by Participant P30:

> There is a situation where actually patients need to go into a care home, but neither social care nor health pays because the patients don't meet any of their criteria. (P30)

In the event that the budgetary control conflict continues to threaten this successful intervention, the hospital may face an increase in DTOC for patients who do not need medical care, but are in need of housing support; hence, beds may be blocked again. However, this situation could be better managed if there was a proactive management mechanism in place that links all parts of the system with the purpose of ensuring that the PDD for each patient is achieved, irrespective of what the patient is waiting for before their discharge (e.g. housing assistance, social care plan, or family decisions). Targeting the causes of variability that threaten timely patient discharge by acting proactively to meet the PDD is an integral part of managing patient flow effectively and of maintaining continuous improvement (Stratton and Knight, 2010).

Looking at Case 3 as an example of LM applications in healthcare and the secondary research in this area, there was little evidence of mechanisms to prevent local optimisation practices and to maintain a continual improvement culture. The LM intervention in Case 3 was associated with a wide range of tools to support the reduction of variability and to improve productivity within the workplace. A few LM tools were used, particularly 5S, to create an organised orderly, clean, and safe workplace, and procedures initiated to maintain it that way. This is evident from the participant who claimed that 'They're using lean to look at how the environment [can] be better' (P10). Although the efforts were based on a national agenda to improve service delivery, the improvements recorded were realised only in some community services and did not reflect on the overall service across the county. This widely applied LM practice has provided significant improvements across the county, but the improvements are inherently local and dependent on local management practice. This was consistent with previous research that identified this issue (Radnor et al, 2012), while others have criticised the outcomes of LM healthcare applications in a different manner (Young and McClean, 2008; Bhasin, 2008; Brandao de Souza, 2009; de Vries and Huijsman, 2011; Poksinska et al, 2016).

However, it should be noted that the majority of the current literature does not discuss how complex healthcare systems can be improved and does not stress the importance of Kanban as a practical mechanism associated with LM practice in managing and maintaining system-wide improvement. There is no mechanism in place, and therefore everything is just optimised around specific improvement activities, such as Kaizen blitz events. Kanban is often perceived to be one of many tools aimed at reducing variability and improving flow in LM; however, its critical importance in sustaining flow improvement is sometimes forgotten.

Since there is no practical mechanism in place to address the requirements of the delivery system, it is not surprising to see the failure of LM applications in the healthcare environment, as most LM efforts end with local optimisation and unsustainable improvement.

Therefore, where the LM redesign is narrowly defined, as in Kaizen blitz activities, the isolated improvements are not linked to the wider delivery system and commonly lack this important means of sustaining the improved flow. Although the Case 3 evidence was limited in this regard, the Productive Ward evidence and the use of 5S is consistent with prior research that identified this problem (Radnor et al, 2012), even though these findings challenge the subsequent evaluation (Radnor et al., 2013).

5.2. Exploring the Assumptions Underpinning the Effective Use of Kanban and TBM in Healthcare Settings

The use of flow mechanisms associated with TOC and LM approaches is a powerful means of managing variation and uncertainty, as they offer a means to identify the causes of disruption to flow and target these causes to create a process of continuous improvement. These approaches operate under different environmental conditions that support their effective use. To demonstrate the assumptions underpinning the effective use of Kanban and TBM in healthcare settings, we refer back to Table 3 in section 2.2.4, to explore how the assumptions associated with TBM and Kanban in their original environments can be reflected in healthcare settings. This is discussed based on the following subsections.

5.2.1. Sensitivity of Medical Flow Path

As discussed in section 2.2.4.1., the use of pooled buffers is a useful way to contrast these two flow mechanism. While Kanban focuses on minimizing variability and is highly susceptible to variability identification, TBM concentrates on variability control and is less sensitive to variability sources. Both Kanban and TBM address variability to manage and improve flow, but their sensitivity to the material/medical flow path is different. Under the Kanban system, process steps, time and transfer paths need to be balanced, to adjust capacity with actual demand (Spear and Bowen, 1999). The majority of TPS designs measure the manufacturing time needed to fulfil the customer's needs and then try to balance resources and equipment accordingly. In contrast, TBM is less sensitive to the material flow path and has been designed for production environments where process times, product mix, and demand are subject to high levels of variation and uncertainty (Chakravorty and Atwater, 1996; Goldratt 2009).

In healthcare, the primary flow is the patient, with a wide mix of instability, indicating a higher degree of pathway complexity. As a result, the opportunity to redesign the process flow is more likely to be complicated and time-consuming. The assumption associated with TBM was reflected in unstable healthcare environments, as previously evident in Case 1 and Case 2 in section 5.1.1., where each patient's conditions are different, requiring different pathways during the care journey. By setting new management rules through the PDD and utilising the function of TBM, patient flow in these environments was successfully managed and improved. Therefore, the assumption associated with TBM is more applicable in complex healthcare settings, as it set rules to focus on the constraints, without the need to focus on all aspects of the system, as with Kanban.

In contrast, the assumption associated with Kanban in TPS environments can be reflected in less complex healthcare settings, where the flow line can be established. This was evident in the literature, with all successful examples of LM being implemented through LM tools to redesign medical pathways so that they deal with particular needs of patients (Fillingham,2007; Kaplan, 2010; Kenney, 2011; Plsek, 2013).

5.2.2. The Choice of Buffering

Section 2.2.5 highlighted that, while reducing and managing variation is a priority in both designs of LM and TOC approaches, the critical difference is that the TBM mechanism is time-based, while the Kanban mechanism is inventory-based. Rather than restricting overproduction by controlling the release of inventories at each workstation or in each individual process, TBM uses time to limit overproduction, by creating an aggregated buffer ahead of the constraints, either market or resource constraints. This enabled a combination of capacity to support the system requirements and more control over variability and uncertainty, ensuring that customer demand was secured and that better performance of the system could be achieved (Dettmer, 2000; Hopp, 2003).

In healthcare, time is critical and needs to be exploited, as enabling patients to be discharged will provide a caring opportunity for another patient. As witnessed in Case 1 and Case 2 through the application of TBM approach, time is the constraint that needs to be exploited and, by setting the PDDs for patients and aggregating them in one buffer, greater visibility and control of the system is obtained. The establishment of these rules leads to centralised management and, thus, efficient use of the available and protective capacity. Controlling the release of inventory, as with Kanban in TPS, can be seen by controlling the release of beds in healthcare. Once the bed is ready, a signal can be sent to send the patient over to the room, as highlighted in notable cases, which leads the application of LM (Grunden, 2007; Kenney, 2011).

5.2.3. Instabilities in Demand

In more stable environments, the stability of demand makes it easier to build well-defined flow paths. This assumption was reflected in the Kanban system through the continual reduction of batch size, and then driven by reducing the set-up time to enable a capacity levelling schedule on a daily basis (Ohno, 1988). This has made it feasible to prioritise jobs through level scheduling between each job centre, as demand in this environment is more predictable. In unstable environments, demand is subject to high degrees of variance and uncertainty, as each customer's requirements differ. In this environment, TBM functions can prioritise and monitor released work and, more importantly, escalate when the system goes out of control (Goldratt 1990; Schragenheim, 2010). Thus, TBM is less sensitive regarding level scheduling.

The same premise can be seen in the healthcare environment. Level scheduling for patients with less complex needs for elective surgery was evident in previous studies (Edwards et al., 2012; Daultani et al., 2015), as bed capacity is adjusted to be equal to demand. So buffering is based on beds, reflecting capacity being adjusted to actual demand. In many healthcare environments, however, demand uncertainty is less controllable and, therefore, establishing a levelling schedule and controlling individual buffers are not easy tasks and not practical in these types of environments. Managing buffers individually did not protect patients' needs, as evident in Case 1 and Case 2, in sections 5.1.3 and 5.1.4. As with complex flow in manufacturing, the TBM approach can significantly encourage patient flow through the use of an aggregated buffer. By setting the PDD for each patient, the TBM mechanism can then direct the resources and activities to be subordinated to the PDD, as well as sending a signal to escalate when the system is not in control.

5.2.4. Detecting Variability and Encouraging Continuous Improvement

As explored in section 2.2.4.4 both Kanban and TBM are committed to supporting the management in targeting the causes of variability and continuous improvement by systematically reducing variation. TBM only accelerates any delays in the process if the red zone of the pooled buffer is breached. With Kanban, any quality problems or delays in the process between individual buffers will not be passed on to the next process. Therefore, variation can be detected immediately.

> To get continuous-flow systems to flow for more than a minute or two at a time, every machine and every worker must be completely capable. That is, they must always be in proper condition to run precisely when needed... By design, flow systems have an everything-works-or-nothing works quality which must be respected and anticipated. (Womack and Daniel, 1994, p. 60)

Continuous improvement is encouraged in Kanban by reducing the inventory to identify problems that can then be addressed (e.g. rocks and river analogy), whereas, in TBM, the causes of delay (e.g. red zone penetration) are targeted and then the buffer can be reduced.

Within healthcare, as witnessed in LM applications, predefined medical pathways enable greater visibility to detect any variation in the processes and any delays that may occur during the patient journey. Standardisation of the procedure and time, for example, at the Pittsburgh Hospital in their operating room, allowed a Kanban mechanism to pull patients when the bed was ready. Thus, any delays in the process can be observed immediately due to the visibility and lower complexity of the pathway*:*

> With stable timing established, the group has been able to focus on other changes aimed at an even more efficient flow through the preoperative pathway (Grunden, 2007, p. 114).

With Kanban in TPS, continual improvement is encouraged by reducing inventory to expose problems; the same assumption can apply in healthcare by reducing or cutting beds so that problems can be exposed. However, a Kanban equivalent to synchronising all parts across the system to maintain flow is not evident within the healthcare setting, when it comes to managing patient flow across the system. Therefore, the failure of LM applications in complex healthcare can be due to the lack of awareness of the environmental conditions associated with the success of LM/TPS, as well as ignorance about the critical role of the Kanban system in improving and sustaining the flow.

By contrast, the TBM approach, as evident in this research offers a means to identify the causes of disruption to flow and to target these causes to create a process of continuous improvement. Any problems or delays that may arise in the patient’s journey can be expedited once the PDD is closer to the red zone of the pooled buffer. Buffer aggregation and zoning enable a focus on control points so that variability and uncertainty can be effectively managed. Continuous improvement is made by identifying the causes of delays which penetrate the red zone; these can then be targeted and eliminated in the future.

Overall, Kanban and TBM play a significant role in the management of flow in production environments, both of which offer ways of identifying the causes of flow disruption and of targeting these causes to build a process for continual improvement. The two mechanisms operate under different environmental conditions which underpin their effective use. In TBM, the number of steps and the path is not generally specified because of the pooled buffer. This reflects the circumstances associated with health and social care where the process steps are uncertain (Litvak and Long, 2000). The aggregated time buffer provides characteristics which are particularly suitable for manufacturing and for healthcare systems with complexities, variability, and uncertainty, whereas Kanban is more sensitive in TBS and performs only under low levels of variation and uncertainty. Table 17 summarises the assumptions of Kanban and TBM in healthcare environments.

Table 17: Assumptions Underlying Kanban and TBM in Healthcare Settings

Kanban assumes	**TBM assumes**
Predefined pathways and procedures	No need for predefined pathways and procedures, only establishing new management rules
Buffering is based on the bed, as it acts as the inventory	Buffering is based on time and is pooled
Level scheduling (e.g. elective surgery)	No level scheduling, as demand may vary, triggering (timely) escalation
Any delay in providing care to the patient can impact on the next patient, such as staff being delayed by a procedure or a bed not being ready for some reason. As the pathway and procedure are cleared, variability can pick up immediately.	Any issues or delays that affect the patient’s journey can be expedited once the PDD of the patient has penetrated the red zone of the pooled buffer.
Continual improvement might be encouraged by reducing beds to expose problems that are then targeted.	Continual improvement is enabled by identifying the causes of delays which penetrate the red zone and can then be targeted and avoided in the future.

5.3. Exploring the Relationship between the Current Development of Healthcare Flow Management and the Manufacturing Flow Mechanism*s*

Comparison with Red2Green Approach

As stated in the literature, the Red2Green approach is currently widely promoted within the NHS and is similar, in some ways, to TBM and LM. This approach act as a visual control method to support in identifying wasted time during the journey of patients and has been adopted in both acute and community settings to minimise internal and external delays and deliver improved patient flow (NHS Improvement, 2017). In conjunction with the Red2 Green approach, the SAFER Patient Flow Bundle combines five elements of recommended practice that need to be enforced together to accomplish cumulative benefits. These five elements are: senior review; expected discharge date and clinical criteria for discharge for all patients; flow of patients at the earliest opportunity; early discharge; and senior review for patients for management and discharge decisions (NHS Improvement, 2017).

The Red2Green approach also acknowledges the importance of the PDD or EDD as the 'patient's time is the most important currency in health and social care' (Prof Brian Dolan, cited in Gordon, 2014, p. 5). As the emphasis of LM and TOC in manufacturing is on shortening the lead time and improving the flow, this approach aims to ensure that the time the patient spends in hospitals is a day well spent, with clear medical goals (Wyatt et al., 2019).

Although Red2 Green approach has shown a significant improvement in reducing LOS and flow in mental health in-patient services (Quinn et al., 2018), it has been argued that, while this approach can provide some indication of what is required, it does not address the issue of high bed occupancy, especially in mental health services. As illustrated in a document published by many authors exploring the NHS Strategy Unit for Mental Health Inpatient Capacity:

> We call for a national programme to support mental health providers to ensure that every day that a patient spends in hospital is a day well spent, with clear clinical

> objectives. Initiatives such as the Red2Green campaign provide some indication of what is required. We must be clear why a patient has been admitted and when a patient is well enough to be discharged. Reaching a clinical consensus on these issues, codifying the results, and introducing mechanisms to assess patients daily will not be straightforward. But it is clear that any proposed solution to the problem of high bed occupancy which does not consider how those beds are used, will be incomplete. (Wyatt et al., 2019, p. 8)

By emphasising the need to turn red days into green days in everyday care activities and to enable patients discharge, it seems there is a lack of central focus across the system that reflects the priority of the PDDs for patients, as this approach is more likely to emphasise expediting and escalating all aspects.

In comparison with TOC and LM, one of the QFI consultants demonstrated this approach during the interviews,

> So, basically, you either have, the patient has a red day or green day. If they have a green day, something happens towards their discharge. If they have a red day, nothing happens, and you're supposed to escalate whatever it is that hasn't happened. So, it's about expediting and escalating, but the problem is it's just escalate everything, it becomes huge. Everybody from each ward has to go down, state their escalation, but if you've got a hospital of 500 patients, you've got 500 escalations. Nothing happens because it's taking all day just to hear them. (P10)

This form of management might end with local improvement, because it does not necessarily provide focus from the perspective of the system. It might generate poor multi-tasking, as staff would be busy doing some other task that does not contribute to the overall goal of the system, rather than the focus being on the priority of the system represented by meeting patients' PDDs. As evidenced in manufacturing, making everything efficient does not necessarily contribute to the overall goal of the system and might be considered wasteful: 'Focus on everything, and you have not actually focused on anything' (Goldratt, 1990, 58).

This was emphasised in manufacturing, where the use of local optimisation measures to keep employees busy all the time would hinder the achievement of the global goal, as it contributes to the unnecessary accumulation of manufacturing and inventories and obscures the importance of this, by generating a misconception that higher rates of efficiency exist (Goldratt, 1990; Jones and Dugdale, 1998).

The Red2Green approach appears to be closer to the LM approach in terms of establishing standard procedures that staff must follow to convert any red sign to green. It may also be classified as a TOC approach by virtue of using similar terms and concepts that originated with the TBM approach in healthcare, such as the review of the progress of the PDDs by MDT and the use of color zoning, but the difference is that TOC is about the focus on addressing the requirements and priorities of the system, rather than focusing on things that do not solve the global constraint of the system.

> I think what people unfortunately tend to do is they tend to look at local constraints as opposed to global constraints and that, and that's why they don't make progress in the whole thing….some places do very well, but they would never ever reference TOC, I mean they do really good at Red2Green but let’s just leave that out. Both Trusts if you went into them and talked about Red2Green they would never, ever reference TOC as part of that process at all. (P31)

The Red2Green approach might be useful and help to improve the performance of the system, however a practical mechanism that acts proactively to ensure that the PDDs are met and helps to identify the causes of delays, enabling continual improvement, seems to be missing. The PDD is directly related to the patient and determines when the patient will be discharged, so providing staff with a shared focus and all related activities are subordinate to satisfying this PDD. To enable this, a practical mechanism should be put in place to prioritise patients based on their PDDs and to ensure that they do not experience a delayed discharge during their journey. The application of DBR-TBM has been shown to be a robust mechanism which provides a central focus by directing resources to ensure that the PDD is met and can track any tasks that are more likely to be delayed. It can, therefore, expedite actions before delayed discharge occurs.

CHAPTER SIX: CONCLUSION

This chapter represents the conclusion of the study and is structured as follows:

- Reviewing the research aim and objectives
- Response to Research Questions
- Research Contribution
- Limitation of Research
- Suggestions for Future Research

6.1. Reviewing the Research Aim and Objectives

The importance of patient flow as a construct is clearly evident today in healthcare institutions and conferences, however, as with manufacturing, there needs to be a mechanism tailored to the instability of needs in this environment. This is reflected in the development of different manufacturing mechanisms, including Ford's physical flow lines, Ohno's Kanban control, and Goldratt's time buffer management (TBM), yet little is known about the factors underpinning these distinctions and how they translate to the underlying healthcare environments. The unique quality of this research was that it aimed to examine how different manufacturing approaches, particularly LM and TOC, can be applied effectively across different healthcare settings. Specifically, it concentrated on investigating the applications of LM and TOC involved in managing and improving patient flow in different healthcare environments.

As detailed in the literature review chapter, the instability and complexity associated with the health and social care system are perceived to be a significant operational challenge. Many healthcare organizations acknowledged the importance of managing patient flow by taking a significant interest in LM and TOC approaches (Olsson and Aronsson, 2015). Within healthcare, there is a tendency to adopt one or more approaches, irrespective of the healthcare environment with little consideration of how these manufacturing approaches might be theoretically developed to meet the needs of different health and social care environments. Therefore, where LM and TOC have been applied, there have been mixed results regarding the success and assumptions underlying the application of such approaches to healthcare

(Stratton and Knight, 2010; Radnor et al., 2012; Blackmore and Kaplan, 2016). The research aim, therefore, was to:

Gain a deeper academic understanding of how established flow management approaches have been developed to meet the distinct conditions within the wider healthcare system, with specific reference to the origins of LM and TOC.

To achieve this aim, a multi-case study approach was conducted, involving over 30 interviews and the selective use of documentation and observation based on four healthcare organisations across three NHS Trusts in the UK. These organisations involved a range of delivery systems where there has been a transformative approach to improving patient flow. The cases incorporate acute and rehabilitative hospital care, social care, out-patient services, and GP-led community care and community mental health.

This adopted case study approach, in line with the best practice advocated by Eisenhardt (1989) and supported by an extensive literature review enabled the researcher to meet the main objectives of this research, which were:

- Building an understanding of the environmental conditions that led to the development of flow mechanisms associated with LM and TOC before considering how they have been interpreted in their transfer to health and social care environments;
- Investigating how the LM and TOC flow mechanisms can be applied effectively in the health and social care environments to manage and improve patient flows; and
- Providing practical guidance on the selection of flow mechanisms that can help healthcare organisations improve patient flow performance.

The combination of the significant bodies of knowledge relevant to this research led to the identification of research gaps as well as the theoretical justification of the research questions. Within case analysis, a more in-depth understanding of the transformative approach to improving patient flow in each individual case was achieved.

In addition, the cross-case analysis and discussion enabled the researcher to explore how the assumptions and associated practices of LM and TOC approaches transfer to healthcare based on Goldratt's principles of managing flow. Utilising these principles as the basis for interpreting the findings helped the researcher examine how the adopted flow mechanisms such as Kanban and TBM have been interpreted and translated in different healthcare environments, hence identifying the common theoretical assumptions and implications.

Overall, the analysis of the collected data contributed to understanding how patient flow can be managed effectively in a more complex medical pathway where there is an interface between different areas across the system (e.g. health and social care system) through the TBM flow mechanism. This emphasises the critical importance of implementing a practical flow mechanism that protects the patient flow from disruption and manages activities proactively to avoid delayed discharge.

6.2. Response to Research Questions

This section serves as a comprehensive summary of the overall research questions to present the main findings and discussions. The three main research questions are used as subheadings in this section.

RQ1: How can LM and TOC approaches be applied effectively to improve patient flow across health and social care?

LM and TOC have developed distinct approaches to managing flow to meet the needs of distinct manufacturing environments, but their success has resulted in flow being widely acknowledged as a proxy for system productivity. Healthcare management similarly acknowledges the importance of flow, but to what extent do the flow mechanisms associated with LM and TOC (Kanban and TBM respectively) support patient flow?

TBM has been shown to be a central feature of TOC across its core manufacturing applications and the same four functions have also been shown to apply in healthcare and social care systems. Kanban is often perceived to be one of many tools aimed at reducing variability and improving flow in LM. However, its critical importance in sustaining flow improvement is sometimes forgotten. This is achieved in manufacturing by establishing clearly defined flow paths, and this need for a clear physical path is a key distinction between TOC and LM.

TBM offers a simple means of managing the instability that undermines flow and focuses on timely accelerating, escalating and continuous improvement strategies (Goldratt, 1990; Chakravorty and Atwater, 1996; Schragenheim, 2010). This is achieved in manufacturing by ensuring that production is set based on market demand (the drum) and using one rope tied to the drum, to protect against sources of variability, and utilising TBM functions to ensure that the drum is never starved of orders on time.

The translation of the TBM approach into healthcare environments was evident in this research, where the PDD was used as a drum which reflects patient need (market demand), the rope was defined as the time to the PDD, and the remaining time for the patient to be discharged was referred to as buffer time. As evident in section 5.1.2., by setting the PDD based on medical needs, the TBM approach provided alignment signals for all support activities to be subordinated to the exploitation of the system constraint (ensuring the PDD is met when the patient is medically fit), through its four functions presented in form of the 'Jonah' software. Establishing an aggregation pooling and using the TBM functions (prioritise, expedite, escalate, and improve) provides a systematic process to avoid localised control and to sustain a continuous improvement culture. Thus, managing variability and balancing the flow within the delivery system is achieved.

The LM approach in healthcare is commonly associated with the creation of a dedicated care pathway that is more akin to a balanced line in the manufacturing environment. This is highlighted in section 5.1.2., where the use of LM tools, such as the 5S and visual management, to standardise work was evident in all the notable cases of successful LM implementation in healthcare, as well as the case observed in this research. Much of this has involved secondary processes, such as consumables, pharmacy, and test procedures, rather than patient flow. Although there have been significant improvements in patient flows, these are typically associated with the redesign of medical pathways (Spear, 2005; Ben-Tovim et al., 2007; Fillingham,2007; Grunden, 2007; O'Connell et al. , 2008; Gubb, 2009; Kenney, 2011; Blackmore et al., 2011; Radnor et al., 2012).

As with the case of Flinders Medical Centre and other cases (section 2.6), the redesign of the medical pathways for particular groups of patients, with dedicated teams of doctors and nurses, facilitated the standardisation of care and enhanced flow improvements. Ensuring central visibility was a key characteristic across all the cases, with process mapping used in all organisations to allow employees to visualise the entire patient journey within the delivery system. These applications are more akin to a flow line environment, as the establishment of TPS LM was based on Ford's flow line model (Ohno, 1988; Fujimoto, 1999). Therefore, these applications are consistent with TPS LM improving flow, through the redesign of the physical flow path (Blackmore et al., 2011; Radnor et al., 2012; Hicks et al., 2015).

Healthcare is a complex system characterised by a range of variability and uncertainty, and time is critical in this environment, as it is in manufacturing, and needs to be exploited. In both environments, the lack of flow management can result in localised control practices. In the two healthcare cases (1 and 2), the translation was associated with managing the level of variation and uncertainty through the establishment of an aggregation buffer and the use of TBM functions to ensure the system requirements were met. This was directly related to establishing the PDD for each patient and directing all resources to focus on the PDD, through one centralised aggregated list as opposed to localised control.

In the LM environment, Kanban was a central component of TPS, and the six rules of Kanban are the means for adopting a system-based approach (Ohno, 1988). With LM implementations in healthcare, there has been evidence of the use of Kanban as a pull system on a dedicated pathway (Black and Miller, 2008; Grunden, 2007; Kenney, 2011). However, evidence of the use of Kanban as a flow mechanism to manage the primary flow of patients across the system, as well as to maintain a continuous improvement process, remains difficult to find.

Hence, where the LM redesign is narrowly defined, as in kaizen blitz activities, the isolated improvements are not linked to the wider delivery system and commonly lack this important means of sustaining the improved flow. Although the case evidence was limited in this regard, the Productive Ward evidence and the use of 5S are consistent with prior research that identified this problem (Radnor et al, 2012), even though these findings challenge the subsequent evaluation (Radnor et al., 2013).

RQ2 Why do LM and TOC approaches work better in different health and social care environments?

The literature review in Section 2.2.2. identified that both LM and TOC shift the emphasis from local performance to system performance by enabling flow across the delivery system. This can be achieved by reducing variability and uncertainty, which is the emphasis of LM, whereas TOC is concerned first with changing the rules concerning the management of flow and the strategic location of buffers. Whereas LM is associated with a wide range of tools to support the reduction of variability and improving flow, TOC essentially has only one, namely TBM.

Although LM is associated with many tools, Kanban was key to the sustained operation of the TPS (Ohno, 1988). However, its application is naturally influenced by relatively low fluctuations in demand and the need for well-defined flow paths and process steps (Hall, 1981; Monden, 1983; Shingo,1989; Spear and Bowen, 1999; Stratton et al., 2008). Whereas TBM uses aggregated buffers, which is more accommodating to variability and uncertainty (Goldratt and Cox, 1984; Umble and Srikanth, 1997; Schragenheim and Dettmer, 2000; Schragenheim, 2010), Kanban is more sensitive to flow disruptions and, therefore, achieves higher performance in a well-defined delivery system.

More varied and uncertain environments benefit from the adoption of aggregated buffers as advocated by TBM in Case 1 (section 5.1.2.), where the move to a central patient queue allowed more effective priority control with the centralised allocation of therapists, as opposed to locally distributed resources. This key feature allows therapists to focus on managing one single queue and adopting a common priority system. The scheduling of the treatment plan at the first consultation (initial assessment appointment) is also consistent with flow management by setting a PDD, so ensuring that the number of patients in the system is kept low.

In Case 2, as highlighted in section 5.1.2., the establishment of a PDD was key to enabling the other support activities to be prioritised around the patients' needs and so proactively managing the threat DTOC, a particular issue for the elderly in their transfer between healthcare and social care providers. In this case, the use of formalised software (Jonah) enables the functions of TBM to be more effectively realised, including Pareto data to support buffer management meetings which seek ways of improving performance.

In both cases, shifting the focus to one aggregated buffer allows efficient use of available and protective capacity (capacity buffer aggregation) to support the requirements of the system and to ensure customer demand (PDD) is protected (exploiting the constraint). Hence, variability and uncertainty can be effectively managed.

By contrast, LM, deriving from the more structured TPS, supports physical flow requirements being structurally entrenched. Less varied and uncertain healthcare environments can derive benefit from the use of LM tools where there is opportunity to define the process flow. As in section 5.1.4, the adoption of 5S in the implementation of the productive toolkit in Case 3 represents efforts to reorganise processes structurally and ensure all resources are at the right place at the right time. This supports the establishment of flow paths through standardised processes. This was also evident in the literature in all successful examples of LM implementation, using LM tools to redesign medical pathways to deal with certain needs of patients (Fillingham, 2007; Kaplan, 2010; Kenney, 2011; Plsek, 2013). Better stability and load levelling resulted from using LM tools to enhance the performance of elective surgery for less complicated patient needs (Edwards et al., 2012; Daultani et al., 2015).

As with manufacturing, healthcare is characterised by a range of variability and uncertainty and, where there is opportunity to define the process flow tightly, LM can be readily applied. This includes patient flow, but these findings suggest it is more suited to elective surgery or emergency pathways that are predefined. Where the pathway is unknown, poorly defined, and involves transfer of care, the flow mechanism is better suited to TBM. This was evident in the TBM solutions witnessed, including the need to subordinate the flow focus to the PDD, thereby synchronising all activities around this patient, whether in out-patient or in-patient environments.

RQ3 What are the assumptions underpinning the effective use of Kanban and TBM in different healthcare environments?

Section 2.2.4. highlighted that both Kanban and TBM are a practical flow mechanism for reducing variation and uncertainty in manufacturing delivery systems, but these two mechanisms work under different environmental conditions that support their effective use. For Kanban to be applied effectively, the predefined flow paths, process steps, and time need to be in place. This was evident in most TPS designs, where the capacity levelling schedule was successfully enabled, indicating less fluctuation in demand and high stability of flow path operations (Hall, 1981; Monden, 1983; Ohno, 1988; Spear and Bowen, 1999; Feld, 2000). While Kanban is more sensitive to load levelling, demand fluctuations, and material flow paths, TBM is less sensitive and has been developed for manufacturing environments where process times, product mix, and demand are subject to high levels of variation and uncertainty (Chakravorty and Atwater, 1996; Dettmer, 2001; Goldratt 2009; Stratton and Knight, 2010; Pacheco, 2018).

In addition to these assumptions, section 2.2.4. demonstrated that the critical difference is that the TBM is time-based, while the Kanban is inventory-based. Rather than restricting overproduction, as with Kanban, by controlling the release of inventory at each workstation, time can be used with TBM to limit overproduction through the creation of a pooled buffer ahead of the system's constraint. Detecting any delays in the process can be expedited through TBM, once the pooled buffer is breached, and, by identifying and targeting causes of buffer penetration, a continual improvement process is established. With Kanban, some rules can detect and address any problems or delays in the process before they pass over to the next job. By addressing these issues, better control of the inventory achieved, resulting in a reduction of inventory over time and balancing the flow across the system.

The same premise associated with Kanban can be seen in the healthcare environment where the drive for standard requirements of patients made it possible to have predefined medical pathways and level scheduling (Edwards et al., 2012; Daultani et al., 2015). Hence, buffering is based on the bed as it acts as the inventory. Given the visibility of the medical pathways and standardisation of the procedure and time, any issues causing delays to the process can be easily detected and then targeted (Grunden, 2007). As with Kanban in TPS, the use of the rocks and water analogy to encourage continual improvement can be seen in healthcare, by reducing beds to expose problems that are then targeted, as illustrated in section 5.1.3.

With highly uncertain patient needs, the TBM approach has shown to have less sensitivity to redesigning the medical pathway and process in the healthcare organisations observed in this research. As with TBM in manufacturing, the use of a pooled buffer was applicable in complex healthcare environments, as TBM set rules exploiting the time constraint without the need to focus on all aspects of the system, as with Kanban. The ability to identify the causes of delays to the patient journey and to address these causes to balance patient flow was also witnessed through the application of TBM in Case 1 and Case 2, outlined in section 5.2.

The aggregated time buffer provides characteristics which are particularly suitable for manufacturing and healthcare systems with complexities, variability, and uncertainty, where Kanban is more sensitive in TBS and performs only under low levels of variation and uncertainty. These assumptions associated with Kanban and TMB are summarised in Table 17 in section 5.3.

6.2. Research Contribution

This research advanced a theoretical understanding of the importance of distinct manufacturing flow mechanisms in managing and improving flow across the entire delivery system and demonstrated how these mechanisms could be used selectively and effectively to fulfil the needs of different healthcare environments.

In manufacturing, there was an awareness of the need to implement a practical flow mechanism to ensure that not only overproduction is limited and a continuous improvement process is in place, but also to avoid local efficiency measures that conflict with this flow focus. This was most clearly demonstrated by the development of Goldratt's (2009) four flow concepts, which highlighted the common and important role of different manufacturing flow mechanisms such as Kanban and TBM in enabling flow while accounting for the various levels of variability and uncertainty associated with the environments where these flow mechanisms originate. Chocking the release of work into the system to avoid overproduction, identifying sources of variability that disrupt the flow, and eradicating local efficiency are the main functions of these flow mechanisms.

In healthcare, little is known about the assumptions underlying the successful implementation of LM and TOC within various healthcare environments, as evidenced by the mixed results of LM and TOC applications (Stratton and Knight, 2010; Radnor et al., 2012; Blackmore and Kaplan, 2016). Similarly, to manufacturing, different healthcare environments can be interpreted in terms of different levels of variability and uncertainty. Therefore, this multi-case study contributed to knowledge by extending Goldratt's four flow concepts framework to healthcare environments. Translating these concepts highlighted the role of the flow mechanisms (particularly TBM) in proactively avoiding delayed transfer of care compared to avoiding overproduction in manufacturing.

As with complex manufacturing environments, the translation of the TBM flow mechanism to manage complex patient flow effectively was evident in this research. Increasing variability and uncertainty in complex pathways necessitate a flow mechanism to ensure that patients are not delayed in receiving care or being discharged from one area of the system to another (e.g. transferring care from the acute path to community hospitals or residential care). This was evident in the TBM solutions witnessed, including the need to subordinate the flow focus to the PDD, thereby synchronising all activities centred on the patient, whether in outpatient or inpatient environments. In stable healthcare environments that encounter less variability and uncertainty, there is an opportunity to define the process flow, and hence LM can be easily applied. This was observed in case study materials investigated in this research, where the flow path is more stable, allowing the standardisation and use of dedicated pathways.

The findings of this thesis reveal significant implications for improving patient flow through developing a better understanding of how the proactive flow mechanism observed in this research successfully reduced the LOS/DTOC and avoided delayed discharge, which is the main issue for the elderly with transfer between healthcare and social care providers. Adopting a proactive flow mechanism that enables patient-centric care is significant in managing patient flow effectively to reduce delayed discharge and frees up additional capacity, avoiding strain on hospital capacity due to reactive policies implemented in response to a sudden unexpected situation as the covid-19 pandemic or winter seasons.

The distinct flow mechanisms are central to improving flow by adopting a system-based perspective through better alignment, involving flow mechanism as a means of offsetting the tendency to optimise locally. In many cases where flow redesign was complicated, this involved pooling time and capacity buffers, implementing a simple method of systematic management and sustaining system stability. Therefore, this research provides valuable and original material for operations and healthcare management literature and provides insights for healthcare administrators.

6.3. Limitations of Research

Although the research used a multiple case study approach to provide a more comprehensive picture and understanding of the transformation of flow mechanisms to healthcare settings, the current study still has limitations. The first limitation concerned limited access to documentary evidence highlighting improvements through a comparison of performance before and after the intervention and, thus, incomplete records of some of the cases evaluated. These case studies involved confidential information regarding the patients. The second limitation was the number of cases. Another limitation of the study is that the findings regarding Case 1 were based on interviews with participants who were no longer involved with the daily activities of the current therapy service. Although the researcher tried to interview people who were currently engaged and working with this service to get their views, those people's schedule was always busy. Thus, the opinions of people with current experience of daily activity in this service were limited. Further study can explore the interpretation of the transformation of the TOC approach from the personal perspective of those who are currently involved in daily service activities.

Finally, the case study approach used in accordance with Eisenhardt's (1989) best practice enabled the researcher to reach saturation and meet the main objectives of this research. Still, the number and diversity of cases were limited, implying that additional research may be required to enhance the findings. The study analysed only four cases, and the inclusion of more cases could provide further insights, especially regarding the translation of LM in healthcare with particular reference to patient flow.

6.4. Suggestions for Future Research

This section highlights briefly some interesting points that might be areas for future research. Although the TBM technique has been used successfully to manage patient flow, the TBM concepts might be more complicated to apply and require a cultural change and stable management to encourage the transformation process. As with Case 4 (see section 5.1.3.), the instability of management (e.g. changing of executive director) was evident, which might have impacted the sustainability of the TBM solution This compares unfavourably to Case 2, where the approach implemented by the same consulting firm (QFI) has had strong support from senior management, resulting in a long-standing solution and sustainable improvement in patient flow for more than nine years, even with the need to duplicate data inputting into the national NHS System One, which does not provide the data in a form suited to managing patient flow.

It is also interesting to note that interviews conducted with senior managers indicated that although TOC concepts and tools are widely used (e.g. colour zoning and PDD), TOC is not understood, resulting in limited benefits.

Compared to the findings from the LM approach, it appears that LM tools can perform better than anticipated in complex cases in terms of redesigning the flow path. As pointed out by an expert in the NHS Improvement, the redesign of the processes was useful during the COVID 19 pandemic to improve services and to protect both patients and staff from viral transmission, and there is certainly a possibility, even in complex cases, to improve the process.

> The theory of constraints is much more about how you view the world, whereas lean is really about how you improve processes. (P31)

From the above discussion, it is presumed that the use of TBM as a practical mechanism to encourage flow has proved to be effective, but the concept of TBM may be too complicated for many organisations, and LM may function better than expected in terms of redesigning the flow. Further research to test the suggestions of this study is therefore recommended, in particular, the suggestion that there is a need for a practical mechanism to manage the flow and why TBM can tend to be more difficult to incorporate in healthcare organisations. The integration between LM tools and TBM in managing patient flow could also be investigated. Developing current hospital practices (e.g. System One) to include features that underlie the practical flow mechanism could be another area for further investigation.

Future research could also focus on institutional culture, which is crucially important, as institutional culture is the shared vision which determines how value to the patient relates to performance, efficiency, safety, equity, and timeliness. Similarly, a reliable healthcare system needs leadership. Quality management is at its essence about transformation, and the vast and changing required improvements in healthcare can only be accomplished by committed, unwavering leadership.

In addition, government-regulated initiatives, such as DTOC or targets, are considered critical for encouraging flow and for fostering cooperation between health and social care providers. Nevertheless, it seems that these measurements are being misused in a way that is contradictory to the overall intent of the system. Instead of addressing system requirements by focusing on discharging delayed patients from the hospital, each department only keeps track of its own success, irrespective of anything else, shifting the global focus of the system to an emphasis on local optimisation. Therefore, further study might investigate the impact of using flow measurements such as DTOC on the overall system performance and whether there is a relationship between these measures and the encouragement of local optimisation practice.

Finally, challenges and conditions facing the health care sector are always growing and never cease. Cancellation of elective surgery during a seasonal crisis might be understood; however, this is still not the optimal solution to any crisis, especially with the advent of the unexpected crisis of the COVID 19 pandemic. This has made the situation worse, suggesting the need for more studies to explore the better management of hospital services in any circumstances.

REFERENCES

Aeenparast, A., Farzadi, F., Maftoon, F. and Yahyazadeh, H., 2019. Patient Flow Analysis in General Hospitals: How Clinical Disciplines Affect Outpatient Wait Times. *Hospital Practices and Research*, 4(4), pp.128-133.

Aherne, J. and Whelton, J. eds., 2010. *Applying lean in healthcare: a collection of international case studies*. CRC Press.

Al-Balushi, S., Sohal, A.S., Singh, P.J., Al Hajri, A., Al Farsi, Y.M. and Al Abri, R., 2014. Readiness factors for lean implementation in healthcare settings–a literature review. *Journal of health organization and management*.

Alderwick, H., Ham, C. and Buck, D., 2015. Population health systems. *Going beyond integrated care, The King's Fund*.

Amonge, A., 2015. Application of Goldratt's Thinking Process to Constraints Within an Emergency Department—A Case Study.

Andersen, H. and Røvik, K., 2015. Lost in translation: a case-study of the travel of lean thinking in a hospital. *BMC Health Services Research*, 15(1).

Andersen, H., Røvik, K. and Ingebrigtsen, T., 2014. Lean thinking in hospitals: is there a cure for the absence of evidence? A systematic review of reviews. *BMJ Open*, 4(1), p.e003873.

Anderson, J. and Aydin, C., 2005. Overview: Theoretical perspectives and methodologies for the evaluation of healthcare information systems. In *Evaluating the organizational impact of healthcare information systems* (pp. 5-29). Springer, New York, NY.

Anderson, P., 1983. Decision Making by Objection and the Cuban Missile Crisis. *Administrative Science Quarterly*, 28(2), p.201.

Anthony, S. and Jack, S., 2009. Qualitative case study methodology in nursing research: an integrative review. *Journal of advanced nursing*, *65*(6), pp.1171-1181.

Antierens, A., Beeckman, D., Verhaeghe, S., Myny, D. and Van Hecke, A., 2018. How much of Toyota's philosophy is embedded in health care at the organisational level? A review. *Journal of Nursing Management*, 26(4), pp.348-357.

Antunes, J., 1998. *Towards A General Theory Of The Process In The Administration Of Production: A Discussion About The Possibility Of Unification Of The Theory Of Constraints And The Theory That Supports The Construction Of Production Systems With Zero Inventory*. PhD thesis. School of Management, Federal University of Rio Grande do Sul.

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y. and Yom-Tov, G., 2015. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, 5(1), pp.146-194.

Atkinson, P. and Coffey, A., 2004. Analysing documentary realities. *Qualitative research*, *3*, pp.77-92.

Bacharach, S., 1989. Organizational Theories: Some Criteria for Evaluation. *The Academy of Management Review*, 14(4), p.496.

Barnable, A., Welsh, D., Lundrigan, E. and Davis, C., 2015. Analysis of the influencing factors associated with being designated alternate level of care. *Home Health Care Management & Practice*, *27*(1), pp.3-12.

Barnas, K., 2011. ThedaCare's Business Performance System: Sustaining Continuous Daily Improvement Through Hospital Management in a Lean Environment. *The Joint Commission Journal on Quality and Patient Safety*, 37(9), pp.387-AP8.

Barratt, M., Choi, T. and Li, M., 2011. Qualitative case studies in operations management: Trends, research outcomes, and future research implications. *Journal of Operations Management*, 29(4), pp.329-342.

Baumann, M., Evans, S., Perkins, M., Curtis, L., Netten, A., Fernandez, J.L. and Huxley, P., 2007. Organisation and features of hospital, intermediate care and social services in English sites with low rates of delayed discharge. *Health & social care in the community*, *15*(4), pp.295-305.

Becker, H., 1986. *Writing for social scientists: How to start and finish your thesis, book, or article.* Chicago: University of Chicago Press.

Ben-Tovim, D., Bassham, J., Bolch, D., Martin, M., Dougherty, M. and Szwarcbord, M., 2007. Lean thinking across a hospital: redesigning care at the Flinders Medical Centre. *Australian Health Review*, *31*(1), pp.10-15.

Berg, B., Denton, B., Nelson, H., Balasubramanian, H., Rahman, A., Bailey, A. and Lindor, K., 2010. A discrete event simulation model to evaluate operational performance of a colonoscopy suite. *Medical Decision Making*, *30*(3), pp.380-387.

Bergen, A. and While, A. 2000. A Case for Case Studies: Exploring the Use of Case Study Design in Community Nursing Research. *Journal of Advanced Nursing*, 31 (4), pp. 926- 934.

Berger, M., 1983. In Defense of the Case Method: A Reply to Argyris. *The Academy of Management Review*, 8(2), p.329.

Bergeron, M. and Audet, N., 2015. One hundred twelve patients above 75 years old with tracheotomy: discharge delayed by 13 days: Our experience. *Clinical Otolaryngology*, 41(1), pp.91-95.

Best, M. and Neuhauser, D., 2006. Walter A Shewhart, 1924, and the Hawthorne factory. *BMJ Quality & Safety*, *15*(2), pp.142-143.

Bhasin, S., 2008. Lean and performance measurement. *Journal of Manufacturing Technology Management*, 19(5), pp.670-684.

Bhasin, S. and Burcher, P., 2006. Lean viewed as a philosophy. *Journal of manufacturing technology management*.

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

Bicheno, J., 2004. *The New Lean Toolbox*. Buckingham, England: Production and Inventory Control, Systems and Industrial Engineering Books.

Bingham, C. and Eisenhardt, K., 2011. Rational heuristics: the 'simple rules' that strategists learn from process experience. *Strategic Management Journal*, 32(13), pp.1437-1464.

Bisognano, M., 2016. So-called "flow failures" are disrespectful to patients. *Institute for Healthcare Improvement, Cambridge, MA*.

Black, J. and Miller, D., 2008. *The Toyota Way to Healthcare Excellence*. Chicago, Ill: Health Administration Press.

Blackmore, C. and Kaplan, G., 2016. Lean and the perfect patient experience. *BMJ Quality & Safety*, 26(2), pp.85-86.

Blackmore, C., Mecklenburg, R. and Kaplan, G., 2011. At Virginia Mason, collaboration among providers, employers, and health plans to transform care cut costs and improved quality. *Health Affairs*, *30*(9), pp.1680-1687.

Bluhm, D., Harman, W., Lee, T. and Mitchell, T., 2011. Qualitative Research in Management: A Decade of Progress. *Journal of Management Studies*, 48(8), pp.1866-1891.

Blumer, H., 1986. *Symbolic interactionism: Perspective and method*. Univ of California Press.

Bowen, G., 2009. Document analysis as a qualitative research method. *Qualitative research journal*, *9*(2), pp.27-40.

Bowen, M. and Neuhauser, D., 2013. Understanding and managing variation: three different perspectives. *Implementation Science*, 8(S1).

Boyer, K. and Swink, M., 2008. Empirical Elephants-Why Multiple Methods are Essential to Quality Research in Operations and Supply Chain Management. *Journal of Operations Management*, 26(3), pp.338-344.

Brandao de Souza, L., 2009. Trends and approaches in lean healthcare. *Leadership in Health Services*, 22(2), pp.121-139.

Brandeau, M., Sainfort, F. and Pierskalla, W., eds., 2004. *Operations Research and Health Care: A Handbook of Methods and Applications*. Kluwer Academic Publishers, London

Bromley, D., 1986. *The Case-Study Method in Psychology And Related Disciplines*. Chichester: Wiley.

Brown, P., 2008. A review of the literature on case study research. *Canadian Journal for New Scholars in Education*, *1*(1), 1-13.

Burgess, N. and Radnor, Z., 2013. Evaluating Lean in healthcare. *International Journal of Health Care Quality Assurance*, 26(3), pp.220-235.

Burrell, G. and Morgan, G., 2017. *Sociological paradigms and organisational analysis: Elements of the sociology of corporate life*. Routledge.

Bygrave, W., 1989. The Entrepreneur ship Paradigm: A Philosophical Look at Its Research Methodologies. *Entrepreneurship Theory and Practice*, 14(1), pp.7-26.

Callard, L., 2009. Working with nurses in the community. *Nursing times*, *105*(9), pp.suppl-15.

Cameron, R. and Molina-Azorín, J., 2011. Mixed methods research in business and management.

Camgöz-Akdağ, H., Camgöz-Akdağ, H., Çalişkan, E., Çalişkan, E., Toma, S. and Toma, S., 2017. Lean process design for a radiology department. *Business Process Management Journal*, *23*(4), pp.779-791.

Campbell, D., Morris, S. and Marsh, S., 2017. *NHS Faces 'Humanitarian Crisis' As Demand Rises, British Red Cross Warns*. [online] the Guardian. Available at: <https://www.theguardian.com/society/2017/jan/06/nhs-faces-humanitarian-crisis-rising-demand-british-red-cross> [Accessed 3 August 2017].

Carman, K., Paez, K., Stephens, J., Smeeding, L., Garfinkel, S., Blough, C. and Mapes, D., 2014. Improving care delivery through lean: Implementation case studies. Rockville, MD: Agency for Healthcare Research and Quality.

Carson, D., Gilmore, A., Perry, C. and Gronhaug, K., 2001. *Qualitative marketing research*. Sage.

Cassell, C., Symon, G., Buehring, A. and Johnson, P., 2006. The role and status of qualitative methods in management research: an empirical account. *Management Decision*, 44(2), pp.290-303.

Cavaye, A., 1996. Case study research: a multi-faceted research approach for IS. *Information systems journal*, *6*(3), pp.227-242.

Cayirli, T., Veral, E. and Rosen, H., 2008. Assessment of patient classification in appointment system design. *Production and Operations Management*, *17*(3), pp.338-353.

Cayirli, T. and Veral, E., 2003. Outpatient Scheduling in Health Care: A review of Literature. *Production and Operations Management*, 12(4), pp.519-549.

CGI, 2014. *Healthcare Challenges and Trends The Patient At The Heart Of Care*. [online] Available at: <https://www.cgi.com/sites/default/files/white-papers/cgi-health-challenges-white-paper.pdf> [Accessed 23 August 2016].

Chakravorty, S. and Brian Atwater, J., 1996. A comparative study of line design approaches for serial production systems. *International Journal of Operations & Production Management*, *16*(6), pp.91-108.

Chan, H., Lo, S., Lee, L., Lo, W., Yu, W., Wu, Y., Ho, S., Yeung, R. and Chan, J., 2014. Lean techniques for the improvement of patients' flow in emergency department. *World journal of emergency medicine*, *5*(1), p.24.

Chidwick, P., Oliver, J., Ball, D., Parkes, C., Hansen, T.L., Fiumara, F., Ferrari, K., Hawkswell, C. and Lumsden, K., 2017. Six Change Ideas that Significantly Minimize Alternate Level of Care (ALC) Days in Acute Care Hospitals. *Healthcare quarterly (Toronto, Ont.)*, *20*(2), pp.37-43.

Chou, Y., Lu, C. and Tang, Y., 2012. Identifying inventory problems in the aerospace industry using the theory of constraints. *International Journal of Production Research*, 50(16), pp.4686-4698.

Churchman, C., 1961. *Prediction and optimal decision*. Englewood Cliffs, NJ: Prentice-Hall.

Cohen, L., Manion, L., and Morrison, K., 2007. *Research Methods in Education*. London: Routledge

Cohen, L. and Manion, L., 1994. Introduction: The nature of enquiry. *Research methods in education*, *4*, pp.1-43.

Collaborative, C., 2020. Elective surgery cancellations due to the COVID-19 pandemic: global predictive modelling to inform surgical recovery plans. *British Journal of Surgery*.

Costa, L., Filho, M., Rentes, A., Bertani, T. and Mardegan, R., 2017. Lean healthcare in developing countries: evidence from Brazilian hospitals. *The International journal of health planning and management*, *32*(1), pp. e99-e120.

Cox III, J., Robinson, T. and Maxwell, W., 2014. Applying the' Theory of Constraints' to solve your practice's most vexing problem. *Family practice management*, *21*(5), pp.18-22.

Cox, J. and Schleier, J., 2010. *Theory of Constraints Handbook*. New York: McGraw-Hill.

CPA. 2014. *The effectiveness of care pathways in health and social care*. [online] Available at: <http://www.cpa.org.uk/information/reviews/CPA-Rapid-Review-Effectiveness-of-care-pathways.pdf> [Accessed 25 February 2021].

CQC. 2015. *Service types | Care Quality Commission*. [online] Available at: <https://www.cqc.org.uk/guidance-providers/regulations-enforcement/service-types#extra-care-housing-services> [Accessed 20 February 2021].

Crawford-Mason, C., 2002. Deming and me. Quality Progress September, 46.

Crema, M. and Verbano, C., 2015. How to combine lean and safety management in health care processes: A case from Spain. *Safety Science*, *79*, pp.63-71.

Creswell, J., 2014. *Research Design: Qualitative, Quantitative, And Mixed Methods Approaches*. 4th ed. Thousand Oaks, California: SAGE Publications Ltd.

Creswell, J., 2003. *Research Design*. Thousand Oaks, Calif.: Sage Publications.

Creswell, J., 1998. *Qualitative Inquiry and Research Design: Choosing Among Five Traditions*, Sage Publications.

Crotty, M., 1998. *The foundations of social research: Meaning and perspective in the research process*. Sage.

Cusumano, M., 1988. Manufacturing innovation: lessons from the Japanese auto industry. *MIT Sloan Management Review*, *30*(1), pp.29-39.

D'Andreamatteo, A., Ianni, L., Lega, F. and Sargiacomo, M., 2015. Lean in healthcare: a comprehensive review. *Health Policy*, *119*(9), pp.1197-1209.

Dannapfel, P., Poksinska, B. and Thomas, K., 2014. Dissemination strategy for Lean thinking in health care. *International Journal of Health Care Quality Assurance*, 27(5), pp.391-404.

Darlington, J., Francis, M., Found, P. and Thomas, A., 2014. Design and implementation of a Drum-Buffer-Rope pull-system. *Production Planning & Control*, 26(6), pp.489-504.

Daultani, Y., Chaudhuri, A. and Kumar, S., 2015. A Decade of Lean in Healthcare: Current State and Future Directions. *Global Business Review*, 16(6), pp.1082-1099.

Dayan, M., 2017. General Election 2017: Getting a Brexit deal that works for the NHS. *Nuffield Trust, Nuffield Trust Briefing*.

Dean, E., 2009. Productive ward measures see hospital's mortality rates plummet. *Nursing Standard*, 24(17), pp.8-8.

Dekier, Ł., 2012. The Origins and Evolution of Lean Management System. *Journal of International Studies*, 5(1), pp.46-51.

De Vos, A., Delport, C., Fouché, C. and Strydom, H., 2011. *Research at grass roots: A primer for the social science and human professions*. Van Schaik Publishers

de Vries, J. and Huijsman, R., 2011. Supply chain management in health services: an overview. *Supply Chain Management: An International Journal*, 16(3), pp.159-165.

Deming, W., 1994. New Economics, MIT Press, Cambridge, MA

Deming, W., 1986. *Out of The Crisis*. Cambridge, MA, Massachusetts Institute of Technology.

Deming, W., 1982. *Quality, Productivity and Competitive Position,* MIT Press, Cambridge, MA.

Deming, W., 1981. Improvement of quality and productivity through action by management. *National productivity review*, *1*(1), pp.12-22.

Dettmer, H., 2001. Beyond Lean manufacturing: Combining Lean and the Theory of Constraints for higher performance. *Port Angeles, US*.

DeWalt, K., and Billie R. DeWalt. 2002. *Participant Observation: A guide for fieldworkers*.

DHSC. 2018. *How the NHS charges overseas visitors for NHS healthcare*. [online] Available at: <https://www.gov.uk/government/publications/how-the-nhs-charges-overseas-visitors-for-nhs-hospital-care> [Accessed 27 March 2021].

Dickson, E., Anguelov, Z., Vetterick, D., Eller, A. and Singh, S., 2009. Use of lean in the emergency department: a case series of 4 hospitals. *Annals of emergency medicine*, *54*(4), pp.504-510.

Dombrowski, U. and Mielke, T., 2013. Lean leadership–fundamental principles and their application. *Procedia cirp*, *7*, pp.569-574.

Drotz, E. and Poksinska, B., 2014. Lean in healthcare from employees' perspectives. *Journal of health organization and management*.

Drupsteen, J., Vaart, T. and Donk, D., 2013. Integrative practices in hospitals and their impact on patient flow. I*nternational Journal of Operations & Production Management, 33(7)*, pp.912-933.

Dunn, P., Maguire, D. and McKenna, H., 2016. *How hospital activity and funding in England have changed over time*. [online] The King's Fund. Available at: https://www.kingsfund.org.uk/publications/hospital-activity-funding-changes [Accessed 21 Dec. 2016].

Dunn, P., McKenna, H. and Murray, R., 2016. Deficits in the NHS. *London: The King's Fund*.

Easterby-Smith, M., Thorpe, R. and Jackson, P., 2015. *Management and Business Research*. 5th ed. London: SAGE Publications.

Easterby-Smith, M., Thorpe, R. and Jackson, P., 1991. *Management and Business Research*. 1st ed. London: SAGE Publications, p.22.

Edwards, K., Nielsen, A. and Jacobsen, P., 2012. Implementing lean in surgery â lessons and implications. *International Journal of Technology Management*, 57(1/2/3), p.4.

Eisenhardt, K. and Graebner, M., 2007. Theory Building from Cases: Opportunities And Challenges. *Academy of Management Journal*, 50(1), pp.25-32.

Eisenhardt, K., 1991. Better stories and better constructs: The case for rigor and comparative logic. *Academy of Management review*, *16*(3), pp.620-627.

Eisenhardt, K., 1989. Building theories from case study research. *Academy of management review*, *14*(4), pp.532-550.

Eisner, E., 1991. The enlightened eye: Qualitative enquiry and the enhancement of aesthetic experience. *New York: Macmillan. research. Educational Researcher*, *22*(7), pp.5-11.

Elliott, V., 2018. Thinking about the coding process in qualitative data analysis. *The Qualitative Report*, *23*(11), pp.2850-2861.

Erickson, F., 1986. Qualitative research on teaching. *Handbook of research on teaching*, *3*.

Everall, A.C., Guilcher, S.J., Cadel, L., Asif, M., Li, J. and Kuluski, K., 2019. Patient and caregiver experience with delayed discharge from a hospital setting: a scoping review. *Health Expectations*, *22*(5), pp.863-873.

Feld, W., 2000. *Lean manufacturing: tools, techniques, and how to use them*. CRC press.

Ferdows, K. and De Meyer, A., 1990. Lasting improvements in manufacturing performance: In search of a new theory. *Journal of Operations Management*, 9(2), pp.168-184.

Fillingham, D., 2007. Can lean save lives?. *Leadership in health services*, *20*(4), pp.231-241.

Fillingham, D., Jones, B. and Pereira, P., 2016. *The challenge and potential of whole system flow*. Health Foundation.

Flick, U., 2009. *An introduction to qualitative research*. London: SAGE.

Flynn, R., Rotter, T., Hartfield, D., Newton, A.S. and Scott, S.D., 2019. A realist evaluation to identify contexts and mechanisms that enabled and hindered implementation and had an effect on sustainability of a lean intervention in pediatric healthcare. *BMC health services research*, *19*(1), pp.1-12.

Ford, H., 1926. *Today and Tomorrow.* Portland, OR: Productivity.

Foster, S., Gordon, P. and McSherry, W., 2009. Rolling out Productive Ward foundation modules across a hospital trust. *Nursing times*, *105*(30), pp.28-30.

Fugimoto, T., 1999. *The Evolution of a Manufacturing System at Toyota*. Oxford University Press, Oxford

George, A., and Bennett, A., 2005. *Case studies and theory development in the social sciences.* Cambridge, MA: MIT Press.

Gephart, R., 2004. Qualitative Research and the Academy of Management Journal. *Academy of Management Journal*, 47(4), pp.454-462.

Gersick, C., 1988. Time and Transition in Work Teams: Toward a New Model of Group Development. *Academy of Management Journal*, 31(1), pp.9-41.

Ghobadian, A., Viney, H. and Redwood, J., 2009. Explaining the unintended consequences of public sector reform. *Management Decision*, 47(10), pp.1514-1535.

Gibbs, G., 2007. The nature of qualitative analysis. *Analyzing qualitative data*, pp.1-9.

Giddings, L. and Grant, B., 2007. A Trojan Horse for Positivism?. A critique of mixed methods research. *Advances in Nursing Science*, 30(1), pp.52-60.

Gioia, D. and Pitre, E., 1990. Multiparadigm Perspectives on Theory Building. *The Academy of Management Review*, 15(4), p.587.

Glaser, B. and Strauss, A., 1967. Grounded theory: The discovery of grounded theory. *Sociology the journal of the British sociological association*, *12*(1), pp.27-49.

Glouberman, S. and Mintzberg, H., 2001. Managing the care of health and the cure of disease—Part I: Differentiation. *Health care management review*, *26*(1), pp.56-69.

Golafshani, N., 2003. Understanding reliability and validity in qualitative research. *The qualitative report*, *8*(4), pp.597-606.

Goldratt, E., 2009. Standing on the shoulders of giants: production concepts versus production applications. The Hitachi Tool Engineering example. *Gestão & produção*, *16*(3), pp.333-343.

Goldratt, E., 1999. *Theory of Constraints.* North River Press

Goldratt, E., 1990. *Theory of constraints*. Great Barrington, MA: North River Press.

Goldratt, E., 1990. *What is this Thing Called Theory of Constraints and How should it be Implemented?*. Croton-on-Hudson, NY: North River Press, Inc.

Goldratt, E. and Cox, J., 1986. *The Goal: A Process of Ongoing Improvement.* rev. ed. Croton-
on-Hudson, NY: North River Press.

Goldratt, E. and Cox, J., 1984. *The Goal: Excellence in Manufacturing.* Croton-on-Hudson, NY: North River Press.

Goldratt, E. and Fox, R., 1986. *The Race.* Croton-on-Hudson, NY: North River Press.

Gordon, P., 2014. *Red2green*. [online] NHS Improvement. Available at: <https://improvement.nhs.uk/documents/1432/Bishipsgate_1_THIRD_Red2Green.pdf> [Accessed 4 October 2018].

Grass, F., Behm, K., Duchalais, E., Crippa, J., Spears, G., Harmsen, W., Hübner, M., Mathis, K., Kelley, S., Pemberton, J. and Dozois, E., 2020. Impact of delay to surgery on survival in stage I-III colon cancer. *European Journal of Surgical Oncology*, *46*(3), pp.455-461.

Grol, R., Bosch, M., Hulscher, M., Eccles, M. and Wensing, M., 2007. Planning and studying improvement in patient care: the use of theoretical perspectives. *The Milbank Quarterly*, *85*(1), pp.93-138.

Grosios, K., Gahan, P. and Burbidge, J., 2010. Overview of healthcare in the UK. *EPMA Journal*, 1(4), pp.529-534.

Grunden, N., 2007. *The Pittsburgh way to efficient healthcare: Improving patient care using Toyota based methods*. CRC Press.

Guba, E. and Lincoln, Y., 1994. Competing paradigms in qualitative research. *Handbook of qualitative research*, *2*(163-194), pp.105-117.

Guba, E. and Lincoln, Y., 1981. *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches*. Jossey-Bass.

Gubb, J., 2009. Have targets done more harm than good in the English NHS? Yes. *BMJ*, 338, p.a3130.

Gupta, M. and Boyd, L., 2011. An Excel- based dice game: an integrative learning activity in operations management. *International Journal of Operations & Production Management,* 31(6), 608-630

Gupta, M. and Boyd, L., 2008. Theory of constraints: a theory for operations management. *International Journal of Operations & Production Management*, 28(10), pp.991-1012.

Gupta, D. and Denton, B., 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, *40*(9), pp.800-819

Gupta, M. and Kline, J., 2008. Managing a community mental health agency: A Theory of Constraints based framework. *Total Quality Management*, *19*(3), pp.281-294.

Hall, R., 2013. *Patient Flow: Reducing Delay in Healthcare Delivery,* Second Edition. 1st ed. Springer.

Hall, R., 1983. *Zero Inventories*. Homewood, Ill.: Business One Irwin.

Hall, R., 1981. Driving the Productivity Machine: Production, Planning and Control in Japan American Production and Inventory Control Society. *Falls Church, Virginia*.

Hallam, C. and Contreras, C., 2018. Lean healthcare: scale, scope and sustainability. *International Journal of Health Care Quality Assurance*, 31(7), pp.684-696.

Hammersley, M., 2012. Troubling theory in case study research. *Higher education research & development*, *31*(3), pp.393-405.

Hancock, D. and Algozzine, R., 2011. *Doing Case Study Research*. New York: Teachers College Press. Columbia University.

Haraden, C. and Resar, R., 2004. Patient flow in hospitals: understanding and controlling it better. *Frontiers of health services management*, *20*(4), p.3.

Harling, K., 2012. An overview of case study. *Available at SSRN 2141476*.

Harris, S. and Sutton, R., 1986. Functions of Parting CEremonies in Dying Organizations. *Academy of Management Journal*, 29(1), pp.5-30.

Harrison, S., 2008. Accident and infection charts put on display to motivate ward team. *Nursing Standard*, 22(25), pp.9-9.

Hartley, J., 2004. Case study research. *Essential guide to qualitative methods in organizational research*, *1*, pp.323-333.

Hassin, R. and Mendel, S., 2008. Scheduling arrivals to queues: A single-server model with no-shows. *Management science*, *54*(3), pp.565-572.

Hayes, R., & Wheelwright, S., 1979. Link Manufacturing Process and Product Life Cycles. *Harvard Business Review, 57*, pp.33–140.

Hewitt, T.A. and Chreim, S., 2015. Fix and forget or fix and report: a qualitative study of tensions at the front line of incident reporting. *BMJ Quality & Safety*, *24*(5), pp.303-310.

Hicks, C., McGovern, T., Prior, G. and Smith, I., 2015. Applying lean principles to the design of healthcare facilities. *International Journal of Production Economics*, *170*, pp.677-686.

Hill, A., and T. Hill., 2012. *Operations Management*. 3rd ed. Palgrave MacMillan.

Hill, T., 1985. *Manufacturing Strategy*. London: Macmillan Education.

Holden, R., Eriksson, A., Andreasson, J., Williamsson, A. and Dellve, L., 2015. Healthcare workers' perceptions of lean: A context-sensitive, mixed methods study in three Swedish hospitals. *Applied Ergonomics*, 47, pp.181-192.

Holden, R., 2011. Lean thinking in emergency departments: a critical review. *Annals of emergency medicine*, *57*(3), pp.265-278.

Hopp, W., 2011. *Supply Chain Science*. Long Grove, Ill.: Waveland Press.

Hopp, W. and Spearman, M., 2004. To pull or not to pull: what is the question?. *Manufacturing & service operations management*, *6*(2), pp.133-148.

Hopp, W. and Spearman, M., 2001. *Factory physics*. 2nd ed. New York, N.Y.: Irwin/McGraw-Hill.

Hostetter, M. and Klein, S., 2020. *In Focus: Improving Patient Flow—In and Out Of Hospitals And Beyond | Commonwealth Fund*. [online] Commonwealthfund.org. Available at: <https://www.commonwealthfund.org/publications/newsletter-article/focus-improving-patient-flow-and-out-hospitals-and-beyond> [Accessed 16 June 2020].

Howells, D. and Smeeton, P., 2015. *Service Improvement Startegy*. Nottinghamshire Healthcare NHS Foundation Trust.

Huberman, M. and Miles, M., 2002. *The qualitative researcher's companion*. Sage.

Hudson, L. and Ozanne, J., 1988. Alternative ways of seeking knowledge in consumer research. *Journal of consumer research*, *14*(4), pp.508-521.

Hughes, C., 2003. *Disseminating qualitative research in educational settings: A critical introduction*. McGraw-Hill Education (UK).

Humphries, R., 2017. Delayed Discharge: 'It's Not Just About the Money'. [online] The King's Fund. Available at: <https://www.kingsfund.org.uk/blog/2017/01/delayed-discharge-not-just-about-money> [Accessed 3 April 2017].

Ihi.org. 2003. *Optimizing Patient Flow: Moving Patients Smoothly Through Acute Care Settings | IHI - Institute for Healthcare Improvement*. [online] Available at: <http://www.ihi.org/resources/Pages/IHIWhitePapers/OptimizingPatientFlowMovingPatientsSmoothlyThroughAcuteCareSettings.aspx> [Accessed 7 November 2018].

Inman, R., Lair Sale, M. and Green, K., 2009. Analysis of the relationships among TOC use, TOC outcomes, and organizational performance. *International Journal of Operations & Production Management*, 29(4), pp.341-356.

Jabareen, Y., 2009. Building a conceptual framework: philosophy, definitions, and procedure. *International journal of qualitative methods*, *8*(4), pp.49-62.

Jacobsson, T., 2012. Operations Management in Healthcare: Principles for Creating Swift Even Patient Flow and Increased Accessibility. *Chalmers University of Technology, Sweden*.

Jarrett, P., 2006. An analysis of international health care logistics: the benefits and implications of implementing just-in-time systems in the health care industry. *Leadership in Health Services*, *19*(1), pp.1-10.

JCAHO, 2004. JCAHO Requirement: New Leadership Standard on Managing Patient Flow for Hospitals. *Joint Commission Perspectives* 24 13–14.

Jick, T., 1979. Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly*, *24*(4), pp.602-611.

Jogulu, U. and Pansiri, J., 2011. Mixed methods: A research design for management doctoral dissertations. *Management research review*.

Johansson, R., 2003. Key note speech at the international conference' Methodologies in Housing Research,' Royal Institute of Technology in cooperation with the International Association of People–Environment Studies, Stockholm, September 22-24, 2003.

Jones, D., 2015. *Four Lessons on Lean Healthcare - Lean Enterprise Academy*. [online] Lean Enterprise Academy. Available at: <https://www.leanuk.org/four-lessons-on-lean-healthcare/> [Accessed 29 August 2017].

Joppe, G., 2000. Testing reliability and validity of research instruments. *Journal of American Academy of Business Cambridge*, *4*(1/2), pp.49-54.

Junior, M. and Filho, M., 2010. Variations of the kanban system: Literature review and classification. *International Journal of Production Economics*, 125(1), pp.13-21.

Kahneman, D. and Tversky, A., 1973. On the psychology of prediction. *Psychological review*, *80*(4), p.237.

Kämäräinen, V., Peltokorpi, A., Torkki, P. and Tallbacka, K., 2016. Measuring healthcare productivity – from unit to system level. *International Journal of Health Care Quality Assurance*, 29(3), pp.288-299.

Kannampallil, T., Schauer, G., Cohen, T. and Patel, V., 2011. Considering complexity in healthcare systems. *Journal of biomedical informatics*, *44*(6), pp.943-947.

Kaplan, G., Patterson, S., Ching, J. and Blackmore, C., 2014. Why Lean doesn't work for everyone. *BMJ Quality & Safety*, 23(12), pp.970-973.

Karakusevic, S., 2016. *Understanding Patient Flow In Hospitals.* [online] Nuffieldtrust.org.uk. Available at: <https://www.nuffieldtrust.org.uk/files/2017-01/understanding-patient-flow-in-hospitals-web-final.pdf> [Accessed 12 January 2017].

Kawulich, B., 2012. Collecting data through observation. *Doing social research: A global context*, pp.150-160.

Kenney, C., 2011. *Transforming health care: Virginia Mason Medical Center's pursuit of the perfect patient experience*. CRC Press.

Kim, C., Spahlinger, D., Kin, J. and Billi, J., 2006. Lean health care: What can hospitals learn from a world-class automaker?. *Journal of Hospital Medicine*, 1(3), pp.191-199.

Kiridena, S. and Fitzgerald, A., 2006. Case study approach in operations management research.

Kiridena, S., 2005. Beyond constructs and linkages: a grounded model of the manufacturing strategy formation process in practice.

Klassen, K.J. and Yoogalingam, R., 2009. Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management*, *18*(4), pp.447-458.

Knight, A., 2014. *Pride and Joy*. Linney Group Ltld.

Knight, A., 2011. The theory of constraints in health and social care. *QFI Consulting*, pp.1-13.

Knight, A. and Stratton, R., 2010. Managing patient flow using time buffers. In *TOCICO International Conference: 8th Annual Worldwide Gathering of TOC Professionals, Las Vegas, NE, Theory of Constraints International Certification Organization*.

Knight, A., 2003. Making TOC the main way of managing the health system, presentation at *TOCICO Upgrade Conference.* Cambridge: Sept. 9.

Knight, A., 2000. Healing the National Health Service. The Ashridge Journal Winter, 8–15.

Kopach, R., DeLaurentis, P.C., Lawley, M., Muthuraman, K., Ozsen, L., Rardin, R., Wan, H., Intrevado, P., Qu, X. and Willis, D., 2007. Effects of clinical characteristics on successful open access scheduling. *Health care management science*, *10*(2), pp.111-124.

Kreindler, S., 2017. The three paradoxes of patient flow: an explanatory case study. *BMC Health Services Research*, 17(1).

Kreindler, S., 2017. Six ways not to improve patient flow: a qualitative study. *BMJ quality & safety*, *26*(5), pp.388-394.

LaGanga, L. and Lawrence, S., 2007. Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences*, *38*(2), pp.251-276.

Lander, E. and Liker, J., 2007. The Toyota Production System and art: making highly customized and creative products the Toyota way. *International Journal of Production Research*, 45(16), pp.3681-3698.

Langley, J., Coyle, J., Gibson, B., Novack, R. and Bardi, E., 2008. *Managing Supply Chains: a logistics approach,* Canada, Cengage Learning.

Laurent, C., 2013. Making time to care. *Nursing Standard*, 28(6), pp.26-27.

Lee, A., 1989. A scientific methodology for MIS case studies. *MIS quarterly*, pp.33-50.

Lee, H., 2019. *NHS England » Making essential and effective integrated care an Always Event*. [online] England.nhs.uk. Available at: <https://www.england.nhs.uk/blog/making-essential-and-effective-integrated-care-an-always-event/> [Accessed 11 May 2021].

Leonard-Barton, D., 1990. A Dual Methodology for Case Studies: Synergistic Use of a Longitudinal Single Site with Replicated Multiple Sites. *Organization Science*, 1(3), pp.248-266.

Levin, D., 1988. *The opening of vision: Nihilism and the postmodern situation*. London: Routledge.

Leviner, S., 2020. Patient Flow Within Hospitals: A Conceptual Model. *Nursing Science Quarterly*, 33(1), pp.29-34.

Lewin, K., 1945. The research center for group dynamics at Massachusetts Institute of Technology. *Sociometry*, *8*(2), pp.126-136.

Lewis, M., 1998. Iterative triangulation: a theory development process using existing case studies. *Journal of Operations Management*, *16*(4), pp.455-469.

Liker, J., 2004. *The Toyota Way: 14 Management Principles from The World's Greatest Manufacturer*. New York: McGraw-Hill.

Lin, J. and Chiu, M., 2015. A Mathematical Model to Evaluate and Improve Lean Management of Healthcare System: A Case Study of Health Examination Center. *Transdisciplinary Engineering: A Paradigm Shift*, p.530.

Lin, C., Chen, F. and Chen, Y., 2013. Knowledge Kanban system for virtual research and development. *Robotics and Computer-Integrated Manufacturing*, *29*(3), pp.119-134.

Lincoln, Y., Lynham, S. and Guba, E., 2011. Paradigmatic controversies, contradictions, and emerging confluences, revisited. *The Sage handbook of qualitative research*, *4*, pp.97-128.

Lindblom, C., 1987. Alternatives to Validity: Some thoughts suggested by Campbell's guidelines. *Knowledge*, 8(3), pp.509-520.

Lipley, N., 2009. Productive wards hailed for transforming care. *Nursing Management*, 16(4), pp.4-4.

Little, J.D., 1992. Tautologies, models and theories: can we find 'laws' of manufacturing?. *IIE transactions*, *24*(3), pp.7-13.

Litvak, E. and Long, M., 2000. Cost and quality under managed care: Irreconcilable differences. *Am J Manag Care*, *6*(3), pp.305-12.

Locke, R. and Thelen, K., 1998. Problems of equivalence in comparative politics: Apples and oranges, again. *Newsletter of the APSA Organized Section in Comparative Politics*, *9*(1), pp.9-12.

Lu, J., Yang, T. and Wang, C., 2011. A lean pull system design analysed by value stream mapping and multiple criteria decision-making method under demand uncertainty. *International Journal of Computer Integrated Manufacturing*, 24(3), pp.211-228.

Lubitsh, G., Doyle, C. and Valentine, J., 2005. The impact of theory of constraints (TOC) in an NHS trust. *Journal of Management Development*, 24(2), pp.116-131.

Luck, L., Jackson, D. and Usher, K., 2006. Case Study: A Bridge across the Paradigms. *Nursing Inquiry*, 13 (2), pp. 103-109.

Mabin, V., Yee, J., Babington, S., Caldwell, V. and Moore, R., 2018. Using the Theory of Constraints to resolve long-standing resource and service issues in a large public hospital. *Health Systems*, *7*(3), pp.230-249.

Mabin, V. and Balderstone, S., 2003. The performance of the theory of constraints methodology: analysis and discussion of successful TOC applications. *International Journal of Operations & Production Management*, *23*(6), pp.568-595.

Macdonald, D., Kirk, D., Metzler, M., Nilges, L., Schempp, P. and Wright, J., 2002. Its all very well, in theory: Theoretical perspectives and their applications in contemporary pedagogical research. *Quest*, *54*(2), pp.133-156.

Mack, L., 2010. The philosophical underpinnings of educational research.

Mackenzie, N. and Knipe, S., 2006. Research dilemmas: Paradigms, methods and methodology. *Issues in educational research*, *16*(2), pp.193-205.

Manning, S., 2011. Working toward improving quality and efficiency in school nursing. *British Journal of School Nursing*, *6*(8), pp.373-374.

Marshall, C. and Rossman, G., 1995. The substance of the study: Framing the research question. *Designing qualitative research*, pp.15-37.

Maskell, B., Baggaley, B. and Grasso, L., 2012. *Practical lean accounting: a proven system for measuring and managing the lean enterprise*. CRC Press.

Mazur, L., McCreery, J. and Rothenberg, L., 2012. Facilitating Lean Learning and Behaviors in Hospitals During the Early Stages of Lean Implementation. *Engineering Management Journal*, 24(1), pp.11-22.

Mazzocato, P., Holden, R., Brommels, M., Aronsson, H., Bäckman, U., Elg, M. and Thor, J., 2012. How does lean work in emergency care? A case study of a lean-inspired intervention at the Astrid Lindgren Children's hospital, Stockholm, Sweden. *BMC health services research*, *12*(1), p.28.

Mazzocato, P., Savage, C., Brommels, M., Aronsson, H. and Thor, J., 2010. Lean thinking in healthcare: a realist review of the literature. *BMJ Quality & Safety*, 19(5), pp.376-382.

McCloskey, R., Jarrett, P., Stewart, C. and Nicholson, P., 2014. Alternate level of care patients in hospitals: what does dementia have to do with this?. *Canadian Geriatrics Journal*, *17*(3), p.88.

McHugh, M. and Ma, C., 2013. Hospital Nursing and 30-Day Readmissions Among Medicare Patients with Heart Failure, Acute Myocardial Infarction, and Pneumonia. *Medical Care*, 51(1), pp.52-59.

McKenna, H., 2016. *Five big issues for health and social care after the Brexit vote*. [online] The King's Fund. Available at: https://www.kingsfund.org.uk/publications/articles/brexit-and-nhs [Accessed 15 Dec. 2016].

McKeon, A., Blunt, I., Edwards, N. and Smith, P., 2014. NHS hospitals under pressure: trends in acute activity up to 2022. *London: Nuffield Trust*.

McLafferty, E. and Farley, A., 2006. Analysing qualitative research data using computer software. *Nursing times*, *102*(24), pp.34-36.

McQueen, M., 2002. Language and power in profit/nonprofit relationships: A grounded theory of inter-sectoral collaboration.

Medeiros, D., Swenson, E. and DeFlitch, C., 2008. Improving patient flow in a hospital emergency department. In *Proceedings of the 40th conference on winter simulation* (pp. 1526-1531). Winter Simulation Conference.

Meredith, J., 1998. Building operations management theory through case and field research. *Journal of operations management*, *16*(4), pp.441-454.

Merriam, S., 2009. *Qualitative case study research*. *Qualitative research: A guide to design and implementation*, pp.39-54.

Merriam, S., 1988. *Case study research in education: A qualitative approach*. Jossey-Bass.

Mertens, D., 1998. *Research Methods in Education And Psychology*. Thousand Oaks, California: SAGE.

Miles, M., and Huberman, A., 1994. *Qualitative Data Analysis*. 2nd ed. London: Sage.

Miller, H., 1998. Variation, innovation and dynamic quality. *The TQM Magazine*, 10(6), pp.447-451.

Mills, J., Harrison, H., Franklin, R. and Birks, M., 2017. Case study research: Foundations and methodological orientations. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* (Vol. 18, No. 1, p. 17).

Mintzberg, H., 1979. An emerging strategy of 'direct' research. *Administrative science quarterly*, *24*(4), pp.582-589.

Monde, Y., 1983. *Toyota Production System*. Norcross, GA.: Industrial Engineering and Management.

Moraros, J., Lemstra, M. and Nwankwo, C., 2016. Lean interventions in healthcare: do they actually work? A systematic literature review. *International Journal for Quality in Health Care*, 28(2), pp.150-165.

Morse, J., 1994. Designing funded qualitative research. In Norman K. Denzin & Yvonna S. Lincoln (Eds.), *Handbook of qualitative research* (pp.220-35). Thousand Oaks, CA: Sage

Murphy, C., Mullen, E., Hogan, K., O'toole, R. and Teeling, S.P., 2019. Streamlining an existing hip fracture patient pathway in an acute tertiary adult Irish hospital to improve patient experience and outcomes. *International Journal for Quality in Health Care*, *31*(Supplement_1), pp.45-51.

Myers, J., Cox, J., Teague, S. and Beck, E., 2016. Transitions of care model inclusive of unplanned care improves the patient experience. *Journal of patient experience*, *3*(1), pp.20-23.

National Audit Office. 2016. *Discharging Older Patients from Hospital - National Audit Office (NAO) Report*. [online] Available at: <https://www.nao.org.uk/report/discharging-older-patients-from-hospital/> [Accessed 20 February 2017].

Nelson-Peterson, D. and Leppa, C., 2007. Creating an Environment for Caring Using Lean Principles of the Virginia Mason Production System. *JONA: The Journal of Nursing Administration*, 37(6), pp.287-294.

Netland, T. and Powell, D. eds., 2016. *The Routledge Companion to Lean Management*. Taylor & Francis.

Neuman, W., 2003. *Social Research Methods: Qualitative and Quantitative Approaches*, Pearson Education.

Newton, N., 2010. Exploring Qualitative Methods: The use of semi-structured interviews. *Exploring Qualitative Methods*, pp.1-11.

Ng, D., Vail, G., Thomas, S. and Schmidt, N., 2010. Applying the Lean principles of the Toyota Production System to reduce wait times in the emergency department. *Canadian Journal of Emergency Medicine*, *12*(1), pp.50-57.

Nguyen, S., Yamamoto, E., Nguyen, M., Le, H., Kariya, T., Saw, Y., Nguyen, C. and Hamajima, N., 2018. Waiting time in the out-patient clinic at a national hospital in Vietnam. *Nagoya journal of medical science*, *80*(2), pp.227-239.

NHS Digital. 2019. *Msitdt Delayed Transfers Of Care (DTOC) Return Data Collection - NHS Digital*. [online] Available at: <https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/msitdt-delayed-transfers-of-care-return-data-collection> [Accessed 12 November 2019].

NHS Datadictionary. n.d. *PATIENT*. [online] Available at: <https://datadictionary.nhs.uk/classes/patient.html> [Accessed 20 February 2021].

NHS England. 2017. *Improvement Leaders' Guide*. [online] Available at: <https://www.england.nhs.uk/improvement-hub/wp-content/uploads/sites/44/2017/11/ILG-2.3-Improving-Flow.pdf> [Accessed 14 February 2021].

NHS England. 2016. *NHS England » Referral to treatment*. [online] Available at: <https://www.england.nhs.uk/rtt/> [Accessed 5 March 2021].

NHS England. 2014. *Statistics » Consultant-led Referral to Treatment Waiting Times Data 2013-14*. [online] Available at: <https://www.england.nhs.uk/statistics/statistical-work-areas/rtt-waiting-times/rtt-data-2013-14/> [Accessed 2 March 2021].

NHS England. n.d. *Guide to NHS waiting times in England*. [online] Available at: <https://www.nhs.uk/nhs-services/hospitals/guide-to-nhs-waiting-times-in-england/> [Accessed 21 February 2021].

NHS Improvement., 2017. *The SAFER Patient Flow Bundle and Red2green Days Approach | NHS Improvement*. [online] Available at: <https://improvement.nhs.uk/resources/safer-patient-flow-bundle-and-red2green-days-approach/> [Accessed 12 March 2019].

NHS Improvement, 2017. *Good Practice Guide: Focus on Improving Patient Flow | NHS Improvement*. [online] Available at: <https://improvement.nhs.uk/resources/good-practice-guide-focus-on-improving-patient-flow/> [Accessed 12 November 2017].

NHS Scotland. n.d. *ISD Scotland | Information Services Division*. [online] Available at: <https://www.ndc.scot.nhs.uk/Dictionary-A-Z/Definitions/index.asp?Search=L&ID=306&Title=Length%20of%20Stay> [Accessed 4 March 2021].

Nisbett, R., and Ross, L., 1980. Human inference: Strategies and shortcomings of social judgment. Englewood Cliffs, NJ: Prentice-Hall.

Noon, C., Hankins, C. and Côté, M., 2003. Understanding the Impact of Variation in the Delivery of Healthcare Services. *Journal of Healthcare Management*, 48(2), pp.82-97.

O'Connell, T., Bassham, J., Bishop, R., Clarke, C., Hullick, C., King, D., Peek, C., Verma, R., Ben-Tovim, D. and McGrath, K., 2008. Clinical process redesign for unplanned arrivals in hospitals. *Medical Journal of Australia*, 188(S6).

Ohno, T., 1988. *Toyota production system*. Cambridge, Mass.: Productivity Press.

Olhager, J., 2013. Evolution of operations planning and control: from production to supply chains. *International Journal of Production Research*, 51(23-24), pp.6836-6843.

Olsson, O. and Aronsson, H., 2015. Managing a variable acute patient flow – categorising the strategies. *Supply Chain Management: An International Journal*, 20(2), pp.113-127.

ONS, 2020. *Healthcare Expenditure, UK Health Accounts - Office for National Statistics*. [online] Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthcaresystem/bulletins/ukhealthaccounts/2018> [Accessed 18 May 2020].

Ozawa, S. and Pongpirul, K., 2013. 10 best resources on … mixed methods research in health systems. *Health Policy and Planning*, 29(3), pp.323-327.

Pacheco, D., Pergher, I., Antunes Junior, J. and Roehe Vaccaro, G., 2018. Exploring the integration between Lean and the Theory of Constraints in Operations Management. *International Journal of Lean Six Sigma*, 10(3), pp.718-742.

Pandit, S. and Naik, G., 2009. Application of theory of constraints on scheduling of drum-buffer-rope system. In *Second International Conference on Emerging Trends in Engineering (SICETE)*.

Pannirselvam, G., Ferguson, L., Ash, R. and Siferd, S., 1999. Operations management research: an update for the 1990s. *Journal of Operations Management*, 18(1), pp.95-112.

Papalexi, M., Bamford, D. and Dehe, B., 2015. A case study of Kanban implementation within the pharmaceutical supply chain. *International Journal of Logistics Research and Applications*, *19*(4), pp.239-255.

Paré, G., 2002. Enhancing the rigor of qualitative research: Application of a case methodology to build theories of IT implementation. *The Qualitative Report*, *7*(4), pp.1-34.

Parkhi, S., 2019. Lean management practices in healthcare sector: a literature review. *Benchmarking: An International Journal*, 26(4), pp.1275-1289.

Patton, M., 1990. *Qualitative evaluation and research methods*. SAGE Publications, inc.

Patton, M., 1986. *Utilization-Focused Evaluation, 2Nd Ed*. Newbury Pk: Sage.

Pearson, D., 2008. *Managing Acute Patient Flows*. [online] Victorian Auditor-General. Available at: <https://www.parliament.vic.gov.au/papers/govpub/VPARL2006-10No157.pdf> [Accessed 14 May 2018].

Pettigrew, A., 1990. Longitudinal field research on change: Theory and practice. *Organization science*, *1*(3), pp.267-292.

Pettigrew, A., 1985. Contextualist research: a natural way to link theory and practice. *Doing research that is useful in theory and practice*, pp.222-273.

Pettigrew, A., 1973. *The politics of organizational decision-making*. Routledge.

Pinfield, L., 1986. A Field Evaluation of Perspectives on Organizational Decision Making. *Administrative Science Quarterly*, 31(3), p.365.

Piore, M. and Sabel, C., 1984. *The second industrial divide*. [New York]: Basic Books.

Plsek, P.E., 2013. *Accelerating health care transformation with lean and innovation: The Virginia Mason experience*. CRC Press.

Poksinska, B., Fialkowska-Filipek, M. and Engström, J., 2016. Does Lean healthcare improve patient satisfaction? A mixed-method investigation into primary care. *BMJ Quality & Safety*, 26(2), pp.95-103.

Porter, L. and Barker, B. (2005), *Using Lean Principles to Increase the Efficiency of Service Delivery in the Public Sector*, Oakland Consulting PLC, Leeds.

Potisek, Nicholas M., Robb M. Malone, Betsy Bryant Shilliday, Timothy J. Ives, Paul R. Chelminski, Darren A. DeWalt, and Michael P. Pignone. "Use of patient flow analysis to improve patient visit efficiency by decreasing wait time in a primary care-based disease management programs for anticoagulation and chronic pain: a quality improvement study." *BMC health services research* 7, no. 1 (2007): 1-7.

Pring, R., 2000. The 'false dualism'of educational research. *Journal of Philosophy of Education*, *34*(2), pp.247-260.

Process Excellence Network, 2012. *Healthcare Case Study: Process Improvement At Bolton NHS Foundation Trust*. [online] Available at: <https://www.processexcellencenetwork.com/lean-six-sigma-business-performance/whitepapers/healthcare-case-study-process-improvement-at-bolto> [Accessed 11 November 2016].

Qu, X., Rardin, R., Williams, J. and Willis, D., 2007. Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European Journal of Operational Research*, 183(2), pp.812-826.

Radnor, Z., Holweg, M. and Waring, J., 2012. Lean in healthcare: the unfilled promise?. *Social science & medicine*, *74*(3), pp.364-371.

Radnor, Z., and Walley, P., 2008. Learning to walk before we try to run: Adapting Lean for the public sector. *Public Money and Management*, 28(1), pp.13-20.

Ragin, C., 1987. *The Comparative Method*. Berkeley [etc.]: University of California.

Raleigh, V., 2019. Trends in life expectancy in EU and other OECD countries: Why are improvements slowing?.

Ravenswood, K., 2011. Eisenhardt's impact on theory in case study research. *Journal of Business Research*, 64(7), pp.680-686.

RCP London. 2018. *Health and social care in England: a quick guide*. [online] Available at: <https://www.rcplondon.ac.uk/projects/outputs/health-and-social-care-england-quick-guide> [Accessed 7 March 2021].

Rees, G. and Gauld, R., 2017. Can lean contribute to work intensification in healthcare?. *Journal of Health Organization and Management*, 31(3), pp.369-384.

Regis, T., Gohr, C. and Santos, L., 2018. Lean Healthcare Implementation: experiences and lessons learned from Brazilian hospitals. *Revista de Administração de Empresas*, *58*(1), pp.30-43.

Remenyi, D., Williams, B., Money, A. and Swartz, E., 1998. *Doing research in business and management: an introduction to process and method*. Sage.

Richards, L., & Richards, T., 1994. From filing cabinet to computer. *Analyzing qualitative data*, pp. 146-172.

Ridder, H., 2017. The theory contribution of case study research designs. *Business Research*, *10*(2), pp.281-305.

Robertson, R., Wenzel, L., Thompson, J. and Charles, A., 2017. Understanding NHS financial pressures. *How are they affecting patient care*.

Robineau, D., 2016. *Ageing Britain: Two-Fifths Of NHS Budget Is Spent On Over-65S*. [online] the Guardian. Available at: <https://www.theguardian.com/society/2016/feb/01/ageing-britain-two-fifths-nhs-budget-spent-over-65s> [Accessed 4 September 2017].

Robinson, F., Cunningham, L., Turner, S., Lindroth, J., Ray, D., Khan, T. and Yates, A., 2016. Improving a Dental School's Clinic Operations Using Lean Process Improvement. *Journal of dental education*, *80*(10), pp.1170-1179.

Rojas, E., Munoz-Gama, J., Sepúlveda, M. and Capurro, D., 2016. Process mining in healthcare: A literature review. *Journal of Biomedical Informatics*, 61, pp.224-236.

Rojas-García, A., Turner, S., Pizzo, E., Hudson, E., Thomas, J. and Raine, R., 2018. Impact and experiences of delayed discharge: A mixed-studies systematic review. *Health Expectations*, *21*(1), pp.41-56.

Rolfe, G., 2006. Validity, trustworthiness and rigour: quality and the idea of qualitative research. *Journal of Advanced Nursing*, 53(3), pp.304-310.

Ronen, B., Pliskin, J.S. and Pass, S., 2012. *Focused operations management for health services organizations*. John Wiley & Sons

Rosenberg, J. and Yates, P., 2007. Schematic representation of case study research designs. *Journal of advanced nursing*, *60*(4), pp.447-452.

Rosman, M., Rachminov, O., Segal, O. and Segal, G., 2015. Prolonged patients' In-Hospital Waiting Period after discharge eligibility is associated with increased risk of infection, morbidity and mortality: a retrospective cohort analysis. *BMC health services research*, *15*(1), pp.1-5.

Sampson, S. and Froehle, C., 2006. Foundations and implications of a proposed unified services theory. *Production and operations management*, *15*(2), pp.329-343.

Sandelowski, M., 1996. One is the liveliest number: The case orientation of qualitative research. *Research in nursing & health*, *19*(6), pp.525-529.

Santibáñez, P., Chow, V.S., French, J., Puterman, M.L. and Tyldesley, S., 2009. Reducing patient wait times and improving resource utilization at British Columbia Cancer Agency's ambulatory care unit through simulation. *Health care management science*, *12*(4), p.392.

Saunders, M., Lewis, P. and Thornhill, A., 2012. *Research Methods for Business Students*. Harlow: Pearson.

Schmenner, R. and Swink, M., 1998. On theory in operations management. *Journal of Operations Management*, 17(1), pp.97-113.

Schonberger, R., 1982. *Japanese manufacturing techniques*. New York: The Free Press.

Schragenheim, E., 2010. From DBR to Simplified-DBR for Make-to-Order. In: Cox, J. F. and Schleier, J. G., eds., *Theory of Constraints Handbook*. United States of America: McGraw Hill, 2010, pp. 211-238.

Schragenheim, E., Dettmer, H. and Patterson, J., 2009. *Supply chain management at warp speed: Integrating the system from end to end*. CRC Press, Boca Raton, FL.

Schragenheim, E., and Dettmer, H., 2000. *Manufacturing at Warp Speed: Optimizing Supply Chain Business Performance*. Boca Raton, FL: St. Lucie Press.

Schragenheim, E. and Ronen, B., 1991. Buffer management: a diagnostic tool for production control. *Production and Inventory Management Journal*, *32*(2), pp.74-79.

Schragenheim, E. and Ronen, B., 1990. Drum-buffer-rope shop floor control. *Production and Inventory Management Journal*, *31*(3), pp.18-22.

Seale, C., 2004. Validity, reliability and the quality of research. *Researching society and culture*, pp.71-83.

ŞENGÜN, H., 2017. Lean hospital approach in healthcare. *International Journal of Current Research, 9(1).* pp. 45032-45037

Shah, R. and Ward, P., 2007. Defining and developing measures of lean production. *Journal of operations management*, *25*(4), pp.785-805.

Shaw, S., Rosen, R. and Rumbold, B., 2011. What is integrated care. *London: Nuffield Trust*, *7*.

Shewhart, W., 1939. Application of statistical method in mass production.

Shewhart, W., 1931. Economic Quality Control of Manufactured Product1. *Bell System Technical Journal*, 9(2), pp.364-389.

Shields, C., 2007. Can case studies achieve the 'Gold Standard'? Or when methodology meets politics. In *Annual Meeting of the American Educational Research Association, Chicago, Illinois*.

Shingo, S., 1989. *A study of the Toyota production system from an industrial engineering viewpoint*. Cambridge, Mass.: Productivity Press.

Shubik, M., 1987. What is an Application and When is Theory a Waste of Time?. *Management Science*, 33(12), pp.1511-1522.

Siha, S., 1999. A classified model for applying the theory of constraints to service organizations. *Managing Service Quality: An International Journal*, 9(4), pp.255-264.

Silverman, D., 2011. *Interpreting Qualitative Data*. London: SAGE Publications Ltd.

Singer, A., Thode Jr, H., Viccellio, P. and Pines, J., 2011. The Association Between Length of Emergency Department Boarding and Mortality. *Academic Emergency Medicine*, 18(12), pp.1324-1329.

Singh, B. and Sharma, S., 2009. Value stream mapping as a versatile tool for lean implementation: an Indian case study of a manufacturing firm. *Measuring Business Excellence*.

Singhal, V., Flynn, B., Ward, P., Roth, A. and Gaur, V., 2008. Editorial: Empirical elephants-Why multiple methods are essential to quality research in operations and supply chain management. *Journal of Operations Management*, 26(3), pp.345-348.

Sipper, D. and Bulfin, R., 1998. *Production*. New York: McGraw-Hill.

Sjøberg, D., Johnsen, A. and Solberg, J., 2012. Quantifying the effect of using Kanban versus scrum: A case study. *IEEE software*, *29*(5), pp.47-53.

Skinner, W., 1974. The Focused Factory. *Harvard Business Review*, pp.113-121

Skinner, W., 1971. The anachronistic factory, *Harvard Business Review*, pp.61-70.

Skinner, W., 1969. Manufacturing-missing link in corporate strategy. *Harvard Business Review*. 47 (3), pp.136–145.

Slack, N., Brandon-Jones, A. and Johnston, R., 2016. *Operations Management*, 8th Ed., Pearson Education Ltd

Slack, N., Brandon-Jones, A. and Johnston, R., 2013. *Operations Management*. 7th ed. Pearson Education Ltd

Smith, J. and Rudd, C., 2010. Implementing the productive ward management programme. *Nursing Standard*, 24(31), pp.45-48.

Snow, T. and Harrison, S., 2009. Ward efficiency scheme frees up. *Nursing Standard*, 23(37), pp.5-5.

Soltani, E., K. Ahmed, P., Ying Liao, Y. and U. Anosike, P., 2014. Qualitative middle-range research in operations management. *International Journal of Operations & Production Management*, 34(8), pp.1003-1027.

Somekh, B. and Lewin, C. eds., 2005. *Research methods in the social sciences*. Sage.

Spear, S., 2005. Fixing health care from the inside, today. *Harvard business review*, *83*(9), p.78.

Spear, S. and Bowen, H., 1999. Decoding the DNA of the Toyota production system. *Harvard business review*, *77*, pp.96-108.

Spearman, M., Woodruff, D. and Hopp, W., 1990. CONWIP: a pull alternative to Kanban. *The International Journal of Production Research*, *28*(5), pp.879-894.

Spencer, L., Ritchie, J., Lewis, J. and Dillon, L., 2004. Quality in qualitative evaluation: a framework for assessing research evidence.

Spencer, M. and Cox, J., 1995. An analysis of the product-process matrix and repetitive manufacturing. *International Journal of Production Research*, 33(5), pp.1275-1294.

Spradley, J.P., 2016. *Participant observation*. Waveland Press.

Søreide, K., Hallet, J., Matthews, J., Schnitzbauer, A., Line, P., Lai, P., Otero, J., Callegaro, D., Warner, S., Baxter, N., Teh, C., Ng-Kamstra, J., Meara, J., Hagander, L. and Lorenzon, L., 2020. Immediate and long-term impact of the COVID-19 pandemic on delivery of surgical services. *British Journal of Surgery*.

Stake, R., 2005. *Qualitative Case Studies. The Sage Handbook of Qualitative Research*. 3rd ed. Thousand Oaks, CA: Sage.

Stake, R., 1998. Case studies. In N. K. Denzin, & Y. S. Lincoln (Eds.), *Strategies of qualitative enquiry* (pp.86−109). London: Sage.

Stake, R., 1995. *The art of case study research*. Thousand Oaks, CA: Sage.

Steed, A., 2012. An Exploration of the Leadership Attributes and Methods Associated with Successful Lean System Deployments in Acute Care Hospitals. *Quality Management in Health Care*, 21(1), pp.48-58.

Stoye, G., 2020. *The Four-Hour NHS Waiting Time Target Saves Lives - Government Should Think Carefully About Changes*. [online] Ifs.org.uk. Available at: <https://www.ifs.org.uk/publications/14677> [Accessed 9 April 2020].

Stratton, R., 2015. Teaching TOC in Context. *In TOCICO International Conference.* Cape Town, South Africa.

Stratton, R., 2012. Buffer management in context. In: *TOCICO International Conference: 10th Annual Worldwide Gathering of TOC Professionals*. Theory of Constraints International Certification Organisation.

Stratton, R., Robey, D. and Allison, I., 2008. Utilising buffer management to manage uncertainty and focus improvement. In *Proceedings of the International Annual Conference of EurOMA*. Groningen, Netherlands, 15-18 June.

Stratton, R., 2008. Theory building: relating variation, uncertainty, buffering mechanisms and trade-offs. In *Proceedings of the 3rd World Conference on Production and Operations Management,* Tokyo, Japan, August.

Stratton, R. and Knight, A., 2010. Managing patient flow using time buffers. *Journal of Manufacturing Technology Management*, *21*(4), pp.484-498.

Stratton, R., Wilson, A. and West, W., 2014. Redesigning community healthcare systems: an action research study. *Proceedings of the international EurOMA conference*. Palermo, Italy, 20-25 June.

Sugimori, Y., Kusunoki, K., Cho, F. and Uchikawa, S., 1977. Toyota production system and kanban system materialization of just-in-time and respect-for-human system. *The international journal of production research*, *15*(6), pp.553-564.

Swisher, J.R. and Jacobson, S.H., 2002. Evaluating the design of a family practice healthcare clinic using discrete-event simulation. *Health Care Management Science*, *5*(2), pp.75-88.

Tabish, S. and Syed, N., 2015. Securing the Future: A Systems Approach to Continuous Improvement in Health Care by Applying the Theory of Constraints. *International Journal of Science and Research*, *438*(1), pp.2674-2695.

Tellis, W., 1997. Application of a case study methodology. *The qualitative report*, *3*(3), pp.1-19.

The King's Fund. 2021. *NHS funding: our position*. [online] Available at: <https://www.kingsfund.org.uk/projects/positions/nhs-funding> [Accessed 4 March 2021].

The King's Fund. 2020. *If the government fails to invest in the NHS workforce, waiting times of over a year could be all too common by the next election*. [online] Available at: <https://www.kingsfund.org.uk/press/press-releases/kings-fund-responds-hospital-activity-figures-november-2020> [Accessed 5 April 2021].

The King's Fund. 2018. *Delayed transfers of care: a quick guide*. [online] Available at: <https://www.kingsfund.org.uk/publications/delayed-transfers-care-quick-guide> [Accessed 23 February 2021].

Thorlby, R., Starling, A., Broadbent, C. and Watt, T., 2018. What's the problem with social care, and why do we need to do better. *The Health Foundation the Institute for Fiscal Studies*.

Tong, A., Sainsbury, P. and Craig, J., 2007. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International journal for quality in health care*, *19*(6), pp.349-357.

Toussaint, J. and Gerard, R., 2010. *On the mend: revolutionizing healthcare to save lives and transform the industry*. Lean enterprise institute.

Trenholm, I. and Wyman, P., 2020. *State of Care*. [online] CQC. Available at: <https://www.cqc.org.uk/publications/major-report/state-care> [Accessed 2 April 2021].

Tsou, C., 2013. On the strategy of supply chain collaboration based on dynamic inventory target level management: A theory of constraint perspective. *Applied Mathematical Modelling*, 37(7), pp.5204-5214.

UK Government. 2014. *Delivering better integrated care*. [online] Available at: <https://www.gov.uk/guidance/enabling-integrated-care-in-the-nhs> [Accessed 23 February 2021].

Umble, M. and Umble, E., 2006. Utilizing buffer management to improve performance in a healthcare environment. *European Journal of Operational Research*, 174(2), pp.1060-1075.

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

Van Calster, J., Willekens, K., Seys, D., Van Elderen, P., Spileers, W. and Vanhaecht, K., 2019. Standardized care by redesign of an intravitreal injection pathway. *European journal of ophthalmology*, *29*(1), pp.92-99.

Van de Ven, A., 1989. Nothing Is Quite So Practical as a Good Theory. *Academy of Management Review*, 14(4), pp.486-489.

Van Herck, P., Vanhaecht, K. and Sermeus, W., 2004. Effects of Clinical Pathways: Do They Work?. *Journal of integrated Care Pathways*, 8(3), pp.95-105.

Vanhaecht, K., 2007. The impact of clinical pathways on the organisation of care processes.

Vermeulen, I.B., Bohte, S.M., Elkhuizen, S.G., Lameris, H., Bakker, P.J. and La Poutre, H., 2009. Adaptive resource allocation for efficient patient scheduling. *Artificial intelligence in medicine*, *46*(1), pp.67-80.

Voss, C., Tsikriktsis, N. and Frohlich, M., 2002. Case research in operations management. *International Journal of Operations & Production Management*, 22(2), pp.195-219.

Walker, H., Chicksand, D., Radnor, Z. and Watson, G., 2015. Theoretical perspectives in operations management: an analysis of the literature. *International Journal of Operations & Production Management*, 35(8), pp.1182-1206.

Walker, D., 2011. *Caring for our aging population and addressing alternate level of care*. Ontario Ministry of Health and Long-Term Care.

Walker, H., Langton, D. and Thomson, L., 2011. 'New to Forensic'; implementing a problem-based introductory educational programme for forensic practitioners in Scotland. *Journal of psychiatric and mental health nursing*, *18*(10), pp.934-942.

Wacker, J., 1998. A definition of theory: research guidelines for different theory-building research methods in operations management. *Journal of Operations Management*, 16(4), pp.361-385.

Walley, P., Silvester, K. and Steyn, R., 2006. Managing Variation in Demand: Lessons from the UK National Health Service. *Journal of Healthcare Management*, 51(5), pp.309-320.

Waring, J.J. and Bishop, S., 2010. Lean healthcare: rhetoric, ritual and resistance. *Social science & medicine*, *71*(7), pp.1332-1340.

Watson, K., Blackstone, J. and Gardiner, S., 2007. The evolution of a management philosophy: The theory of constraints. *Journal of operations Management*, *25*(2), pp.387-402.

West, M., Bailey, S. and Williams, E., 2020. The courage of compassion: Supporting nurses and midwives to deliver high-quality care. *London: The King's Fund*.

Westbrook, R., 1995. Action research: a new paradigm for research in production and operations management. *International Journal of Operations & Production Management*, 15(12), pp.6-20.

Westwood, N., James-Moore, M. and Cooke, M., 2007. *Going Lean in the NHS*. NHS Institute for Innovation and Improvement.

White, D., Froehle, C. and Klassen, K., 2011. The Effect of Integrated Scheduling and Capacity Policies on Clinical Efficiency. *Production and Operations Management*, 20(3), pp.442-455.

Wijewickrama, A.A. and Takakuwa, S., 2006, December. Simulation analysis of an outpatient department of internal medicine in a university hospital. In *Proceedings of the 2006 Winter Simulation Conference* (pp. 425-432). IEEE.

Winch, S. and Henderson, A., 2009. Making cars and making health care: a critical review. *Medical Journal of Australia*, 191(7), pp.414-415.

Womack, J., Byrne, A., Flume, O., Kaplan, G. and Toussaint, J., 2005. Going lean in health care-Institute for healthcare improvement. *Boston: MA*.

Womack J and Jones D., 1996. *Lean Thinking: Banish Waste and Create Wealth in Your Corporation.* New York: Simon and Schuster.

Womack, J., Jones, D. and Roos, D., 1990. *The Machine That Changed the World*. New York: Rawson Associates.

Woodside, A. and Wilson, E., 2003. Case study research methods for theory building. *Journal of Business & Industrial Marketing*, 18(6/7), pp.493-508.

Yin, R., 2014. *Case study research: Design and methods.* Los Angeles, CA: Sage

Yin, R., 2013. Validity and generalization in future case study evaluations. *Evaluation*, *19*(3), pp.321-332.

Yin, R., 2003. *Applications of Case Study Research*, Sage Publications.

Yin, R., 1994. *Case Study Research: Design and Methods*, Sage Publications.

Yin, R., 1981. The case study crisis: Some answers. *Administrative science quarterly*, *26*(1), pp.58-65.

Yin, R., 1981. The case study as a serious research strategy. *Knowledge*, *3*(1), pp.97-114.

Young, T. and McClean, S., 2008. A critical look at Lean Thinking in healthcare. *BMJ Quality & Safety*, *17*(5), pp.382-386.

Young, T., Brailsford, S., Connell, C., Davies, R., Harper, P. and Klein, J.H., 2004. Using industrial processes to improve patient care. *Bmj*, *328*(7432), pp.162-164.

Younis, J., Salerno, G., Fanto, D., Hadjipavlou, M., Chellar, D. and Trickett, J.P., 2012. Focused preoperative patient stoma education, prior to ileostomy formation after anterior resection, contributes to a reduction in delayed discharge within the enhanced recovery programme. *International journal of colorectal disease*, *27*(1), pp.43-47.

APPENDIX 1

Semi-Structured Interview Questions

Theoretical focus: The questions focus on how variability and uncertainty has been reduced and managed in improving patient flow. Of particular interest is the use of management signaling tools at the local and wider system level to prioritise, expedite and focus the ongoing improvement, as typified by sustained applications of quality, Lean Management (LM) and the Theory of Constraints (TOC) activity in manufacturing delivery systems.
Structure: These interview questions have been sectioned to reflect the structure of the selected service evaluation case write-up and only some of these sections may be relevant depending on the interviewee's case experience. The questions will explore the rationale for the intervention together with any reasons for not achieving the predicted improvement and subsequent performance. As interventions commonly involve external consultants they will also be included in assessing how this intervention relates to their wider case experience.

Selection of service in need of intervention

- What stimulated the need and area of focus for the intervention?
    - Were there any reports or particular performance gaps identified?
    - Why the particular case was studied selected? (Making particular reference to sources of variability and use of buffering mechanisms)

Selection of intervention approach

- What alternative approaches to intervention were considered?
- What was the basis of the selected intervention / consultancy leading the change?
    - Variability and flow management approaches?

Original system design and performance

- What was the process to be changed?
    - Clarify the process and the issues with its original design and execution.
    - Process map? Sources of variability and how flow is prioritized, controlled and systematically improved.
- How was this measured and how was it performing?
    - Reports/run chart performance
    - Particular interest in variability and flow measures (LOS, causes of delay, etc)

Intervention

- How was the intervention managed (extent and timing)?
    - Consultancy team, education/training, software?
- Was it formally allied to a particular approach (e.g. lean or TOC).
- What was the focus of the intervention process?

    - E.g. kaizen blitz process redesign, Lean productivity series or TOC buffer management.
- How did the intervention actively manage flow
    - E.g. priority control, expediting, system escalation, targeted improvement activity.
- What performance improvements were predicted over what time?
    - e.g. LOS reductions, delayed discharge, Externally reported performance measures, bed closure, cost reduction
- Software support
    - If software was used in the solution how was it introduced and used?
    - What was the key purpose and how was it used to enable performance improvement.

Subsequent performance

- Did the implementation proceed as planned?
- What were the results and how has this been sustained and improved on over?
- Were specific aspects of the intervention more effective and long lasting?
- What factors contributed to the initial and ongoing performance and why?

APPENDIX 2

Excerpt of the Coding System

<u>Motivational conditions for intervention</u>

Motivation for change

An indicator that presents evidence of poor patient flow and the need for improvement (e.g., increase of patients' length of stay, waiting times, delayed transfer of care, patients' demand, patients' complaints, etc.)

Gaps identified

An indication implies if any reports or specific gaps have been found that drive the need for improvement.

Selection of this particular service

An indicator that identifies any causes of variability and the way in which they were managed.

Approaches applied before

A segment denotes the consideration of any alternative intervention strategies.

Foundation of these approaches

An indication that refers to the basis of previous flow mechanism approaches used in the organisation (e.g., using a flow mechanism to control variability or a unique approach led by consulting).

Issues and challenges

A segment which illustrates the system and the difficulties encountered during its original design and implementation.

Design and performance

A segment explains how the original system was designed and operated, highlighting how patient flow was measured, managed, prioritised, and systematically improved. Examples of measuring the system's performance include:

- The length of stay.
- Delayed transfer of care.

- Waiting time
- Identifying the causes of delay to target improvement.

Nature and techniques of intervention

Nature of the intervention

A segment that describes the basis of the adoption of intervention and how the intervention was managed and introduced to the healthcare organization (e.g. timing, consultancy team, training, software etc.)

The adopted approach

A segment that explores the intervention/approach used within the organization and illustrates the focus of the intervention process (e.g. LM, TOC, others).

Effectiveness of managing the flow

A segment that explains the intervention and its associated flow mechanism (kaizen blitz process redesign, Lean productivity series, or TOC buffer management) used to improve the flow. This includes an illustration of the intervention process, the extent to which the implemented approach actively manages the flow, performance improvements (LOS reductions, delayed discharge, externally reported performance measures, bed closure, cost reduction), and an evaluation of the solution's usage and performance if any software is used.

Criteria for intervention success

Intervention progress

A segment that explains how the intervention/approach adapted or progressed over time or as planned. Typically, new interventions or changes are evaluated in terms of user acceptance and its success in meeting the organisation's goals.

Results/performance sustainability

A segment that concludes to what extent the influence of the intervention is successful in managing and improving patient flow and the long-term sustainability of improvements. Examples of measuring the system's performance include:

- The length of stay.
- Delayed transfer of care.
- Waiting time
- Identifying the causes of delay to target improvement

Overall satisfaction with the intervention

A segment which suggests intervention's effectiveness and reliability. The efficacy and dependability of the intervention are not the only reasons for success; other factors may have contributed to the initial and ongoing performance.

APPENDIX 3

Examples of reflective quotations from transcripts

Cases	Extracts of transcripts	Reflective quotations
Case 1	"From each meeting will come actions the different individuals were going to take, which was going to drive down the waiting time. It was a visual representation of the waiting times and capacity"	The effectiveness of the flow mechanism on managing the patient flow as it provides the resources with the opportunity to take effective and proactive action to protect the delivery system from being destabilised.
Case 2	"Eight or nine years ago, the average length of stay would have been 60 plus days, and I'm delighted to say that now we report exceptions on the basis of any length of stay over 20 days"	Clear evidence that the intervention has made a significant improvement on patient flow performance.
Case 3	"Stock room organisation was terrible, and people would reorder because they couldn't find something because it was buried in the bottom of other things and things would go out of date"	Effects of unorganized work (waste) on system productivity and efficiency
Case 4	"There's a situation where actually patients need to go into a care home, but neither social care nor health pay because the patients don't meet any of their criteria"	Evidence of common issues in healthcare management practices. Focusing on cost-saving and local optimisation might contribute to poor patient flow as it does not serve the system-wide improvement.

APPENDIX 4

Examples of anonymized transcript

Selection of service in need of intervention

- ***What stimulated the need and area of focus for the intervention?***

This was identified as an area where a paediatric speech, and language therapy had **a very long waiting list**, and there have been a number of different interventions before, and they hadn't really stuck. There's a, a large population of staff in a relatively large population of staff in there. And it was seen as a service that was underperforming. So we didn't choose to work with the pediatric speech and language team, we were asked to work with them by a senior manager who felt that there were improvements that could be made in that.

- ***Were there any reports or particular performance gaps identified?***

Yes. Well, **the waiting lists were all over 13 weeks.** And there was an understanding that the nature of these patients timeliness with which we serve them has a big impact on their life outcomes because they had already caught the boundaries of the service. So there weren't really seeing people in secondary schools. It tended to be up to about the age of seven. So they've already got this kind of artificial barrier there. And if you got a child, for instance, that was preschool, and they were having trouble communicating through the lack of speech or speech issues, **they were waiting so long to receive a service from us**. There might be a primary school patient by time, we got to them and then their issues of just multiplied. They've grown exponentially because they've got speech issues and, and now they're in a social educational environment and often, we knew that there were connections between future comes referrals. So, child mental health with issues like Autism, attention deficit and those kinds of things will go in hand in hand with speech issues because if they didn't get the early years stuff right, it was having an impact, and we will contribute into not in the early years **because we will not be addressing it in a timely fashion**.

- ***Why the particular case was studied selected? (Making particular reference to sources of variability and use of buffering mechanisms)***

Because of those things, so I think we selected that **because of the length of waiting times and because of the fact that in not addressing it,** we will probably generating more work for ourselves and for the patient throughout multiple services in the future. So it seemed to do that earlier stuff seem like a sensible thing to do. And also it's a very expensive service to run. The staff were better paid generally than other clinicians. There was an awful lot of part-time work a largely because a lot of the staff would do some private work as well. So rather than having a full-time job with us, they will be working part-time with us and doing some private consultations with people as well. So what that means is that a lot of the mandatory training and things like that and comes out of the part-time hours. So the fixed cost element stays the same but then the added value element is reduced with the, with the part-time work.

Selection of intervention approach

- ***What alternative approaches to intervention were considered?***

Well, we'd already done work in other services **using the productive community model, which was a more traditional lean model really but written in an NHS language**. So in that Kind of setting, we'd done things like **5S,** we done group working, we done goal alignment. But actually what we saw with the speech and language team was that it wasn't really a team, it was a service made of a **lot of blocks** little bits, almost like **independent units,** and they were very driven by whoever was leading that unit and what their specialism was or what their passion was rather than being coherent. Nice. So, **the theory of constraints was chosen because it gave us the opportunity to look along the whole service path and look at the different ways in which people would move through the service.**

So we saw a lot of multiple assessments because they have a first assessment and that would identify that they might need to have a different assessment or more specialist assessment. Then they'd have a specialist assessment, and they would do, wouldn't be chosen as appropriate for that service. So, we knew that there **was a lot of delays in those process**. So that's why we chose the **theory of constraints.**

- ***What was the basis of the selected intervention / consultancy leading the change?***

So we didn't really have in house experience of **the theory of constraints**. My personal view is, is that actually, they're all very connected. But, as a division, the **theory of constraints** was seen as a different approach to any other approach as they would select one of them really. And I think because **it was a service environment and there was lots of complexity** in it, **the theory of constraints gave us the opportunity to sort of where's the bottleneck and let's drive on the bottleneck and then was the next bottleneck and drive on that** rather than it be a kind of put everything through one big sausage machine of this is the new way we work, you know, and that was seen as something that may be speech and language as a service will be more open to.

Original system design and performance

- ***What was the process to be changed?***
    - ***Clarify the process and the issues with its original design and execution.***

- ***Process map? Sources of variability and how flow is prioritized, controlled and systematically improved.***

- Well, at that stage we didn't really know what needed to be changed. We knew there was lots of waste in the system and lots of handoffs from one part to another part. So the view was that we could streamline some of that, but there was so much of it that we needed to be quite targeted in the way that we did that.
- So we mapped out on the process and we understood where a lot of the issues were and I'm the bit that was consistent across all, we call them **packages of care**. The bit that was consistent across all packages of care was this front end or from **referral, triage and assessment to the start of the package of care**. And then a way that was easier to address because it had less clinical specialists challenge in it than picking off each of the packages of care. And there were hundreds of different packages. We did it **under the guidance of QFI.** So we would have a consultant with us. I had a team with myself and then I had three facilitators.
- so we did most of the interventions, and a lot of the data gathering, and some of the education, but that was led by a consultant within QfI who was feeding off our knowledge as well to interpret where the next steps on.
- So in our sort of hierarchy of decision making, we said, well, the **patient need comes first** and actually that hadn't really been the case because really people were organizing their work around **their own personal needs** for a long long time.
- And so what we are a big part of the work that we did was to **put everybody's appointments onto an electronic calendar** because before people were put in their appointments in their **own paper diaries** and that was **not visible** to people. And so it was very difficult to understand the utilization and the value that was being got from staff time, how many appointments and certain types of intervention might have all of that kind of **data was hidden**.

- ***How was this measured and how was it performing?***
  - Reports/run chart performance
  - Particular interest in variability and flow measures (LOS, causes of delay, etc)

So what we did was we **measured the times between different stages. So a referral to treatment** or each of those stages. So **from receiving a referral to registered on the system, to triage, to book an appointment, to attending the first assessment: each of those steps were measured, but we monitored what was waiting each of those stages. And we set up buffers and understood whether they were in kind of red, amber, green in terms of where they should've been** so that we can see if something was waiting more than a, whatever it was at that stage. There were boundaries at each stage then, it would **go amber and red** and you can see that we graphed all that and we will feed in that back constantly to the teams in terms of service report.

Intervention

- ***How was the intervention managed (extent and timing)?***
    - ***Consultancy team, education/training, software?***

Well, they had a head of service who worked quite closely with us and understood what it was we were trying to achieve. but they themselves were quite locked into the way that they work and also I'm not overly challenging staff like trying to keep the peace which is part of their role. So, there was a balance to be had in that, but the program full delivery was worked out between QFI and largely the quality improvement team, but with some input from the operations management team for that service as well.

- ***Was it formally allied to a particular approach (e.g. lean or TOC).***

Yes, it was allied to the **theory of constraints.**

- ***What was the focus of the intervention process?***
    - ***E.g. kaizen blitz process redesign, Lean productivity series or TOC buffer management.***

The focus was **the Toc buffer system**. So, it was very much about being able to monitor something regularly, understand where it was examined that with the team, and then for people to identify actions that would **minimize the waiting times and reduce it**. So we would treat them as **a bottleneck**, and I'm minimized that **bottleneck to improve the flow**.

- ***How did the intervention actively manage flow***
    - ***E.g. priority control, expediting, system escalation, targeted improvement activity.***

- By the **measurement of waiting times** at the set stages.

- it was a **weekly report** we reviewed with the management team, and it did the referral to assessment, and then it did the different teams. So, you have by service, and then you have by team underneath, and it was a weekly dialing. So, **a telephone meeting** where **the zs were scrutinized.** So, **from each meeting will come actions the different individuals were going to take, which was going to drive down the waiting time. It was a visual representation of the waiting times and capacity.**

- ***What performance improvements were predicted over what time?***
    - ***e.g. LOS reductions, delayed discharge, Externally reported performance measures, bed closure, cost reduction***

I think there was a commitment. I'd have to have a look to see what that commitment was, but a lot of it really was about the process we were going to go through. So, the education process, the mapping, the making things transparent, the engaging leaders in that challenge process. There was more emphasis put on that. There was no kind of the **waiting times** will be this. It wasn't that clearly defined not overall. But we definitely saw it at that front end and we definitely got that. That turned around that.

- ***Software support***
  - ***If software was used in the solution how was it introduced and used?***

- **we never got the real software Jonah**. We never got Jonah and we never got our system one solution. They did want us to use a particular piece of software called **"Jonah"** because it cannot link or integrated to system one. and **there was reluctance from the it department to embrace another type of software on top of system as well. So we tried to work with TPP, who are the system one providers that proved very challenging because they wanted one system, the whole of the country houses and so**. They did kind of work with it a little bit, but we didn't get where we wanted to on that from. **So we ended up there. A lot of the measurement was offline, so it was Excel spreadsheet type things, which wasn't ideal**.
- **we were producing all the monitoring graphs around delays, to what extent are people stacked up, to what extent are things delayed, are we so many days over, are we in the Green Zone, the amber zone, the red zone, the black zone.**
- So we did, we did have stuff, but it wasn't the software thing was an issue in hindsight, **maybe it would have been better with Jonah because it is simple to use.**

  - ***What was the key purpose and how was it used to enable performance improvement.***

Once you got understanding, once we know they got capacity, then you start asking the questions of what if we got the capacity why you were not seeing more patients and getting them of the **waiting list.**

<u>**Subsequent performance**</u>

- ***Did the implementation proceed as planned?***
- It **went on longer than planned**, so I think initially we contract QFI for three months for phase one and then it was meant to be another three months or four months for phase two, but that phases went on a lot longer than that.
- **So phase one was about the baselining, the initial education, and the understanding the opportunity/ Phase two was those four elements around**

measurement, the education, the packages of care and the process improvement to make things more visible.

- ***What were the results and how has this been sustained and improved on over?***

- We got pulled off the project at some point. I think we're on it **for about 18 months** in total. And the **QFI finished**, we had some really **serious resistance** from some members of the pediatric speech and language team.

- I think we **shifted the culture**. I think the **transparency around the appointment** was a **big and sustained thing**. I think **it's now normal to offer assessment appointments out of the area and it depends on the area, the more affluent area, the easier it tends to be for people to travel out of the area**. They probably got cars, it can be as simple as that. There's **more flexibility in the system now to adapt to variation** in **demand across the area** and, and **this visibility** of who's doing what. So if someone does go off sick, somebody else can pick it up. It doesn't just sit there until they come back.

- So you are getting a fairer service where you've got **more equity in waiting times system one diary is absolutely key of all the things we did that thing was probably the most beneficial.** I'm looking at the packages of care. An estimating how long the package of care should take was probably the most challenging thing we did with them because they didn't want to say it takes this long, and there were so many **variations of packages of care.**

- **Were specific aspects of the intervention more effective and long lasting?**

- Yes. Because if you look at the **weekly review** that's dropped off. So that was the less of it was effective at the time, but **it hasn't been sustained.** And the **diary thing** that set up and get people compliant, **but it has been sustained.** And also, managers within the service and above the service have a really embraced that as a **very transparent way of understanding what the forecast availabilities** so that they **can actively manage balancing the availability to the forecast demand** so they can forecast demand of what's gone before. And it's quite, it's a bit seasonal, but actually it's not a seasonal as they felt it was because they were driving a lot of seasonality in the system. And I think that **understanding is as been sustained** as well.

- The system one diary, the packages of care and the **management view**, reviewing the management report, the measurement reporting and interrogation of it. These were **the output of TOC implementation**.

- **What factors contributed to the initial and ongoing performance and why?**

- Okay. Well I guess there was this cultural history of the service being difficult and that definitely was a factor that contributed to the initial work because there was a long time in that a and a lot of resistance in the education process. It wasn't a case of teach them the theory and then let them go and then support them in applying that. There was lots of push back on that. So, those cultural factors implement influenced it. I think it was system technology impact. We didn't go with **Jonah,** we didn't go with Jonah for good reasons, but then we didn't get an effective automated system one solution either. System one did upgrade a couple of years later and some of that upgrade did reflect a lot of that workflow staff that we'd done with them to some extent. But it was a national roll out of this is what you've got sort of thing.

- So I'm sure it contributed to it, but it wasn't those were things that the impacted on us on that initial performance , and some management resistance, there was a lot of managers at the end of their career, there was a lot of managers, they just wanted it to stay steady for the next couple of years because I know what I'm doing and I'm not overly challenged and I can't cope with anymore than this. And I'm the expert. I've worked in speech and language therapy all my life.