Use of network analysis, and fluid and diffusion approximations for stochastic queueing networks to understand flows of referrals and outcomes in community health care

by

Ryan Palmer MMath

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy

Clinical Operational Research Unit

Department of Mathematics

Faculty of Mathematical & Physical Sciences

University College London

April, 2018

Disclaimer

I, Ryan Palmer, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature _____

Date _____

Abstract

Community services are fundamental in the delivery of health care, providing local care close to or in patient homes. However, planning, managing and evaluating these services can be difficult. One stand out challenge is how these services may be organised to provide coordinated care given their physical distribution, patients using multiple services, and the increasing use of these services by patients with differing needs. This is complicated by a lack of comparable measures for evaluating quality across differing community services. Presented in this thesis is work that I conducted, alongside the North East London Foundation Trust, to understand referrals and the use of outcome data within community services through data visualisation and mathematical modelling.

Firstly, I applied several data visualisations, building from a network analysis, to aid the design of a single point of access for referrals into community services - helping to understand patterns of referrals and patient use. Of interest were concurrent uses of services, whether common patterns existed and how multiple referrals occurred over time. This highlighted important dynamics to consider in modelling these services.

Secondly, I developed a patient flow model, extending fluid and diffusion approximations of stochastic queueing systems to include complex flow dynamics such as

re-entrant patients and the use of multiple services in sequence. Patient health is also incorporated into the model by using states that patients may move between throughout their care, which are used to model the differential impact of care. I also produced novel methods for allocating servers across parallel queues and patient groups.

Finally, I developed the concept of "the flow of outcomes" - a measure of how individual services contribute to the output of patients in certain health states over time - to provide operational and clinical insight into the performance of a network of services.

This thesis was completed under the supervision of ***Professor Martin Utley, Professor Naomi Fulop, Dr Christina Pagel and Dr Nora Pashayan****.*

Acknowledgments

Thank you to the Health Foundation for their generous funding and support as part of their Improvement Science PhD programme.

Thank you to the North East London Foundation Trust for their collaboration throughout this project. In particular, I thank Dr Stephen O'Connor and Geraldine Rogers for their knowledge, support and help in getting this work started. Your help in connecting me with the relevant staff and resources was invaluable.

Thank you to Professor Martin Utley for your support, encouragement and supervision during my PhD. Thank you for all that you have taught me and for giving me this opportunity to participate in mathematical research. Thank you for helping me to improve as a researcher and as a mathematician.

Thank you Professor Naomi Fulop for your help and insight into working with, and within, health care. Thank you for helping me to improve how I communicate my research to various audiences.

Thank you to Dr Christina Pagel and Dr Nora Pashayan for your support and encouragement throughout my PhD. Thank you also to Dr Sonya Crowe for your insights and help.

Thank you to all those who were, and those who still are, a part of CORU. Thank you for your friendship and support. Thank you to the UCL's Department of Mathematics for this opportunity to pursue a PhD in Mathematics. And thank you Professor Frank Smith for helping to "sanity" check various parts of my work.

Thank you Anna Schüle for your keen-eye and help in proof reading my thesis.

Thank you Rachel Palmer, my loving wife, for your continued support, prayers and dinners. Thank you for loving me and looking after me during this time, even when I have been "stressy" and difficult. Without you, I would not have completed this PhD nor survived the last three years. Thank you to my parents, Derek and Hazel Palmer, for your love and care, and for insisting that I do my homework. Thank you Luke Palmer for your brotherly love and competition. Without my family, I would not have been able to begin a PhD, let alone finish one.

Thank you to my church family (Christ Church Balham), and to Alex, Brian, David, Paco and Seb. Without your continued support, prayers and encouragement, over many years, this would have been far harder and I would have lost sight of all that counts.

Finally, I praise God for His continued grace, kindness and sustaining love. *Soli Deo Gloria.*

Ryan Palmer, *University College London*, April 2018

To my wife Rachel Palmer.

Our adventure began with maths, now it is so much more.

But maths is still awesome.

Contents

Disclaimer **2**

Abstract **3**

Acknowledgments **5**

List of Tables **12**

List of Figures **14**

1 Introduction **18**

1.1 Health care policy and community services 20

1.2 Operational research 21

1.2.1 Patient flow modelling 21

1.2.2 Queueing theory 22

1.3 North East London Foundation Trust 25

1.4 Evaluating community health care 26

1.5 Purpose and aim of this thesis 27

1.6 Structure of this thesis 30

2 Literature Review **32**

2.1 Introduction 33

2.2 Method of review 34

2.3 Results of literature searches 38

2.4 Analysis of papers 40

2.4.1 "Patient flow within community care" 40
2.4.2 "Patient flow and outcomes" papers 46
2.5 Summary and discussion of findings 62
2.5.1 Limitations 64
2.6 Conclusions and directions for work 65

3 Understanding referral data through data visualisation and analysis 69
3.1 Introduction 70
3.2 Research landscape and original contribution 72
3.3 Initial steps 74
3.3.1 Understanding patient referrals - learning from care leads . . . 74
3.3.2 Routine patient data - content and cleaning 78
3.4 Methods 79
3.4.1 Analysis of individual patient pathways 79
3.4.2 Network map 81
3.4.3 Chains of referrals and concurrent uses of multiple services . . 86
3.4.4 Subsequent uses of community services 88
3.5 Application to NELFT referral data 89
3.5.1 Network Map 89
3.5.2 Chains of referrals and concurrent uses of multiple services . . 91
3.5.3 Subsequent uses of community services 96
3.6 Summary and Discussion 99
3.6.1 Limitations 101
3.6.2 Possible avenues for future work 102
3.7 Conclusions 102

4 Measuring patient outcomes within community services 104
4.1 Introduction 105
4.2 Sources used for understanding outcome measurement 107
4.3 Important themes and measures across the sources 109
4.3.1 NELFT quality themes and outcome measures 112
4.4 Summary and Discussion 116
4.5 Conclusions 118

5 Fluid and diffusion approximations for modelling the flow of heterogeneous patients within a network of queues 120

5.1 Introduction . . . 121

5.2 Introduction to fluid and diffusion approximations for stochastic systems 125

5.2.1 Applications of fluid and diffusion approximations within health care . . . 127

5.2.2 Contribution . . . 128

5.3 Description of the stochastic system . . . 130

5.3.1 Dynamic multi-class server allocations . . . 135

5.4 Set up for fluid and diffusion approximations . . . 138

5.5 Fluid and diffusion approximations for stochastic queueing networks with heterogeneous patients . . . 142

5.5.1 Definitions . . . 142

5.5.2 Mathematical foundation for limiting theorems and the fluid approximation . . . 143

5.5.3 Diffusion approximation . . . 152

5.5.4 Virtual waiting time . . . 158

5.5.5 Production of outcomes . . . 164

5.6 Summary and Discussion . . . 165

5.6.1 Limitation . . . 166

5.6.2 Possible avenues for future work . . . 167

5.7 Conclusions . . . 168

6 Application of fluid and diffusion approximations to patient flow in community health care 170

6.1 Introduction . . . 171

6.2 Application to community health care . . . 173

6.3 Exploration of the accuracy of fluid and diffusion approximations for modelling community health care . . . 178

6.3.1 Description of the simulation model . . . 178

6.3.2 Single service and a single health state . . . 182

6.3.3 Single service and multiple health states . . . 200

6.3.4 Extending to multiple services 212
6.4 Summary and Discussion 216
6.4.1 Limitations 219
6.4.2 Possible avenues for future work 221
6.5 Conclusions 222

7 Conclusions **223**

Bibliography **228**

A Chapter 3 **243**
A.1 Details of data cleaning process 243

B Chapter 5 **245**
B.1 Proof of Theorem 5.5.1 245

C Chapter 6 **250**
C.1 Code for discrete event simulation of stochastic system 250
C.2 Code for discrete event simulation of virtual waiting time 263
C.3 Code for the fluid and diffusion approximation 267
C.4 Code for fluid and diffusion approximation of virtual waiting time . . 274
C.5 Code for implementing comparison between the models 275
C.6 Parameters used in the fluid and diffusion approximation of section 6.3.4 280

List of Tables

1.1 Examples of community services 19

2.1 Final terms for literature searches. The terms in bold are those used within the initial search and the non-bold terms are those added through an iterative process. 36

2.2 Inclusion and exclusion criteria for assessing papers presenting models of patient flow 37

2.3 Reasons for exclusion at full text assessment 40

2.4 Papers included from "Patient flow within community care" search only 50

2.5 Papers included from both "Patient flow within community care" search and "Patient flow and outcomes" search 57

2.6 Papers included from "Patient flow and outcomes" search only 58

3.1 Variables contained in the dataset used for producing visualisations . 79

3.2 Sources for community referrals included within the dataset. Community services included in this table represent those that did not feature as a specialty in the dataset. 83

3.3 NELFT community services, known as specialties, included within the data 84

3.4 Seven high activity referral edges form the bulk of activity in the specialty to specialty network, Figure 3.15 92

3.5 Table of chains that occur more than 20 times in the data, noting how many second referrals occurred in the first 14 days, 28 days, and length of the data 95

4.1 Table of key outcome measures as identified from Quality Accounts 2013-2017, conversations with NELFT staff, commissioning datasets and routine patient data 111

5.1 Parameter definitions for stochastic system $t \in [0, T],\ k \in H,\ i \in Ser$ 132

6.1 Total computation time for the simulation as the number of runs increases 181

6.2 Parameters used to assess the accuracy of the approximations - steady state analyses of the effect of c, λ and q 184

6.3 Errors between the approximations and the simulation during the "formation period" as a percentage of the simulated solution - effect of q, c and λ 187

6.4 Errors between the approximations and the simulation after the "formation period" as a percentage of the simulated solution - effect of q, c and λ 188

6.5 Parameters used to assess the accuracy of the approximations - steady state analyses of the effect of θ, δ_R and δ_U 190

6.6 Error between the approximations and simulation as a percentage of the simulated solution - effect of parameters $\theta, \delta_R, \delta_U$ 191

6.7 Parameters used to assess the accuracy of the approximations - steady state analysis of a single service and two health states 201

6.8 Error between the approximations and simulation as a percentage of the simulated solution 202

6.9 Parameters used to assess the accuracy of the approximations - time-varying analysis of a single service and two health states 204

6.10 Time taken to solve the fluid approximation for different sizes of dt . 211

A.1 Details of the data cleaning process. 243

C.1 Parameters used in the fluid and diffusion approximation of section 6.3.4 281

List of Figures

1.1 Diagram of a queueing system and its fundamental processes 23

2.1 Flow chart of literature search results 39

3.1 Diagram of a simple network representing referrals between services . 73
3.2 Example of service categorisation task 75
3.3 Example of referral mapping task for physical health services 76
3.4 Map produced by CCG of possible referral routes through community physical health services 77
3.5 Hypothetical patient level referral data for community health care . . 80
3.6 A histogram showing the distribution of the total number of referrals per patient in the dataset - maximum of 64 81
3.7 A histogram showing the distribution of the maximum number of concurrent referrals, per patient, within the date range of the dataset - maximum of 12 82
3.8 A histogram showing the distribution of referral lengths, per referral, in days - maximum $>$ 800 days. Each bar represents a span of 20 days 82
3.9 Example network diagram 85
3.10 Timelines showing how the number of patients in their 2nd, 3rd, 4th, 5th and 6th+ referrals changes over time. Time = 0 corresponds to the start date of a patient's index referrals. 86
3.11 Timelines showing how the number of patients involved in 2, 3, 4, and 5+ referrals at the same time changes over time. Time = 0 corresponds to the start date of a patient's index referrals. 87

3.12 Network map of referrals within NELFT's community health care services. All sources, specialties and edges. 90
3.13 Network maps of referrals within NELFT's community health care services. High activity network: edges with > 2 per month. 91
3.14 Network map of referrals within NELFT's community health care services. Low activity network: edges with ≤ 2 per month. 92
3.15 Network map of referrals within NELFT's community health care services. Specialty only network with all specialty to specialty referrals. 93
3.16 Chord plot for chains consisting of a source and two specialties. Only instances with > 20 occurrences are shown to improve interpretability. 94
3.17 Example of joint uses of each community service, for instances of $>$ 20 occurrences 96
3.18 Example of sunburst plot's interactive capability 97
3.19 Timelines of subsequent referrals for patients whose index referral was to the District Nursing Service 98

5.1 Diagram of patient flow between services within the queueing network 131
5.2 Diagram of a single service within the stochastic queueing network . . 131

6.1 (a) Stability issues with forward Euler calculation for the variance when time step is too large, $dt = 0.6$; (b) Improved stability and accuracy of solution for smaller time step, $dt = 0.5$. 180
6.2 Example of increased accuracy in the variance of the simulated waiting time as the number of runs increases 181
6.3 Comparison of results from the approximations and the simulation for a system with $\lambda = c = 20, q = 0.5$ and $\widehat{\rho} = 2$ 184
6.4 Error in the average value of each process state. After time $t = 2.8$ the accuracy of the fluid approximation improves greatly. 185
6.5 Graph of error in the variance of each process state. After time $t = 2.8$ the accuracy of the diffusion approximation improves greatly. 185
6.6 Error in the approximation when $\widetilde{\rho} = 1.\dot{1}$ 189
6.7 Example of an initial spike in the variance of the VWT 189

6.8 Example of the accuracy of the approximations in modelling a fall in the rate of arrival - process states 193
6.9 Example of the accuracy of the approximations in modelling a fall in the rate of arrival - the VWT 193
6.10 Example of the accuracy of the approximations in modelling a fall in the rate of arrival - the variance of the VWT 194
6.11 Example of the accuracy of the approximations in modelling a system that moves from an overloaded phase to an underloaded phase - process states 195
6.12 Example of the accuracy of the approximations in modelling a system that moves from an overloaded phase to an underloaded phase - VWT 195
6.13 Example of a seasonal spike in arrivals - the number of patients in each process state and the variance of each 196
6.14 Example of a seasonal spike in arrivals - the error in the number of patients in each orbit 197
6.15 Example of a seasonal spike in arrivals - the error in the variance . . 197
6.16 Example of a seasonal spike in arrivals - the VWT 198
6.17 Example of a seasonal spike in arrivals - the variance of the VWT . . 198
6.18 Two health state system with dynamic server allocation - number of patients in each process state and their variance 205
6.19 Two health state system with dynamic server allocation - number of servers available to each queue over time as they gain/lose servers . . 206
6.20 Two health state system with dynamic server allocation - the VWT . 207
6.21 Two health state system with dynamic server allocation - variance of the VWT 208
6.22 Two health state system with dynamic server allocation - the production of outcomes 209
6.23 Example of small errors in the solution of the VWT when modelling competing queues. These errors are of order dt. 211
6.24 Three service and three health state system - number of patients in each process state 214
6.25 Three service and three health state system - dynamic server allocation 214

6.26 Three service and three health state system - the VWT and variance 215
6.27 Three service and three health state system - the production of outcomes 215

Chapter 1

Introduction

Community health care is formed of clinically diverse and geographically dispersed services that provide local health care close to or in patient homes. These services are fundamental in the delivery of care and important within global policy [1], helping to maintain and improve patient health through a range of clinical activities. This includes potentially recurring, long-term care (such as diabetes services) and shorter, potentially one-off episodes of care (such as an orthopaedic service), see Table 1.1 for examples of community services. The breadth of their role, diversity of care and complexity of patient use present a challenge in the planning and operation of community services. This is where the application of operational research methods to community health care may contribute, which is the focus of this thesis.

In beginning this work, I shadowed several community services, including: a prosthetics and wheelchair service; an integrated community care team (a service for patients with complex, long-term conditions who may require health and social care); an acute admission avoidance service (known as the "falls ambulance" consisting of a paramedic and a nurse); and district nursing.

In observing these services, both the care offered and patients treated were diverse. Some patients presented with a single condition, simply requiring a check up; whilst others presented with on-going problems, such as persistent leg ulcers, and

Service name	**Description**
Adult speech and language therapy	Supports patients with communication and swallowing difficulties - such as those with stroke, brain injury, dementia, or voice disorders.
Cardiac service	Provides support to patients with heart failure.
Community rehabilitation service	Provides assessment and rehabilitation to patients who have been diagnosed with a neurological condition.
Community treatment team	Aims to prevent unnecessary hospital admissions for people experiencing a physical health crisis.
District nursing	Provides 24 hour care for people with an identified nursing need, including those who are chronically sick or terminally ill.
Orthopaedic service	Assesses patients' needs - such as physiotherapy - before and after hip or knee replacements.
Podiatry service	Provides foot care services to those with a clinical need, treating patients with long-term conditions - such as diabetes and rheumatoid arthritis.

Table 1.1: Examples of community services

required regular treatment. Other patients were more complex, presenting multiple co-morbidities that stemmed from a single long-term condition, such as diabetes, and required several services.

A final observation was the potential impact of these services on the wider health care system. For example, the "falls ambulance" responded to new emergency calls that would result in an acute admission since patients required treatment that a paramedic could not provide, but a nurse/paramedic team could. Having treated a patient and kept them out of acute care, they then sought to refer them to community services that could meet their needs and provide ongoing support. Thus, their aim was to reduce acute demand and enable patients to receive the appropriate care in the community.

1.1 Health care policy and community services

Community services are seen as crucial in meeting the current and future challenges facing health care [2]. They help to ensure that care: is person-centred, coordinated and closer to patients' homes; maintains patient health and independence; and, minimises hospital stays wherever possible [3]. Several high-priority national policies, such as the Better Care Fund and the NHS Forward View [4], require a larger role for the community sector. Thus, there has been an emphasis within NHS policy towards moving services out of acute settings and into the community. This is often motivated by the perceived benefits that increased community care may lead to reduced health care costs, improved access to services, improved quality of care, a greater ability to cope with an increasing number of patients, and improved operational performance in relation to a patient's health and time [1].

A scoping review analysed the evidence for the impact that moving services out of acute settings may have on the quality and efficiency of care [5]. It found that under certain conditions, moving services into the community may help to increase patient access and reduce waiting times. However, across multiple types of care (minor surgery, care of chronic diseases, outpatient services, and GP access to diagnostic tests), the quality of care and health outcomes may be compromised if a patient requires competencies - such as minor surgery - for which acute services are better equipped.

On the evidence for the effect on the monetary cost of services, [5] found that it was generally expected that community care would be cheaper when offset against acute savings. However, increases in the overall volume of care [6] and reductions in economies of scale [7, 8] may lead to an increase in overall cost in certain instances.

Given the importance of these services, the growing emphasis in delivering more

care within the community sector, and the questions around how best to develop and manage these services, a clearer understanding is required. This is where applying *operational research* (OR) methods to community care services can contribute.

1.2 Operational research

OR is a discipline in which analytical methods from mathematics, statistics, engineering and systems thinking are used to inform and understand decision related problems. Traditionally, it is used to provide insight into processes that occur in complex systems and organisations, given the system's purpose, the available resources and the key measures used to evaluate the system. For example, within health care, systems may be modelled to understand how goals (such as improved access) may be achieved when constraints (such as fixed capacity) and objectives (such as reduced operational costs) are considered. An example of one such method is *patient flow modelling*, which is the focus of this thesis.

1.2.1 Patient flow modelling

In a model of flow, a system is viewed as comprising a set of distinct *compartments* or *states* through which continuous matter or discrete entities move. A key characteristic is that the set of states and the set of transitions between them comprise a complete description of the modelled system.

Within health care applications, the entities of interest are commonly patients (alternatives include blood samples or information). In [9], two viewpoints for understanding patient flow are identified, an *operational perspective* and, less commonly, a *clinical perspective*. From an operational perspective, the states that patients enter, leave and move between are defined by clinical and administrative activities, and how patients interact with a care system. Thus, states may represent, for example,

a consultation with a physician, being on a waiting list for surgery, or specific care settings.

From the clinical perspective, states are defined by some aspect of patient health; for instance, whether a patient has symptomatic heart disease, or the clinical stage of a patient's tumour. A more generic view is that the states represent an amalgam of activity, location, patient health and changeable demographics [10].

Within the modelling process, characteristics of the patient population and the system states are incorporated to evaluate how such *factors influence flow*. Examples of the former include patient demographics or care requirements, whilst examples of the latter include capacity constraints relating to staffing, resources, time or budgets. The characteristics used depend upon the modelled system, modelling technique, and questions being addressed.

Furthermore, the performance of a system may be evaluated by *output measures*, such as resource utilisation [11], average physician overtime [12] and waiting time [13]. The measures used depend upon the modelled problem, modelling technique and the factors considered to influence flow.

1.2.2 Queueing theory

One way to model patient flow is through the use of *queueing theory* - the mathematical study of how queues form, grow and disperse as potential service users (e.g. customers, patients) seek to access a service. Here I briefly discuss some of the basics of queueing theory. For a more comprehensive treatment see Advances in Queueing: Theory, Methods, and Open Problems [14] and Applied Probability and Queues [15].

To analyse a queueing system, several processes must be understood and modelled, see Figure 1.1. Firstly, service users from a *population* of potential users arrive to the service. This population is considered to be all those who may possibly use the service or a group of service users who are of interest. It may be modelled as

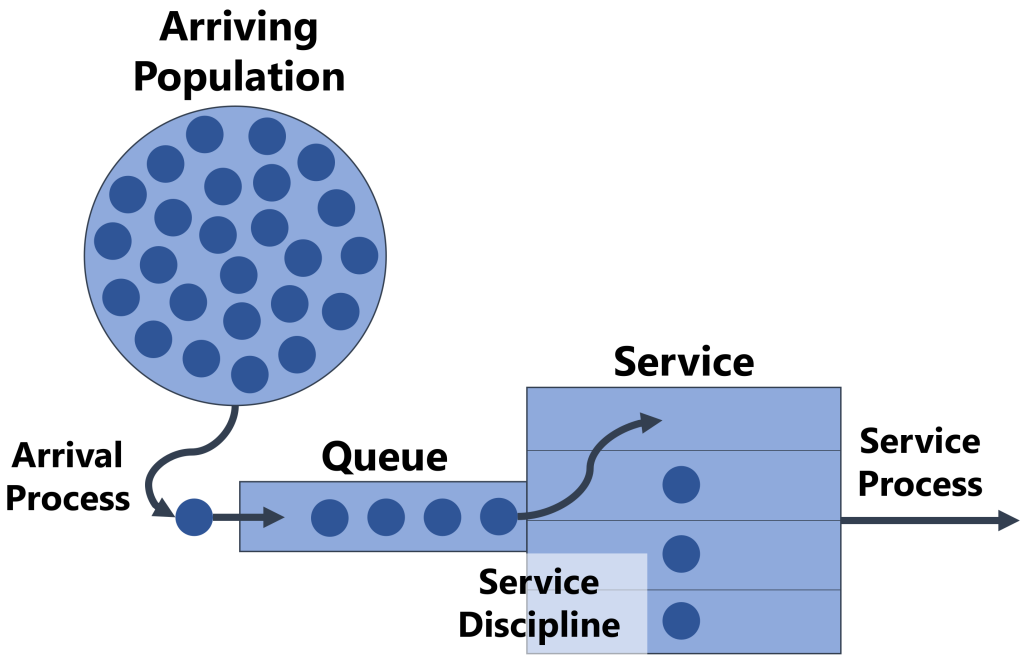

Figure 1.1: Diagram of a queueing system and its fundamental processes

infinite or finite (e.g. a pool of repeat users), and homogeneous or heterogeneous (e.g. when several distinguishable types of service user exist). Furthermore, how people arrive at the service must be understood and is modelled by an *arrival process*. This considers the characteristics of their arrivals (e.g. whether they arrive individually or in batches), possible influences (e.g. limited waiting space), and the timing of arrivals (e.g. whether they can be modelled stochastically or deterministically; the time distribution).

A Poisson arrival process is commonly assumed, where service users arrive according to some mean rate λ, such that, in a time interval $[0, t]$ the probability that n service users arrive is given by:

$$\mathbb{P}(i = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}$$

A benefit of this distribution is that *inter-arrival times* (the time between consecutive arrivals) are exponentially distributed and thus *Markovian* (possessing a "memory-less" property).

Secondly, the configuration and behaviour of the *queue* need to be defined. The configuration relates to whether there is a single queue or multiple parallel queues, and if there is limited waiting space. The queue's behaviour includes whether service users may: *renege* (when someone leaves the queue having waited too long), *balk* (when someone does not join the queue based on its length) or *jockey* (for parallel queues, someone moves to another queue to try and reduce their waiting time).

Finally, the *service process* must be defined. This consists of: a service discipline (how service users are selected for service e.g. first come first served, last in first out, by priority, etc.), the number of servers (single or multiple) and the distribution of service time. Similar to the arrival process, there is a probabilistic distribution that governs the time within which patients complete service. Commonly, service times are considered to be independent and exponentially distributed with a mean service rate of μ. Thus, the Markovian property holds for service times. For T_n, the time when the $n-$th service user completes service:

$$\mathbb{P}(T_n > t_0 + t | T_n > t_0) = \mathbb{P}(T_n \geq t)$$

By describing arrival processes as Poisson processes and treating service times as exponentially distributed, the model is viewed as a random process and called a *Markovian model*. Thus, the future movement of an entity is dependent only upon its present state, and independent of the time spent in that state or the pathway previously travelled. Whilst few systems (especially in health care) are truly Markovian, such systems may be understood using *steady state analysis*. This allows for the calculation of long run averages of system metrics.

A key concept within a queueing system is *capacity*. This relates to the limitations that a system's service processes and available resources place on how service users are served and the number that can be served. In a real world system, this may be interpreted as the number of available beds, working hours of staff, or the constraints

of physical space. Within a model, these may be represented by *capacity constraints* such as multiple servers, time variable dynamics and limited queue lengths.

Models of queueing systems may become large and complex depending on the application and the analysis carried out. In this thesis, I develop methods for modelling networks of services with heterogeneous patients and complex flow dynamics.

1.3 North East London Foundation Trust

During this project I have collaborated with the North East London Foundation Trust (NELFT). Covering a large area of north east London (Waltham Forest, Redbridge, Barking and Dagenham, Havering) and Essex, NELFT provide health care to a population of almost 2.5 million patients.

In recent years, NELFT adopted a community based model of health care for some of their services such as their provision of care for long-term illnesses, mental health and elderly patients. In accordance with national developments and changes in policy [16], this followed "the trend to deliver more care out of hospital" [17].

To begin this work, I held scoping meetings with clinicians and care managers from several community services. This included meeting with staff from the integrated care management service; the COPD, heart failure and respiratory service; and older adult care. From these meetings, I decided to focus this research on community based physical health services for patients aged 65 and over based in Havering. This choice was made due to the operational difficulties of these services and mix of service users. For example, the diverse profile of services makes planning and organising difficult, as do the multiple points of access for each service. These difficulties were apparent from conversations with NELFT clinical staff; thus, the insight gained during this work may be beneficial to NELFT.

1.4 Evaluating community health care

Typically, OR methods are used to quantitatively understand systems in terms of their operational capability to deliver service. Within health care, this is the ability of services to provide care efficiently given the available resources, demand of patients and the service they provide. Thus, outputs often focus on *process outcomes* - measurable, operational performance indicators such as waiting times, queue lengths and throughput, that can be compared to the real system.

In practice, health care services are also assessed according to other measures that relate to several domains (for example: clinical effectiveness, patient safety and patient experience [18]). The measurement of these is known as *quality measurement*.

In [3], the authors found that despite the crucial role of community health care, these services had been overlooked within the general national agenda around assuring and improving quality. In contrast to acute and primary care, they found that approaches for measuring and improving quality in these services were less developed. Moreover, they found that nationally, data on quality, quality measures for community services, and systems for capturing such data had been underdeveloped. This was especially true for national comparability and comparability between services. Thus, they noted that as challenges arise within the coming years, there is a risk that poor or declining quality would not be identified promptly.

Furthermore, they noted that outcome measures similar to those used within acute services, namely short-term clinical outcomes, may be ineffective within the community setting. This was due to the focus of many community services towards treating long-term conditions.

Finally, the authors suggested that the implementation of national and local quality measures is complicated by:

- the diversity of services provided within the community sector;

- multiple service providers;
- the complexity introduced by the mix of patients and how they use services;
- weak information infrastructure in community care;
- the difficulty of monitoring the quality of care provided in patient homes.

1.5 Purpose and aim of this thesis

Historically, community services have received less attention within the OR literature than acute and primary services. Given the challenges noted above, the aim of this project was to inform how patient flow in community health care may be modelled, and to explore how *patient outcomes* are used by these services. Informed by the literature [19] and clinical practice, in this thesis patient outcomes are considered as: *measurable aspects of patient health that are potentially influenced by care.* Notably, they may be used to track patient health and evaluate the quality of care.

Incorporating patient outcomes within patient flow modelling is increasingly pertinent. For example, improved patient outcomes and satisfaction are often used justifications for the increased provision of community care [1]; thus, the combination of outcomes and patient flow modelling may help evaluate this assertion. Furthermore, this combination may help to inform the organisation of health care services according to operational capability and clinical impact on the patient population; unifying two concerns of providers and patients in a single modelling framework.

The ultimate goal of this thesis is to begin to develop methods for modelling the "flow of outcomes" - the perspective as to how individual services contribute to the output of outcomes within a network of services where patients may participate in multiple care interactions. I achieve this through the combination of patient flow modelling and patient outcome progression.

The rationale for this work is partly motivated by how outcome measures are used at a population level. Often, they are understood as proportions of patients from a given population with common clinical characteristics post care.

When a time frame is considered, its length is often significantly greater than that of a care interaction e.g. monthly. Thus, these measures represent a somewhat static view in relation to the daily operation of services, which may be misleading. Since care is dynamic, static measures fail to capture variability in the daily "output" of outcomes from services. Furthermore, these measures fail to capture how changes in "outcome production" occur as patient flow problems arise and disperse. By incorporating patient outcomes into models of patient flow, the timely impact that services have on patient health may be understood.

The combination of patient flow modelling and patient outcomes may also be used as a novel method for measuring the performance of a system. In practice, services are often evaluated individually; however, when patients are able to use several services and have multiple care interactions, the performance of services is inherently linked. Thus, through the development of these methods, a holistic view of a system's performance may be gained; viewing the services as components within an interrelated network of services that work together to produce good outcomes.

Given the above focus, the presented work is formed of two sections. Chapters 2, 3 and 4 present my work towards understanding the operation and clinical impact of community health care services and how patients use these services. This is achieved by: a systematic review of OR literature (chapter 2); analysis of referral data for NELFT community services through visualisation to understand common referral dynamics (chapter 3); and, an exploration of what outcome measures are collected and used to evaluate different community services, and how they may be used in a patient flow model (chapter 4).

Of note, the data analysis and visualisations (chapter 3) had two major contri-

butions within this body of work. Firstly, it enabled the identification of important patient flow dynamics such as patients reusing services multiple times and the possibility for patients to use several services either sequentially or concurrently. This was important for the development of the theoretical model in chapter 5 since these are key dynamics to model. Furthermore, this was the first (to my understanding) application of such methods in community health care, that focussed on the dynamics seen in this setting. Secondly, it provided an opportunity to analyse NELFT's services and share important information and insights that may otherwise have been difficult to identify or communicate. An example of this is the complexity and diversity of patient referrals and use of services in the system, especially when reuse and common pathways were considered.

The second section of work is presented in chapters 5 and 6 and consists of my work in developing a theoretical method for modelling the combination of patient flow and patient outcomes within a network of queues. Notably, there is a disconnect between the applied body of work and the theoretical. Whilst the work in chapters 5 and 6 is informed by the prior chapters and was originally intended to be implemented within the Trust, only a theoretical modelling approach could be presented due a lack of usable data on patient use and outcomes.

The methods developed are extensions of fluid and diffusion approximations for stochastic queueing networks (chapter 5), which are evaluated to understand the parameter space for which they are accurate, given the extensions (chapter 6). This method provides a framework for quickly modelling large time-dependent complex systems due to its scalability. Through an analysis of smaller systems, understanding of the accuracy and use of these methods is highlighted in chapter 6.

Notably, these methods differ from those seen in chapter 2 since they have not previously been applied to community health care, can be applied to several diverse services, used for time and health dependent analysis of a system, and may be used

to model patient mix in the system. The main contribution of this model is the emphasis on and ability to evaluate systems by the contribution of services to the clinical (outcomes) and operational (flow) performance of the system. Furthermore, the methods produced in chapter 5 may be used to model the types of problem and provide the types of analysis stated above. This is explained in detail in chapter 6.

1.6 Structure of this thesis

To begin, I systematically review two types of relevant literature. By considering applications of patient flow models in community services and how outcomes have been previously incorporated into flow models, I identify any possible gaps in the literature. Based on the conclusions of this chapter, I position my work in chapter 5 to address some of the gaps that arise - such as time dependent modelling, networks of multiple services and the possibility of multiple care interactions.

In chapter 3, I explore the referral processes and dynamics of patient use in NELFT community services. Through visualisations of patient data, I seek to understand how the care provided by multiple community services may be modelled. This work focuses on how services are connected by referrals and patient use, informing the key dynamics and patient flow mechanisms of the model. By identifying the reuse of services and use of several services as key dynamics, these are incorporated into the theoretical framework of chapter 5.

In chapter 4, I explore what outcome measures NELFT use to monitor the clinical effectiveness of their community services by surveying several sources of information. The aim is to understand how and by what measures community services are evaluated in order to inform how the progression of patient outcomes may be modelled when patients attain care and use several services. In particular, I identify whether there are any measures currently used by NELFT that may be incorporated into

patient flow modelling and the method produced in chapter 5.

Having learned about what happens within community health care in practice, in chapter 5 I develop a theoretical model of such systems. This method is a fluid and diffusion approximation of a stochastic network of services which incorporates patient health. Here I present the mathematical framework and proofs for this method, and produce a basis for modelling the "flow of outcomes". I also introduce a dynamic multi-class server allocation for parallel queues - where servers are assigned to parallel queues and continuously adjusted in response to changes in demand for service and patient mix.

In chapter 6, I evaluate the appropriateness of the fluid and diffusion methods developed in chapter 5 for modelling community health care services, considering how they may relate to real world systems. Through a series of applications to hypothetical scenarios, I explore the parameter space of these methods, assessing when they are most accurate and identifying the types of analyses they may be used for. Beginning with small systems, I address how the accuracy of the model is affected by the extensions introduced in chapter 5, from which understanding is gained for application to a larger system. This chapter culminates with a discussion on modelling the "flow of outcomes".

To conclude this thesis, in chapter 7 I discuss the contributions made in each chapter and the possible directions for future work.

Chapter 2

Literature Review

In this chapter, I present a systematic literature review of operational research methods for modelling patient flow. This is formed of a review I published with my supervisors [20], which I have updated for use within this thesis.

Papers are assessed for inclusion at three levels, with the references of included papers also assessed for inclusion. I make comparisons between each paper's setting, definition of states, factors considered to influence flow, output measures and implementation of results. The discussion focuses on the common complexities and characteristics of the models, from which I suggest possible directions for future work and discuss how this informs the rest of this thesis. The aims of this chapter are to:

1. Explore applications of patient flow models in community services;
2. Understand how outcomes have been previously incorporated into flow models;
3. Identify any possible gaps that exist within the literature and position my work to address them.

2.1 Introduction

For many decades operational research methods have been applied to several settings and problems within health care to better understand the challenges facing health care systems and inform decision making through useful analysis. Whilst the focus of this thesis is patient flow, it should be noted that operational research methods have been applied to a wide range of systems and problems including scheduling, resource and capacity management, clinical and administrative modelling, logistics, economic analysis, risk management, treatment evaluation, and design and layout modelling [21].

Due to the range of problems that may be modelled and the number of available techniques, many reviews of operational research literature focus on specific applications or methods. Thus, to inform the work presented in this thesis, I systematically reviewed two types of relevant literature. The first type were publications that presented models of operational patient flow within a community health care context, denoted "Patient flow within community care". The second type were publications that presented combinations of patient outcomes and patient flow modelling in any setting, denoted "Patient flow and outcomes". No specific setting was sought in the latter to identify potentially transferable methods.

For a broad and comprehensive review of operational research methods in application to health care see [21] and [22]. In [21] the authors present a literature review of operational research within UK health care, covering a broad range of methods and applications. From their findings they categorise the types of technique employed, and analyse the applications and publication trends since 2000. Due to its breadth, their review also contains several methods and applications not featured in this chapter and serves as a good basis for understanding recent developments in the literature. This review also includes problems typical of other health care setting such as overcrowding and resource allocation problems often found in acute

and emergency care settings. In [22] the authors produce a taxonomy of the types of planning decisions that may be modelled, the suitable methods for understanding a given scenario and the possible benefits of the analysis. Their review considers several care settings and services (ambulatory, emergency, surgical, inpatient, residential and home care) and highlights the breadth of both the field of operational research and the problems to which such methods may be applied. They also identify several types of planning decision that do not feature in this chapter (for example, access policy, staff scheduling and facility layout).

Throughout this thesis I will refer back to the findings detailed in this chapter. Furthermore, by using a systematic approach, I ensure that this review is reproducible and rigorous.

Structure and aims of the chapter

In section 2.2, the method of review is discussed, noting how the search was conducted, the process used to assess the literature for inclusion and the framework used for reviewing the literature. Following this, the results of the search and inclusion assessment are presented in section 2.3. In section 2.4.1, the "Patient flow with community care" papers are analysed, and in section 2.4.2 "Patient flow and outcomes" papers are analysed. In section 2.5, I summarise and discuss the findings from both searches, drawing out key themes from across the two literatures. This chapter ends in conclusions drawn from the findings and suggestions of future avenues of work.

2.2 Method of review

To analyse this literature, I conducted a configurative systematic review - an approach for gathering and understanding a heterogeneous literature, to identify patterns and develop new concepts [23]. I performed two searches to find peer-reviewed

operational research (OR) publications, as previously noted. Seeking papers published in English before September 2017 (previously November 2016 in the published review) with no lower bound publication date, I conducted this search within the electronic databases Scopus, PubMed and Web of Science.

For each search I used a combination of search terms listed in Table 2.1. To find papers related to "Patient flow within community care" I sought records with at least one operational research method term in the article title, journal title or keywords AND at least one patient flow term in the article title, journal title, keywords or abstract AND at least one community health setting term in the article title, journal title, keywords or abstract. To find papers related to "Patient flow and outcomes" I sought records with at least one operational research method term in the article title, journal title or keywords AND at least one patient flow term in the article title, journal title, keywords or abstract AND at least one outcome term in the article title, journal title, keywords or abstract. Operational research method terms were not sought in abstracts since this greatly increased the number of redundant papers.

Initial sets of search terms relating to community health care settings and operational research methods were informed by [22], with synonyms added prior to the preliminary searches. For patient flow terms and outcome terms, I formed initial lists that I considered to be relevant (original search terms are in bold in Table 2.1). The first batch of papers found using these terms were examined for further applicable search terms, and these were subsequently added to form an updated list.

Papers obtained from the final searches were assessed for inclusion for full review at three levels. If a paper was not a literature review, it was required to meet all the inclusion and none of the exclusion criteria outlined in Table 2.2. For each included paper, references were assessed using the same inclusion and exclusion process to find any papers that may have been missed in the searches.

Literature reviews were included at each level if they were concerned with opera-

OR method terms	Patient flow terms	Setting terms	Outcome terms
Computer simulation	**Access time**	**Community based**	**Outcome**
Discrete event simulation	**Bed occupancy**	**Community clinic**	**Patient class**
Heuristics	**Capacity allocation**	**Community facility**	**Patient type**
Markov chain	**Capacity management**	**Community level**	**Quality of life**
Markov decision	**Capacity planning**	**Diagnostic facilities**	**Readmission**
Markov model	**Care management**	**Health care center**	**Referral**
Mathematical model	**Patient flow**	**Health care centre**	
Mathematical programming	**Patient pathway**	**Health care clinic**	*Disease progression*
Metaheuristics	**Patient process**	**Health care practice**	*Health status*
Operational management	**Patient route**	**Health care service**	
Operational research	**Patient throughput**	**Health center**	
Operations management	**Process flow**	**Health centre**	
Operations research	**Wait time**	**Health clinic**	
Optimisation	**Waiting list**	**Health clinic**	
Optimization	**Waiting time**	**Health facility**	
Queueing		**Healthcare center**	
Queuing	*Care access*	**Healthcare centre**	
Simulation model	*Demand management*	**Healthcare clinic**	
System dynamics	*Flow of patients*	**Healthcare facility**	
	Patients' flow	**Healthcare practice**	
Integer programming	*Flow of care*	**Healthcare service**	
Linear programming		**Home care**	
Modelling patient		**Home health care**	
Network analysis		**Long term care**	
Stochastic analysis		**Mental health**	
Stochastic modelling			
Stochastic processes		*Care facility*	
Visual simulation		*Community care*	
		Community health	
		Community healthcare	
		Homecare	
		Medical center	
		Medical centre	
		Multi facility	
		Multiservice	
		Residential care	
		Walk in	

Table 2.1: Final terms for literature searches. The terms in bold are those used within the initial search and the non-bold terms are those added through an iterative process.

<table>
<tr><th>Assessment level</th><th>Criteria</th><th>Patient flow within community care</th><th>Patient flow and outcomes</th></tr>
<tr><td rowspan="3">Title and journal</td><td rowspan="2">Inclusion</td><td>At least one operational research method term in the article title, journal title or keywords; AND

At least one patient flow term in the article title, journal title, keywords or abstract; AND

At least one community health setting term in the article title, journal title, keywords or abstract.</td><td>At least one operational research method term in the article title, journal title or keywords; AND

At least one patient flow term in the article title, journal title, keywords or abstract; AND

At least one outcome term in the article title, journal title, keywords or abstract.</td></tr>
<tr><td colspan="2">English language; published before September 2017 in peer–reviewed journals.</td></tr>
<tr><td>Exclusion</td><td colspan="2">Title or journal of publication had no relevance to operational research, health care or patient flow.</td></tr>
<tr><td rowspan="3">Abstract</td><td>Inclusion</td><td colspan="2">Abstract suggested that the paper focussed on operational processes of health care and that operational research methods were used to model patient flow.</td></tr>
<tr><td rowspan="2">Exclusion</td><td colspan="2">Papers based within management settings other than operational management;

The delivery of health care was not evaluated;

Only different scheduling policies were evaluated.</td></tr>
<tr><td>Abstract indicated that the paper was not based in community care.</td><td>Abstract indicated that the paper did not use patient outcomes.</td></tr>
<tr><td rowspan="2">Full text</td><td>Inclusion</td><td colspan="2">Abstract level inclusion criteria met in the full text;

A model was presented using mathematical concepts and language;

The model was well specified and reproducible;

Quantitative analysis of a health care system was conducted within the paper.</td></tr>
<tr><td>Exclusion</td><td colspan="2">Criteria for exclusion at abstract level met in the full text;

A model was viewed only in terms of its inputs and outputs without knowledge of its internal workings;

A model was formulated as a composition of concepts that could not be used for analysis;

A model was not rooted in analysis.</td></tr>
</table>

Table 2.2: Inclusion and exclusion criteria for assessing papers presenting models of patient flow

tional research methods for evaluating patient flow, were focussed on the operational processes of health care, and had no equivalent systematic review already included. Additionally, within the "Patient flow within community care" literature, review pieces were included if they solely focussed on community settings, whilst within the "Patient flow and outcome" literature, review pieces were included if they focussed on uses of patient outcomes in modelling processes.

Data tables were constructed to collate key characteristics of the literature and shape the analysis, noting each paper's setting, definition of states, factors considered to influence flow, output measures and implementation of results. The factors that were considered to influence flow included key flow dynamics (e.g. reuse of services), patient behaviours (e.g. reneging), important constraints (e.g. resource or time), and differences within the patient population (e.g. stratified groups or variable health). In this review, implementation refers to any actions taken by the researchers to share and use the results of the work within the modelled setting. Examples include feedback to stakeholders or changes within the operation or organisation of a system.

Informed by initial readings, papers were grouped into five categories based on analytical method: Markovian methods, non-Markovian methods, system dynamic approaches, analytical methods featuring time dependence and simulation approaches. The findings from these papers are tabulated and shown in Tables 2.4, 2.5 and 2.6.

2.3 Results of literature searches

The combined results of the original and updated searches, and selection of papers, are shown in an adapted PRISMA flow chart [24], Figure 2.1. Reasons for the exclusion of papers at full text assessment are shown in Table 2.3.

The published search and inclusion assessment provided 25 "Patient flow within community" papers, 23 "Patient flow and outcomes" papers and five papers in the

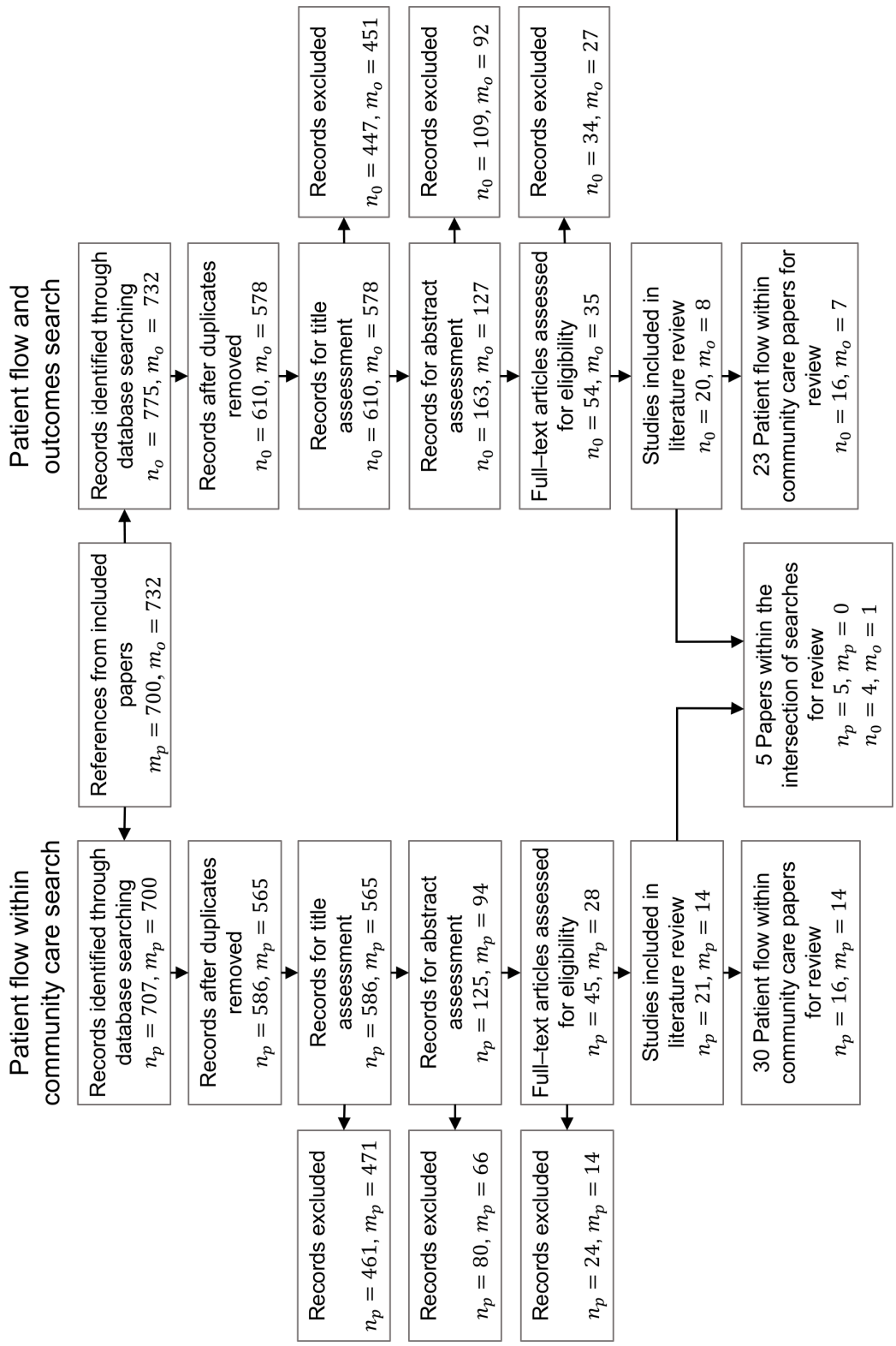

Figure 2.1: Flow chart of literature search results

58 papers were eligible for review: 30 "Patient flow within community care" papers; 23 "Patient flow and outcomes" papers; and five papers within the intersection.

intersection for full review. The updated search provided an extra five "Patient flow within community" papers, one newly published with four found from its references.

Number of papers excluded at full text assessment	Reason for exclusion				
	No OR/patient flow modelling	Non–community settings	Model not reproducible/ specified/ quantitative	Analysis of different scheduling policies	No patient outcomes
24 "Patient flow within community care" literature	5	9	7	3	N/A
14 "Patient flow within community care" references	2	8	3	1	N/A
34 "Patient flow and outcomes" literature	10	N/A	2	7	15
27 "Patient flow and outcomes" references	4	N/A	–	1	22

Table 2.3: Reasons for exclusion at full text assessment

In the next section, a synthesised analysis of publications within their respective search group is presented. Papers in the intersection are included in the "Patient flow within community care" section. I highlight three papers that are particularly informative in developing the methods presented in this thesis, providing further methodological insight into their work.

2.4 Analysis of papers

2.4.1 "Patient flow within community care"

Markovian models

The settings of these publications were long-term care [25, 26], residential mental health care [27], post-hospital care pathways [28], flow between community services and hospital care [29] and services for elderly patients with diabetes [30].

Within these models, states were defined as different services or stages of care,

with two papers also defining states of future care requirements [28, 30]. In [28] these states included patient mortality, admission to long-term care and re-hospitalization, whilst in [30] states of subsequent health progression were defined.

Two main factors were considered to influence flow within these models: the effect of congestive blocking caused by limited waiting space [27, 29] and the diversity of patients, defined by demographics [25, 26, 28] and severity of disease [30]. In considering blocking, flow was influenced by the available capacity and the average occupancy of each service.

The output measures used within these papers were: queue lengths and wait times for each state (with and without congestive blocking) [27, 29]; forecasts of demand [25, 26]; and the probability that patients would be in a given post-care outcome state [28, 30]. An analysis of different scenarios was undertaken in the latter two papers to identify how alternative pathways may help improve post-care outcomes.

None of the papers explicitly reported implementation of their results.

Spotlight: Modeling patient flows using a queuing network with blocking

N. Koizumi, E. Kuno and T. E. Smith used an open queueing network to analyse the effect of limited waiting capacity on wait times within community services for progressive, residential, mental health care [27]. Progressive care is configured so that patients move from higher intensity care to lower intensity care as their health improves and their care progresses. Three service states were considered within this paper: extended acute hospitals, residential facilities and supported housing, each with different service times and different capacities. Source states (general community care and acute hospital care) were modelled with infinite capacity.

The analysis in this paper focused on how limited queueing capacity within a queueing network may cause a congestive phenomenon known as blocking. This occurs when a patient, in a service i, is ready to move to the next level of care, in service $i + 1$, but no spaces are available. Since each service state represents

residential care, patients remain in service i until there is a free space in $i + 1$, potentially blocking incoming patients to service i.

The outputs for this method were average queue length and waiting times, evaluated for each service, first without blocking, and then with. In the latter, an algorithm was used to calculate the effect of blocking by an effective service time. For patients moving to a service with limited capacity, this was a combination of their service time and the expected wait time for the next service.

This method could be used to analyse how flow problems for one service affect flow throughout the system. For example, congestion in the system was greatly reduced if bottlenecks at supported housing were reduced, perhaps through changes in capacity. However, the authors also found that increasing the capacity of other stations would cause the problem to intensify at supported housing. Overall, this paper introduced: complex flow dynamics (how limited queueing capacity affected flow across the system), a mix of services, and progressive health care where patients have multiple care interactions.

Non-Markovian steady state models

Producing an optimisation approach for resource allocation, [31] defined states as services within specified pathways. The aim was to minimise overall costs whilst maintaining a desired level of care as measured by metrics, such as wait time targets, given the capacity constraints of the system, such as the number of beds. There was no indication of implementation.

System dynamics

In system dynamic approaches complex organizations are modelled using a system of coupled ordinary differential equations to analyse and design effective policies and process structures [32]. Five applications were found for modelling systems of

markedly different sizes and setting; these included an evaluation of the UK's NHS [33], community services used to bolster acute cardiac services [34], and long-term care [35, 36, 37].

States were defined as different services, such as community or acute services [34], types of residential and long-term care [35, 36, 37], or different sectors of care [33] (namely primary, acute, NHS continuing care and community care). In each model and setting, capacity and transition rate variables, such as waiting list size and clinical referral guidelines were considered, with [35, 36, 37] also including ageing populations. Furthermore, a feedback mechanism was used in [34] to evaluate how changes in the input variables affected future demand.

The main metrics used in each model related to demand and access, namely waiting times and patient activity such as the long run use of services and the length of queues [33]. In all of the papers, scenario analysis was performed to evaluate how changes within the model input parameters affected their outputs.

Implementation was reported in two papers [33, 35], both noting that results had been shared with clinical partners.

Analytical methods including time dependence

Applications included long-term institutional care [38, 39], home/community care [40], community mental health services [10, 41], care after discharge from an acute stroke unit [42], and specialist clinics [43, 44].

The state definitions within these models related to stages of care/different services [10, 38, 39, 40, 41, 42], whether patients were "waiting" or "in service" [43, 44], and health states, in particular stages of health progression [43] or post care outcomes [42]. The factors considered to influence flow included capacity of services [41, 44], patient demographics and care requirements [38, 39, 40, 42], and patient health between recurrent appointments [43].

Commonly, the system metrics related to the time a patient spent interacting with parts of the system - such as expected length of stay, waiting times and time spent in states. Other measures included the daily cost of care and likely post care outcome states for patients in different demographic groups [42]. In [41] appointment allocations were "optimised" to meet desired levels of queue lengths and wait times across multiple types of care interaction. Similarly, in [43] an "optimised" timing for sequential appointments was sought given variable patient health. The possible future demand for services, under different scenarios, was evaluated in [40].

Of these applications, three reported implementation, this included the creation of a tool [41], the sharing of findings with stakeholders [10] and the use of a model for care planning [40].

Spotlight: Improving Health Outcomes Through Better Capacity Allocation in a Community-Based Chronic Care Model

S. Deo, S. Iravani, T. Jiang, K. Smilowitz, and S. Samuelson sought to combine both clinical and operational aspects of health care in modelling the treatment of school children with asthma [43]. Using disease progression modelling alongside finite horizon stochastic dynamic programming, they modelled a scenario where patients periodically required repeated care from a single service. A model was created to allocate care appointments in order to maximise aggregate health outcomes subject to resource constraints. Each visit had the potential to improve their health, whilst longer times between visits meant patient health was more likely to decline. Thus, the method was designed to provide an "optimal" duration between visits for patients with varying health care needs.

The state of the system was tracked by an amalgam of each patient's information consisting of the time since their last appointment and their health between visits. Patient health was assumed to progress according to a Markov process defined over K discrete health states, 0 being the best and $K-1$ the worst. Given the modelling

of health in response to care, and in the absence of it, two health transition matrices were considered, representing positive and negative progression respectively.

This method included: time varying patient health that changed in response to care or in the absence of it; a patient's future use of service in light of health and limited capacity; and the effect of multiple care episodes on patient health.

Simulation methods

The settings of these papers included long-term care [45, 46, 47], outpatient services [48, 49, 50, 51, 52, 53], primary care and ambulatory clinics [54, 55, 56], and provisions of integrated acute and community services [57, 58, 59].

States were defined as different services, clinics, sectors of care or health care tasks within single clinics. In two papers the flow of patient information was modelled alongside patient flow [51, 53]; thus, state definitions also included stages of information flow.

The factors considered to influence flow were the health care requirements and demographics of patients [48, 49, 51, 55, 56], constrained capacity and rates of no show/reneging [48, 49, 56]. Monetary influences such as budgetary constraints, cost of care and profitability were considered in four papers [45, 52, 57, 59]. Also, the variability of time in completing care tasks was considered [51].

Common metrics related to the time that a patient spent in a state or in the system as whole. "Optimised" capacity levels relating to key performance measures were also widely considered [46, 47, 52]. In one paper [50], a single, composite system metric was calculated as an aggregate of multiple performance measures (such as average throughput, average system time and average queue time) and were stratified by day, facility routing and patient group.

Several papers noted implementation of suggested changes [46, 48, 50, 51, 54, 56].

2.4.2 "Patient flow and outcomes" papers

Markovian models

The modelled settings were transplant waiting lists [60, 61, 62], intensive care units [63] and emergency care [64]. In these models, states related to whether patients were "waiting" or had obtained a service/transplant. Patient priority states were also defined to reflect health deterioration [62].

The factors that influenced flow related to patient health (such as levels of severity, organ type required or probability of survival), with groups or states used to assign priorities within the patient population [61, 62], or to represent different demographics and care requirements. In each transplant paper, the reneging characteristics of different patient groups were considered with patients modelled as leaving the waiting list due to death or for other reasons [60, 62].

The output measures of these papers commonly related to the wait time faced by patients. Other metrics included the probability of reneging per patient group [62], the expected number of deaths for waiting patients [61], and lives saved under an admission policy [63]. In one paper [60], the average time spent in the system and in the queue for each demographic group was calculated, alongside the proportion of patients from each group who received a transplant.

None of the papers reported an implementation of their results.

Non-Markovian steady state models

The modelled settings and applications included an emergency department [11] and two waiting lists; one for hospital care [65], the other for transplant patients [66]. States were defined as stages of hospital care [11] and as "waiting" or "in service"[65, 66].

The factors considered to influence flow were seasonality [11], resource availability

and patient health (groups relating to care requirements [11, 66] or the wellness of patients [65]). Each model used metrics relating to the amount of time a patient spent within different parts of the system.

Implementation of results was noted by one paper [11], having developed software for use by clinicians and care managers. In addition, they provided feedback and educational sessions to help stakeholders understand the work.

System dynamics

For system dynamic approaches, a single paper was found [67], presenting an evaluation of patient flow between states of acute care and home care for patients with chronic disease. The factors considered to influence flow related to patient groups (based on their care requirements), whether they possessed insurance and potential improvements in their health outcomes given the care they received. Congestion and total resources were also considered. A scenario analysis was performed to evaluate the impact of different patient routes and resource allocations on the level of demand for services and the cost of providing care.

Analytical methods including time dependence

The modelled settings included care for chronic diseases [68], two intensive care models [69, 70], two radiotherapy models [71, 72] and two transplant waiting lists [73, 74].

States were defined as "in service" or "waiting", different services or appointment slots [71, 72] and multiple health states [68, 69, 73, 74]. Additionally, the factors considered to influence flow commonly related to differences within the patient population pertaining to: variable health [68, 73, 74]; care requirements/health related groups [73]; and the availability of resources such as organs [73, 74] or appointment slots [68, 71, 72].

Metrics produced by these methods commonly focussed on the amount of time a patient spent waiting for a service - for example, the optimal timing of appointments [68] or transplants [74], subject to changes in patient health. In one paper [73], output measures were calculated for different groups of patients to evaluate equity within organ allocation. Forecasts of capacity requirements and optimal allocations of resources, based on patient groups, were also common.

Two papers noted that their suggestions had influenced decision making [71, 68].

Spotlight: Dynamic Allocation of Kidneys to Candidates on the Transplant Waiting List

Following [60], where a simple queueing model was used to analyse the dynamics of multiple patient classes on a waiting list, S. A. Zenios, G. M. Chertow and L. M. Wein produced a deterministic model for a similar transplant waiting list system [73]. Modelled using a system of ordinary differential equations, they used a continuous representation of the state space instead of the usual discrete representation. The aim was not to produce an exact model but rather to represent the dynamics of the waiting list to assess transplant allocation policies.

Patients were grouped into several classes, representing their health status and demographics, with organs also classified into several groups representing demographic, immunological and physiological characteristics. Upon receiving a transplant, a patient could change health state, representing improved health. Furthermore, patients on the waiting list could leave due to death.

In this model, both patients and organs were assumed to arrive according to independent Poisson distributed processes. Thus, upon arrival, a patient would wait until an appropriate organ arrived (became available), which was then immediately transplanted i.e. no service time. They also modelled the potential for graft failure with patients who had received a transplant potentially experiencing a negative change in health and rejoining the queue.

By including the different demographics of patients and organs, and changes in health, the method was used to assess the factors that affect the equity-efficiency trade off in managing transplant waiting lists.

It was stated throughout the paper that the fluid model had multiple flaws in terms of accurately reflecting the system. For instance, in modelling stochastic flow processes, the deterministic model lacked variance. However, the intention was to stylistically represent the system so that new policies could be formed and assessed.

Overall, this paper included: deterministic, fluid representations of patient flow; several health related patient classes that changed in response to care; and, the possibility for the reuse of a service based on patient health.

Simulation methods

Applications included a cardiac catheterization clinic [75], transplant waiting lists [76, 77, 78], an emergency department [79], neonatal intensive care [80] and a health care resource allocation model [81].

Within these papers, states were defined as the number of beds; "waiting" or "in service", and health care tasks [75, 81].

The factors considered to influence flow within these models included patient demographics or care requirements [75, 77, 78, 81]; the health, mortality and survival rates of patients [77, 78, 81]; and capacity such as available resources and beds.

Several metrics were calculated within these methods, with the time patients spent interacting with or waiting within parts of the system a common measure. Other outputs of interest included capacity allocation [75, 76, 80], the cost of care [80], health benefits of service [81], and expected the survival rate of patients [77, 78].

Two papers noted the adoption of some of their suggested changes [75, 79].

Title	Authors	Setting	States	Factors considered to influence flow	Output measures	Implementation of results
Markovian Models						
Forecasting demand for long-term care services	Lane et al. (1985) [25]	Long-term care	Levels of care - 5 levels of intensity	Age Care required	Forecast of patients requiring care	Not explicitly stated
Forecasting Client Transitions in British Columbia's Long-Term Care Program	Lane et al. (1987) [26]	Long-term care	Levels of care - 5 levels of intensity - 2 services: Home, facility	Age Care required Gender	Forecast of patients requiring care Comparison of cohort transitions in system	Not explicitly stated
Modeling patient flows using a queuing network with blocking	Koizumi et al. (2005) [27]	Community care - mental health - Physical queues	Multiple residential services	Service capacity Traffic intensity per service Congestive blocking	Queue lengths and wait times - with and without blocking	Not explicitly stated
A block queueing network model for control patients flow congestion in urban healthcare system	Song et al. (2012) [29]	Community and hospital pathways - Physical queues	Community services Hospital registration General hospitals	Service capacity Traffic intensity per service Congestive blocking Batch arrival process	Queue lengths and wait times - with and without blocking	Not explicitly stated
Non-Markovian Steady State Analysis						
A model for planning resource requirements in health care organizations	Bretthauer & Côté (1998) [31]	General approach, examples: blood bank, health maintenance organisation - Physical queues	Different services Stages of care	Resource constraints e.g. Number of clinicians Performance constraints e.g. Wait time Multiple time periods	Optimised total capacity costs	Not explicitly stated

Table 2.4: Papers included from "Patient flow within community care" search only

Title	Authors	Setting	States	Factors considered to influence flow	Output measures	Implementation of results
System Dynamics Analysis						
A patient flow perspective of U.K. health services: exploring the case for new "immediate care" initiatives	Wolstenholme (1999) [33]	UK health service - Physical and non-physical queues	Primary care Secondary care Community care NHS continuing care	Volume of patients arriving Service capacity	Queue lengths Waiting times Bed occupation Long run use of services Scenario analysis	Some insights shared with NHS staff
Simulation analysis of the consequences of shifting the balance of health care: A system dynamics approach	Taylor et al. (2005) [34]	Community and acute - Non-physical queues	Cardiac services in community	Wait time Size of waiting list Demand feedback mechanism Clinical guidelines Service capacity	Average wait times Cumulative patient referrals and activity Overall cost of care Scenario analysis	Collaboration noted
Simulating the Impact of Long-Term Care Policy on Family Eldercare Hours	Ansah et al. (2013) [35]	Long-term care	Place of residence	Population age Capacity: - Beds	Effect of changes to capacity, demand and population dynamics on: - Eldercare hours - Eldercare leads	Developed with collaboration and feedback from care
Implications of long-term care capacity response policies for an aging population: A simulation analysis	Ansah et al. (2014) [36]	Long-term care - Physical queues	Types of care: - Acute - LTC	Capacity: - Clinicians - Beds Distribution of clinicians across care venues	Effect of changes to capacity, demand and population dynamics on: - Patient demand - Required capacity	Not explicitly stated
Policy choices in dementia care - An exploratory analysis of the Alberta continuing care system (ACCS) using system dynamics	Cepoiu-Martin and Bischak (2017) [37]	Home/residential care services - Non-physical queues	Types of residential care Waiting list for types of residential care	Varying age and health: - Progression of dementia Capacity: - Staffing and beds	Sensitivity analysis - Effect of changes in system parameters on: waiting list, staffing levels Scenario analysis	Not explicitly stated

Table 2.4 (Continued): Papers included from "Patient flow within community care" search only

Title	Authors	Setting	States	Factors considered to influence flow	Output measures	Implementation of results
Analytical methods featuring time dependence						
A continuous time Markov model for the length of stay of elderly people in institutional long-term care	Xie et al. (2005) [38]	Long-term care - Physical queues	Residential home care Nursing home care - Long stay - Short stay	Maximum likelihood estimation (MLE) of model parameters	Sojourn time Estimation of LOS Patterns of care usage	Not explicitly stated
A model-based approach to the analysis of patterns of length of stay in institutional long-term care	Xie et al. (2006) [39]	Long-term care - Physical queues	Residential home care Nursing home care - Long stay - Short stay	MLE of parameters Left truncated data Right censored data Patient characteristics: - Previous care, gender	Sojourn time Estimation of LOS Patterns of care usage	Not explicitly stated
Analytical methods for calculating the distribution of the occupancy of each state within a multi-state flow system	Utley et al. (2009) [10]	Community mental health care - Uncapacitated	General states Illustrated with states as different stages of care	Time spent in state	Time dependent distribution for occupancy of states	Suggestions made to stake holders
A deterministic model of home and community care client counts in British Columbia	Hare et al. (2009) [40]	Long-term care - Uncapacitated	Different aspects of LTC: - Home care - Accommodation Care environment - Publicly funded/ non-publicly funded	Time varying population characteristics: - Patient age - Wealth - Health status Initial conditions	Future demand for each aspect of LTC	Model used for planning future care
A mathematical modelling approach for systems where the servers are almost always busy	Pagel et al. (2012) [41]	Community mental health care - Non-physical queues	Different services	Capacity constraints e.g. Appointment slots Servers always busy	Optimal appointment allocation subject to wait time and capacity	Formulation of a tool
Appointment capacity planning in specialty clinics: a queueing approach	Izady (2015) [44]	Specialty clinics - Physical queues	Waiting In service	Abandonment - Fixed, capacity dependent Rejoins to queue Appointment type	Patient wait time Queue lengths No-show probability Referral variance	Not explicitly stated

Table 2.4 (Continued): Papers included from "Patient flow within community care" search only

Title	Authors	Setting	States	Factors considered to influence flow	Output measures	Implementation of results
Simulation Analysis						
Improving outpatient clinic efficiency using computer simulation	Clague et al. (1997) [48]	Outpatient - genito urinary medical clinic - Physical queues	Stages of care	Patient groups: - Clinical staff required - New or returning Mixed arrivals No shows Staffing constraints	Patient wait time Doctor wait time Clinic overtime Scenario analysis	Application of method in response to a feedback survey
Evaluating the design of a family practice healthcare clinic using discrete-event simulation	Swisher & Jacobson (2002) [49]	Family Practice Healthcare clinic - Physical queues	Stages of care Locations in the clinic	Patient groups: - Health Mixed arrivals No shows Staffing constraints	Patient wait time Staffing costs Revenue Clinician overtime Staff/ facility utilisation Scenario analysis	Not explicitly stated
Improving patient flow at an outpatient clinic: Study of sources of variability and improvement factors	Chand et al. (2009) [51]	Outpatient clinic - Physical queues	Stages of care Stages of patient information flow	Variability in task times Patient characteristics: - New or returning - Administrative markers	Patient wait time Physician overtime: - AM and PM Scenario analysis	Some suggested changes have been implemented
Reducing patient wait times and improving resource utilization at British Columbia Cancer Agency's ambulatory care unit through simulation	Santibáñez et al. (2009) [54]	Community care - ambulatory care unit - Physical queues	Stages of care process	Shared resources Appointment type Capacity constraints Scheduling policy	Patient wait time Appointment duration Resource/clinician utilisation Time in system Scenario analysis	Suggestions made to senior management

Table 2.4 (Continued): Papers included from "Patient flow within community care" search only

Title	Authors	Setting	States	Factors considered to influence flow	Output measures	Implementation of results
Simulation Analysis						
Facilitating stroke care planning through simulation modelling	Bayer et al. (2010) [57]	Stroke services - Physical and non-physical queues	Stages of a stoke pathway - Acute - Community	Patient groups: - Health related Probabilistic: - Death rate, length of stay Capacity constraints	Predicted bed days - Acute - Care home Cost of providing resource Scenario analysis	Not explicitly stated
Using discrete event simulation to compare the performance of family health unit and primary health care centre organizational models in Portugal	Fialho et al. (2011) [55]	Primary healthcare - Non-physical queues	Stages of care	Administrative characteristics Consultation type Opening hours Appointment duration Routes of care	Days to arrange a GP consultation Annual number of different consultations Waiting time Financial costs	Not explicitly stated
Modeling the demand for long-term care services under uncertain information	Cardoso et al. (2012) [45]	Long-term care - Uncapacitated	Different aspects of LTC - Home based - Ambulatory - Institutional	Patient groups: - Demographics - Chronic disease - Level of dependency Mortality rates Capacity	Future demand Resources required to meet demand for each aspect of LTC Cost Scenario analysis	Not explicitly stated
A Simulation Optimization Approach to Long-Term Care Capacity Planning	Zhang et al. (2012) [46]	Long-term care - Uncapacitated	Waiting In service	Patient characteristics: - Age and gender - Arrival rate - LOS Initial conditions	Optimised capacity relating to waiting time targets Future demand Scenario analysis	Collaboration, training and feedback highlighted

Table 2.4 (Continued): Papers included from "Patient flow within community care" search only

Title	Authors	Setting	States	Factors considered to influence flow	Output measures	Implementation of results
Simulation Analysis						
Applying discrete event simulation (DES) in healthcare: the case for outpatient facility capacity planning	Ponis et al. (2013) [52]	Outpatient clinics - Non-physical queues	Different services	Patient characteristics: - Administrative; medical Budget constraints Capacity constraints Appointment types Abandonment Distance from clinic	Resource utilisation Cost of care Optimised service provision	Not explicitly stated
Developing an adaptive policy for long-term care capacity planning	Zhang and Puterman (2013) [47]	Long-term care - Uncapacitated demand	Waiting In service	Patient characteristics: - Age and gender - Arrival rate - LOS Initial conditions Achievement of wait time targets in previous year	Adaptive policy for capacity planning Optimised capacity relating to waiting time targets Future demand Scenario analysis	Not explicitly stated
Patient flow improvement for an ophthalmic specialist outpatient clinic with aid of discrete event simulation and design of experiment	Pan et al. (2015) [53]	Specialist outpatient clinic - Physical queues	Stages of care Stages of information flow Waiting	Patient characteristics: - Services required - Punctuality/no show Layout of clinic Resource capacity: - Staffing levels - Shared resource Inter-relation of patient flow and info flow Batched info arrival	Turnaround time Waiting time Allocation of appointment slots Scenario analysis	Implementation of results

Table 2.4 (Continued): Papers included from "Patient flow within community care" search only

Title	Authors	Setting	States	Factors considered to influence flow	Output measures	Implementation of results
Simulation Analysis						
Simulation analysis on patient visit efficiency of a typical VA primary care clinic with complex characteristics	Shi et al. (2014) [56]	Primary healthcare clinic - Physical queues	Stages of care	Patient groups: - Arrival type - Care requirements No shows Number of double booked appointments	Service utilisation Wait time Factor study	Suggestions made to management
A simulation model for capacity planning in community care	Patrick et al. (2015) [58]	Acute care Long-term care - Physical queues	Different services	Patient groups: - Care requirements - Priority - Preference Capacity Reneging	Required capacity: - Wait time/list size - Percentage of patients in preferred facility Scenario analysis	Not explicitly stated
A simulation optimisation on the hierarchical health care delivery system patient flow based on multi-fidelity models	Qiu et al. (2016) [59]	Community care General hospitals - Physical queues	Community services General hospitals Stages of care	Patient groups: - Care requirements Profit Priority Inter-hospital flow	Queueing network: Optimised resources to achieve maximum profit Simulation: Evaluation of feasible solutions regarding: - Profit, service use, cured patients	Not explicitly stated

Table 2.4 (Continued): Papers included from "Patient flow within community care" search only

Title	Authors	Setting	States	Factors considered to influence flow	Output measures	Implementation of results
Markovian models						
An analytical framework for designing community-based care for chronic diseases	Kucukyazici et al. (2011) [28]	Community care - post acute services - Non-physical queues	Different services Post care outcomes	Demographics of inter service flow	Likely post care outcomes for pathways Scenario analysis	Not explicitly stated
The long-term effect of community-based health management on the elderly with type 2 diabetes by the Markov modelling	Chao et al. (2014) [30]	Community services for diabetes	Health states	Treatment pathway Results of randomized controlled trial Variable severity of disease	Probability of a patients belonging to a given outcome state as time progresses	Not explicitly stated
Analytical methods featuring time dependence						
Intelligent patient management and resource planning for complex, heterogeneous, and stochastic healthcare systems	Garg et al. (2012) [42]	Integrated care system i.e. acute, social, and community - Non-physical queues	Post hospital services	Patient groups: - Demographics - Care requirements - Length of stay	Forecast number of patients in post care outcome Forecast daily/total cost	Not explicitly stated
Improving health outcomes through better capacity allocation in a community-based chronic care model	Deo et al. (2013) [43]	Community care - for asthmatic patients - Non-physical queues	In service - appointment Waiting Health states	Variable health Time between appointment Service capacity Health benefit of treatment	Optimised appointment allocation subject to health benefit and capacity	Not explicitly stated
Simulation Analysis						
Evaluating multiple performance measures across several dimensions at a multi-facility outpatient center	Matta & Patterson (2007) [50]	Outpatient services - Physical queues	Different services	Day of week Patient groups: - Care requirements Patient pathway/ throughput Clinician overtime	Single parameter for analysing multiple, stratified performance measures Scenario analysis	Some suggested changes have been implemented

Table 2.5: Papers included from both "Patient flow within community care" search and "Patient flow and outcomes" search

Title	Authors	Setting	States	Factors considered to influence flow	Output measures	Implementation of results
Markovian models						
Modeling the transplant waiting list: A queueing model with reneging	Zenios (1999) [60]	Waiting list - transplant - Non-physical queues	Waiting list Obtained transplant	Patient groups: - Demographic - Transplant type Organ groups Reneging - death	Wait time in system and until transplant - per group Fraction of patients who receive transplant per group	Not explicitly stated
Optimizing admissions to an intensive care unit	Shmueli et al. (2003) [63]	Intensive Care Unit - Physical queues	ICU beds Waiting for service In service	Variable health: - survival probability Capacity - beds Loss model	Expected number of statistical lives saved through outcome based admission policy	Not explicitly stated
Modeling and analysis of high risk patient	Wang (2004) queues [61]	Waiting list - transplant - Non-physical queues	Waiting list Obtained transplant	Patient priority: - Health related Risk of death List size	Queue lengths and wait time - per group Expected number of deaths	Not explicitly stated
Differentiated waiting time management according to patient class in an emergency care center using an open Jackson network integrated with pooling and prioritizing	Kim and Kim (2015) [64]	Emergency care centre - Physical queues	Waiting for service In service	Patient groups: - Acuity level Admission policy Patient group pooling Infinite waiting space	Waiting time - FCFS - Hybrid (FCFS and priority) - Hybrid with pooled groups	Not explicitly stated
A model for deceased-donor transplant queue waiting times	Drekic et al. (2015) [62]	Waiting list - transplant - Non-physical queues	Waiting list Obtained transplant Patient priority - Health related	Variable health Prioritisation Reneging List size and blocking	Queue length Wait time Reneging probabilities - per group	Not explicitly stated

Table 2.6: Papers included from "Patient flow and outcomes" search only

Title	Authors	Setting	States	Factors considered to influence flow	Output measures	Implementation of results
Non-Markovian Steady State Analysis						
Efficiency and welfare implications of managed public sector hospital waiting lists	Goddard & Tavakoli (2008) [65]	Waiting list - hospital care - Non-physical queues	Number of people on the waiting list	Service capacity Rationing system Proportion of sick patients admitted	Wait time - All patients - For least ill patients	Not explicitly stated
A multi-class queuing network analysis methodology for improving hospital emergency department performance	Cochran & Roche (2009) [11]	Emergency department - Physical queues	Stages of care	Patient group: - Care requirements Seasonality Number of beds	Queue length/ wait time Service utilisation Requirements for a desired level of utilisation	Software made available to EDs Feedback to clinicians and ED managers
A queueing model to address wait time inconsistency in solid-organ transplantation	Stanford et al. (2014) [66]	Waiting list - transplant - Non-physical queues	Waiting list Obtained transplant	Patient groups: - Care requirements Organ groups Compatibility	Wait time per patient type	Not explicitly stated
System Dynamics Analysis						
Modeling chronic disease patient flows diverted from emergency departments to patient-centered medical homes	Diaz et al. (2015) [67]	Care for chronic disease	Stages of care - Emergency departments - Ambulatory services	Patient groups: - Insured and uninsured - Care requirements Resource capacity Death Congestion	Impact on demand for services and required capacity Resource utilisation Cost Health impact Scenario analysis	Not explicitly stated
Analytical methods with time dependence						
Dynamic allocation of kidneys to candidates on the transplant waiting list	Zenios et al. (2000) [73]	Waiting list - transplant - Non-physical queues	Transplant queue Obtained transplant	Variable health Patient demographic Organ groups Availability of organ Transplant failure/re-join Quality of life measure	Wait time and time until transplant - per group Fraction of patients who receive transplant per group	Not explicitly stated

Table 2.6 (Continued): Papers included from "Patient flow and outcomes" search only

Title	Authors	Setting	States	Factors considered to influence flow	Output measures	Implementation of results
Analytical methods with time dependence						
The optimal timing of living-donor liver transplantation	Alagoz et al. (2004) [74]	Waiting list - transplant - Non-physical queues	Waiting list Obtained transplant Health states - Waiting in time period	Variable health Organ quality Post-transplant survival rate	Optimal timing of transplant	Not explicitly stated
A model for managing patient booking in a radiotherapy department with differentiated waiting times	Thomsen & Nørrevang (2009) [71]	Radiotherapy - Non-physical queues	Radiotherapy slots	Patient groups: - Care requirements - Waiting time guarantee Capacity	Lower and upper limits for slot allocation per group	Suggested use within department
Investigating hospital heterogeneity with a multi-state frailty model: application to nosocomial pneumonia disease in intensive care units	Liquet et al. (2012) [69]	Intensive Care Unit	Admission Infection Death Discharge	Patient groups: - Frailty - Type of admission - Infection	Number of patients with infection - Death - Discharge	Not explicitly stated
Optimizing intensive care unit discharge decisions with patient readmissions	Chan et al. (2012) [70]	Intensive Care Unit - Non-physical queues	ICU beds Number of people in the system	Variable health Demand driven discharge - Cost such as loss in QUALY Congestion	Optimisation of cost incurred by demand dependent discharge Readmission and mortality rates	Not explicitly stated
Planning for HIV screening, testing, and care at the veterans health administration	Deo et al. (2015) [68]	Community care - for HIV patients - Non-physical queues	Stages of care Health states	Variable health Allocation of screening Budgetary constraints Service constraints	Optimal screening policy with regards to health benefit, budget, capacity, and staffing levels	Several suggestions influenced decision making
Radiation Queue: meeting patient waiting time targets	Li et al. (2015) [72]	Radiotherapy - Non-physical queues	Types of treatment slot for radiotherapy machines	Patient groups: - Care requirements - Service times Capacity Patient pooling	Required capacity to meet wait time targets Optimal allocation of capacity for different patient groups	Not explicitly stated

Table 2.6 (Continued): Papers included from "Patient flow and outcomes" search only

Title	Authors	Setting	States	Factors considered to influence flow	Output measures	Implementation of results
Simulation Analysis						
Simulating hospital emergency departments queuing systems: (GI/G/m(t)):(IHFF/N/∞)	Panayiotopoulos & Vassilacopoulos (1984) [79]	Emergency department - Physical queues	Waiting list In service	Variable clinician capacity Waiting capacity Variable patient priority: - Health related	Average number of patients and average time - in system and queue	Some suggested changes have been implemented
Development of a Central Matching System for the Allocation of Cadaveric Kidneys: A simulation of Clinical Effectiveness versus Equity	Yuan et al. (1994) [76]	Transplant waiting list - Non-physical queues	Waiting list Received transplant	Patient groups Organ groups Compatibility Availability of organs Time spent waiting	Assessment of different allocation algorithms - Time until transplant/ waiting time Unused organs	Not explicitly stated
Patient flows and optimal healthcare resource allocation at the macro-level: a dynamic linear programming approach	van Zon & Kommer (1999) [81]	General method for resource allocation	Stages of care Health states	Variable health Duration of medical activity Patient pathway	Optimisation of resources: patient health/wait time Scenario analysis	Not explicitly stated
A simulation model to investigate the impact of cardiovascular risk in renal transplantation	McLean & Jardine (2005) [77]	Waiting list - transplant - Non-physical queues	Waiting list Obtained transplant	Transplant failure Patient characteristics: - Demographics - Health risk/mortality	Post-transplant survival rate Scenario analysis	Not explicitly stated
A clinically based discrete-event simulation of end-stage liver disease and the organ allocation	Shechter et al. (2005) [78]	Waiting list - transplant - Non-physical queues	Waiting list Obtained transplant	Patient characteristics: - Demographics/care requirements/health	Post-transplant survival rate after - 1 year and 3 years	Not explicitly stated
Capacity planning for cardiac catheterization: a case study	Gupta et al. (2007) [75]	Cardiac catheterization clinic - Physical queues	Stages of care	Patient group: - Care requirements Clinician case load	Optimised capacity allocation subject to desired wait times Scenario analysis	Some suggested changes have been implemented
A discrete event simulation tool to support and predict hospital and clinic staffing	DeRienzo et al. (2016) [80]	Neonatal Intensive Care Unit - Physical queues	ICU beds	Patient groups: - Admission type/health Resource capacity	Estimated staffing allocation and cost Forecast of demand	Not explicitly stated

Table 2.6 (Continued): Papers included from "Patient flow and outcomes" search only

2.5 Summary and discussion of findings

I will now discuss findings from across the literature, drawing together common themes and key characteristics as presented in Tables 2.4, 2.5 and 2.6. Overall, I reviewed 58 papers presenting models of patient flow. 35 applied to community care services, which included mental health services, physical health services, long-term care, outpatient care, and patient flow between acute and community settings. 34 applications used, in some form, queue lengths or the amount of time that a patient spent within states as output measures. The second most common metrics were monetary costs and the allocation of capacity related resources.

Within the "Patient flow and community care" literature a range of flow characteristics were considered. For instance, patient access and arrivals to community services were modelled as unscheduled [34], by appointment [43, 68], by external referral [27], or a mixture of the above [29, 51]. Furthermore, multiple care interactions were modelled as either sequential visits to different services [27, 29] or as single visits where multiple tasks were carried out [51]. In either instance, patients were sometimes able to recurrently visit the same service over time with some patients using the service more frequently [43, 56].

As per Tables 2.5 and 2.6, within the "Patient flow and outcome" literature there were ten models of transplant/waiting lists; eight of community, ambulatory and outpatient services; three of emergency departments; four for intensive care; two for radiotherapy and one general model of resource allocation. Outcome measures were incorporated within the outputs of these models in three broad ways: 1) system metrics were stratified by outcome related groups; 2) variable patient or population level health was used as an objective or constraint within a model to influence resource allocation; or 3) health outcomes were used as system metrics themselves - such as patient mortality or future use of care. Notably, 15 papers stratified patients into groups based on differing health/outcomes in which they remained; whilst 13

papers incorporated health/outcomes that could change during a course of care. By modelling changes in patient health/outcomes, a model's output was informed by the clinical effect of a care interaction, or absence of a care interaction, on patients and on the operation of the system (e.g. [43, 68]).

Patient groups relating to health/outcomes were used in models of each method, and were commonly used in resource and service capacity allocations (e.g. [43, 45, 50, 70]). Notably, their application within steady state methods may be limited since it is difficult to model health/outcome dependent variables, such as service times, because the order of patients within the queue is unknown in these methods. Furthermore, health/outcomes that could change during a course of care were commonly used within time dependent methods (e.g. analytically [68, 74], simulation [50]). They were often used to model the effect of care on a population where the modelled time period was large or where multiple interactions were considered (e.g. [43, 72]).

Across both literatures queues could be categorised as either physical or non-physical. Physical queues form when patients wait for service within a fixed physical space such as a clinic or emergency department [11, 51, 54, 56], or when moving between residential care and waiting within a service [27, 38, 39]. Thus, these queues may be constrained by a fixed physical capacity, with demand modelled from the point when patients physically arrive at a service.

The most common analysis of physical queues in this review related to the operation of single type of service (such as long-term care services), to gain insight into the delivery of care (such as flow between multiple treatments/consultations in a single visit e.g. [11]). Notably, studies of physical queues were carried out using each type of method, with the choice of method dependent on the desired insight, the factors considered to influence flow and the size of the system. Steady state methods were often sufficient if queue lengths and wait times were of primary concern over long periods of time (comparative to service length). However, if variability in input

parameters or periodic influences were important, time variable methods were more common.

Alternatively, non-physical queues occur when patients may wait in any location away from the service such as their own place of residence [52, 55]. An example of when these queues are modelled includes when a patient's wait is potentially long and unknown for multiple care interactions [43]. Non-physical queues may represent unconstrained demand since there is no physical limit to waiting space; however, there may be set sizes such as capped waiting lists [61].

The most common analysis of non-physical queues related to waiting lists and multiple uses of a single service or several services [43, 73]. When modelling demand and access at a system level, steady state analysis or time dependent methods were typically used (e.g. steady state [25, 26], time dependent [10]). In scenarios of resource allocation, time variable methods were increasingly used [41]. Within these models, variable health/outcomes were widely considered over longer time frames of care and when multiple interactions were possible.

As a final observation, the reporting of implementation and collaboration varied greatly within each group of analytical method.

2.5.1 Limitations

It should be noted that the work presented in this chapter is limited due to the difficulty of systematically reviewing this literature. In particular, I found two main difficulties. Firstly, papers were published within a wide range of journals, some within health care journals, others in operational research (OR) journals; whilst a proportion were found within journals that were neither health specific nor OR specific. Secondly, I found that within the literature, patient flow was described and referred to in many ways. No clear standards were found; thus, locating these papers was particularly difficult.

Due to these complexities, I cannot claim that these findings are exhaustive. However, by following an iterative process of literature searching the findings are representative of the research landscape, allowing for meaningful conclusions to be drawn in the next section.

2.6 Conclusions and directions for work

The factors that are considered to influence patient flow within community health care are often markedly different to acute services, and can vary from one service to another. Considering the characteristics discussed in this review, it is common for a mixture of complex dynamics to be modelled within community care applications. Thus, modelling these services can become complicated, requiring innovative methods to include all or some of these dynamics - as highlighted by the breadth of methods presented in this review.

I now draw out some possible directions for future patient flow modelling in community care. These conclusions are formed in light of known challenges for community care, gaps found within the literature and any transferable knowledge between the two sets of literature.

Few models considered patient flow within systems of differing community services with many studies focussing on single services/single types of service. Likewise, few considered the mix of patients. A significant challenge in managing community health care is how to co-ordinate and deliver care within physically distributed services, that are used by a mix of patients (with differing frequency and care needs), who may use a range of services [3]. With a shift of focus in the NHS towards care for the increasing number of patients with multiple long-term illnesses [4], the patient mix within each service further exacerbates this challenge. Given that it is often difficult to measure the impact that changes within one part of the system have on

the whole system [27], it would be beneficial to develop methods for modelling the flow of heterogeneous patients through multiple services.

Another useful direction is to develop time dependent analytical methods. Whilst often analytically difficult, there are important benefits in using these methods as shown by the wide range of applications within this review - such as faster speed of calculation compared to simulation methods. Given the characteristics of community services previously discussed, methods for modelling time varying capacity, demand and timing of patient (e.g. seasonal spikes) would be a helpful addition. As would methods for modelling systems where steady state assumptions do not hold (e.g. systems that become heavily loaded during seasonal changes in demand). The development of these methods would be beneficial in analysing the time variable impact of changes in the immediate, short-term and long-term for the whole system.

Finally, 13 papers used variable health/outcomes, of which five applied to multiple care interactions. Again, considering the purpose and nature of community care, a useful direction for future study would be towards methods that use measures of health that may change throughout a care process. In particular, those that allow for the improvement and decline of patient health throughout several care interactions. A good example of these methods is presented by [43, 68]. Having otherwise not been widely explored, methods that quantify and evaluate the quality of care and include an interaction between patient outcomes, care pathways and flow within the system would be valuable and appropriate for community care modelling. I note here that health states need not only represent the severity of illness, rather they may also represent a patient's "capacity to benefit" from care - a concept that I develop further throughout this thesis. This incorporates notions of the extent to which a patient's outcomes may possibly be impacted by the receipt of care, the likelihood of a change in their health as a result of service (or lack of it), and the potential impact on their future need/use of services. With the goal of maintaining and improving patient

health whilst preventing any negative impact, modelling the capacity to benefit of patients provides a mechanism for analysing the trade off in/effect of providing care to patients with different needs.

In considering OR methods for community services that combine patient flow modelling and patient outcomes, there may be some transferable knowledge from transplant models and radiography models. These models may provide a useful basis for modelling non-physical queues because they share some distinct similarities to community care services - such as time varying demand, limited resources and in some cases re-entrant patients [73]. Furthermore, they may be informative for both scheduled care [72] and unscheduled care [60].

The identified gaps and the above literature motivated the work contained in the remainder of this thesis. In particular, the theoretical framework developed in chapter 5 can be used to model patient flow through several services and multiple service interactions. The method is also time dependent and health dependent, with a patient's health able to improve, decline or stay the same throughout a course of care, further meeting a gap in the literature for analysing networks of care.

Whilst constructed for a Markovian service network, the method differs from the reviewed literature since it is a deterministic fluid and diffusion approximation of a stochastic process. The method is unseen within the review and has not previously been applied to community health care. Only system dynamic approaches or the fluid method in [73] are similar since they represent the modelled services as a system of coupled ODEs. However, the above give stylistic representations of the system where as the method in chapter 5 is intended to give an approximation of a stochastic service system to provide efficient and accurate analysis. Likewise the variance of the process is considered in chapter 5, providing a further difference.

Finally, the focus towards evaluating systems from the novel perspective, denoted the "flow of outcomes", has not been widely seen in the literature. Thus, the methods

in chapter 5 help to emphasise the potential of such a view in evaluating systems and introduce both a language and framework to better understand, communicate and apply this perspective.

Chapter 3

Understanding referral data through data visualisation and analysis

In this chapter I present work that I carried out to inform the development of patient flow models for community health services. I present my key findings about the dynamics of referrals within community services, gained through collaborative work and through the analysis of patient level referral data.

To understand this data, I applied several visualisation methods to NELFT referral data, each focussing on a different characteristic of community referrals. At the time of carrying out this analysis, managers within Havering community health services were beginning to design a single point of access (SPA) for managing referrals within their services. This presented an opportunity to make a timely contribution by using data visualisations to inform their thought process in designing this service through the possible identification of important referral paths, patterns of patient use and key services. The main benefit of these visualisations to NELFT was the opportunity to explore and discuss the nature of referrals within their community services as informed by the data.

In this chapter I:

1. Identify key characteristics of how patients use community services relevant to patient flow modelling;
2. Present data visualisation methods for understanding complex referral data in community health care;
3. Discuss how these data visualisations may inform the planning of services;
4. Inform the development of patient flow models for community health care.

3.1 Introduction

In developing methods for modelling patient flow, it is important to understand how patients interact with health care services and the key issues in providing care. As identified in chapter 2, there is a range of referral and flow dynamics that may be considered when modelling patient flow within community health care. Thus, I worked collaboratively with care leads from the North East London Foundation Trust (NELFT) to learn about the key dynamics seen in their services.

Focusing on community services for patients aged 65 and over in Havering (a London borough), I sought to understand how patients used these services and whether there were any common characteristics of patient use. In beginning this work, I found that NELFT possessed a wealth of data, yet did not have the capacity or resources to analyse it in order to learn more about their community referrals and inform service planning. As a result, I sought to provide helpful insight by producing accessible methods for analysing their data; in particular, informative methods for visualising patient referral data.

Methods for visualising referral data are important because they can help both researchers and care managers to ask and answer questions that might otherwise be unclear, or difficult to interpret, from the raw data. They can help to communicate complex information in a "digestible" and understandable way. This makes them easier to share with collaborators, creating an engaging format for delivering information that promotes questions and discussion.

At the time of carrying out my analysis, care managers within Havering community services were beginning to design a single point of access (SPA). The SPA is a service that manages referrals between community services, seeking to streamline the process and reduce inappropriate referrals. The plans to implement this service presented an opportunity to make a timely contribution. In applying data visualisation methods to their referral data, I helped to inform their thought process in designing this service. This will be discussed later in the chapter.

Structure of chapter

In the following section, I briefly explore the literature on visualising electronic patient records and referral data across health care settings. In section 3.3, I detail the initial thought process gained from scoping conversations with Havering community care leads. I also discuss the data obtained for this work and the process taken to clean it. This is followed by a descriptive analysis of the data and a discussion of the methods in section 3.4.

In section 3.5, I detail the visualisations and present an application of them to NELFT community health care data. Firstly, I present a network map, using filters and simple network statistics to help identify patterns of patient use and significant groupings of services. Secondly, I conduct an analysis of referral pathways looking at the concurrent use of services by patients and chains of referral from one service to another. Thirdly, I produce an aggregated patient pathway plot, looking at how

the subsequent care of patients, who used a given service, develops over time.

This chapter culminates in a discussion of these visualisations, the insight gained in designing a SPA and the limitations of this work. In my concluding remarks I note how this work informs the development of the patient flow model in chapter 5.

3.2 Research landscape and original contribution

For an introduction and overview of visualisation methods and their use for understanding data, see Graphical Perception and Graphical Methods for Analyzing Scientific Data [82]. In this article, the authors briefly identify and summarise some of the key principles behind data visualisation and comment on some of the key features of graphical representations that may lead to more effective graphical perception.

Within health care, several studies have used methods to visualise electronic health records using a range of methods and to insight into different settings. These studies range from visualisations of a single patient's data [83, 84] to larger sets of multiple patients [85]. Overall, the applications of simple visualisation methods highlight how presenting complex data in a straightforward and digestible manner can provide valuable insight. Additionally, the more complex methods may help to identify key patterns within the data [85]. Across the range of methods, it is common and helpful to use different colours and sizes of both text and objects; visual filters; and interactive methods to explain differences in data [86].

A common visualisation method is to represent data using a network. In network representations, the pairwise relationships between sets of entities are described visually by a collection of shapes, usually circles, (nodes) and lines connecting them (edges), see Figure 3.1. Nodes are connected by an edge if they relate to each other,

e.g. if nodes represent services, an edge connecting them may represent a referral between them. This structure is both visually informative and creates a means for mathematical analysis.

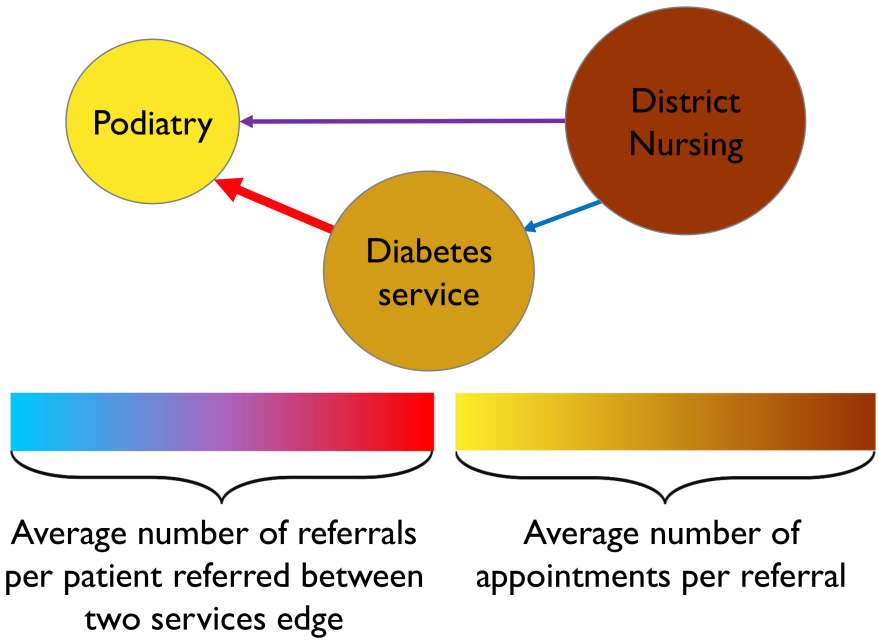

Figure 3.1: Diagram of a simple network representing referrals between services

Network representations may be used to understand a variety of health care processes, such as the clinical pathways of patients [87] or how patient records are used by care providers [88]. By producing a network, characteristics of the nodes and edges may be represented by different sizes or colours. For example, the size of a node may represent the number of patients in a specific service [87], whilst the width of an edge may represent the number of patient referred between two services. In addition, node and edge colouring may be used to highlight key information about the nodes or how they relate to each other, as shown in Figure 3.1.

Contributions

In this chapter I use several visualisation methods to explore different characteristics of community referrals, analyse the data and aide the communication of the results. The aims of this work were to understand this large and complex system

of community services, identify how patients used multiple services and to explore whether common patterns of referrals existed. Furthermore, these maps help to understand how these services could be modelled, and help NELFT by providing insight into how patients used their services, aiding discussion as to how these services may be organised. In particular, these methods helped to inform these conversations by highlighting and communicating key characteristics of patient referrals and uses of service as seen within the data.

This work differs from the existing literature since it considers community health care and the referral dynamics of this sector, in particular the reuse of services, concurrent uses of different services and potential patterns of subsequent referrals. Furthermore, the combination of methods alongside the network representation and the dynamics which they are used to analyse has not been seen within the literature before. Whilst applied to a single provider (NELFT), the methods are generalisable and easy to use in other boroughs, trusts and organisations. These maps are visually impactful, informative and simple to create, increasing their scope for use in practice.

Regarding the contribution to the work in this thesis, the dynamics identified through this analysis are considered when developing the theoretical model in chapter 5. In particular, the potential for patients to reuse services and the potential for the sequential use of multiple services.

3.3 Initial steps

3.3.1 Understanding patient referrals - learning from care leads

This visualisation work was carried out in collaboration with NELFT clinical staff to understand better the structure of NELFT community services and how

patients were referred between them. A particular example of this was a meeting held with leads from physical health services and mental health services during which they participated in a mapping exercise to identify: the types of community services delivered by NELFT; how patients were referred into and between them; whether any common pathways existed (theoretically and in practice); whether there were any significant characteristics that should consider from the outset; and, whether any previous attempts had been made to visualise this system.

This involved two tasks. The first was to categorise services into mental or physical health services, and into home based, clinic based or inpatient based services. The sources of referral were also identified. An example of this task is given in Figure 3.2. The second task was to organise services into possible referral pathways in order to identify theoretical referral pathways and to learn about what they expected to occur in practice and why. This informed the methods I could use to visualise their data. An example of this task is presented in Figure 3.3.

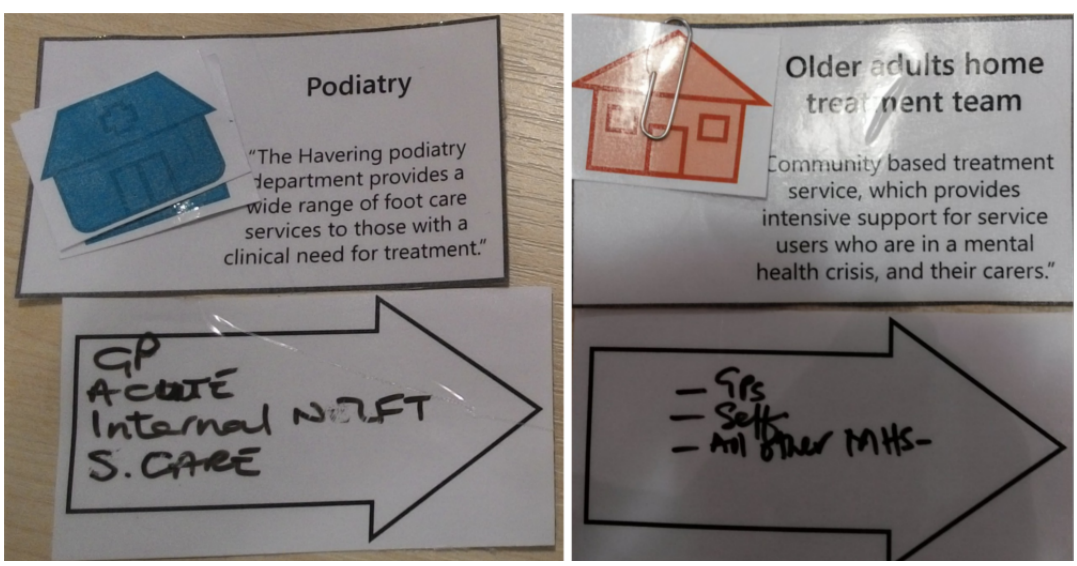

Figure 3.2: Example of service categorisation task

Blue symbols indicate physical health services, whilst orange indicate mental health. The symbol paired with Podiatry shows that it is clinic based, whilst the symbol paired with the Older Adults Home Treatment Team represents a home based service. The arrow cards were used to note the routes of referral into services.

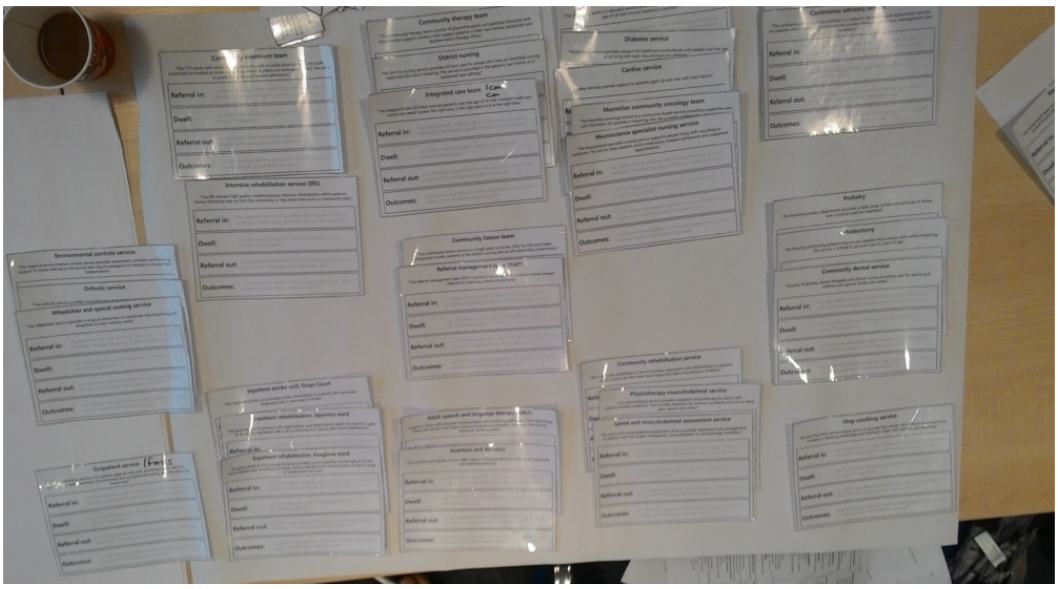

Figure 3.3: Example of referral mapping task for physical health services

Care leads noted that no defined pathway existed; thus, all services could refer to each other. However, they suggested groups of services based on patient use.

From these meetings, three characteristics were identified as important. Firstly, it was theoretically possible for all physical health services to refer to each other, creating a complete network of more than thirty services. Secondly, in practice, this network would be highly connected since patients could use any combination of services, with more defined pathways seen amongst mental health services both theoretically and practically. Thirdly, it was indicated that few routes for referral existed between mental health services and physical health services. A further key detail was the possibility that patients could reuse certain community services multiple times. This was a feature of both mental and physical health.

Furthermore, the CCG had previously attempted to visualise this system, see Figure 3.4. The map begins with access points in primary, community and acute care, and presents possible pathways through community services for patients experiencing medical "crisis" - although the term crisis was not defined. The map includes details of opening times and possible routes for discharge; however, it lacks notions of service capacity and patient activity, and represented theoretical pathways.

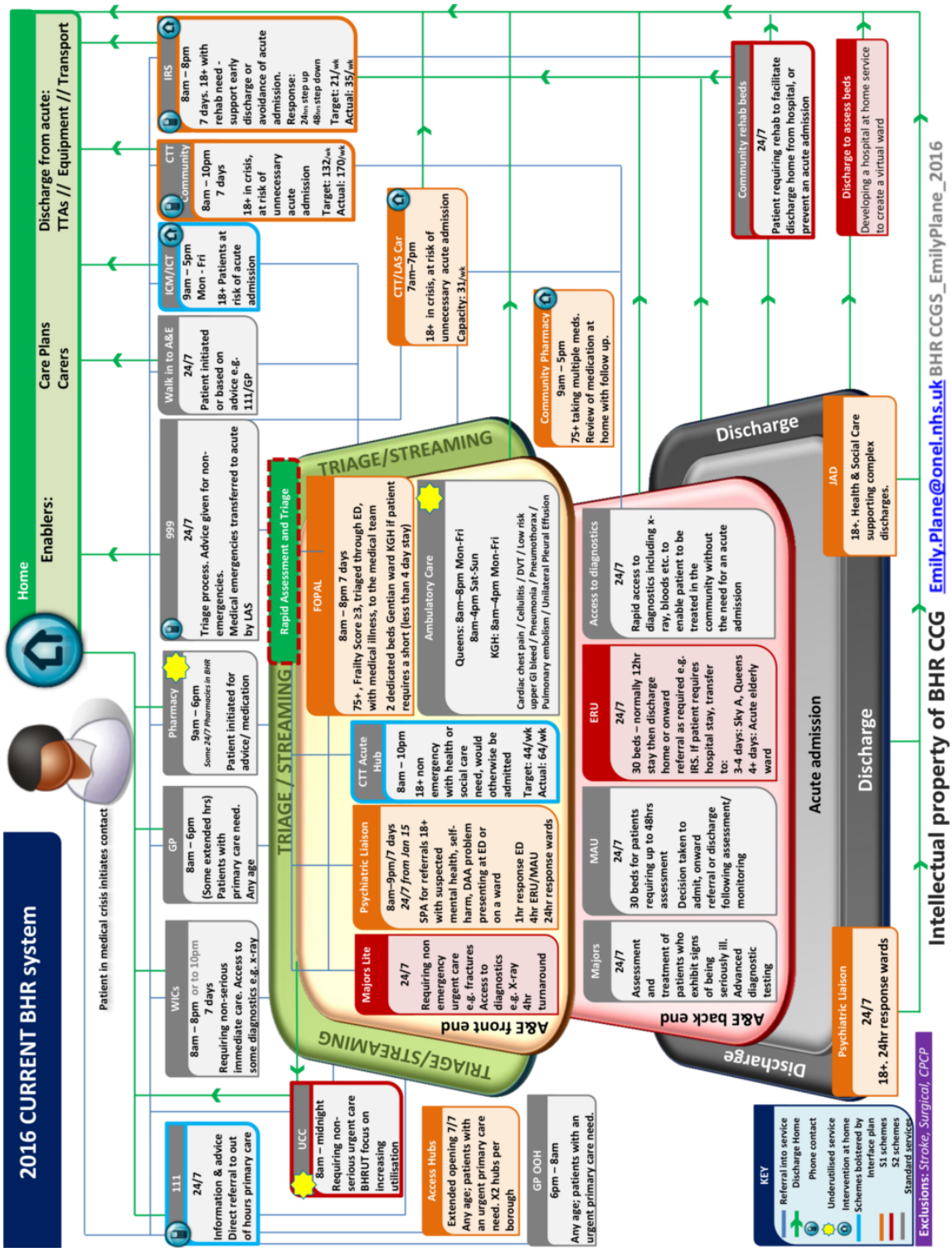

Figure 3.4: Map produced by CCG of possible referral routes through community physical health services

Since there are few referral paths between physical and mental health services I decided to only focus on physical health services. Reasons for this include the difficulties in accessing data for both sets of services because patient data were recorded on

different systems. Furthermore, by working with physical health services there was an opportunity to provide greater insight due to the lack of definable pathways, the possibility of patients reusing services and the potentially highly connected nature of these services.

To produce the visualisations, I obtained the relevant data after discussions with NELFT's performance team and data managers. The scope and structure of this data is discussed in the next section, presenting some initial observations and analysis.

3.3.2 Routine patient data - content and cleaning

I obtained a non-identifiable routine dataset of patient level data, extracted from RiO (an electronic patient record system), which was stored securely within UCL's Data Safe Haven [89]. This is a technical solution for storing, handling and analysing sensitive data, certified to the ISO27001 information security standard and conforming to the NHS Information Governance Toolkit.

Working within this Safe Haven introduced limitations in processing and using the data. For example, the mapping software was installed locally on a laptop and was not available within the Safe Haven. Furthermore, in using the Safe Haven, the data had to be first cleaned, aggregated and processed within the secured setting and then extracted for use. Other limitations are discussed later.

The data consisted of all referrals for patients aged 65 and over to NELFT community services in Havering from 1 April 2014 to 31 August 2016. Lasting from the date of referral until discharge or loss, referrals consisted of one or more appointments - with each appointment represented as a row in the data. Thus, a patient's community care history could span multiple rows, representing the combination of their referrals to community services. See Table 3.1 for key variables within the data.

Prior to its use the data were cleaned. Due to the size of the dataset size (over 1,200,000 rows and over 30 columns), I used the statistical software Stata. There

Variable name	**Description**	**Level**
Client ID	Non-identifiable patient reference. 20293 unique IDs	Patient
Ref ID	ID for referral. 65306 unique IDs	Referral
Referral Datetime	Date and time of referral	Referral
Source	Description of where patients were referred from. 61 sources (Processed)	Referral
Specialty Description	Description of medical discipline patients were referred to. 34 specialties (Processed)	Referral
Discharge Datetime	Date and time of discharge	Referral
Length of Stay	Length of referral - days	Referral
Appointment Date	Date and Time of contact	Appointment

Table 3.1: Variables contained in the dataset used for producing visualisations

were two main reasons for cleaning the data. Firstly, dates and times were stored as a concatenated string; thus, these variable needed to be assessed for accuracy and split for use. Secondly, some services were recorded under multiple names, such as the Nutrition and Dietetics Service which was also recorded as: *Adult Nutrition & Dietetic Service, Nutrition & Dietetic Service, Acute - Nutrition and Dietetic Service.* Further details of the cleaning process see Appendix A.1.

3.4 Methods

3.4.1 Analysis of individual patient pathways

The data were left truncated - i.e. some referrals beginning before 1 April 2014 - and right censored - i.e. not all referrals had concluded by 31 August 2016. For all referrals, only appointments from 1 April 2014 onwards were included in the data. Figure 3.5 illustrates the above, presenting an example of hypothetical referral data for an individual patient.

Due to left truncation, a patient's first referral was unknown; therefore, I refer to their first referral in the data as their *index referral*. To overcome the difficulties presented by incomplete data, only referrals that began on or after 1 April 2014 and finished on or by 31 August 2016 were analysed, the limitations of which is discussed later. Notably however this does not impact the use of the methods since they are generalisable and do not depend on the data's structure.

From the inspection of individual patient plots, there were four clear dynamics of patient use. Firstly, there was a large variation in the number and range of services used by patients. Some used a single service once, whilst others had multiple referrals, shown in Figure 3.6. Notably, patients sometimes used the same service multiple times, a characteristic I explore later.

Secondly, referrals to different services commonly overlapped. Figure 3.7 shows the distribution of the maximum number of services used at the same time per patient.

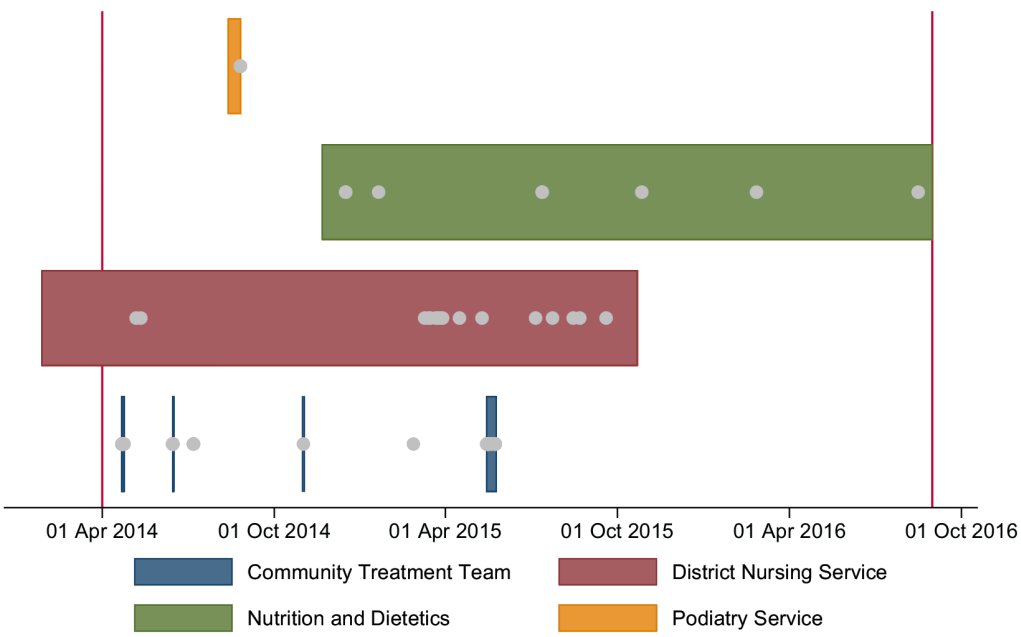

Figure 3.5: Hypothetical patient level referral data for community health care

The grey dots each represent an appointment. The vertical red lines represent the start date, 1 April 2014, and the end date of the data, 31 August 2016.

Thirdly, the number of appointments, timing of subsequent referrals and length of referral varied. Figure 3.8 shows the distribution of referral length. Whilst the majority of patients experienced short length of stays, there is a long tail to this distribution with many lasting over 200 days.

Fourthly, there was variability in the services used by patients and how they used them. Informed by the above observations, I developed methods for exploring whether common patterns of referral existed in the data.

3.4.2 Network map

Motivated by the potential complexity of referral pathways and size of the system, I produced a network map of NELFT referral data using Gephi [90] - an open-source network analysis and visualization software package.

The networks produced in this work are formed of two types of node: specialties and sources (terms used within the data) shown in Tables 3.2 and 3.3 respectively. Specialties represent NELFT's community services, whilst sources are the services

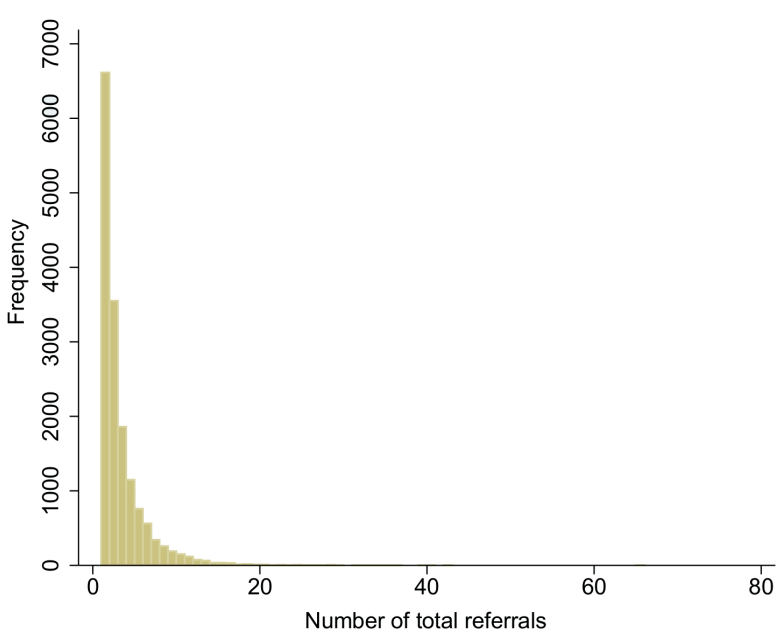

Figure 3.6: A histogram showing the distribution of the total number of referrals per patient in the dataset - maximum of 64

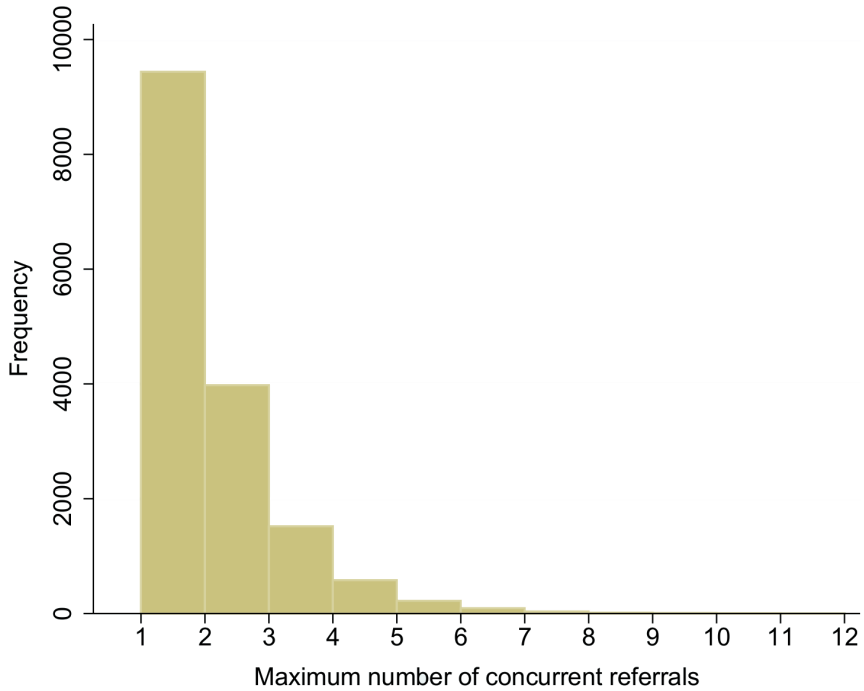

Figure 3.7: A histogram showing the distribution of the maximum number of concurrent referrals, per patient, within the date range of the dataset - maximum of 12

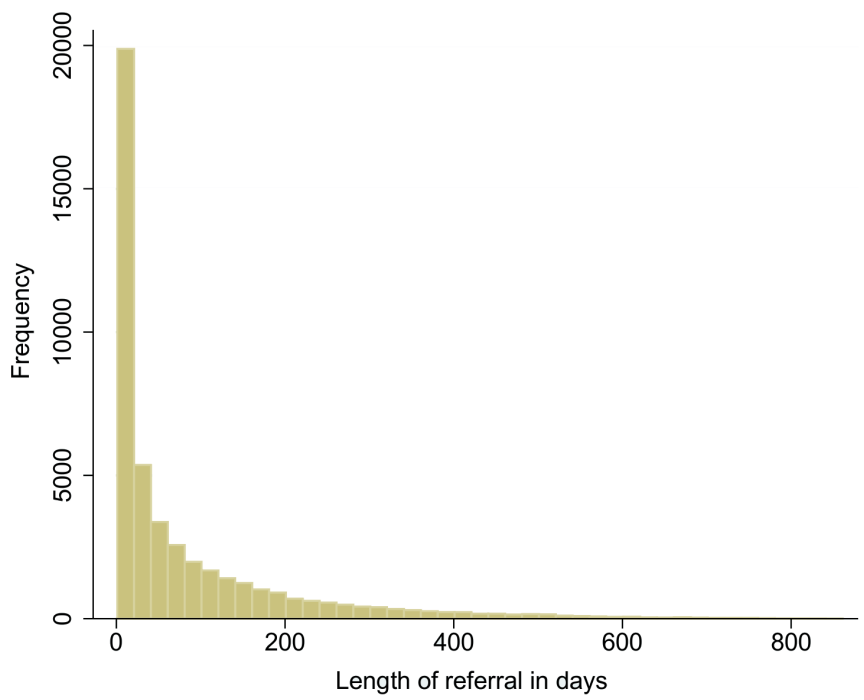

Figure 3.8: A histogram showing the distribution of referral lengths, per referral, in days - maximum $>$ 800 days. Each bar represents a span of 20 days

outside of these that refer to them. Some specialties refer to other specialties and edges are directed from the referring service to the receiving. Specialties may consist

Referral Sources	**Total referrals made**	**Referral Sources**	**Total referrals made**
Acute	12,586	Occupational Health	1
Age Concern	2	Occupational Therapy Service	315
Care Co-ordination	1	Other Community Health provider	1,477
Care Home	1,868	Other Community Professional	108
Carer/Relative	2,434	Other Medical Referral	24
Clinical Assessment Service	1	Other Source of Referral	3,777
Community Inpatient Service	162	Palliative Care Service	8
Community Nursing Service	182	Physiotherapy Service	807
Community Psychiatric Nursing Service	3	Practice Nurse	121
Community Specialist Nurse	318	Private Care Provider	1
Day Centre	1	Private Hospital	1
Day Hospital	91	Psychiatry Service	1
Discharge Liaison Service	91	Rapid Assessment Team	38
GP	12,601	Residential Home Staff	9
Health Visiting Service	10	Self Referral	4,877
Heart Failure Service	25	Sheltered Accommodation Service	3
Hospice	75	Social Services	146
Intermediate Care Service	68	Specialist Neurology Nursing	1
Key Worker	46	Specialist Nursing	18
Mental Health Support Worker	2	Stroke Services	304
Night Nursing Service	15	Transfer In	4
Non-Emergency Health Care Service (111)	27	Walk-in Centre	1

Table 3.2: Sources for community referrals included within the dataset. Community services included in this table represent those that did not feature as a specialty in the dataset.

of several teams that each provide different types of care (for example the District Nursing Service provides both rehabilitative care and palliative care); thus, specialities may refer patients to themselves, representing a referral between teams. Notably, in Table 3.2, only a handful of sources are services within mental health, each initiating a small number of referrals.

Referral Specialty	**Total referrals received**	**Referral Specialty**	**Total referrals received**
Community Beds	331	Neurological Specialist Nursing	63
Community Cardiology Services	380	Nutrition and Dietetics	2,575
Community Matron	1,328	Oncology Specialist Nursing	107
Community Rehabilitation Services	1,409	Orthopaedics	1,031
Community Therapy Service	2,064	Orthotics	544
Community Treatment Team	12,598	Phlebotomy	2,251
Continence Service	375	Podiatry Service	288
Continuing Healthcare	1	Prosthetics Service	10
Diabetes Service	568	Psychological Services	38
District Nursing Service	8,461	Respiratory Service	1,014
Falls Service	976	Specialist Palliative Care	8
Integrated Care Liaison	971	Specialist Seating	4
Intensive Rehabilitation	1,562	Speech and Language Services	912
Intermediate Care Service	102	Tissue Viability Services	942
Leg Ulcer Service	2	Wheelchair Service	414
Musculoskeletal Service	4,179		

Table 3.3: NELFT community services, known as specialties, included within the data

Volume of activity and frequency of reuse is represented by the network in four ways: edge width, node size, edge colour and specialty node colour. Edge width represents the total number of referrals between two nodes - ranging from 1 to 5,810. Edge colour represents the average number of times each unique patient who used the edge was referred from the given edge's source to its specialty. Moreover, source node size represents the total number of referrals initiated, and specialty node size represents the total number received. Source nodes are uniformly coloured white and specialty nodes are coloured according to the average number of appointments per referral. Other informative metrics may be used for colouring such as monetary cost or aggregate patient outcomes.

Notably, this network can only be used to interpret pairwise relationships. Whilst

visually connected, empirically the data may not contain continuous paths between more than two nodes. For example, consider four nodes A, B, C and D connected as in Figure 3.9.

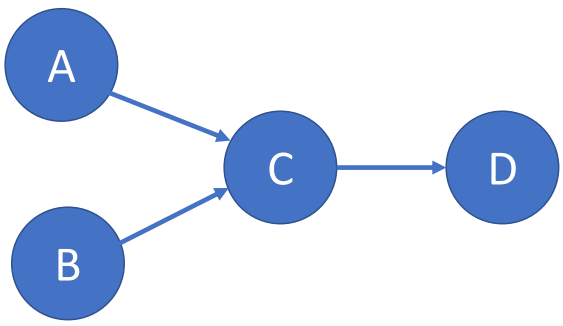

Figure 3.9: Example network diagram

Firstly, whilst the network shows that C received referrals from both A and B, it does not show whether patients referred to D originated from A or B. Rather, it is possible that all those referred from C to D originated from either A only, B only, or neither; we do not know from the network alone due to multiple sources.

The case when this is neither introduces the second limitation, caused by left truncation. It could be the case that all patients referred from C to D were originally referred from either A or B; however, if these referrals occurred before the start date of the data, these referrals are not included in the analysis. Thus, I only know that, during the timeframe of the data, patients were referred from C to D. A similar case could be true for every edge in the network.

Finally, since the data is aggregated, specific patient pathways cannot be followed within the network; hence, it is not known whether two or more referrals relate to a unique patient. Given these limitations, I developed visualisations that could help provide this information, presented next.

3.4.3 Chains of referrals and concurrent uses of multiple services

Figure 3.10 shows how the number of patients in their 2nd, 3rd, 4th, 5th and 6th+ referral changes over time, where time $t = 0$ is the start date of each patient's index referral. Additionally, Figure 3.11 shows how the number of patients in 2, 3, 4 and 5+ concurrent referrals changes since the start of their index referral.

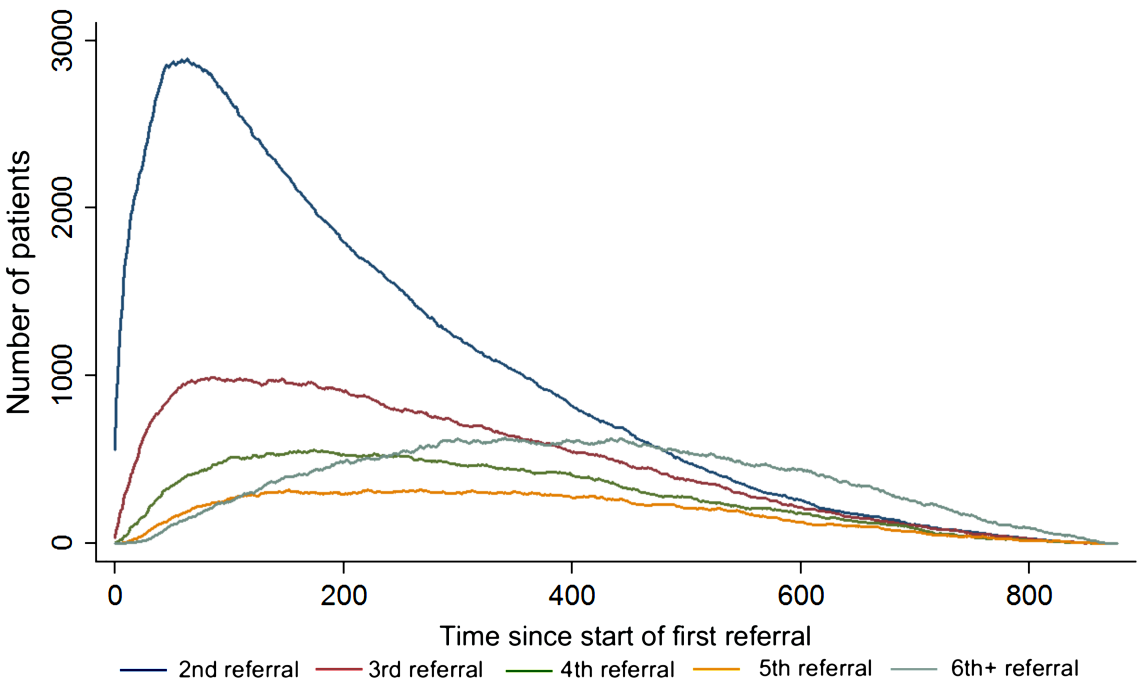

Figure 3.10: Timelines showing how the number of patients in their 2nd, 3rd, 4th, 5th and 6th+ referrals changes over time. Time = 0 corresponds to the start date of a patient's index referrals.

Due to right censoring, these plots bias towards shorter referrals. However, they highlight the potential for subsequent referrals to overlap. Thus, I examined the data to identify chains of referrals and services that were commonly used at the same time.

A chain occurs when patients are first referred to a specialty that then refers them to another specialty. These chains are visualised in a plot created using the R package "alluvial" [91]. Flowing from left to right, the plot begins with the chain sources. Lines are plotted from the source to the first specialty that patients are

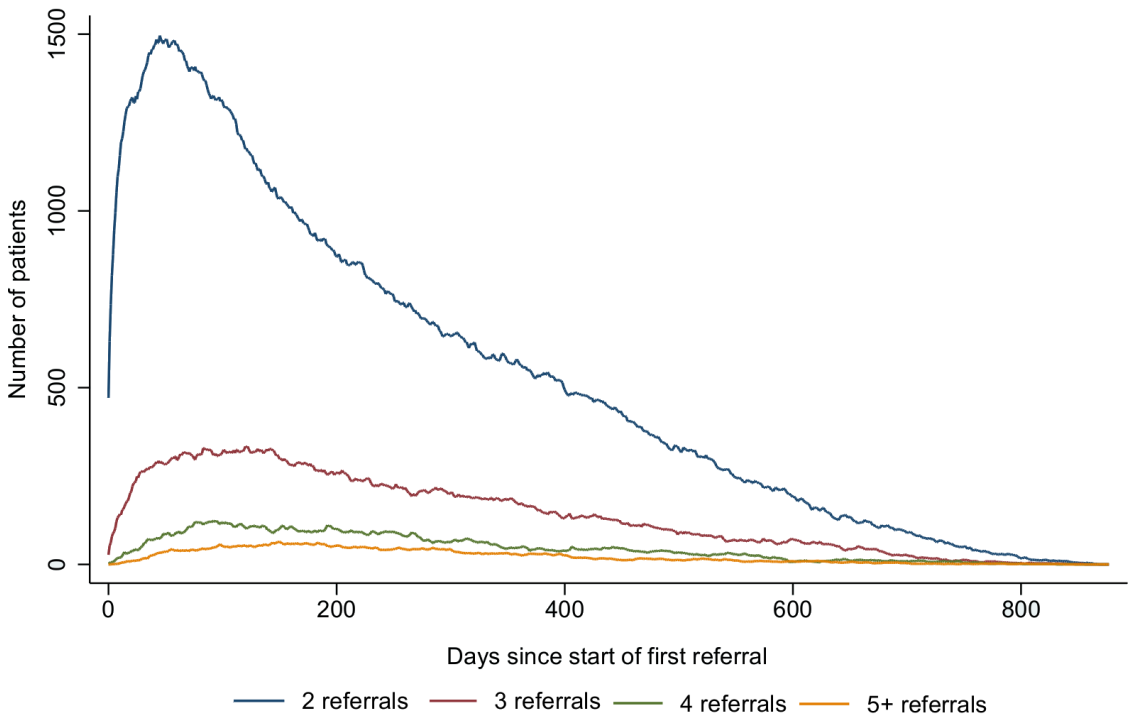

Figure 3.11: Timelines showing how the number of patients involved in 2, 3, 4, and 5+ referrals at the same time changes over time. Time = 0 corresponds to the start date of a patient's index referrals.

referred to. The line's width indicates the total number of referrals and its colour indicates the first specialty that is referred to in the chain. Similarly, lines are drawn from the first specialty to the second specialty, maintaining the same colour. This plot is ordered so that chains from the source, through the first specialty, to the second are clearly distinguishable.

Similarly, I used an interactive sunburst plot, produced using R package "sunburstR" [92], to visualise how patients concurrently used services. This is a hierarchical plot showing the number of patients using different combinations of services.

The plot consists of layers of rings, each divided into segments that represent different services, indicated by colour. The inner most ring contains parent segments, representing all services used concurrently with at least one other service. The size of each segment shows the total number of times this service was used concurrently with other services. In the next ring, the parent segments are divided into sub-segments. Considered in combination with their parent segment, these segments

represent pairs of concurrently used services, with size indicating how many times they were used together. Each subsequent ring follows this pattern, dividing into further sub-segments, increasing the number of services used together.

This plot is also orderless to aid navigation. This means that the size of a segment for some service A in the second ring with some parent service B, is equal to the size of a segment for service B in the second ring when service A is its parent. Joint uses can therefore be examined starting from a service of interest. In creating the graph in this manner, each possible order of service combination is included. Thus, combinations of three unique services will have 6 segments in the third ring, whilst combinations of four unique services will have 24 in the fourth ring.

3.4.4 Subsequent uses of community services

From conversations with NELFT service managers, they wanted to understand how patients' subsequent referrals developed over time and what services they used. Therefore, I visualised the future community care referrals of patients with a common index service by plotting each of their referrals as horizontal lines, each beginning at the start date of referral and finishing on the date of discharge. This plot displays aggregated patient data, grouping referrals by service, and is ordered by referral date, then referral length. Again, time 0 represents the start date of a patient's index referral.

I now present an application of these methods to NELFT's data, discussing how they may inform the design of their single point of access.

3.5 Application to NELFT referral data

3.5.1 Network Map

Beginning with the complete network, Figure 3.12, there are 75 nodes comprising 44 sources and 31 specialties (11 referring to other specialties) with 386 edges representing 45,506 referrals. Whilst this network is visually complex, the relationships between nodes can be explored interactively within Gephi by highlighting individual nodes to reveal their nearest connecting neighbours. When sharing the complete map with collaborators, this was useful for highlighting key information, exploring the network and identifying services of interest.

Working through the network, there are two levels of activity. Figure 3.13 is a high activity network, containing edges with > 2 referrals per month, which represents the bulk of activity within the system. Featuring 36 nodes - 14 sources, 22 specialties (seven referring to other specialties) - and 81 edges this network represents 93.1% of all patient referrals. Figure 3.14 is the low activity network, containing edges with ≤ 2 referrals per month. Representing the remaining 6.9% of referrals, there were a total of 74 nodes - 44 sources, 30 specialties (11 referring to other specialties) - and 305 edges, highlighting the large number of low activity services and edges.

To explore this system further, I calculated several network statistics. For the complete network, Figure 3.12, the average number of edges connecting each node is 5.15. Furthermore, the directed network density with loops - the number of edges in the network divided by the total number of possible edges, where specialties may refer to themselves - is 0.17. Due to the source-specialty structure of the network, I had to formulate an adapted density formula for this network:

$$G_D = \frac{|E|}{|N| \times |N_{spec}|}$$

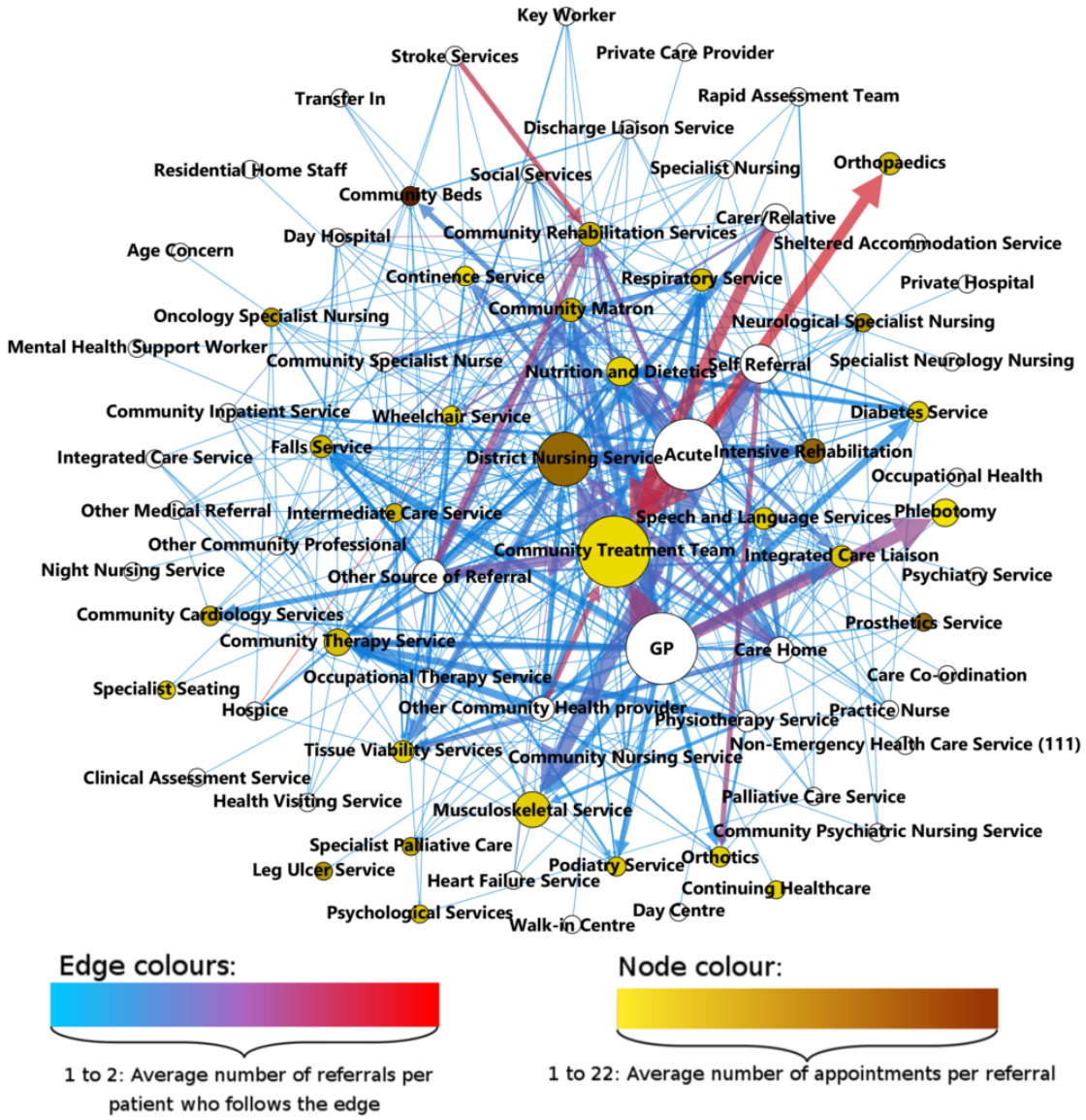

Figure 3.12: Network map of referrals within NELFT's community health care services. All sources, specialties and edges.

where, $|E|$ is the total number of edges in the network, $|N|$ the total number of nodes, and $|N_{spec}|$ the total number of specialties.

These measures are reasonably low given my collaborator's suggestion that because all specialties could refer to all specialties, this network would be highly connected. However, given this reasonably low density, this may not be the case.

To investigate further, I filtered the network to consider specialty to specialty referrals only, Figure 3.15. Containing 28 specialties and 74 edges, this network

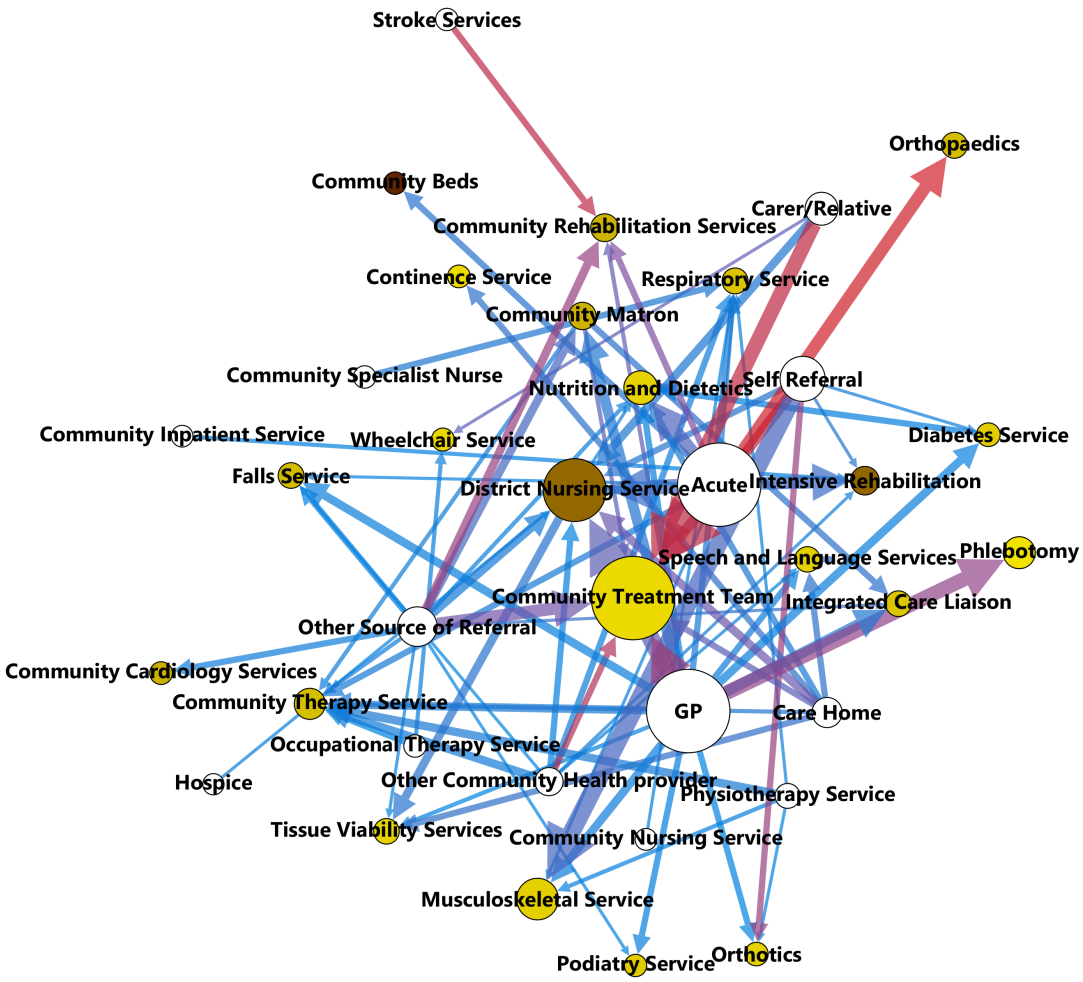

Figure 3.13: Network maps of referrals within NELFT's community health care services. High activity network: edges with > 2 per month.

represented 2,919 referrals with a directed graph density of 0.094. Interestingly, eight nodes and seven edges represented 2,100 of these referrals, Table 3.4. This confirmed that specialties did not refer to each other as much as initially suggested.

3.5.2 Chains of referrals and concurrent uses of multiple services

Next, I analysed chains of referrals consisting of a source, first specialty and second specialty, Figure 3.16. To begin, I only consider the first and second specialties.

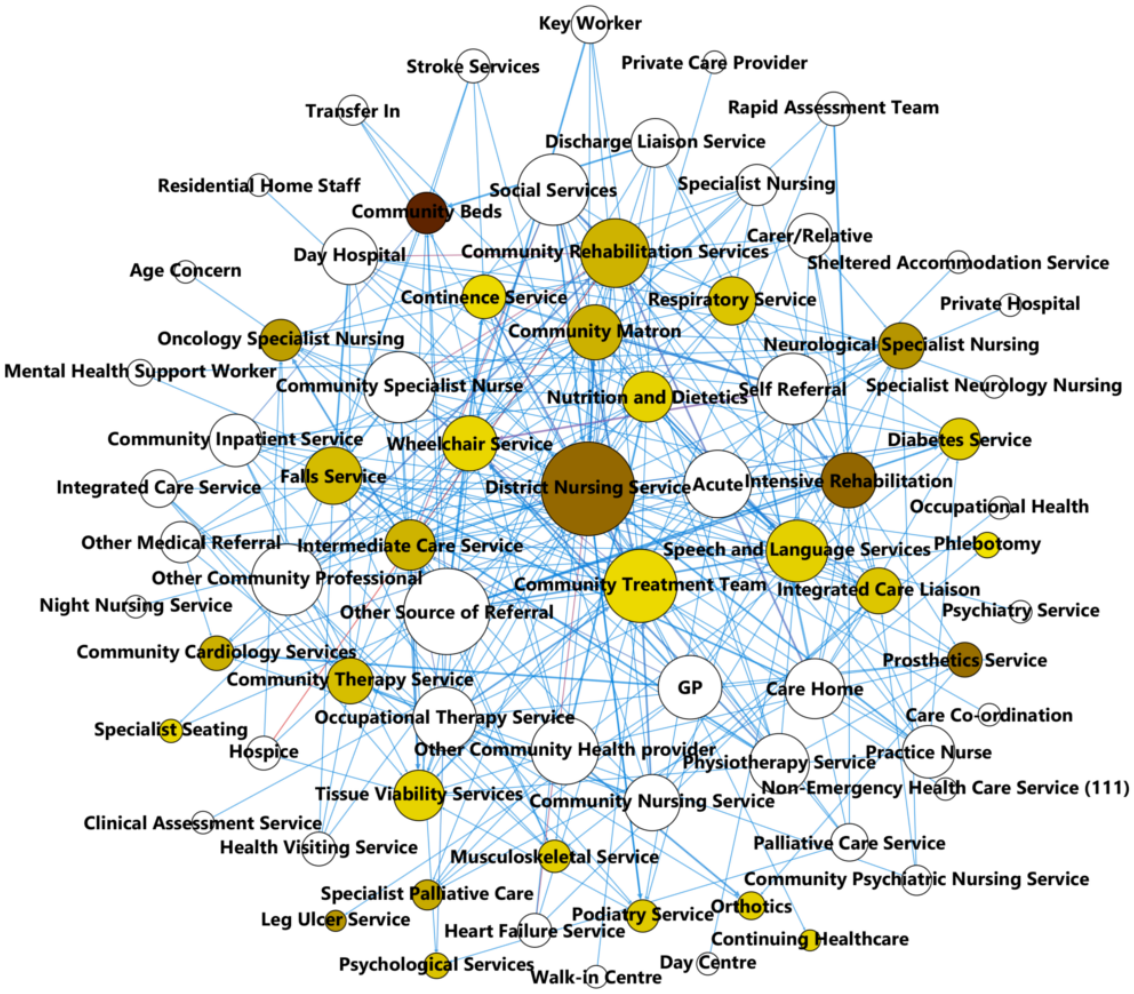

Figure 3.14: Network map of referrals within NELFT's community health care services. Low activity network: edges with ≤ 2 per month.

Referring specialty	Receiving specialty	Total referrals
Community Matron	Community Treatment Team	117
Community Matron	Integrated Care Liaison	179
Diabetes Service	Nutrition and Dietetics	189
Speech and Language Services	Speech and Language Services	223
Nutrition and Dietetics	Nutrition and Dietetics	398
District Nursing Service	District Nursing Service	483
District Nursing Service	Tissue Viability Services	511

Table 3.4: Seven high activity referral edges form the bulk of activity in the specialty to specialty network, Figure 3.15

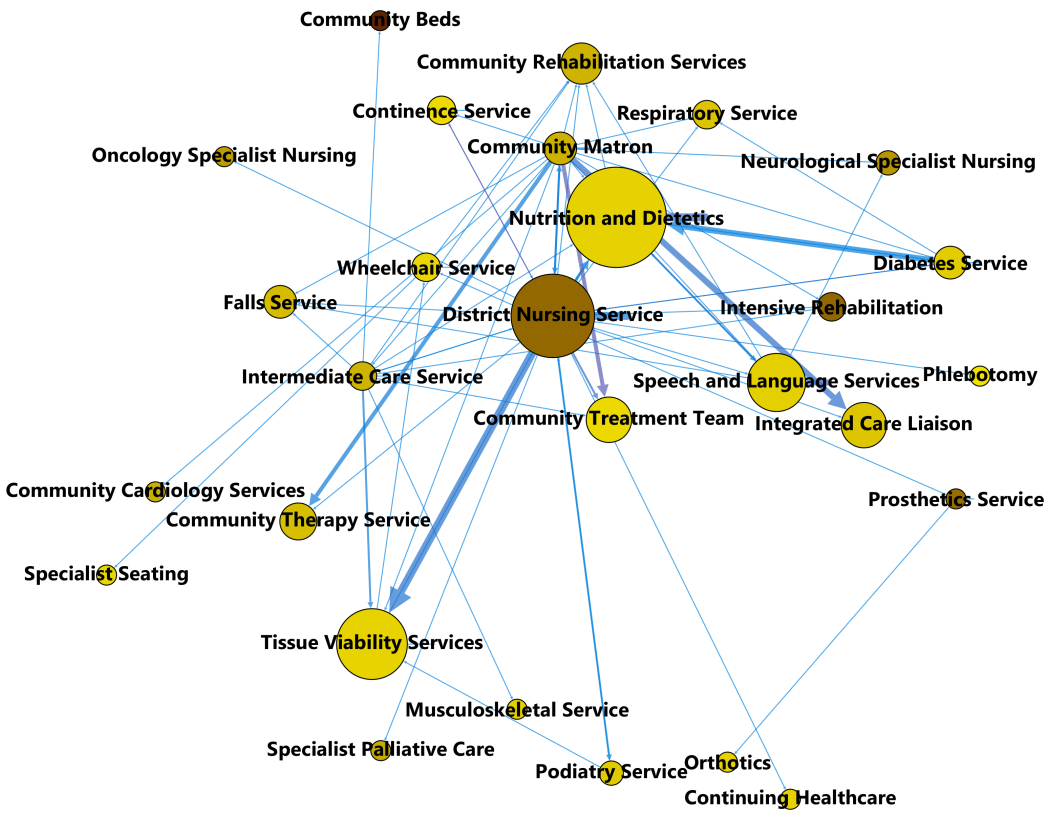

Figure 3.15: Network map of referrals within NELFT's community health care services. Specialty only network with all specialty to specialty referrals.

I found 47 combinations of specialties representing 814 patient uses. Altogether, there were nine different first specialties and 23 different second specialties, eight of which were common between the two sets.

Of the first specialties, District Nursing Service (DNS) and Community Matron services were the most common, featuring in 17 and 15 chains, accounting for 470 and 270 total patient uses, respectively. In comparison, the next most common were Nutrition and Dietetics, and the Diabetes Service featuring in three and two chains, amounting to 114 and 161 patient uses, respectively.

The maximum number of chains that a second specialty featured in was five, with a mean of 2.04 appearances. Furthermore, there was large range in the number of patient uses for these second specialties: Nutrition and Dietetics, 325; Tissue Viability service, 227; DNS, 154; Community Therapy Service, 77; Community Treatment

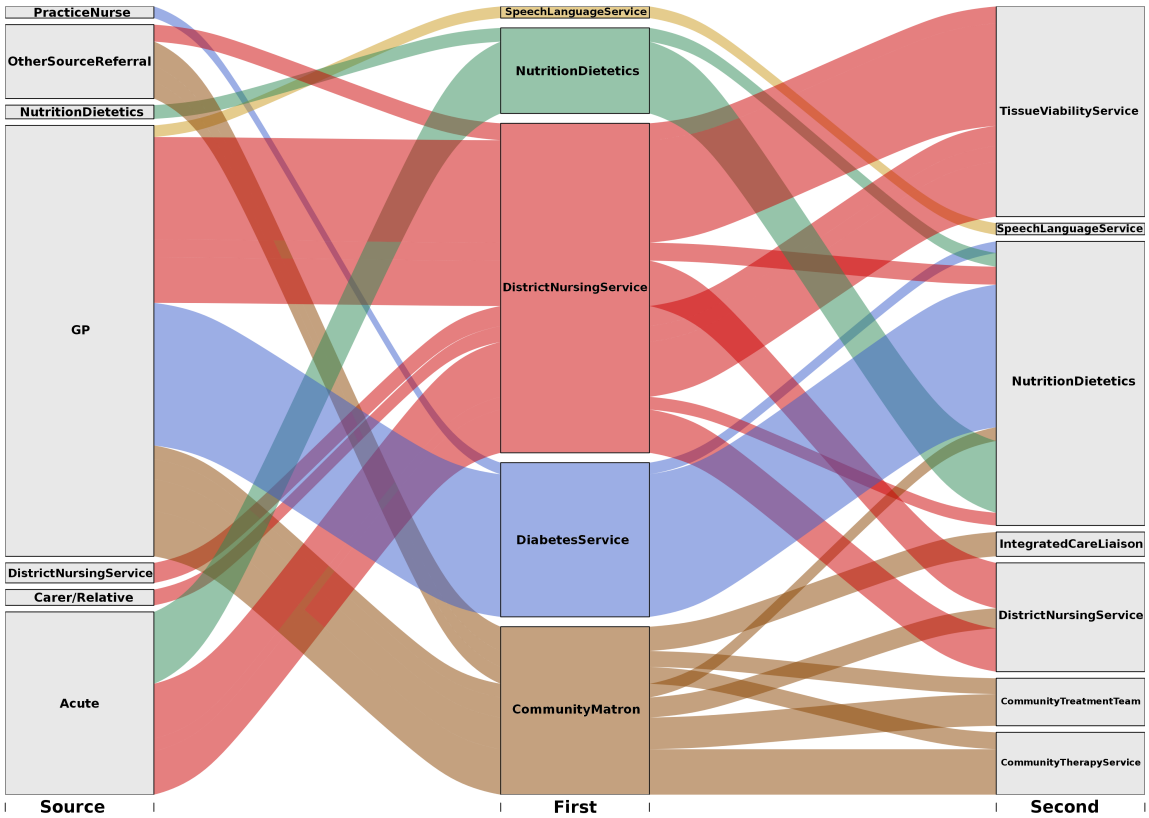

Figure 3.16: Chord plot for chains consisting of a source and two specialties. Only instances with > 20 occurrences are shown to improve interpretability.

Team, 77; and Speech and Language Services, 62. Significantly, 260 patient uses represented loops where the first and second specialty were the same.

In response to this information, care leads suggested that it would be useful to explore the timing of these onward referrals, in particular instances of quick referral and discharge. This could help identify inappropriate referrals e.g. instances where patients were referred to an incorrect service. Table 3.5 presents chains with more than 20 occurrences, and how many second referrals occurred in the first 14 days, 28 days, and time span of the data.

Investigating chains of a source and three specialties would be a natural next step; however, I cannot present an extensive analysis here since few of these chains existed in the dataset. In particular, there were 66 such chains representing 196 patients uses, yet only two had more than 10 occurrences.

Referring specialty	Receiving specialty	Onward referrals		
		14 days	28 days	Overall
District Nursing Service	Tissue Viability Service	69	91	225
Diabetes Service	Nutrition and Dietetics	107	121	159
District Nursing Service	District Nursing Service	66	73	122
Nutrition and Dietetics	Nutrition and Dietetics	29	51	97
Community Matron	Community Therapy Service	34	44	65
Community Matron	Community Treatment Team	19	25	62
District Nursing	Nutrition and Dietetics	14	20	44
Community Matron	Integrated Care Liaison	41	42	42
Speech and Language Service	Speech and Language Service	7	14	37
Community Matron	District Nursing Service	7	9	31
Community Matron	Nutrition and Dietetics	6	9	22

Table 3.5: Table of chains that occur more than 20 times in the data, noting how many second referrals occurred in the first 14 days, 28 days, and length of the data

I also visualised joint uses of service, see Figure 3.17. It is immediately clear from the plot that the DNS and the Community Treatment Team are the most concurrently used services. Whilst this map is visually complex, its interactive capability helps overcome this for nuanced analysis. Furthermore, as noted by [82] the identification of angle sizes is often inaccurate; however, the ability interact with the sunburst plot helps to overcome this and improve its usability.

For example, to explore how the Respiratory Service is used concurrently, one starts at its parent segment in the inner ring. Highlighting the segment indicates how many times it has been used concurrently. Moving through the outer rings, highlighting a sub-segment reveals a step by step chain of the services used and the number of occurrences, as in Figure 3.18.

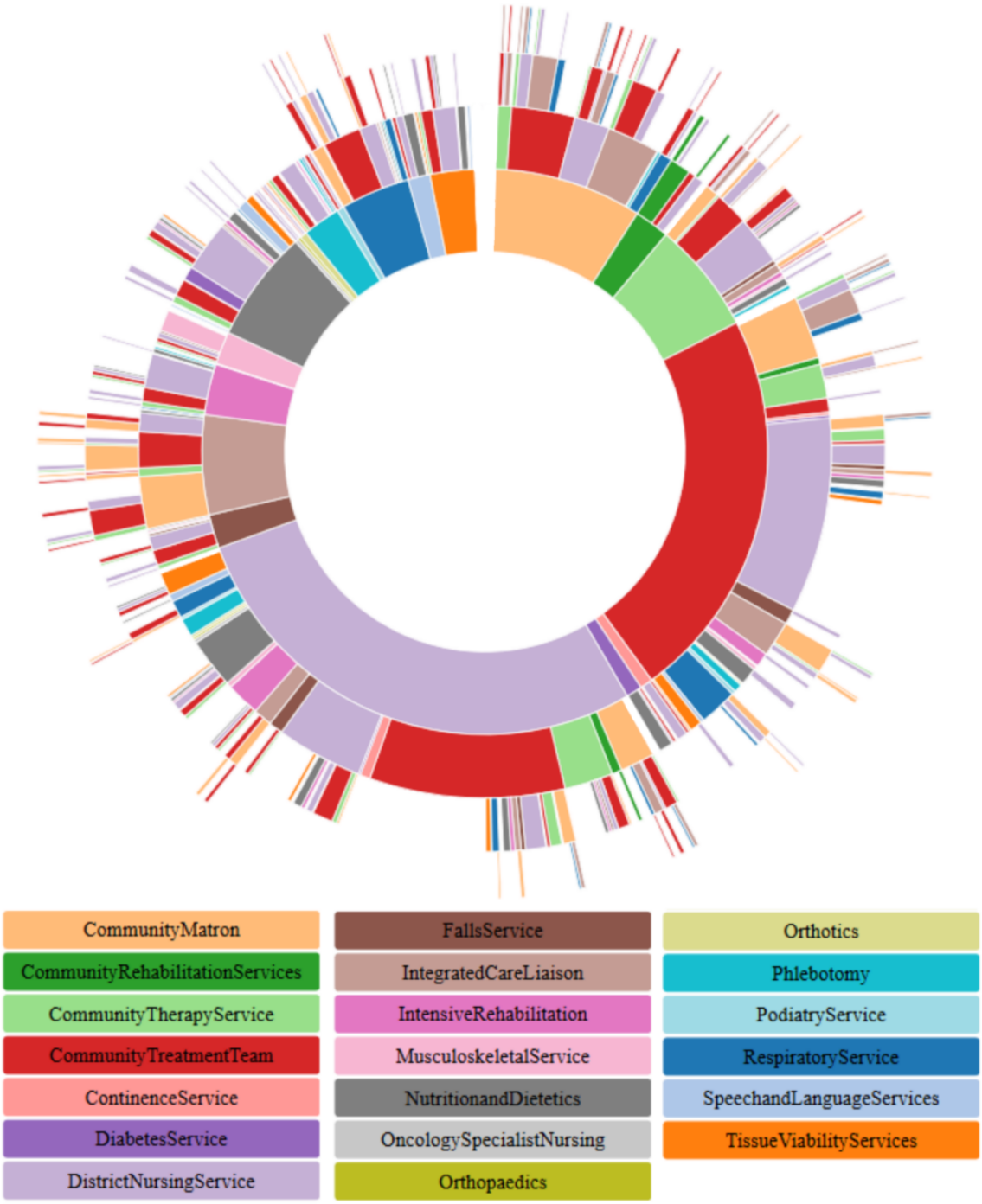

Figure 3.17: Example of joint uses of each community service, for instances of > 20 occurrences

3.5.3 Subsequent uses of community services

The previous analyses show that the DNS is highly connected, frequently appearing in several sequences of referrals and concurrent uses of service. Considering patients with an index referral to the DNS, I created timeline plots to visualise their

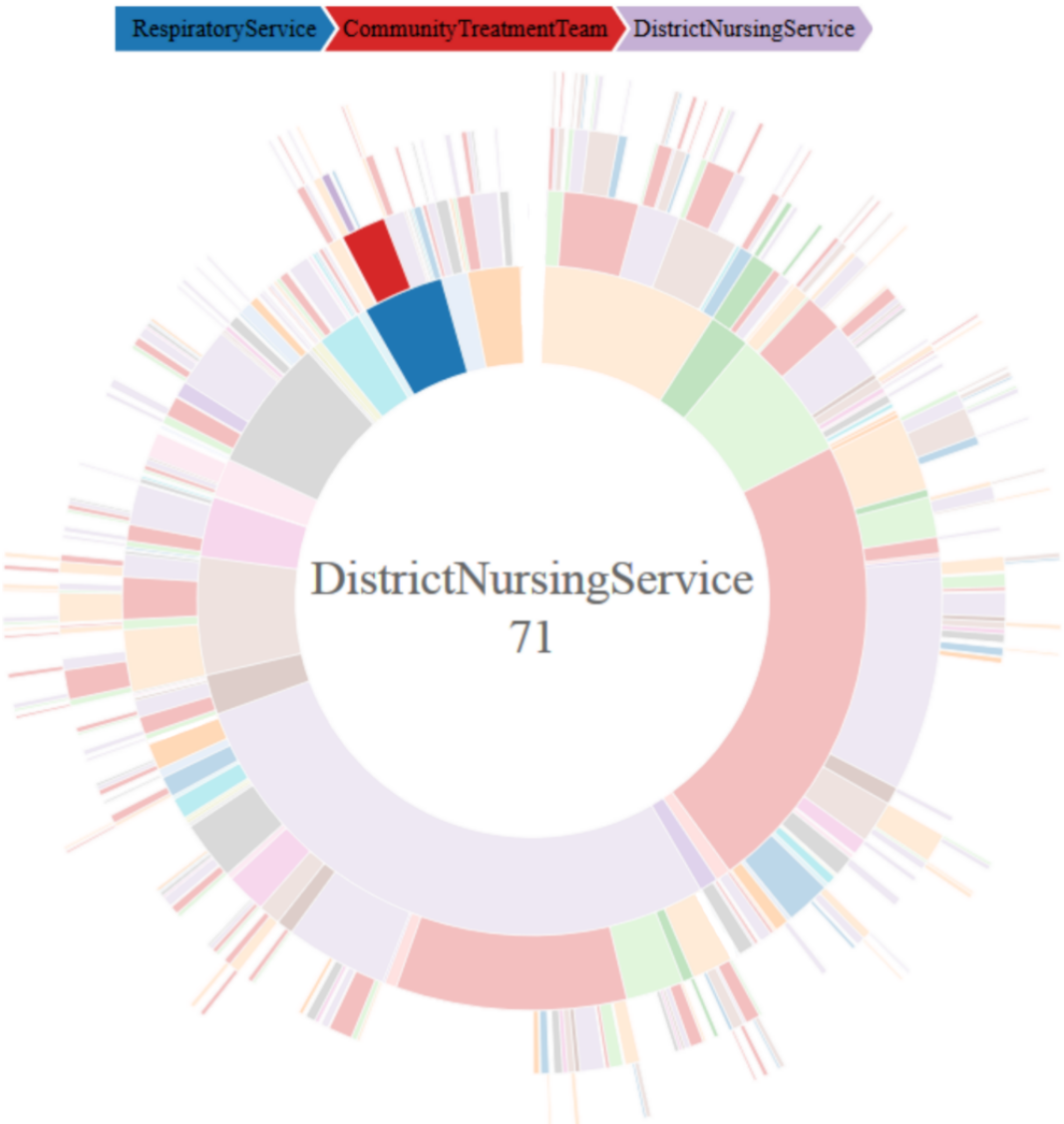

Figure 3.18: Example of sunburst plot's interactive capability

This example shows how highlighting a segment in the third ring causes all segments other than the concurrently used services to fade out. It also provides a helpful chain of concurrently used services at the top, and the number of occurrences in the centre.

subsequent uses of community services, Figure 3.19. This plot adds context to the chains presented above and may help to identify patterns in patient use.

Patient use within each service looks markedly different due to the variation in the care each service provides and why patients use them. For example, the many Community Treatment Team referrals only last for short periods of time since this service aims to prevent unnecessary acute admissions by providing responsive care

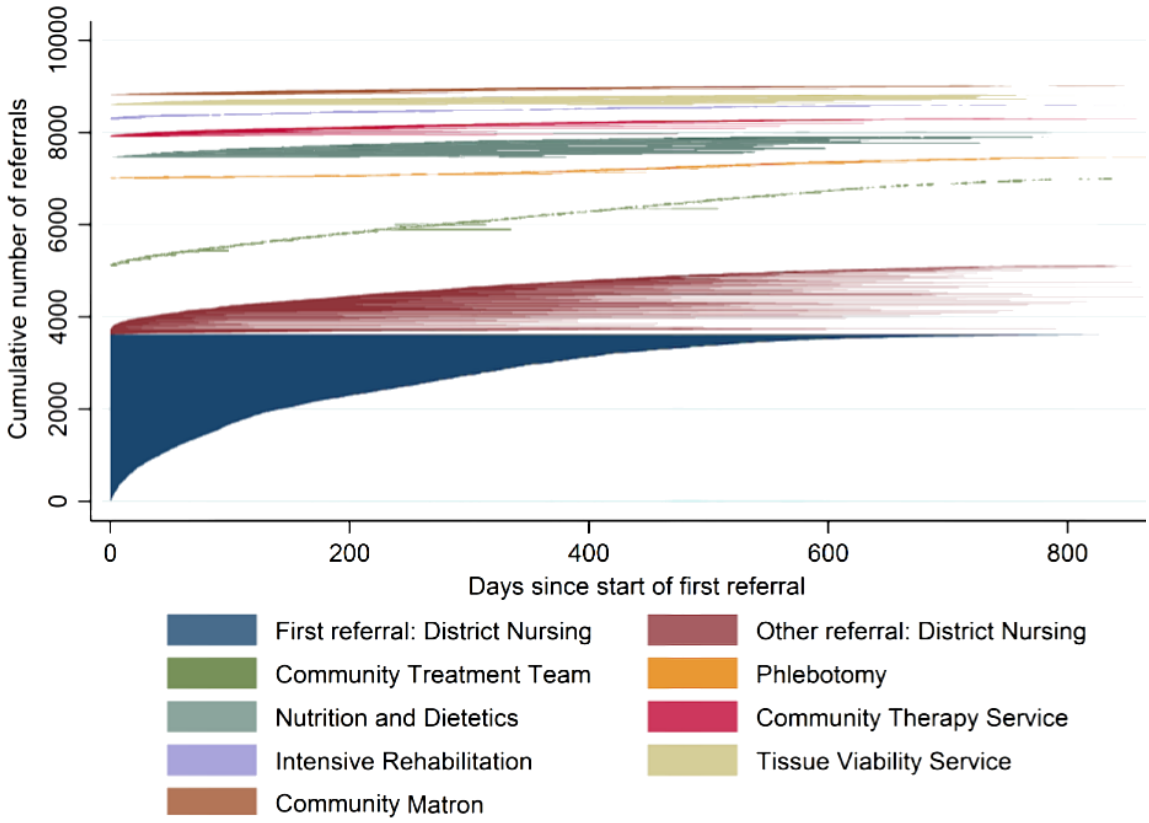

Figure 3.19: Timelines of subsequent referrals for patients whose index referral was to the District Nursing Service

Each referral is represented by a horizontal line, starting at the date of referral, ending at the date of discharge. Referrals are grouped by service and sorted by referral date and length.

within the patient's home. By comparison, the referral lengths of patients in the DNS vary greatly, with some patients staying for a few days, others for years. This reflects the diversity of care offered within this service. Furthermore, patients whose index referral was to the DNS also had subsequent referrals to the DNS, highlighting the potential for reuse.

Another interesting feature is the lack of significant clusters or patterns. This may occur due to the limitation of using an index referral i.e. I potentially lack all prior information regarding a patient's use of community services, limiting the insight gained from this analysis. This may also highlight my collaborators' observations that clear pathways are hard to define amongst these services due to the scope for any patient to use any range of services and the diversity of patients.

3.6 Summary and Discussion

Accessible methods for visualising referral data are useful for understanding how patients use health care and for informing the management and organisation of health care services. Through collaboration with care leads and the exploration of patient level data, I applied several visualisations methods to NELFT referral data in order to analyse the key dynamics of referrals within their community services. These visualisations helped to understand multiple uses of service, the timing of patient use and quantify repeated use of services.

Having shared this work with care managers throughout, the main benefit of these methods to NELFT was the opportunity to ask and investigate more refined questions around the nature and patterns of referrals as they designed their single point of access (SPA). Notably, answers to these questions could be gained from exploring the raw data and further analysis of more complete data. Hence, the visualisations help to identify dynamics and patterns that should be further explored within the data. Whilst the sharing of the visualisations led to more detailed discussion, at the time of concluding this work I do not know how they may have helped to inform their decision making; thus, I can only detail the resulting discussion topics and points raised.

Using a network representation, node and edge colouring, and network filtering helped to provide greater understanding of NELFT's community referral data. This helped inform NELFT's thoughts around the design of a SPA and helped to identify directions for further investigation. This included questions around what level of activity is appropriate within the system and whether what is seen within the network map was expected. Furthermore, the range of activity seen in this system was of interest. For instance, the sharing of the map led to discussions as to whether the high number of low activity referral pathways is appropriate for the range of services and whether the high volume of low activity unnecessarily complicated the referral

process or reflected a positive characteristic of the system.

When interacting with the network in Gephi, service managers began to be identify possible services for inclusion within the SPA, discussing what sort of activity it should handle, e.g. only external referrals from GP, social care and acute? In particular, whether this would help to reduce some of the low level activity and help avoid inappropriate referrals. Considering high activity, questions arose about whether natural groups of services existed and how referrals between them could be handled by a SPA and how the introduction of a point of triage may affect the structure of these referrals. For example, if the SPA only handled NELFT to NELFT referrals would this have a positive affect on patient access through improved handling of multiple referrals for single patients and whether the SPA would help streamline referrals so that patients are referred directly to the appropriate services.

Evaluating chains and concurrent uses of community services enabled the analysis of sequences and the identification of common pathways. This information may be used to inform referral guidelines and service planning. For example, after sharing this work with my collaborators, they suggested this method may help identify cases of inappropriate referrals in instances where the initial referral is short and the specialty acts as a "point of triage", rather than a point of care. Thus, the SPA could be used to prevent this. Similarly, such a referral may indicate that patients who are referred to a particular service, say the Diabetes service, often require another service, such as the Nutrition and Dietetics service, which could have been made alongside the initial referral, potentiality improving the ease and speed of access.

Finally, plotting uses of community services after a patient's index referral helped care leads to understand what services patients used, when they used them, and how long their referrals lasted. This visualisation could provide information that may be useful for the initial triage of referrals within a SPA. For example it could give an indication of whether there is a likely pathway for certain patients, aiding the

future planning of their care. Notably however, the presented example (given the limitations noted below) indicated no patterns or clear insight to further motivate this discussion. Depending on the index service, there may be instances where explicit patterns occur, further informing SPA referral practice and guidelines.

3.6.1 Limitations

Through collaborative working, several limitations of this work were identified alongside areas for future research. Limitations in processing and using the data were introduced by working within the Safe Haven, for example, Gephi was not available within the Safe Haven. This increased the complexity of the visualisation processes because the data had to be first processed in the secured setting and then extracted in aggregated form. A further limitation occurred when sharing this work. When exploring the visualisations, collaborators would ask questions that could only be answered by patient level data. Not having this available introduced a time lag in the information I could provide and stifled useful conversations.

Furthermore, I obtained a single extraction of data, limiting the work because I did not have complete information for every patient, and did not know their entry points to community care. It would have been insightful to apply these methods to include each patient's first referral, but I was limited to using index referrals. The end date of the data added a further limitation, since patients who entered the system later would use fewer services, introducing bias towards shorter referrals. This could potentially be overcome by using Kaplan-Meier curves or multi-state models to evaluate referral lengths; however, this does not overcome the limitations introduced in identifying overlapping and subsequent referrals.

A solution to these issues is to work physically within the organisation, where data is easily accessed and updated. To this end, I ran a seminar for care leads within mental health, physical health and social care to teach them how to implement the

network methods. However, it should be noted that access to more data may not improve the work. Community services change rapidly with new configurations and referral guidelines regularly introduced. As a result, datasets that span large time periods may include multiple configurations of the system and lead to inaccurate conclusions or a misrepresentation of the system.

3.6.2 Possible avenues for future work

Some questions that arose from this mapping work that were not directly addressed include:

- Can groups of service be identified in the system where patients "bounce between" them?
- Can data visualisation help to identify inappropriate referrals?
- Can a patient's total care be described by including services outside of physical community care, e.g. acute care, social care and mental health?

These would be good directions for the future of data visualisation for health care planning. Each addresses key difficulties in the provision of community care that are hard to identify from the raw data alone.

3.7 Conclusions

In this chapter I used several visualisations to aid the interpretation of complex referral data. The primary aim was to learn about how patients are referred into community services, how they use these services and whether there are any key dynamics to be included within the patient flow model I develop in chapter 5.

Each analysis focussed on different referral characteristics of community health care and provided insight into how patients used community services, considering

the interface with external points of referral such as acute care. The network map helped to portray the vastness and complexity of the system, to identify groups of services according to patient activity, and to quantify the reuse of services. Analysing chains and concurrent uses of services provided insight into the progression of patient care and common combinations of services. Finally, plotting subsequent referrals indicated the timing, length and patterns of future community care referrals for patients with a common index referral.

Together, these visualisations give a broad understanding of the referral dynamics in the system, providing informative analysis in three ways. Firstly, they help to present and understand complex data. Secondly, they are accessible and may aid the identification of individual or groups of services which patients commonly use. Thirdly, they stimulate conversation around what information is beneficial in planning these services.

As mentioned, in applying these maps several important dynamics of how patients use community services were identified. These were: uses of multiple services; the potential for patients to have overlapping referrals with different services; and the potential for patients to reuse services. As a result, the work presented in this chapter has directly informed the development of the patient flow model presented in chapter 5. A final benefit of this work is that all the visualisation are produced using open source software.

Chapter 4

Measuring patient outcomes within community services

In this chapter, I explore how outcome measures are currently used by the North East London Foundation Trust (NELFT) within their community services, and how they may be used within patient flow modelling. The aim of this work is to identify the range of outcome measures currently considered important for evaluating a diverse collection of services, and what measures are collected and used within services. This is achieved by surveying four sources of information (NELFT Quality Accounts, conversations with staff, routine patient data and commissioning data) from which a range of measures that are identified and presented.

In addition, I discuss the use, suitability and limitations of these measures for evaluating clinical performance and discuss whether there are any measures currently collected by NELFT that are suitable for use in a patient flow model. Furthermore, I note whether any measures not currently collected by NELFT would be useful to incorporate into a patient flow model. The aims of this chapter are to:

1. Survey several sources to understand the outcome measures that are important for monitoring quality across NELFT community services;

2. Suggest measures that may be incorporated within models of patient flow.

4.1 Introduction

Maintaining and improving quality is a major priority for health care providers around the world. In the UK, the quality of health care has been of increasing importance, featuring heavily in both health care policy [18, 93, 94, 95] and research (see work by the King's fund [96] and Health Foundation [97]).

Within the health care literature, there a many definitions of "quality". The World Health Organisation (WHO) define quality as "the extent to which health care services, provided to individuals and patient populations, improve desired health outcomes". [98] The Institute of Medicine (IoM) uses a similar definition, adding that this is dependent upon "current professional knowledge"[99]. The Health Foundation extend this further still stating that quality is multi-dimensional, encompassing several aspects of care, summarising quality as "the degree of excellence within health care" [97].

The dimensions of quality, commonly called domains, relate to the aims and purpose of health care. Defining quality in terms of domains helps to communicate the meaning and relevance of quality in health care, and provides a framework for measuring that quality. Throughout the literature various domains have been considered. In Lord Darzi's 2008 review [18] three domains were suggested: patient safety, clinical effectiveness and patients' experience. Alternatively, the Care Quality Commission (CQC) use five domains in their inspection framework, assessing whether care is: safe, effective, caring, responsiveness to patient needs, and well led [100]. The Institute of Medicine define six domains, establishing whether care is: safe, effective, patient-centred, timely, efficient and equitable [99].

As noted in the definitions of quality by the WHO and IoM, quality may be measured by changes in the health outcomes of individuals and populations of patients. These are measurable aspects of patient health that can be influenced by care interactions. With clinical effectiveness a fundamental component of health care quality,

the measurement of health outcomes provides a means for assessing whether care has its intended effect - to maintain and improve the welfare of patients. Therefore, health care providers routinely collect data on outcome measures to quantify and assess the quality of their care. This is important for the evaluation of both individual services and systems of multiple services as they help to track trends in quality and identify areas for improvement. With this in mind, I explore what outcome measures the North East London Foundation Trust (NELFT) use for measuring clinical effect across multiple services in this chapter.

Structure and contribution of the chapter

Seeking to inform how outcomes may be used within models of patient flow through multiple services, I focus on measures that are used across a range of services. The aim is to identify what measures community services are used to evaluate, after which I discuss how useful these measures may be used in evaluating the clinical impact of services. Having discussed and identified any suitable measures, I then discuss how such measures may be used in a model of patients attaining care and using several services.

As in chapter 3, I concentrate on the measures relevant to community services located in Havering and used by patients aged 65 and over, drawing this information from several sources: NELFT Quality Accounts 2013-2017, informal conversations with NELFT staff, commissioning data and routine patient data. These sources are discussed further in the next section.

In section 4.3, I identify the range of measures used in community services and present a synthesis of findings from across the sources, drawing out common themes and measures. The themes relate to the overarching goals identified within the sources and are discussed in turn, referencing the relevant measures used to evaluate them. I the discuss their appropriateness and limitations in evaluating clinical

performance across different services.

The chapter concludes with a discussion of how outcome measures may be incorporated into models of patient flow. This includes suggestions as to whether any measures currently collected by NELFT may be useful in modelling their services. Additionally, I indicate types of measures that are not currently collected by NELFT that could be informative in a flow model. These suggestions will be made with a system's view in mind - how measures may be used to evaluate quality across diverse services, and how they may be used to assess the contribution of multiple services to improvements in patient outcomes.

4.2 Sources used for understanding outcome measurement

NELFT Quality Accounts 2013-2017

Quality Accounts are produced annually by each NHS care provider to report on the quality of their services according to patient safety, clinical effectiveness and patient experience. They comment upon the progress made over the year, the achievement of goals set the previous year, and outline new targets for the coming year. This includes the identification of services and streams of care where quality is an organisational or national priority for the next year. Additionally, the reports highlight key outcome measures for monitoring the progress of services in meeting their future quality goals. By surveying five years of NELFT reports, I gain a comprehensive account of NELFT's past, present and future organisational priorities and the types of measures deemed important.

Learning from conversations with NELFT staff

To understand outcome measurement at both service level and patient level, I spoke to several of NELFT's clinical and managerial staff. This included clinicians from various community services, community care managers and performance managers, with conversations carried out both in person and over the telephone. During these conversations, I asked a range of questions to understand how each service operated, what types of care they provided, what outcome measures they collected, and how they used them to both monitor patient health and evaluate the quality of their service.

Data - workbook of service level performance measures

For commissioning purposes, a workbook of performance measures for community services is produced annually by NELFT. It contains several key performance indicators for each service, ranging from generic measures to service specific measures. Notably, the data contained in this workbook does not differentiate between patient groups and represents patients of all ages who used the services. I therefore refer to this data to understand the outcome measures that are important for commissioning community services, but recognise that it does not specifically inform the care of elderly patients.

Data - patient level

Finally, I use the same data as in chapter 3 to explore which performance measures were recorded during care interactions within community services. In addition to the information noted in chapter 3, there are several other fields in the dataset, including: the reason for referral, the number of contacts, the length of stay, discharge date and urgency of referral.

At appointment level, the data consists of appointment dates, cancellation dates

(if relevant), outcome of appointment, and the activity that occurred during the contact. The outcome field contains information about whether the appointment was attended and whether follow up was suggested or changes were made to the patient's care plan. In addition, the activity field details the medical tasks carried out within the contact, ranging from physical interventions to the administration of medication or advice. Notably, neither of these fields has any direct relation to patient health or the evaluation of clinical outcomes.

4.3 Important themes and measures across the sources

From the sources, several measures were identified as key for monitoring clinical effectiveness within NELFT community services, collated in Table 4.1. Each of these measures was reported at one or more levels: patient level, service level or organisational level, as noted in the Table 4.1.

At patient level, measures are used to track, monitor and evaluate the clinical quality of care that service users receive. These measures may be used to inform the future care of a patient, whilst helping services to understand the impact of their care on individuals.

At service level, patient level measures are often aggregated to evaluate the quality of care across the population of patients. This includes the use of clinical, experiential and operational measures.

At an organisational level, measures are collected to inform the commissioning, management and evaluation of services both locally and nationally. Measures that directly related to the running, costing and budgetary capacity of services were important at this level, in particular those relating to the operational effectiveness of the system.

Performance measures	**Type of measure**	**Level of reporting**
5 × 5 surveys	Patient experience	Patient, Service, Organisation
Friends and family test	Patient experience	Patient, Service, Organisation
Length of appointments	Process outcome	Patient, Service
Length of referral	Process outcome	Patient, Service
Number of acute admissions amongst patient population	Process outcome	Service, Organisation
Number of cancelled appointments	Process outcome	Service, Organisation
Number of contacts per referral	Process outcome	Patient, Service
Number of discharges from community services	Process outcome	Service, Organisation
Number of non-arrivals (DNA)	Process outcome	Service, Organisation
Number of patients using community services	Process outcome	Service, Organisation
Number of referrals to community services	Process outcome	Patient, Service, Organisation

Performance measures	**Type of measure**	**Level of reporting**
Number of referrals to community services accepted	Process outcome	Service, Organisation
Number of reuses of community services	Process outcome	Service, Organisation
Proportion of avoidable acute admissions amongst hospital frequent attendees	Process outcome	Service, Organisation
Waiting times	Process outcome	Service, Organisation
Presence/absence of harm	Clinical measure	Patient, Service, Organisation
Volume of adverse events due to staff absence	Clinical measure	Patient, Service, Organisation
Condition specific measures	Clinical measure	Patient
Self-management/stability of condition	Clinical measure	*Not explicitly recorded*

Table 4.1: Table of key outcome measures as identified from Quality Accounts 2013-2017, conversations with NELFT staff, commissioning datasets and routine patient data

Notably, not all the measures in Table 4.1 are explicit clinical measures, noted in the second column. Rather in some instances patient experience measures and

process outcomes were suggested as helpful indicators of clinical effectiveness. Process outcomes relate to the operational capability of care services to treat patients efficiently, such as waiting times and length of stay.

Examples of condition specific measures are the Timed Up and Go test, a physiotherapy test, assessing the time it takes patients to leave their seat and get moving, and blood sugar levels for diabetes patients. I will now discuss these measures and their limitations in measuring the clinical effectiveness across community services.

4.3.1 NELFT quality themes and outcome measures

Patient experience and satisfaction

Patient experience and satisfaction were important themes at multiple levels (as identified within the Quality Accounts, conversations with staff and the commissioning workbook). Primarily measured by surveys, experience was considered to be linked to: ease of access, staff communication with patients and whether the service met patient expectations.

During conversations, it was frequently suggested that experience was a key service level priority, since patients would often use and reuse community services throughout their lifetime. As a result, improving patient experience was important to ensuring that patients continued to engage with these services.

The main measure for gauging patient experience and satisfaction was the 5 $\times$ 5 survey. Undertaken by five service users a month, these surveys assessed multiple aspects of a patient's care experience through a set of simple questions. Included within this was the friends and family test (FFT), in which patients were asked whether they would recommend the service to friends or family. This was widely noted as an important measure since patients would only recommend services they

considered to be beneficial and where they had a positive experience. It was also indicated that clinicians and care managers regularly used this information to understand how their service is perceived by patients, and learn where they could focus improvement initiatives.

Survey results are helpful measures as they provide a framework for comparing diverse services against a service's patient experience priorities. However, 5×5 surveys are limited in their use for evaluating the experiences of specific patient groups such as elderly patients. This is because the five monthly respondents represent a small sample of the entire population of service users and not just those of specific demographics. Thus, the results are dependent on the patients who are willing to take part in the survey. The experience data for elderly patients may therefore be missing from the results or only consist of a small sample that is not representative of the population. This is problematic since results may skew towards more positive or negative results.

Whilst patient experience is helpful for understanding how patients engage with (and may continue to engage with) a service, it does not give any explicit clinical insight as to how their health is affected as a result of care.

Service integration and patient activity

A recurring theme throughout my conversations with NELFT staff and the Quality Reports was the integration of care within community services. Written after the publication of the Francis report [94], the 13-14 report stressed the importance of quality measurement and improvement within NELFT services, identifying the integration of community services as a way of achieving this. The 14-15 report further noted their intention to develop community services with integration in mind.

Whilst not explicitly defined within the accounts, integration is often thought to improve quality through: better communication between services, a reduction in task

duplication (where appropriate), information sharing, and improved access between services. These themes came through in my conversations with staff, where measures of patient activity and service use were suggested for evaluating the integration of services and the quality of care they provided. Examples included process outcomes such as: time between referrals to different services, length of appointments, number of referrals between community services, and the activities carried out by different teams within each contact.

The tracking of patients after discharge was also noted as important. For instance, the number of patients who use community health care in the future may help measure whether these services are meeting their goals, e.g. acute admission avoidance and patients reusing community services.

In practice, tracking patients after discharge can be difficult since different services and sectors of health care use different electronic systems to record patient information. Hence, information about which services a patient used, in-between and after community referrals, is not always easily available.

For the most part only process outcomes relating to activity within individual services were suggested as a means to measure improved quality in relation to integration. Clinical improvements resulting from integration were often mentioned; however, no explicit measures for these were given.

Acute admission avoidance

The Quality Reports, conversations with clinicians, and the commissioning workbook indicated that the effectiveness of community services could be measured by the number of patients requiring urgent services and by any changes in avoidable demand for acute services. An avoidable acute admission occurs if a patient is admitted to an acute service when a non-acute service exists that could have met their needs. Several services were identified as important in achieving this, including those

that support patients with long-term conditions (e.g. diabetes service, occupational therapy, health visiting and district nursing); and community services providing care usually accessed within hospital (e.g. phlebotomy). Outcome measures for evaluating the impact of these services were noted as the number of patients using these services, the number of patients re-referring to these services, time between acute admissions of their service users, and the number of acute admissions amongst service users - all of which are process outcomes.

Using avoidable admissions to measure the clinical effectiveness of community services is limited since it can be difficult to identify when an avoidable admission occurred. Furthermore, in seeking to reduce avoidable admissions one must be careful to not reduce or avoid necessary admissions. Again, whilst linked to the clinical impact of services, I only found process measures.

Discharge, self management and health improvement

Within both sets of data and from conversations, discharge from service was an important measure since it was considered to be a marker of patient stability and independence. This was especially true for patients with long-term conditions. These patients will engage with community services throughout their lifetime to maintain their health and manage a condition. Thus, a discharge in this setting is representative of them attaining positive clinical outcomes.

As a measure on its own, the number of discharges should be used with caution since early or inappropriate discharges typically associate with negative outcomes, both clinically and experientially. Through improvements in clinical effectiveness - such as increased uptake and completion of personalised care plans - improved quality may be marked by increased discharge. However, discharge is a process outcome and is linked to the operational capabilities of the service to meet the current and future care needs of patients. Thus, an increased discharges does not necessarily indicate

improved quality since this may occur due to increased early discharges and capacity driven decisions, which may associate with poor quality service.

It was also emphasised that self management was a useful measure. This occurs when a patient is able to monitor and maintain their condition outside of the service by themselves. This may be used to gauge the effect of a single service or an amalgam of services in helping patients manage their own conditions. Currently, self management is implicitly recorded as part of a patients clinical notes, and is not explicitly recorded as an outcome within NELFT's routine datasets. In practice, it represents an amalgamation of measures, some of which are service/condition specific, and is the result of progress according to the range of measures. Whilst the clinical definition of self management will differ between services, it is a concept that translates across services.

Again, as with discharge, caution must be taken when considering self management since it is often measured as a process outcome relating to a lack of service use, e.g. the patient no longer using the service or requiring the service. Thus, to consider this as a positive clinical measure, the absence of service use has to associate with positive clinical outcomes, which may not always be the case.

4.4 Summary and Discussion

The collection and use of outcome measures are important for the evaluation of health care. They help to quantify quality and provide understandable information that can be used to monitor, maintain and improve care. By using a range of measures, clinicians and care managers are able to assess the impact of their care on individuals and populations of patients, and identify where and how improvements may be made. In practice, this can be difficult since quality measurement is multifaceted, consisting of multiple domains and many levels of use.

From the exploration above, I found several themes and priorities for quality measurement within NELFT's community services and a range of measures for evaluating them. In particular, three types of measure stood out: experience measures, clinical measures and process measures. However, the use of these measures is not easily compartmentalised into discrete categories since many correlate with several aspects of care.

The most common measures were for process outcomes relating to patient activity and the operation of services. They were used to evaluate the clinical and operational quality of a wide range of services since they are easily collected and informative.

To use process outcomes as a gauge of clinical effectiveness, they must be assumed to associate with positive clinical outcomes, but this may not always be true. Whilst process measures evaluate the operational capability of a service, how care is delivered and how patients use it; clinical outcomes measure how a patient's health and well-being is affected by the receipt (or non-receipt) of care. Whilst a process outcome may reflect positive clinical outcomes, this is not wholly the case since they are inextricably linked to the operation of services and patient activity. This includes non-clinical factors such as staffing levels, the time of year and referral guidelines.

Moreover, process outcomes that link to a cessation or non-use of service - such as discharge from service or reduction in acute admission - do not always correlate with positive clinical outcomes. Whilst a positive link is clear when a discharge occurs for clinical reasons; a positive link is not so clear when the decision is made due to capacity related issues. In the latter case, prematurely discharging patients from service may result in negative clinical and experiential outcomes. Since process measures are not always clinically driven, they cannot provide the same information as clinical outcomes. To assume that they indicate clinical effectiveness in the absence of other information is inappropriate.

In contrast to process measures, there was a lack of clinical outcome measures for

use across services. In obtaining data for this work, information relating to clinical outcome measures was unavailable. The only such measures widely collected were service or condition specific and were recorded in the text fields of patients notes that were not available to me due to confidentiality reasons.

Of most interest to this work are clinical outcome measures for comparison across multiple services; however, these were neither routinely collected, nor was the required infrastructure or framework in place for this to happen. Whilst more generic and comparable clinical measures - such as independence and self management - were referred to, they were not explicitly recorded.

4.5 Conclusions

In chapter 2, I identified three ways in which outcome measures may be incorporated into models of patient flow. These were: 1) stratifying system metrics by outcome related groups; 2) as objectives or constraints within models of resource allocation; or 3) as system metrics themselves. From this survey of available measures and data, I found no clinical outcome measures - either explicitly recorded or available to us - that could be used informatively within patient flow modelling.

With a view towards patient flow modelling, several of the process measures and dynamics found in this chapter will be included within the model developed in chapter 5; namely, patients reusing services and uses of multiple community services. In addition, uses of services other than community services will be considered.

Notably, solely modelling these dynamics and process outcomes is not new and the collection of measures that I have found above do not help us in seeking to incorporate health outcomes in patient flow models. My rationale for including health outcomes is to move away from assuming a positive correlation between changes in process and health, and produce a model for understanding how individual services affect patient

health and contribute positive health impact.

To this end, what is needed are clearly defined and recordable outcomes that provide insight into the clinical quality of a diverse range of services. In [3], this is summarised as "a framework for understanding and measuring quality that accurately and fully covers the whole range of community service activities and impact".

A clinical example within NELFT where this sort of measure is considered is their Integrated Care Management service (ICM). The ICM manages the care of patients with complex long-term conditions, whose care may require several health and social care services. To monitor a patient's health and response to care, they use three categories (red, amber and green) to reflect the level of support a patient requires and the stability of their health. Patients may progress through different stages throughout their care, improving and declining as they progress. A similar marker may be helpful if introduced across services and morbidities.

In a patient flow model, such measures may be incorporated by defining states of patient health that represent categories of outcome that patients may move between in response to multiple care interactions. An example of a publication which uses such health states is [43] where health states are used to model improvement in health as a result of care or a decline according to natural progression.

By representing outcomes in this way, the effect of patients with different health care requirements on the operation of the system can be modelled, alongside how different patients are affected by care (or the lack of it). This adds further insight into process measures, such as discharge, providing a clinical perspective that would be informative in the quality evaluation of multiple services. Such measures may be used in systems of single and multiple services, in systems with loss or the potential for patients to reuse services, and in situations where patients with diverse backgrounds and care requirements use the same service. In the next chapter I will begin to explore how such measures could be used in patient flow modelling.

Chapter 5

Fluid and diffusion approximations for modelling the flow of heterogeneous patients within a network of queues

In this chapter, I present fluid and diffusion approximations for a network of stochastic queues with heterogeneous patients. These methods feature some of the flow dynamics identified in chapters 3 and 4, namely, the potential for patients to reuse services and for patients to sequentially use different services. For generality these methods also include the potential for patients to abandon the queue, and the possibility for them to subsequently rejoin the queue, or use an alternative service. Furthermore, I incorporate clinical outcomes in the form of transitions between health states that patients may move between, as a result of service, or lack of it.

For tractability, patients are modelled as being served in parallel queues according to their health state. To overcome some of the difficulties that this may introduce, a dynamic multi-class server allocation is applied to the system. Here parallel queues share servers from a single pool and are continuously reassigned across parallel queues

in response to fluctuations in the demand for service and patient attributes. By using parallel queues, the outputs from this method (the average and variance of the number of patients in the system, waiting time estimate and the production of outcomes) can be calculated for each health state, process state and may be aggregated for use at a service level.

The aims of this chapter are to:

1. Describe the stochastic system and the key patient flow dynamics of the model;
2. Develop the fluid and diffusion approximations for describing the expected behaviour of each queue, process state and health state;
3. Present methods for calculating system outputs and their variance.

5.1 Introduction

In this chapter, I develop a method for informing the planning and operation of networks of queues through a combination of outcome and operational measures. This framework has been informed by the work of chapters 2, 3 and 4; however, this model provides a theoretical method that may be applied to community services as well as other settings (such as telecommunications). Given the purpose and work presented within this thesis the formulation in this chapter is presented in reference to community health care.

In producing a modelling framework that includes these two perspectives on quality, new avenues for understanding and analysing systems of care are created such as the "flow of outcomes". This includes notions of how the health of patients with different capacities to benefit from care is affected by receiving, or not receiving care, and how patients with different care needs affect the operation of the system.

In chapters 3 and 4, several referral and patient flow dynamics were identified for community health care, namely uses of multiple services (either sequentially or concurrently), reuse of services and the use of services outside of the community setting. Whilst each needs to be considered when developing a completely descriptive model, this is not a simple task. Notably, the dynamics of re-entrant patients and concurrent uses of service can be methodologically difficult to model.

A range of methods could be used to model such systems including simulation, system dynamic approaches and Markov chain approaches. However, as the system becomes larger in considering multiple services and patient health, the problem becomes computationally expensive. Furthermore, including service reuse and health transitions within these models can be impractical due to the large number of events that need to be considered. Thus, given the multiplicative effect on the state space, the system can become very time consuming to model. There is a similar limitation in using traditional queueing methods since the linear algebra involved quickly becomes complex and even intractable [101]. Ultimately, computation time may not be an issue depending on the desired analysis and requirements of the model; however, this can reduce how and when the model may be used. For example, an analysis of a wide range of scenarios may be beneficial for an optimisation approach or scenario analysis, yet the running time of the above methods may limit such approaches.

One way to quickly model these types of system is by fluid and diffusion approximation, which I develop in this chapter. Notably, these methods are valid only under strict Markovian assumption that may not correspond with real-life data.

By extending current methods, I develop a model that includes several complex dynamics, namely the potential for patients to: reuse services; use different services sequentially; abandon the queue; and, having abandoned, rejoin the queue or use an alternative service. These methods will give insight into the impact of multiple care interactions on patient health. Furthermore, they may be used to inform resource

allocation and referral protocols within complex systems through new metrics that provide clinical and operational insight.

In chapter 4, I found that there is both a lack of data and the required mechanisms for measuring clinical outcomes across diverse community services. As a result, to include clinical outcomes within the model, I use a theoretical framework similar to that of [43]. I define multiple distinct health states, which patients may move between throughout a course of care, to represent the severity of patient health and their capacity to benefit from service.

To overcome some of the limitations introduced by multiple health states, I use a process for continuously allocating servers across multiple queues. This is a form of server allocation for multiple classes of patient where servers are allocated to queues depending on specific characteristics - for example, a proportional allocation that reflects the proportion of total demand each patient group represents. An important distinction of this process is that the server allocation updates continuously throughout the modelled time period, with queues gaining and losing servers to one another in response to changes in patient demand. Thus, this approach is a dynamic method for modelling the demand driven operational response of services to fluctuations between different streams of arrivals.

As I develop these methods, I will proceed without application to community health care. For now, I will focus on how these approximations are formulated, their mathematical validity and how they may be used to produce informative performance measures. Having established the above, in chapter 6 I will explore how these methods may be used to model patient flow in community health care.

Structure of the chapter

In section 5.2, I provide an introduction to fluid and diffusion approximations, noting what they are, how they are formulated and why they are used. I also provide

a brief survey of previous applications of fluid and diffusion approximations in health care. In section 5.3, I describe the stochastic system for which I develop the fluid and diffusion approximations and define the key flow parameters. In section 5.4 and 5.5, I set up the system and produce the fluid limit, the diffusion limit and a method for calculating virtual waiting time (VWT) - an estimate of waiting time. Throughout, I provide proof and mathematical argument as to why the extensions hold. This chapter ends with a discussion of the developed methods, highlighting the limitations of the approach and areas for future work.

Key terms

Before continuing, I will first define several terms that I use throughout the remainder of this thesis. Firstly, I reserve the term *rejoin* to refer to patients who abandon a queue and subsequently join it again. Secondly, I reserve the term *reuse* to refer to patients who, after completing service, seek to use the same service again. The term *re-enter* may be used to describe either process. Thirdly, I will discuss when a system is *heavily loaded* - informally, when demand exceeds the capability of a service to serve patients, which, if sustained, may lead to infinitely long average queue lengths. Finally, I will discuss instances when the system is *underloaded*, *overloaded* or *critically loaded*. Notably, these are widely used within the queueing literature and are used here with respect to the fluid scale. A system is *underloaded* if there is no queue and there are servers available to serve new arrivals; whilst a system is *overloaded* when all of the servers are busy and a queue has formed. Furthermore, the system is *critically loaded* when all servers are busy, but no one is queueing.

5.2 Introduction to fluid and diffusion approximations for stochastic systems

In chapter 2, I discussed [73] in which a fluid model was used to represent a transplant waiting list. Thus, a continuous, deterministic representation of the system's state variables was considered rather than the traditional, discrete, stochastic representation.

The authors used transitions in health states to model the effect of service, such that, upon receiving a transplant, a patient could change health state as they departed the waiting list. Additionally, the model included a mechanism of reuse where patients could experience a graft failure causing them to re-enter the waiting list. These dynamics are similar to some of those I seek to model; in particular, that a patient's use of service may result in a change of health, and the possibility for patients to reuse a service depending on their health. However, this method is not entirely appropriate for modelling community health care since there is no consideration of the stochastic variation within the arrival and service processes, which are important features of patient flow in this setting.

Exploring the merits of fluid models further, I found several methods and applications where fluid approximations had been formed for stochastic queueing systems. A fluid approximation is the limit in distribution for a stochastic process that can be calculated by scaling the size of the system (number of servers and new arrivals) and applying the law of large numbers. The system can then be modelled by a coupled system of ordinary differential equations (ODEs) that tracks the average state of the network i.e. the number of patients in the system over time [102]. This produces a continuous approximation of the discrete process and overcomes some of the computational difficulty of traditional methods. For example, in traditional state-based probabilistic methods the probability distribution is calculated over the entire state

space; however, in a fluid approximation, the modelling process is simplified through an abstract state representation based on state variables [103].

Fluid approximations can be used to model a variety of dynamics and produce several useful output measures [104]. However, at first order, they are a deterministic approximation and do not capture the variance seen within the system. To gain information about the variance, a diffusion approximation can be produced through an application of the functional central limit theorem to the scaled process [102]. The variance within the queueing process can then be calculated using a system of ODEs that results from the limit, adding insight into the system's stochastic variability.

By describing the system in this way, fluid and diffusion approximations provide a means for efficient calculation of measures of complex queueing processes. The approach avoids state space explosion in the analysis of large systems [101], which is a useful property as I come to consider multiple services and health states. Noted within the literature, fluid and diffusion limits are particularly accurate for large and heavily loaded systems due to the scaling process used [102, 105]. Finally, these approximations have been shown to be appropriate for analysing the behaviour of time-varying systems, and for understanding the finite-horizon evolution of systems in steady-state [106].

Work by Mandelbaum A et al. [104, 107], has been significant in developing my thoughts throughout this chapter. In these articles, methods are established for a range of systems that exhibit complex flow dynamics such as abandonment and rejoin. Their work highlights that under certain conditions, fluid and diffusion approximations can accurately model the queueing processes of stochastic systems.

In addition, the work of S Ding et al. [105] has informed my work in this chapter. They produce a first order fluid approximation to describe the operation of a call centre, represented as a single service with multiple servers. In this system customers could: call and be served if a server is free; wait within a queue until a server

is free; abandon the queue due to impatience; potentially rejoin the queue after abandoning; or reuse the service again having completed service. Notably, this model introduces the concept of *process orbits* - the flow of patients through *process states* that represent a customer's pre-service/post-service/post-abandonment flow, e.g. the states in which customers who are seeking to rejoin the queue or reuse the service wait within before re-entering the queue. I will continue to use this terminology as I present my method.

5.2.1 Applications of fluid and diffusion approximations within health care

To illustrate previous applications of fluid limits to health care services, I briefly discuss four papers. Each paper applies to acute settings, with two focusing on emergency departments [106, 108], and the other two producing more general methods [101, 109].

Firstly, in [106] several methods were presented for modelling queues with reuse. Applied to an emergency department during a mass casualty event, fluid and diffusion limits were produced to understand scenarios where services temporarily became heavily loaded. The focus in this paper was to determine the required staffing levels for meeting the demand of new and reusing patients within the heavily loaded intervals. The approximations were appropriate for modelling this system since they captured the time varying nature of arrivals and service, as seen within their data.

Secondly, again in application to emergency care [108], a fluid approximation was used to inform staffing levels within an overcrowded emergency department. The authors devised new rules and algorithms for staffing policies, seeking to minimise patient delays and staffing costs. In the model, patients were split into two states representing those who were "sick" and those who were "well", with "sick" patients

entering the "well" state after receiving care. In addition, after completing service, patients waited to be discharged, placing a further constraint on resources since this discharge process was carried out by servers who would otherwise treat "sick" patients. Whilst this method introduced a concept of health states, wellness in this model was measured by a process outcome since all patients who received service became "well". Hence, there was no variation in, or measurement of, clinical benefit. Rather, the process outcome of "receiving care" was assumed to have a perfect, positive effect on patient health.

Thirdly, in [101] a fluid limit was used to model the flow of patients within a hospital in order to inform the creation of a dynamic scheduling policy that could improve patient flow. In using a fluid approximation, the authors noted that their solution was scalable, aiding its application to, and analysis of, different systems.

Finally, in [109] a fluid approximation was used to assess a system where patients chose which queue to join based on waiting time information. Waiting times were provided to patients upon arrival; however, a delay existed in providing this information. Thus, the waiting time information they were given reflected the state of the queue some time before their arrival. Using theory from dynamical systems to assess how the queues interacted, the analysis focussed on the behaviour of these queues in response to delays. The aim was to understand how services could provide customers with waiting time information in order to avoid unwanted system dynamics.

5.2.2 Contribution

As far as I am aware, this is the first piece of work to consider the application of fluid and diffusion approximations to community health care and their key flow dynamics. In seeking to produce a method for modelling some aspects of patient flow within community health care, I extend the methods set out in [105] and [107]. In particular, I combine the approaches to capture all the dynamics considered by [105],

whilst producing the range of outputs given by [107]. Thus, I formulate fluid and diffusion limits to calculate the average and variance of system outputs, including the virtual waiting time (an estimation of waiting time) for the system.

An additional distinction of these methods is the extension to multiple services, the inclusion of transitions between health states, and a generalisation of the rejoin process where patients may use an alternative service after abandonment. Thus, the methods make a methodological contribution to the way in which networks of queues and complex flow dynamics may be modelled, and how the performance of a queueing system may be understood given the combination of patient flow and outcomes. In particular, this framework can be used to conduct analysis of time varying systems, where parameters are dependent on both time and patient health. By using health states the flow of patients with differing resource/service requirements and different capacities to benefit from care. Thus, it makes a contribution to the possible uses and applications for understanding the "flow of outcomes". The model's output is informed by the effect of care, or absence of it, on patient health and the effect of patients with different health care requirements, e.g. service times, on the operation of the system. This is highlighted by the production measure, section 5.5.5.

Throughout this chapter I provide formal proof and argument to show that these extensions are mathematically valid. The method I produce provides a framework for modelling community services quickly, whilst capturing key dynamics and producing informative outputs. The speed of computation is a significant benefit of these methods. If the methods can provide results comparable to others, such as simulation and Markov chain approaches, they may be used in to perform analyses that is otherwise potentially too time consuming, such as optimisation and scenario analysis, for large complex systems. This will be addressed further in chapter 6 where the parameter space is examined to understand when these methods are accurate and valid.

5.3 Description of the stochastic system

Consider a network consisting of J services, each with multiple servers. During a time interval $[0, T]$, patients may arrive to, and be served within, any one of the services, as in Figure 5.1. In each service patients may: arrive as a new patient; abandon the queue and potentially rejoin it, seek to use an alternative service or leave the system as a loss (L); or, receive service and potentially reuse the same service, use another service within the network, or leave as a discharge (D), see Figure 5.2. Thus, each service consists of five process orbits: $m \in \{Q, R, U, A, O\}$, representing the service and queue, the rejoin process, the reuse process, the alternative service process, and the other service process, respectively. Note that the term *alternative service* always refers to a use of service after abandonment, and that *other service* refers to a use of service having previously received care.

Suppose that at any one time, a patient belongs to a health state denoted $k \in \{1, ..., K\} = H$, the set of all health states. Each represents a category/amalgam of progressive outcome measure that patients move between as they proceed through the system. To denote each state that a patient may occupy, I use the following notation. For a service $i \in \{1, ..., J\} = Ser$, health state $k \in H$, at time $t \in [0, T]$:

$Z_{k,Q,i}(t) \coloneqq$ number of patients in the queue or service

$Z_{k,R,i}(t) \coloneqq$ number of patients in the rejoin orbit

$Z_{k,U,i}(t) \coloneqq$ number of patients in the reuse orbit

$Z_{k,A,i}(t) \coloneqq$ number of patients in the alternative service orbit

$Z_{k,O,i}(t) \coloneqq$ number of patients in the other service orbit

$Z_{k,L,i}(t) \coloneqq$ number of patients lost due to abandonment

$Z_{k,D,i}(t) \coloneqq$ number of patients discharged

I now describe the flow process in the system and the parameters that govern it. Explicitly defined in Table 5.1, each parameter is required to be locally integrable throughout the modelled time period. Each parameter is also required to be continuous throughout the modelled time period if the dynamic server allocation introduced in 5.3.1 is used and the VWT is calculated. This will become clearer in section 5.5.4.

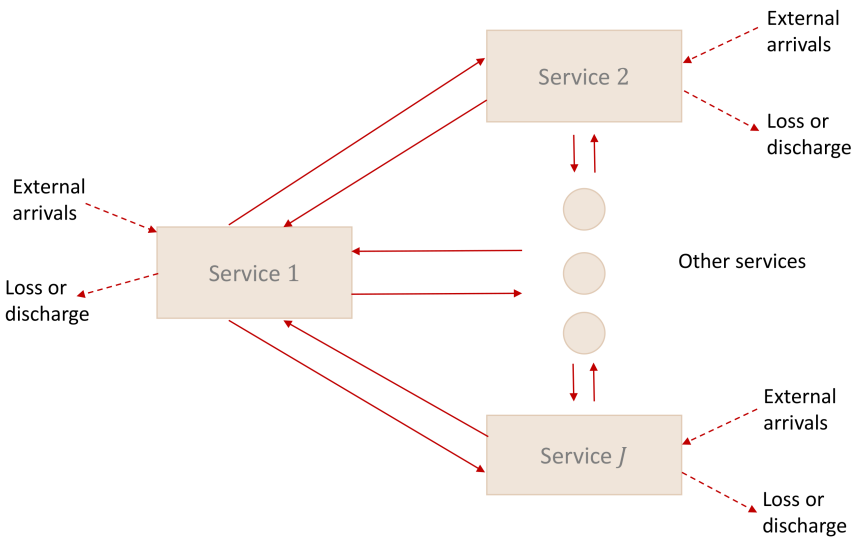

Figure 5.1: Diagram of patient flow between services within the queueing network

New patients enter services from external sources, and may reuse the same service, use another service, or leave the system due to abandonment or discharge

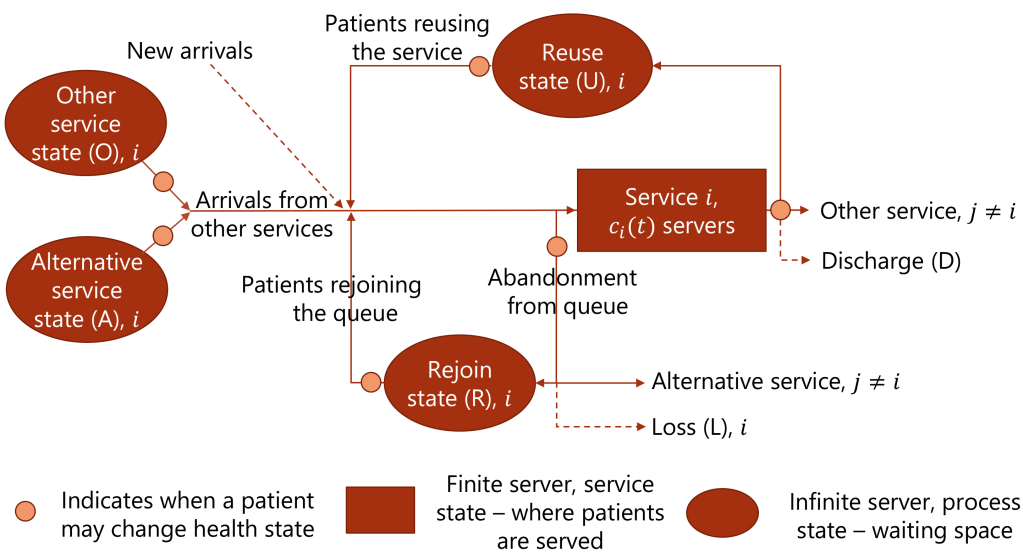

Figure 5.2: Diagram of a single service within the stochastic queueing network

$c_i(t)$	Number of servers available for service i.
$C_{k,i}(t)$	The number of servers allocated to each queue. $\sum_{k\in H} C_{k,i}(t) = c_i(t)$
$\lambda_{k,i}(t)$	Arrival rate of new patients, time-inhomogeneous Poisson process.
$\mu_{k,i}(t)$	Service rate of patients, where $\Gamma_{k,S,i}(t)$ is a time-inhomogeneous exponentially distributed service time with mean $\mathbb{E}\Gamma_{k,S,i}(t) = 1/\mu_{k,i}(t) < \infty$.
$\theta_{k,i}(t)$	Abandonment rate of patients, where $\Gamma_{k,L,i}(t)$ is a time-inhomogeneous exponentially distributed amount of time with mean $\mathbb{E}\Gamma_{k,L,i}(t) = 1/\theta_{k,i}(t) < \infty$.
$s_{k,l,m,i}(t)$	Probability that a patient moves from a health state $k \in H$ to a health state $l \in H$ given that they are leaving the orbit $m \in \{S, R, U, A, O\}$ or queue $m = L$.
$r_{k,L,i,j}(t)$	Probability that a patient, having abandoned the queue for service $i \in Ser$, enters the alternative service orbit for $j \in Ser, j \neq i$. For $j = i$ a patient enters the rejoin orbit of i. $r_{k,L,i,J+1}(t)$ denotes a loss from service $i \in Ser$
$r_{k,s,i,j}(t)$	Probability that a patient completes service within $i \in Ser$ and enters the orbit for arrivals from other services for $j \in Ser, j \neq i$. For $j = i$ a patient enters the reuse orbit of i. $r_{k,S,i,J+1}(t)$ denotes a discharge from service $i \in Ser$
$\delta_{k,R,i}(t)$	Rate of rejoin, where $\Gamma_{k,R,i}(t)$ is a time-inhomogeneous exponentially distributed amount of time until patients re-enter the queue with mean $\mathbb{E}\Gamma_{k,R,i}(t) = 1/\delta_{k,R,i}(t) < \infty$.
$\delta_{k,A,i}(t)$	Rate of alternative service use, where $\Gamma_{k,A,i}(t)$ is a time-inhomogeneous exponentially distributed amount of time until patients enter the queue with mean $\mathbb{E}\Gamma_{k,A,i}(t) = 1/\delta_{k,A,i}(t) < \infty$.
$\delta_{k,U,i}(t)$	Rate of reuse, where $\Gamma_{k,U,i}(t)$ is a time-inhomogeneous exponentially distributed amount of time until patients re-enter the queue with mean $\mathbb{E}\Gamma_{k,U,i}(t) = 1/\delta_{k,U,i}(t) < \infty$.
$\delta_{k,O,i}(t)$	Rate of other service use, where $\Gamma_{k,O,i}(t)$ is a time-inhomogeneous exponentially distributed amount of time until patients enter the queue with mean $\mathbb{E}\Gamma_{k,O,i}(t) = 1/\delta_{k,O,i}(t) < \infty$.

Table 5.1: Parameter definitions for stochastic system $t \in [0, T],\ k \in H,\ i \in Ser$

Servers

For each service $i \in Ser$, a time-varying number of servers, $c_i(t)$, are available. For analytical tractability, patients are assumed to be served on a first come first served (FCFS) basis within their respective health states, forming up to K parallel queues per service. Notably, health states can be defined such that patients have equal priority within their group; thus, FCFS is a natural approach to take. The number of servers allocated to each queue is denoted $C_{k,i}(t)$ such that, for all $k \in H$ at time t, $\sum_{k \in H} C_{k,i}(t) = c_i(t)$. Several definitions of $C_{k,i}(t)$ are given in section 5.3.1.

Since the number of servers for each service (and queue) may vary with time, it is possible that the number of servers may drop below the number of patients in service. Within the model, this situation is handled by using pre-emptive resumption [107]. That is, the number of patients in service is reduced to equal the number of servers by placing arbitrary "excess" patients into an infinite buffer space. The service of these patients is assumed to be paused and later resumed once a server becomes available, with priority ahead of the queue.

New arrivals

New patients, in a health state $k \in H$, arrive at a service $i \in Ser$ according to a time-inhomogeneous Poisson process of rate $\lambda_{k,i}(t)$.

Queue and service

If a server is free, patients enter service and are served according to a time-inhomogeneous exponentially distributed process of rate $\mu_{k,i}(t)$. If no servers are available, the arriving patient will wait within an infinite buffer space, forming a non-physical queue. Whilst waiting, a patient may lose patience and abandon the queue at a time-inhomogeneous exponentially distributed rate $\theta_{k,i}(t)$.

Abandonment

Upon abandoning, one of three events may occur. A patient may rejoin the queue (seeking to access the service again) with probability $r_{k,L,i,i}(t)$, where the subscript L denotes a patient leaving the queue. Such patients enter the rejoin orbit, where they spend a time-inhomogeneous exponentially distributed amount of time, rejoining the queue at rate $\delta_{k,R,i}(t)$. Alternatively, a patient may seek to use an alternative service with probability $r_{k,L,i,j}(t),\ i \neq j,\ j \in Ser$. These patients enter the alternative service orbit for j, where they spend a time-inhomogeneous exponentially distributed amount of time, joining the queue at a rate $\delta_{k,A,j}(t)$. The third possibility is that a patient will leave the system as a loss with probability $r_{k,L,i,J+1}(t) > 0$. Notably: $\sum_{j=1}^{J+1} r_{k,L,i,j}(t) = 1,\ \ \text{for all} t \in [0,T]$.

Completing service

Similarly, after completing service, one of three events may occur. Having used a service i, a patient may seek further service within i with probability $r_{k,S,i,i}(t)$, entering the reuse orbit. Again, patients spend a time-inhomogeneous exponentially distributed amount of time in this state, arriving to the service at rate $\delta_{k,U,i}(t)$. Alternatively, with probability $r_{k,S,i,j}(t),\ i \neq j,\ j \in Ser$ a patient may seek to use another service, entering the orbit of arrivals from other services for j, remaining in this state for a time-inhomogeneous exponentially distributed amount of time, and join the queue at a rate $\delta_{k,O,j}(t)$. Lastly, there is a probability $r_{k,S,i,J+1}(t) > 0$ that a patient will not require any further service and leave the system as a discharge. Notably: $\sum_{j=1}^{J+1} r_{k,S,i,j}(t) = 1,\ \ \text{for all } t \in [0,T]$.

Changes in health state

A patient's health state may change throughout their interaction with the system and is modelled to occur at: the completion of service; the point of abandoning the

queue; or, upon joining the queue as a rejoin, reuse, alternative service arrival or other service arrival as in Figure 5.2. $s_{k,l,m,i}(t)$ is the probability that a patient moves from a health state $k \in H$ to a health state $l \in H$ given that they are leaving a process state $m \in \{S, R, U, A, O\}$ at time t, or abandoning the queue, $m = L$.

The stochastic process

Given the above, the stochastic process for this system, $\{\mathbf{Z}(t), t \geq 0\}$, can be defined as a vector of length $7KJ$ (since there are seven process orbits, K health states and J services) such that:

$$\mathbf{Z}(t) \coloneqq (\mathbf{Z}_{1,1}(t), \mathbf{Z}_{2,1}(t), ..., \mathbf{Z}_{K,1}(t), \mathbf{Z}_{1,2}(t), ..., \mathbf{Z}_{K,2}(t), ..., \mathbf{Z}_{K,J}(t))^T \tag{5.1}$$

where, for $k \in H, i \in Ser$:

$$\mathbf{Z}_{k,i}(t) \coloneqq (Z_{k,Q,i}(t), Z_{k,R,i}(t), Z_{k,U,i}(t), Z_{k,A,i}(t), Z_{k,O,i}(t), Z_{k,L,i}(t), Z_{k,D,i}(t))$$

This is a Markov process since the inter-arrival rates, service duration and orbit durations are exponentially distributed, and health/service state transitions are Markovian. The state space for this process is $\mathbb{Z}_+^{7KJ}$, which is the space of length $7KJ$ vectors whose entries are 0 or positive integers.

5.3.1 Dynamic multi-class server allocations

To ensure analytical tractability whilst modelling the differentiated service of patients in different health states, each service is modelled using parallel queues pertaining to each health state. Thus, patients in each queue seek service from a single pool of servers. To maintain the FCFS assumption, servers must be allocated to each queue.

The simplest allocation is to divide the number servers across parallel queues, assigning a constant number to each. Considering the simplest, yet unrealistic, illustration of assigning servers equally highlights an issue, that care must be taken when K is not a factor of $c_i(t)$. To overcome this scenario, define: $C_{k,i}(t) = \left\lfloor \frac{c_i(t)}{K} \right\rfloor$, the floor of the fraction. If $\sum_{k=1}^{K} C_{k,i}(t) < c_i(t)$, assign $c_i(t) - \sum_{k=1}^{K} C_{k,i}(t)$ servers, one at a time, to arbitrary queues until all servers are assigned. Notably, in scenarios when $C_{k,i}(t)$ does not depend on the output of the stochastic system (e.g. a constant function) the input parameters may be defined as piecewise continuous [105].

In using constant allocations, the only interaction between the queues is through the health state transitions of patients, otherwise the queues act autonomously. However, in real world systems, a further way in which the queues may affect one another is through a patient's use of servers, i.e. by using a server, a patient denies others the opportunity to be served by that server.

One way to model this is through a dynamic server allocation. There is a wide and extensive literature that is relevant to this type of allocation such as that on allocating servers in multi-class queues e.g. [110, 111, 112] and scenarios of server sharing e.g. [113, 114]. Here I apply an allocation which continually updates according to the changes in the overall demand for service, the attributes of different patient groups and the mix of patients. Thus, a queue with more "dominant" attributes, e.g. queues with higher proportion of overall demand or a longer proportional service times, may require more servers throughout the modelled time period than other queues. Within the stochastic system the number of servers allocated to each queue is updated each time an event occurs that changes the size of $Z_{k,Q,i}(t)$ e.g. an arrival (new or from a process state), a completion of service or an abandonment. Thus, the fluid approximation will provide a deterministic approximation of this process.

One method for allocating servers in this way is to assign them to each queue according to the proportion of patients in each health state k and in the queue or

service for a service i:

$$C_{k,i}(t) = C_{k,i}(\mathbf{Z}(t)) = \left\lfloor \frac{c_i(t) Z_{k,Q,i}(t)}{\sum_{l=1}^{K} Z_{l,Q,i}(t)} \right\rfloor \tag{5.2}$$

Again, the method introduced earlier is implemented when the number of servers does not divide into an integer. Furthermore, to ensure that this allocation may always be calculated, the system must never become empty; thus, the initial condition must be non-empty. As seen in the next chapter, this is not an issue.

Alternatively, a continuous weight or cost function, $B_{k,i}(t)$, could be used to favour patients in certain health states. For example, $B_{k,i}(t)$ may be defined as $1/\mu_{k,i}(t)$. In this case, servers are allocated to the queues that will take the longest time to serve. Furthermore, if $B_{k,i}(t) = \theta_{k,i}(t)$, servers are allocated based on the potential for patients to abandon, seeking to mitigate losses in the system. Thus, for the stochastic process, one may allocate servers by:

$$C_{k,i}(\mathbf{Z}(t)) = \left\lfloor \frac{c_i(t) B_{k,i}(t) Z_{k,Q,i}(t)}{\sum_{l=1}^{K} B_{l,i}(t) Z_{l,Q,i}(t)} \right\rfloor, \text{ for all } t \in [0, T] \tag{5.3}$$

Again, the method introduced above may be implemented to ensure that all the servers are allocated; the system must also never become empty. A further limitation of these allocations is that their equivalent result from fluid approximation must be continuously differentiable, a limitation introduced by the calculation of the VWT. Notably each of the above allocations depends on $Z_{k,Q,i}(t)$ which is dependent on $C_{k,i}(t)$ by definition. This is not a limitation however since a fall in the number of allocated servers leads to patients formerly in service re-entering the queue such that $Z_{k,Q,i}(t)$is unchanged. Furthermore, allocations may be defined based on different orbits or process states as long as the definition remains continuously differentiable.

These allocations may be used to understand how the service requirements of patients in different health states and fluctuations in demand may affect the operation

of the system. Likewise, this method helps to overcome the limitations that parallel queues introduce; namely, the possibility that patients in some health states may be waiting, yet servers assigned to other queues are inactive. By modelling a dynamic allocation process, servers are reallocated to queues as changes occur in the system such that if there are any patients waiting in the system, it is not possible for servers to become inactive. Furthermore, this method may be used to understand how the allocation of servers may help to mitigate negative process outcomes (such as the abandonment of patients in poor health states).

5.4 Set up for fluid and diffusion approximations

To proceed, conservation equations must be formulated for the stochastic system (5.1). These are constructed in a similar manner to [105], extending their definitions to include multiple health states, dynamic server allocations, multiple services and the new orbits these introduce.

Aside 5.4.1 (A sketch to show how the fluid approximation is constructed) *Having outlined the system and flow dynamics, the mathematical mechanisms are defined for modelling how patients move between process states and health states - the flux terms. These are Poisson processes of rate 1, where "time" is the number of patients flowing through the process state (the reason for this will become clear).*

A scaled process is produced for this system by scaling the number of servers and arrivals for each service by a factor $\eta > 0$, and dividing the number of patients in each state by η. By the definition of the flux terms, as $\eta \to \infty$, the law-of-large-numbers gives the limit of these terms as the mean number of patients flowing through each part of the system.

Informally, as η grows large, the system becomes more deterministic. This is because the arrival rate grows larger, increasing the probability of patients arriving, causing the system to behave more predictably (e.g. it is more likely that all servers are busy, that patients will abandon the queue, etc.).

This approximation is continuous and the limit gained from this scaled process is a deterministic approximation for the stochastic system, which is shown to be mathematically valid and unique.

The flux terms for modelling the movement of patients must first be defined for each health state $k, l \in H$ and for each service $i, j \in Ser$. The arrival process of new patients is $\Pi_{\lambda_{k,i}(t)}$, a Poisson process of rate $\lambda_{k,i}(t)$. The number of patients leaving process states - service (S), abandonment from queue (L), rejoin (R), reuse (U), alternative service (A) and arrivals from other services (O) - are defined as $\Pi_{k,m,i}(.),\ m \in \{S, L, R, U, A, O\}$ independent Poisson processes of rate 1 such that:

$$D_{k,S,i}(\mathbf{Z}(t)) = \Pi_{k,S,i}\left(\int_0^t \mu_{k,i}(u) \min\left(Z_{k,Q,i}(u), C_{k,i}(\mathbf{Z}(u))\right) du\right) \tag{5.4}$$

$$D_{k,L,i}(\mathbf{Z}(t)) = \Pi_{k,L,i}\left(\int_0^t \theta_{k,i}(u)(Z_{k,Q,i}(u) - C_{k,i}(\mathbf{Z}(u)))^+ du\right) \tag{5.5}$$

$$D_{k,R,i}(\mathbf{Z}(t)) = \Pi_{k,R,i}\left(\int_0^t \delta_{k,R,i}(u) Z_{k,R,i}(u) du\right) \tag{5.6}$$

$$D_{k,U,i}(\mathbf{Z}(t)) = \Pi_{k,U,i}\left(\int_0^t \delta_{k,U,i}(u) Z_{k,U,i}(u) du\right) \tag{5.7}$$

$$D_{k,A,i}(\mathbf{Z}(t)) = \Pi_{k,A,i}\left(\int_0^t \delta_{k,A,i}(u) Z_{k,A,i}(u) du\right) \tag{5.8}$$

$$D_{k,O,i}(\mathbf{Z}(t)) = \Pi_{k,O,i}\left(\int_0^t \delta_{k,O,i}(u) Z_{k,O,i}(u) du\right) \tag{5.9}$$

Note: $(x)^+ \coloneqq \max(0, x)$. Proof of these statements may be produced along the lines of Lemma 2.1 in [115].

Multinomial random variables - a generalisation of the binomial distribution (see [116] for more information) - are used to model the movement of patients between health states. For patients who transition to a new process state according to

$D_{k,m,i}(\mathbf{Z}(t)),\ m \in \{S, L, R, U, A, O\}$, a change in health is modelled by:

$$\mathbf{MS}_{k,m,i}(\mathbf{Z}(t)) \sim \text{Mult}(D_{k,m,i}(\mathbf{Z}(t)), \mathbf{s}_{k,m,i}(t)) \tag{5.10}$$

where $\mathbf{MS}_{k,m,i}(\mathbf{Z}(t))$ is a vector of length K. Its l-th element, denoted $MS^{(l)}_{k,m,i}(\mathbf{Z}(t))$, gives the number of patients who were in health state k before moving to health state l, having entered a new process state according to $D_{k,m,i}(\mathbf{Z}(t))$. This process is governed by health state transition parameters:

$$\mathbf{s}_{k,m,i}(t) = (s_{k,1,m,i}(t), s_{k,2,m,i}(t), ..., s_{k,K,m,i}(t))$$

where $\sum_{l=1}^{K} s_{k,l,m,i}(t) = 1$ such that $\sum_{l=1}^{K} MS^{(l)}_{k,m,i}(\mathbf{Z}(t)) = D_{k,m,i}(\mathbf{Z}(t))$.

Again, multinomial random variables are used to model the movement of patients after abandoning the queue. For patients who, upon abandoning the queue for $i \in Ser$, have moved to a health state k, $\sum_{l=1}^{K} \mathbf{MS}^{(k)}_{l,L,i}(\mathbf{Z}(t))$, their post abandonment movement is modelled by:

$$\mathbf{MR}_{k,L,i}(\mathbf{Z}(t)) \sim \text{Mult}\left(\sum_{l=1}^{K} \mathbf{MS}^{(k)}_{l,L,i}(\mathbf{Z}(t)), \mathbf{r}_{k,L,i}(t)\right) \tag{5.11}$$

where $\mathbf{MR}_{k,L,i}(\mathbf{Z}(t))$ is a vector of length $J+1$. Its j-th element, denoted $MR^{(j)}_{k,L,i}(\mathbf{Z}(t))$, gives the number of health state k patients who enter the alternative service orbit for $j \in Ser,\ j \neq i$; or, for $j = i$, enter the rejoin orbit of i; or, for $j = J + 1$, leave the system. This process is governed by post-abandonment transition parameters:

$$\mathbf{r}_{k,L,i}(t) = (r_{k,L,i,1}(t), r_{k,L,i,2}(t), ..., r_{k,L,i,J}(t), r_{k,L,i,J+1}(t))$$

where: $\sum_{j=1}^{J+1} r_{k,L,i,j}(t) = 1$ such that $\sum_{j=1}^{J+1} MR^{(j)}_{k,L,i}(\mathbf{Z}(t)) = \sum_{l=1}^{K} \mathbf{MS}^{(k)}_{l,L,i}(\mathbf{Z}(t))$.

Finally, the movement of patients after completing service is modelled in a similar

way. For patients who, upon completing service in $i \in Ser$ have moved to health state k, $\sum_{l=1}^{K} \mathbf{MS}_{l,S,i}^{(k)}(t)$, their post-service movement is modelled by:

$$\mathbf{MR}_{k,S,i}(\mathbf{Z}(t)) \sim \text{Mult}\left(\sum_{l=1}^{K} \mathbf{MS}_{l,S,i}^{(k)}(\mathbf{Z}(t)), \mathbf{r}_{k,S,i}(t)\right) \tag{5.12}$$

where $\mathbf{MR}_{k,S,i}(\mathbf{Z}(t))$ is a vector of length $J+1$. Its j-th element, denoted $MR_{k,S,i}^{(j)}(\mathbf{Z}(t))$, gives the number of health state k patients who enter the orbit for arrivals from others services for $j \in Ser,\ j \neq i$; or, for $j = i$, enter the reuse orbit of i; or, for $j = J+1$, are discharged after service in i. This process is governed transition parameters:

$$\mathbf{r}_{k,S,i}(t) = (r_{k,S,i,1}(t), r_{k,S,i,2}(t), ..., r_{k,S,i,J}(t), r_{k,S,i,J+1}(t))$$

where $\sum_{j=1}^{J+1} r_{k,S,i,j}(t) = 1$ such that $\sum_{j=1}^{J+1} MR_{k,S,i}^{(j)}(\mathbf{Z}(t)) = \sum_{l=1}^{K} \mathbf{MS}_{l,S,i}^{(k)}(\mathbf{Z}(t))$.

Given the flux terms detailed above, the conservation equations for patient flow in the stochastic system (5.1), for $t \in [0, T)$, are for $k, l \in H$ and $i, j \in Ser$:

$$\begin{aligned} Z_{k,Q,i}(t) = Z_{k,Q,i}(0) + \Pi_{\lambda_{k,i}(t)} &+ \sum_{l=1}^{K} MS_{l,R,i}^{(k)}(\mathbf{Z}(t)) + \sum_{l=1}^{K} MS_{l,U,i}^{(k)}(\mathbf{Z}(t)) \\ &+ \sum_{l=1}^{K} MS_{l,A,i}^{(k)}(\mathbf{Z}(t)) + \sum_{l=1}^{K} MS_{l,O,i}^{(k)}(\mathbf{Z}(t)) \\ &- D_{k,S,i}(\mathbf{Z}(t)) - D_{k,L,i}(\mathbf{Z}(t)) \end{aligned} \tag{5.13}$$

$$Z_{k,R,i}(t) = Z_{k,R,i}(0) + MR_{k,L,i}^{(i)}(\mathbf{Z}(t)) - D_{k,R,i}(\mathbf{Z}(t)) \tag{5.14}$$

$$Z_{k,U,i}(t) = Z_{k,U,i}(0) + MR_{k,S,i}^{(i)}(\mathbf{Z}(t)) - D_{k,U,i}(\mathbf{Z}(t)) \tag{5.15}$$

$$Z_{k,A,i}(t) = Z_{k,A,i}(0) + \sum_{j=1;j\neq i}^{J} MR_{k,L,j}^{(i)}(\mathbf{Z}(t)) - D_{k,A,i}(\mathbf{Z}(t)) \tag{5.16}$$

$$Z_{k,O,i}(t) = Z_{k,O,i}(0) + \sum_{j=1;j\neq i}^{J} MR_{k,S,j}^{(i)}(\mathbf{Z}(t)) - D_{k,O,i}(\mathbf{Z}(t)) \tag{5.17}$$

$$Z_{k,L,i}(t) = Z_{k,L,i}(0) + MR_{k,L,i}^{(J+1)}(\mathbf{Z}(t)) \tag{5.18}$$

$$Z_{k,D,i}(t) = Z_{k,D,i}(0) + MR_{k,S,i}^{(J+1)}(\mathbf{Z}(t)) \tag{5.19}$$

5.5 Fluid and diffusion approximations for stochastic queueing networks with heterogeneous patients

To compute the fluid and diffusion approximations, an appropriate metric space is required to find the limits of the stochastic processes. Thus, the set in which the approximations are contained, and the metric used to measure distances between each set members need to be defined - the Skorokhod space and the J_1 topology respectively. Through careful definition important mathematical principles such as the convergence of sequences, continuity of functions and completeness can be formulated for the specific set.

5.5.1 Definitions

Definition 5.5.1 (Skorokhod space) ***Skorokhod space****, denoted* $D([0,T],\mathbb{R}^n)$*, where* $[0,T] \subset \mathbb{R}$ *and* $n \in \mathbb{N}$*, consists of right continuous functions* $x : [0,T] \to \mathbb{R}^n$ *that admit left limits* $x(t^-)$ *at each point* $t \in (0,T]$*.*

In [117] four metric topologies are proposed of which, for the fluid limit (as in [105]), I use the J_1 metric. In particular it provides a natural and convenient formalism for describing trajectories of stochastic processes that admit discontinuities, such as the trajectories of Poisson processes [118].

Definition 5.5.2 (Skorokhod J_1 topology) *Within* $D([0,T],\mathbb{R}^n), x_\eta \in D$ *converges to* $x_0 \in D$ *under* ***Skorokhod*** J_1 ***topology*** *if, for the family of increasing homeomorphisms* $\Lambda : E \to E$*,* $\exists$ *a sequence of functions* $\lambda_\eta \in \Lambda$ *such that:*

$$\sup_{t\in[0,T]} |\lambda_\eta(t) - t| \to 0 \text{ and } \sup_{t\in[0,T]} |x_\eta(\lambda_\eta(t) - x_0(t))| \to 0; \text{ as } \eta \to \infty.$$

Definition 5.5.3 (Converges in distribution) *Let* $\left\{X^{(\eta)}\right\}_{\eta=1}^{\infty}$ *be a sequence of random variables and denote the distribution function of each* $X^{(\eta)}$ *by* $F^{(\eta)}(t)$. $\left\{X^{(\eta)}\right\}_{\eta=1}^{\infty}$ ***converges in distribution*** *if and only if there exists a distribution function* $F(t)$ *such that the sequence* $\left\{F^{(\eta)}(t)\right\}_{\eta=1}^{\infty}$ *converges to* $F(t)$ *for all* $t \in [0, \infty)$*, where* $F(t)$ *is continuous. If a random variable* x *has distribution function* $F(t)$*, then* x *is called the* ***limit in distribution*** *of the sequence, denoted* $X^{(\eta)} \xrightarrow{d} x$*, with convergence:*

$$\lim_{\eta \to \infty} X^{(\eta)}(t) \stackrel{d}{=} x(t), \quad \textit{for all } t \in [0, \infty)$$

Definition 5.5.4 (Almost surely) *Let* $(\Omega, F, \mathbb{P})$ *be a probability space. An event* $E \subset F$ *occurs* ***almost surely*** *if* $\mathbb{P}(E) = 1$*, and* $\mathbb{P}(E^c) = 0$*, where* E^c *is the compliment of* E *(the event that* E *does not occur). This is denoted a.s.*

Note: The difference between "almost sure" and "sure" events is the same as the difference between an event that happens with probability 1 and one that always happens. An event that is "sure" will always happen, and any outcome not in this event cannot occur. For an event that is "almost sure", outcomes not in the event space are theoretically possible; however, the probability of them occurring is smaller than any fixed positive probability. One cannot definitively say that these outcomes will never occur, but for most purposes this can be assumed to be true.

5.5.2 Mathematical foundation for limiting theorems and the fluid approximation

Having described the stochastic system in section 5.4, I now formulate the fluid limit for equations (5.13)-(5.19). First, consider a sequence of models where the η-th model - denoted by the superscript (η) - has a scaled arrival rate $\eta\lambda_{k,i}(t)$ for new patients and scaled number of servers $\eta c_i(t)$ for all $k \in H$ and $i \in Ser$. The

scaled fluid process is defined as: $\overline{Z}^{(\eta)}_{k,m,i}(t) := \frac{Z^{(\eta)}_{k,m,i}(t)}{\eta}$, for $k \in H,\ i \in Ser,\ m \in \{S, R, U, A, O, L\}$ and $n \in \{S, L\}$. This is similar to (5.13)-(5.19), where (5.4)-(5.12) are replaced by:

$$\overline{D}^{(\eta)}_{k,S,i}(t) = \Pi_{k,S,i}\left(\eta \int_0^t \mu_{k,i}(u) \min\left(\overline{Z}^{(\eta)}_{k,Q,i}(u), C_{k,i}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right) du\right) \Big/ \eta \tag{5.20}$$

$$\overline{D}^{(\eta)}_{k,L,i}(t) = \Pi_{k,L,i}\left(\eta \int_0^t \theta_{k,i}(u) \left(\overline{Z}^{(\eta)}_{k,Q,i}(u) - C_{k,i}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)^+ du\right) \Big/ \eta \tag{5.21}$$

$$\overline{D}^{(\eta)}_{k,R,i}(t) = \Pi_{k,R,i}\left(\eta \int_0^t \delta_{k,R,i}(u)\overline{Z}^{(\eta)}_{k,R,i}(u) du\right) \Big/ \eta \tag{5.22}$$

$$\overline{D}^{(\eta)}_{k,U,i}(t) = \Pi_{k,U,i}\left(\eta \int_0^t \delta_{k,U,i}(u)\overline{Z}^{(\eta)}_{k,U,i}(u) du\right) \Big/ \eta \tag{5.23}$$

$$\overline{D}^{(\eta)}_{k,A,i}(t) = \Pi_{k,A,i}\left(\eta \int_0^t \delta_{k,A,i}(u)\overline{Z}^{(\eta)}_{k,A,i}(u) du\right) \Big/ \eta \tag{5.24}$$

$$\overline{D}^{(\eta)}_{k,O,i}(t) = \Pi_{k,O,i}\left(\eta \int_0^t \delta_{k,O,i}(u)\overline{Z}^{(\eta)}_{k,O,i}(u) du\right) \Big/ \eta \tag{5.25}$$

$$\overline{\mathbf{MS}}^{(\eta)}_{k,m,i}(t) = \frac{\mathbf{MS}^{(\eta)}_{k,m,i}(t)}{\eta},\ \mathbf{MS}^{(\eta)}_{k,m,i}(t) \sim \text{Mult}\left(\eta \overline{D}^{(\eta)}_{k,m,i}(t), \mathbf{s}_{k,m,i}(t)\right) \tag{5.26}$$

$$\overline{\mathbf{MR}}^{(\eta)}_{k,n,i}(t) = \frac{\mathbf{MR}^{(\eta)}_{k,n,i}(t)}{\eta},\ \mathbf{MR}^{(\eta)}_{k,n,i}(t) \sim \text{Mult}\left(\sum_{l=1}^{K} \mathbf{MS}^{(\eta)^{(k)}}_{l,n,i}(t), \mathbf{r}_{k,n,i}(t)\right) \tag{5.27}$$

This emits the scaled fluid process defined as:

$$\overline{\mathbf{Z}}^{(\eta)}(t) \coloneqq \left(\overline{\mathbf{Z}}^{(\eta)}_{1,1}(t), \overline{\mathbf{Z}}^{(\eta)}_{2,1}(t), ..., \overline{\mathbf{Z}}^{(\eta)}_{K,1}(t), \overline{\mathbf{Z}}^{(\eta)}_{1,2}(t), ..., \overline{\mathbf{Z}}^{(\eta)}_{K,2}(t), ..., \overline{\mathbf{Z}}^{(\eta)}_{K,J}(t)\right)^T \tag{5.28}$$

where, for $k \in H,\ i \in Ser$:

$$\overline{\mathbf{Z}}^{(\eta)}_{k,i}(t) \coloneqq \left(\overline{Z}^{(\eta)}_{k,Q,i}(t), \overline{Z}^{(\eta)}_{k,R,i}(t), \overline{Z}^{(\eta)}_{k,U,i}(t), \overline{Z}^{(\eta)}_{k,A,i}(t), \overline{Z}^{(\eta)}_{k,O,i}(t), \overline{Z}^{(\eta)}_{k,L,i}(t), \overline{Z}^{(\eta)}_{k,D,i}(t)\right)$$

From the definitions above, a fluid limit can be produced for (5.13)-(5.19). The formulation and proofs follow those set out by [105]; however, I make the non-trivial extensions to include health states, multiple services, the alternative service orbit and the orbit for arrivals from others services. The proofs differ from [105] by the

introduction of health states which add a complication when establishing the limiting behaviour; particularly the introduction of the multinomial distributions. Hence, the bounds in section (b) of the following proof differs and requires careful attention in order to show the limiting behaviour of the sequence. Furthermore the proofs have been extended to additional services and process orbits. Fundamentally, the proofs closely mirror those in [105], and are given to show the validity of the extensions.

Definition 5.5.5 (Fluid limit) *Let $D([0,\infty),\mathbb{R}^{7KJ})$ be the Skorokhod space of right continuous functions with the left limits in $\mathbb{R}^{7KJ}$ having the domain $[0,\infty)$, endowed with Skorokhod J_1 topology. Suppose $\left\{\mathbf{Z}^{(\eta)}\right\}_{\eta=1}^{\infty}$ is a sequence of stochastic processes within $D\left([0,\infty),\mathbb{R}^{7KJ}\right)$. If there exists a limit in distribution, for the scaled process $\left\{\overline{\mathbf{Z}}^{(\eta)}(.)\right\}_{\eta=1}^{\infty}$ such that $\overline{\mathbf{Z}}(.) \xrightarrow{d} \mathbf{z}(.)$, then $\mathbf{z}(.)$ is called the **fluid limit** of the original stochastic model.*

To formulate the limit for a time period $[0,T]$, $k \in H$ and $i \in Ser$, initial conditions are required: $(z_{k,Q,i}(0), z_{k,R,i}(0), z_{k,U,i}(0), z_{k,A,i}(0), z_{k,O,i}(0), z_{k,L,i}(0), z_{k,D,i}(0))$.

Theorem 5.5.1 *For a time period $[0,T]$, $k \in H$ and $i \in Ser$, assume that for $m = \{Q,R,U,A,O,L,D\}$, $\overline{Z}^{(\eta)}_{k,m,i}(0) \xrightarrow{d} z_{k,m,i}(0)$ as $\eta \to \infty$. Then the fluid limit of (5.1) is the unique solution to the following system of equations where $t \in [0,T)$:*

$$\begin{aligned} z_{k,Q,i}(t) = z_{k,Q,i}(0) &+ \int_0^t \lambda_{k,i}(u) + \sum_{l=1}^{K} s_{l,k,R,i}(u)\delta_{l,R,i}(u)z_{l,R,i}(u)\ du \\ &+ \int_0^t \sum_{l=1}^{K} s_{l,k,U,i}(u)\delta_{l,U,i}(u)z_{l,U,i}(u)\ du \\ &+ \int_0^t \sum_{l=1}^{K} s_{l,k,A,i}(u)\delta_{l,A,i}(u)z_{l,A,i}(u)\ du \\ &+ \int_0^t \sum_{l=1}^{K} s_{l,k,O,i}(u)\delta_{l,O,i}(u)z_{l,O,i}(u)\ du \\ &- \int_0^t \theta_{k,i}(u)(z_{k,Q,i}(u) - c_{k,i}(\mathbf{z}(u)))^{+}\ du \\ &- \int_0^t \mu_{k,i}(u)\min\left(z_{k,Q,i}(u), c_{k,i}(\mathbf{z}(u))\right)\ du \end{aligned} \tag{5.29}$$

$$z_{k,R,i}(t) = z_{k,R,i}(0) + \int_0^t r_{k,L,i,i}(u) \sum_{l=1}^{K} s_{l,k,L,i}(u)\theta_{l,i}(u)(z_{l,Q,i}(u) - c_{l,i}(\mathbf{z}(u)))^+ \, du - \int_0^t \delta_{k,R,i}(u) z_{k,R,i}(u) \, du \tag{5.30}$$

$$z_{k,U,i}(t) = z_{k,U,i}(0) + \int_0^t r_{k,S,i,i}(u) \sum_{l=1}^{K} s_{l,k,S,i}(u)\mu_{l,i}(u) \min\left(z_{l,Q,i}(u), c_{l,i}(\mathbf{z}(u))\right) du - \int_0^t \delta_{k,U,i}(u) z_{k,U,i}(u) du \tag{5.31}$$

$$z_{k,A,i}(t) = z_{k,A,i}(0) + \int_0^t \sum_{j=1;j\neq i}^{J} \sum_{l=1}^{K} r_{k,L,j,i}(u) s_{l,k,L,j}(u)\theta_{l,j}(u)(z_{l,Q,j}(u) - c_{l,j}(\mathbf{z}(u)))^+ du - \int_0^t \delta_{k,A,i}(u) z_{k,A,i}(u) \, du \tag{5.32}$$

$$z_{k,O,i}(t) = z_{k,O,i}(0) + \int_0^t \sum_{j=1;j\neq i}^{J} \sum_{l=1}^{K} r_{k,S,j,i}(u) s_{l,k,S,j}(u)\mu_{l,j}(u) \min\left(z_{l,Q,j}(u), c_{l,j}(\mathbf{z}(u))\right) du - \int_0^t \delta_{k,O,i}(u) z_{k,O,i}(u) du \tag{5.33}$$

$$z_{k,L,i}(t) = z_{k,L,i}(0) + \int_0^t \sum_{l=1}^{K} r_{k,L,i,J+1}(u) s_{l,k,L,i}(u)\theta_{l,i}(u)(z_{l,Q,i}(u) - c_{l,i}(\mathbf{z}(u)))^+ \, du \tag{5.34}$$

$$z_{k,D,i}(t) = z_{k,D,i}(0) + \int_0^t \sum_{l=1}^{K} r_{k,S,i,J+1}(u) s_{l,k,S,i}(u)\mu_{l,i}(u) \min\left(z_{l,Q,i}(u), c_{l,i}(\mathbf{z}(u))\right) \, du \tag{5.35}$$

To prove this theorem the following results are required. Applying the law of large numbers to (5.20)-(5.27) for $t \in [0,T)$; $i,j \in Ser$; $k,l \in H$; $m \in \{R,U,A,O\}$ as $\eta \to \infty$:

$$\frac{\Pi_{\lambda_{k,i}(t)\eta}}{\eta} \xrightarrow{d} \int_0^t \lambda_{k,i}(u) \, du < \infty \tag{5.36}$$

$$\overline{D}^{(\eta)}_{k,m,i}(t) \xrightarrow{d} \int_0^t \delta_{k,m,i}(u) z_{k,m,i}(u) \tag{5.37}$$

$$\overline{D}^{(\eta)}_{k,S,i}(t) \xrightarrow{d} \int_0^t \mu_{k,i}(u) \min\left(z_{k,Q,i}(u), c_{k,i}(\mathbf{z}(u))\right)\, du \tag{5.38}$$

$$\overline{D}^{(\eta)}_{k,L,i}(t) \xrightarrow{d} \int_0^t \theta_{k,i}(u)(z_{k,Q,i}(u) - c_{k,i}(\mathbf{z}(u)))^+\, du \tag{5.39}$$

$$\overline{MS}^{(\eta)^{(l)}}_{k,m,i}(t) \xrightarrow{d} \int_0^t s_{k,l,m,i}(u)\delta_{k,m,i}(u) z_{k,m,i}(u)\, du \tag{5.40}$$

$$\overline{MR}^{(\eta)^{(j)}}_{k,S,i}(t) \xrightarrow{d} \int_0^t \sum_{l=1}^{K} s_{l,k,S,i}(u) r_{k,S,i,j}(u)\mu_{l,i}(u) \min\left(z_{l,Q,i}(u), c_{l,i}(\mathbf{z}(u))\right)\, du \tag{5.41}$$

$$\overline{MR}^{(\eta)^{(j)}}_{k,L,i}(t) \xrightarrow{d} \int_0^t \sum_{l=1}^{K} s_{l,k,L,i}(u) r_{k,L,i,j}(u)\theta_{l,i}(u)(z_{l,Q,i}(u) - c_{l,i}(\mathbf{z}(u)))^+\, du \tag{5.42}$$

Furthermore, the following lemma must be proved, as noted in [105].

Lemma 5.5.2 *The sequence of scaled processes* $\left\{\overline{\mathbf{Z}}^{(\eta)}(.)\right\}_{\eta=1}^{\infty}$ *is relatively compact and all weak limits are a.s. continuous.*

Proof 1 (Lemma 5.5.2) *Following the method in [105], but adapting to the extended system (5.1), I need to show that* $\left\{\overline{\mathbf{Z}}^{(\eta)}(.)\right\}_{\eta=1}^{\infty}$ *is relatively compact with continuous limits. To do this, it is sufficient to show the following two properties, which together show that the process is a.s. continuous. These are from Corollary 7.4 and Theorem 10.2 of [119].*

(a) *for any* $T \geq 0, \epsilon > 0$*, there exists a compact set* $\Gamma_T \subset \mathbb{R}^{7KJ}$ *such that:*

$$\mathbb{P}\left(\overline{\mathbf{Z}}^{(\eta)}(t) \in \Gamma_T, t \in [0,T]\right) \to 1, \text{ as } \eta \to \infty;$$

i.e. as the system scales with η*, the solution remains contained in some set* Γ_T*, for* $t \in [0,T]$ *with probability 1.*

(b) *for any* $\epsilon > 0$*, and* $T \geq 0$*, there exists a* $\delta > 0$*, such that:*

$$\limsup_{\eta\to\infty} \mathbb{P}\left(\omega\left(\overline{\mathbf{Z}}^{(\eta)}(t), \delta, T\right) \geq \epsilon\right) \leq \epsilon$$

where:

$$\omega\left(\overline{\mathbf{Z}}^{(\eta)}(t),\delta,T\right) := \sup_{\substack{\tau,t\in[0,T]\\|t-\tau|<\delta}} \max_{\substack{m\in\{Q,R,U,A,O,L,D\}\\k\in\{1,\ldots,K\}\\j\in\{1,\ldots,J\}}} \left|\overline{Z}^{(\eta)}_{k,m,j}(t)-\overline{Z}^{(\eta)}_{k,m,j}(\tau)\right|$$

Proof of (a): *Firstly, for the scaled fluid process, establish an upper bound for the total number of patients in the system for $t\in[0,T]$. A simple bound is the total arrivals in time t with no departures - this is equivalent to the bound used in [105]:*

$$\begin{aligned}
&\sum_{j=1}^{J}\sum_{k=1}^{K}\Big(\overline{Z}^{(\eta)}_{k,Q,j}(t)+\overline{Z}^{(\eta)}_{k,R,j}(t)+\overline{Z}^{(\eta)}_{k,U,j}(t)+\overline{Z}^{(\eta)}_{k,A,j}(t)\\
&\qquad+\overline{Z}^{(\eta)}_{k,O,j}(t)+\overline{Z}^{(\eta)}_{k,L,j}(t)+\overline{Z}^{(\eta)}_{k,D,j}(t)\Big)\leq\\
&\sum_{j=1}^{J}\sum_{k=1}^{K}\Big(\overline{Z}^{(\eta)}_{k,Q,j}(0)+\overline{Z}^{(\eta)}_{k,R,j}(0)+\overline{Z}^{(\eta)}_{k,U,j}(0)+\overline{Z}^{(\eta)}_{k,A,j}(0)\\
&\qquad+\overline{Z}^{(\eta)}_{k,O,j}(0)+\overline{Z}^{(\eta)}_{k,L,j}(0)+\overline{Z}^{(\eta)}_{k,D,j}(0)+\Pi_{\lambda_{k,j}(T)\eta}/\eta\Big)
\end{aligned}$$

Since $\Pi_{\lambda_{k,j}(t)\eta}$ is a time inhomogeneous Poisson process of rate $\lambda_{k,j}(t)\eta$, by (5.36) and the assumption of Theorem 5.5.1, for $\eta\to\infty$:

$$\begin{aligned}
&\sum_{j=1}^{J}\sum_{k=1}^{K}\Big(\overline{Z}^{(\eta)}_{k,Q,j}(0)+\overline{Z}^{(\eta)}_{k,R,j}(0)+\overline{Z}^{(\eta)}_{k,U,j}(0)+\overline{Z}^{(\eta)}_{k,A,j}(0)+\overline{Z}^{(\eta)}_{k,O,j}(0)\\
&\qquad+\overline{Z}^{(\eta)}_{k,L,j}(0)+\overline{Z}^{(\eta)}_{k,D,j}(0)+\Pi_{\lambda_{k,j}(T)\eta}/\eta\Big)\xrightarrow{d}\\
&\sum_{j=1}^{J}\sum_{k=1}^{K}\Big(z_{k,Q,j}(0)+z_{k,R,j}(0)+z_{k,U,j}(0)+z_{k,A,j}(0)+z_{k,O,j}(0)\\
&\qquad+z_{k,L,j}(0)+z_{k,D,j}(0)+\int_0^T\lambda_{k,j}(t)dt\Big)
\end{aligned}$$

Hence: $\mathbb{P}\left(\overline{\mathbf{Z}}^{(\eta)}(t) \in \Gamma_T, t \in [0,T]\right) \to 1$ *as* $\eta \to \infty$, *where:*

$$\Gamma_T = \Bigg\{(x_1, x_2, ..., x_{7KJ}) \Bigg| \sum_{n=1}^{7KJ} x_n \leq \sum_{j=1}^{J}\sum_{k=1}^{K}\Bigg(z_{k,Q,j}(0) + z_{k,R,j}(0) + z_{k,U,j}(0) + z_{k,A,j}(0) + z_{k,O,j}(0) + z_{k,L,j}(0) + z_{k,D,j}(0) + \int_0^T \lambda_{k,j}(t)dt\Bigg); \; x_1, x_2, ..., x_{7KJ} \geq 0\Bigg\}$$

Proof of (b): *The aim here is to show that, for any positive* ϵ, *the probability of the maximum difference between two points in each of the states (that relate to two points in time within a carefully chosen interval) exceeding* ϵ, *is negligible. i.e. as time progresses, there are no large jumps in the scaled fluid process.*

First consider the differences in each of the states of the scaled fluid process and find a sensible bound for each. Following the method set out by [105], but extending to include multiple health states and services, it follows from (5.13)-(5.19) that, for all $\tau, t \in [0,T]$, $k \in H$ *and* $j \in Ser$*:*

$$\left|\overline{Z}^{(\eta)}_{k,Q,j}(t) - \overline{Z}^{(\eta)}_{k,Q,j}(\tau)\right| \leq \frac{\left|\Pi_{\lambda_{k,j}(t)\eta} - \Pi_{\lambda_{k,j}(\tau)\eta}\right|}{\eta} + \sum_{m\in\{S,L,R,U,A,O\}}\left|\overline{D}^{(\eta)}_{k,m,j}(t) - \overline{D}^{(\eta)}_{k,m,j}(\tau)\right|$$

$$\begin{aligned}\left|\overline{Z}^{(\eta)}_{k,R,j}(t) - \overline{Z}^{(\eta)}_{k,R,j}(\tau)\right| &\leq \left|\overline{D}^{(\eta)}_{k,R,j}(t) - \overline{D}^{(\eta)}_{k,R,j}(\tau)\right| + \left|\overline{MR}^{(\eta)(j)}_{k,L,j}(t) - \overline{MR}^{(\eta)(j)}_{k,L,j}(\tau)\right| \\ &\leq \left|\overline{D}^{(\eta)}_{k,R,j}(t) - \overline{D}^{(\eta)}_{k,R,j}(\tau)\right| + \sum_{l=1}^{K}\left|\overline{D}^{(\eta)}_{l,L,j}(t) - \overline{D}^{(\eta)}_{l,L,j}(\tau)\right|\end{aligned}$$

$$\begin{aligned}\left|\overline{Z}^{(\eta)}_{k,U,j}(t) - \overline{Z}^{(\eta)}_{k,U,j}(\tau)\right| &\leq \left|\overline{D}^{(\eta)}_{k,U,j}(t) - \overline{D}^{(\eta)}_{k,U,j}(\tau)\right| + \left|\overline{MR}^{(\eta)(j)}_{k,S,j}(t) - \overline{MR}^{(\eta)(j)}_{k,S,j}(\tau)\right| \\ &\leq \left|\overline{D}^{(\eta)}_{k,U,j}(t) - \overline{D}^{(\eta)}_{k,U,j}(\tau)\right| + \sum_{l=1}^{K}\left|\overline{D}^{(\eta)}_{l,S,j}(t) - \overline{D}^{(\eta)}_{l,S,j}(\tau)\right|\end{aligned}$$

$$\begin{aligned}\left|\overline{Z}^{(\eta)}_{k,A,j}(t) - \overline{Z}^{(\eta)}_{k,A,j}(\tau)\right| &\leq \left|\overline{D}^{(\eta)}_{k,A,j}(t) - \overline{D}^{(\eta)}_{k,A,j}(\tau)\right| + \sum_{i\neq j}\left|\overline{MR}^{(\eta)(j)}_{k,L,i}(t) - \overline{MR}^{(\eta)(j)}_{k,L,i}(\tau)\right| \\ &\leq \left|\overline{D}^{(\eta)}_{k,A,j}(t) - \overline{D}^{(\eta)}_{k,A,j}(\tau)\right| + \sum_{i\neq j}\sum_{l=1}^{K}\left|\overline{D}^{(\eta)}_{l,L,i}(t) - \overline{D}^{(\eta)}_{l,L,i}(\tau)\right|\end{aligned}$$

$$\left|\overline{Z}^{(\eta)}_{k,O,j}(t)-\overline{Z}^{(\eta)}_{k,O,j}(\tau)\right| \leq \left|\overline{D}^{(\eta)}_{k,O,j}(t)-\overline{D}^{(\eta)}_{k,O,j}(\tau)\right| + \sum_{i\neq j}\left|\overline{MR}^{(\eta)^{(j)}}_{k,S,i}(t)-\overline{MR}^{(\eta)^{(j)}}_{k,S,i}(\tau)\right|$$
$$\leq \left|\overline{D}^{(\eta)}_{k,O,j}(t)-\overline{D}^{(\eta)}_{k,O,j}(\tau)\right| + \sum_{i\neq j}\sum_{l=1}^{K}\left|\overline{D}^{(\eta)}_{l,S,i}(t)-\overline{D}^{(\eta)}_{l,S,i}(\tau)\right|$$
$$\left|\overline{Z}^{(\eta)}_{k,L,j}(t)-\overline{Z}^{(\eta)}_{k,L,j}(\tau)\right| \leq \left|\overline{MR}^{(\eta)^{(J+1)}}_{k,L,j}(t)-\overline{MR}^{(\eta)^{(J+1)}}_{k,L,j}(\tau)\right|$$
$$\leq \sum_{l=1}^{K}\left|\overline{D}^{(\eta)}_{l,L,j}(t)-\overline{D}^{(\eta)}_{l,L,j}(\tau)\right|$$
$$\left|\overline{Z}^{(\eta)}_{k,D,j}(t)-\overline{Z}^{(\eta)}_{k,D,j}(\tau)\right| \leq \left|\overline{MR}^{(\eta)^{(J+1)}}_{k,S,j}(t)-\overline{MR}^{(\eta)^{(J+1)}}_{k,S,j}(\tau)\right|$$
$$\leq \sum_{l=1}^{K}\left|\overline{D}^{(\eta)}_{l,S,j}(t)-\overline{D}^{(\eta)}_{l,S,j}(\tau)\right|$$

By (5.37)-(5.42) limits can be found for the above. For this, note that by (a) there exists a finite constant V such that, for the event:

$$\Gamma^{(\eta)}_T = \left\{\overline{\mathbf{Z}}^{(\eta)}(t)\leq V; t\in[0,T]\right\},\ \mathbb{P}\left(\Gamma^{(\eta)}_T\right)\to 1 \text{ as } \eta\to\infty$$

Therefore, on the event $\Gamma^{(\eta)}_T$ each system state variable is bounded by V and the following inequalities hold for all $\tau, t\in[0,T]$ such that $|t-\tau|\leq\delta$ and $m\in\{R,U,A,O\}$:

$$\gamma_{k,S,j}\delta \coloneqq \max_{u\in[\tau,t]}\left(\mu_{k,j}(u)C_{k,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)\delta$$
$$\geq \int_\tau^t \mu_{k,j}(u)\min\left(\overline{Z}^{(\eta)}_{k,Q,i}(u), C_{k,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)du$$
$$\gamma_{k,L,j}\delta \coloneqq \max_{u\in[\tau,t]}\theta_{k,j}(u)V\delta \geq \int_\tau^t \theta_{k,j}(u)\left(\overline{Z}^{(\eta)}_{k,Q,j}(u)-C_{k,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)^+du$$
$$\gamma_{k,m,j}\delta \coloneqq \max_{u\in[\tau,t]}\delta_{k,m,j}(u)V\delta \geq \int_\tau^t \delta_{k,m,j}(u)\overline{Z}^{(\eta)}_{k,m,j}(u)du$$

Since $Z_{k,L,i}(t)$ and $Z_{k,D,i}(t)$ are absorbing states for patients who are lost to the system or have been discharged, their processes are captured by $\gamma_{k,S,j}$ and $\gamma_{k,L,j}$, respectively.

Let $\Gamma^{(\eta)^c}_T$ be the compliment of the event $\Gamma^{(\eta)}_T$ - the event $\left\{\overline{\mathbf{Z}}^{(\eta)}(t)>V : t\in[0,T]\right\}$ - such that $\mathbb{P}\left(\Gamma^{(\eta)^c}_T\right)\to 0$. To prove $\mathbb{P}\left(\omega\left(\overline{\mathbf{Z}}^{(\eta)},\delta,T\right)\geq\epsilon\right)\leq\epsilon$, find an upper

bound for $\mathbb{P}\left(\omega\left(\overline{\mathbf{Z}}^{(\eta)}, \delta, T\right) \geq \epsilon\right)$ *and show that it diminishes as* $\eta \to \infty$. *A suitable bound can be formed by considering* $\omega\left(\overline{\mathbf{Z}}^{(\eta)}, \delta, T\right)$ *in the case of each state variable i.e. finding the conditions for which each* $\left|\overline{D}^{(\eta)}_{k,m,j}(t) - \overline{D}^{(\eta)}_{k,m,j}(\tau)\right| \geq \epsilon, m \in St = \{S, L, R, U, A, O\}$ *and the probability that this occurs:*

$$\begin{aligned}
\mathbb{P}\Big(\omega\left(\overline{\mathbf{Z}}^{(\eta)}, \delta, T\right) \geq \epsilon\Big) &\leq \mathbb{P}\Big(\Gamma_T^{(\eta)^c}\Big) + \sum_{j=1}^{J}\sum_{k=1}^{K} \mathbb{P}\left(\omega\left(\frac{\Pi_{\lambda_{k,j}(t)\eta}}{\eta}, \delta, T\right) \geq \frac{\epsilon}{7KJ}\right) \\
&\quad + \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{m\in St} \mathbb{P}\left(\omega\left(\overline{D}^{(\eta)}_{k,m,j}(t), \gamma_{k,m,j}\delta, \gamma_{k,m,j}T\right) \geq \frac{\epsilon}{7KJ}\right) \\
&\xrightarrow{d} \sum_{j=1}^{J}\sum_{k=1}^{K}\left(\mathbb{P}\left(\max_{u\in[t,\tau]} \lambda_{k,j}(u)\delta \geq \frac{\epsilon}{7KJ}\right) + \sum_{m\in St}\mathbb{P}\left(\gamma_{k,m,j}\delta \geq \frac{\epsilon}{7KJ}\right)\right)
\end{aligned}$$

Convergence holds due to the earlier results gained using the law of large numbers (5.36)-(5.42), and by the continuity of $\omega(x(t), \gamma_{k,m,j}\delta, \gamma_{k,m,j}T)$ *and* $\omega(x(t), \delta, T)$ *with respect to* $x(t)$. *Hence, (b) holds with any* δ *such that* $\max_{u\in[t,\tau]} \lambda_{k,j}(u)\delta < \epsilon/7KJ$ *and* $\gamma_{k,m,j}\delta < \epsilon/7KJ$, *for all* $k \in H,\ j \in Ser$ *and* $m \in St$. ■

Now to prove Theorem 5.5.1. by extending the proof from [105] to (5.1), which involves applying the law of large numbers to more cases. Fundamentally, the proof is the same.

Aside 5.5.1 (Sketch of proof for Theorem 5.5.1) *Re-write the system in vector form, for* $t \in [0, T]$*:*

$$\overline{\mathbf{Z}}^{(\eta)}(t) = \overline{\mathbf{Z}}^{(\eta)}(0) + \mathbf{G}^{(\eta)}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) + \int_0^t \mathbf{H}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right) du$$

by adding and subtracting (5.29)-(5.35) to their respective counterpart in (5.13) - (5.19). This can be rearranged to show that the right hand side tends to zero (because of the law of large numbers) and that a unique limit exists for (5.13)-(5.19) i.e. the solution to (5.29)-(5.35).

Proof 2 (Theorem 5.5.1) *The proof closely follows that of [105]; therefore, it has been included within Appendix B.1. for reference.* ■

Analytical expressions cannot be found for (5.29)-(5.35); however, these equations can be solved using common numerical schemes. In the following chapter I solve them iteratively using the trapezium rule.

By Theorem 5.5.1, fluid approximations are gained for server allocations (5.2) and (5.3). For (5.2), the continuous fluid approximation is:

$$c_{k,i}(\mathbf{z}(t)) = \frac{c_i(t) z_{k,Q,i}(t)}{\sum_{l=1}^{K} z_{l,Q,i}(t)}, \text{ for all } t \in [0,T] \tag{5.43}$$

For (5.3), the weighted allocation, the continuous fluid approximation is:

$$c_{k,i}(\mathbf{z}(t)) = \frac{c_i(t) B_{k,i}(t) z_{k,Q,i}(t)}{\sum_{l=1}^{K} B_{l,i}(t) z_{l,Q,i}(t)}, \text{ for all } t \in [0,T] \tag{5.44}$$

Notably, the server allocation algorithm does not need to be implemented for the continuous case since the fluid approximation is capable of modelling non-integer allocations.

As previously noted, to calculate the VWT (section 5.5.4), $c_{k,i}(\mathbf{z}(t))$ needs to be continuously differentiable. Thus, $z_{k,Q,i}(t)$, for all $k \in H, i \in Ser$ must be continuously differentiable throughout $[0,T]$, giving the requirement that all input parameters are continuous. These time varying server allocations are a further output for the model.

5.5.3 Diffusion approximation

Following the method set out by A Mandelbaum et al. in the Appendix of [107] and proved in [104], I apply their method to (5.1) to formulate a diffusion limit for a system which includes health states, multiple services, arrivals to other services,

uses of alternative services after abandonment and reuse. The method is largely unchanged and is presented here to provide a more in-depth understanding of how the method works in application to (5.1).

The diffusion limit quantifies deviations from the first order fluid approximation [104], providing a second order limit that gives a system of ODEs for calculating the mean and covariance of the diffusion process. As a result, information is gained about the variance seen in the stochastic system. In the next chapter I will show that, under the right conditions, the variance of the diffusion process closely matches the sample variance from a simulated stochastic system.

Since all of the flow functions are continuous, the assumptions stated in Theorem 2.4 of [104] are maintained. Thus, the diffusion approximation may be formulated by using the method in [107].

By Theorem 5.5.1, $\lim_{\eta\to\infty} \overline{\mathbf{Z}}^{(\eta)}(t) = \mathbf{z}(t)$ a.s. with uniform convergence on compact sets of t [107]. Thus, the diffusion limit is gained by applying the functional central limit theorem to $\widehat{\mathbf{z}}(t) = \{\widehat{\mathbf{z}}(t)|T > t \geq 0\}$ [104]. That is, if $\lim_{\eta\to\infty} \sqrt{\eta}(\overline{\mathbf{Z}}^{(\eta)}(0) - \mathbf{z}(0)) = \widehat{\mathbf{z}}(0)$ holds, where $\widehat{\mathbf{z}}(0)$ is a constant, then:

$$\lim_{\eta\to\infty} \sqrt{\eta}\left(\overline{\mathbf{Z}}^{(\eta)}(t) - \mathbf{z}(t)\right) \stackrel{d}{=} \widehat{\mathbf{z}}(t)$$

This is a convergence in distribution of the processes [104]. If the set of time points $\{t \in [0,T)\,|z_{k,Q,i}(t) = c_{k,i}(\mathbf{z}(t))\}$ has zero measure, $\widehat{\mathbf{z}}(t)$ is a Gaussian process [107]. Thus, the mean vector and covariance matrix for the diffusion process are the unique solutions to autonomous differential equations. Furthermore, for a service $j \in Ser$ both $\min(z_{k,Q,j}(t), c_{k,j}(\mathbf{z}(t)))$ and $(z_{k,Q,j}(t) - c_{k,j}(\mathbf{z}(t)))^+$, are everywhere continuous. Also, they are everywhere differentiable, except when $z_{k,Q,j}(t) = c_{k,j}(\mathbf{z}(t))$.

Noting that $\widehat{\mathbf{z}}(t)$ is a column vector, for $0 \leq t < T$ and for all $\mathbf{x} \in \mathbb{R}^{(7KJ)}$, define:

$$\alpha_t(\mathbf{x}(t)) \equiv \sum_{i \in I} \alpha_{t,i}(\mathbf{x}(t))\mathbf{v}_i$$

such that:

$$\frac{d}{dt}\mathbb{E}\left[\widehat{\mathbf{z}}(t)\right] = \mathbf{A}_t^T \mathbb{E}\left[\widehat{\mathbf{z}}(t)\right]$$

and:

$$\frac{d}{dt}\mathbf{Cov}[\widehat{\mathbf{z}}(t)] = \mathbf{Cov}[\widehat{\mathbf{z}}(t)]\mathbf{A}_t^T + \mathbf{A}_t\mathbf{Cov}[\widehat{\mathbf{z}}(t)] + \mathbf{B}_t$$

where $\mathbf{A}_t = D\alpha_t(\mathbf{z}(t))$ is the Jacobian of $\alpha_t(\mathbf{z}(t))$ when differentiated at $\mathbf{z}(t)$ and $\mathbf{B}_t = \sum_{i \in I} \alpha_{t,i}(\mathbf{z}(t))\mathbf{v}_i \otimes \mathbf{v}_i$ is the tensor product of two vectors forming a symmetrical matrix. For $0 \leq t < T$, the matrices, $\mathbf{A}(t), \mathbf{B}(t)$ and $\mathrm{Cov}[\widehat{\mathbf{z}}(t)]$ are of dimension $7KJ \times 7KJ$.

Instead of working with the index notation as in (5.45) and (5.46), I use a more explicit notation to highlight how this method applies to the extended system. Beginning with the rate functions, for $k, l \in H$ and $i, j \in Ser$:

$$\begin{aligned}
&\alpha_{k,i,1}(\mathbf{z}(t))\ \lambda_{k,i}(t) \\
&\alpha_{k,l,i,2}(\mathbf{z}(t)) = s_{k,l,R,i}(t)\delta_{k,R,i}(t)z_{k,R,i}(t) \\
&\alpha_{k,l,i,3}(\mathbf{z}(t)) = s_{k,l,U,i}(t)\delta_{k,U,i}(t)z_{k,U,i}(t) \\
&\alpha_{k,l,i,4}(\mathbf{z}(t)) = s_{k,l,A,i}(t)\delta_{k,A,i}(t)z_{k,A,i}(t) \\
&\alpha_{k,l,i,5}(\mathbf{z}(t)) = s_{k,l,O,i}(t)\delta_{k,O,i}(t)z_{k,O,i}(t) \\
&\alpha_{k,l,i,6}(\mathbf{z}(t)) = s_{k,l,L,i}(t)r_{l,L,i,J+1}(t)\theta_{k,i}(t)(z_{k,Q,i}(t) - c_{k,i}(\mathbf{z}(t)))^+ \\
&\alpha_{k,l,i,7}(\mathbf{z}(t)) = \ s_{k,l,S,i}(t)r_{l,S,i,J+1}(t)\mu_{k,i}(t)\min\left(z_{k,Q,i}(t), c_{k,i}(\mathbf{z}(t))\right) \\
&\alpha_{k,l,i,j,8}(\mathbf{z}(t)) = \ s_{k,l,L,i}(t)r_{l,L,i,j}(t)\theta_{k,i}(t)(z_{k,Q,i}(t) - c_{k,i}(\mathbf{z}(t)))^+ \\
&\alpha_{k,l,i,j,9}(\mathbf{z}(t)) = \ s_{k,l,S,i}(t)r_{l,S,i,j}(t)\mu_{k,i}(t)\min\left(z_{k,Q,i}(t), c_{k,i}(\mathbf{z}(t))\right)
\end{aligned}$$

$$z_{k,Q,i}(t) = z_{k,Q,i}(0) + \int_0^t \alpha_{k,i,1}(u) + \sum_{l=1}^{K}\Bigg(\alpha_{l,k,i,2}(u) + \alpha_{l,k,i,3}(u) + \alpha_{l,k,i,4}(u) + \alpha_{l,k,i,5}(u) - \alpha_{k,l,i,6}(u) - \alpha_{k,l,i,7}(u) - \sum_{j=1}^{J}\left(\alpha_{k,l,i,j,8}(u) + \alpha_{k,l,i,j,9}(u)\right)\Bigg)\, du$$

$$z_{k,R,i}(t) = z_{k,R,i}(0) + \int_0^t \sum_{l=1}^{K}\Big(\alpha_{l,k,i,i,8}(u) - \alpha_{k,l,i,2}(u)\Big)du$$

$$z_{k,U,i}(t) = z_{k,U,i}(0) + \int_0^t \sum_{l=1}^{K}\Big(\alpha_{l,k,i,i,9}(u) - \alpha_{k,l,i,3}(u)\Big)du$$

$$z_{k,A,i}(t) = z_{k,A,i}(0) + \int_0^t \sum_{l=1}^{K}\left(\sum_{i=1;j\neq i}^{J}\alpha_{l,k,j,i,8}(u) - \alpha_{k,l,i,4}(u)\right)du$$

$$z_{k,O,i}(t) = z_{k,O,i}(0) + \int_0^t \sum_{l=1}^{K}\left(\sum_{i=1;j\neq i}^{J}\alpha_{l,k,j,i,9}(u) - \alpha_{k,l,i,5}(u)\right)du$$

$$z_{k,L,i}(t) = z_{k,L,i}(0) + \int_0^t \sum_{l=1}^{K}\alpha_{l,k,i,6}(u)du$$

$$z_{k,D,i}(t) = z_{k,D,i}(0) + \int_0^t \sum_{l=1}^{K}\alpha_{l,k,i,7}(u)du$$

Continuing with this notation, I now form a basis of transition vectors of length $7KJ$. Denoting the m-th element of each vector as $\mathbf{v}^{(m)}_{k,i,1}$, the transition vectors are defined as:

$$\mathbf{v}^{(m)}_{k,i,1} = \begin{cases} 1, & \text{if } m = 7K(i-1) + 7(k-1) + 1 \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{v}^{(m)}_{k,l,i,2} = \begin{cases} 1, & \text{if } m = 7K(i-1) + 7(l-1) + 1 \\ -1, & \text{if } m = 7K(i-1) + 7(k-1) + 2 \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{v}^{(m)}_{k,l,i,3} = \begin{cases} 1, & \text{if } m = 7K(i-1)+7(l-1)+1 \\ -1, & \text{if } m = 7K(i-1)+7(k-1)+3 \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{v}^{(m)}_{k,l,i,4} = \begin{cases} 1, & \text{if } m = 7K(i-1)+7(l-1)+1 \\ -1, & \text{if } m = 7K(i-1)+7(k-1)+4 \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{v}^{(m)}_{k,l,i,5} = \begin{cases} 1, & \text{if } m = 7K(i-1)+7(l-1)+1 \\ -1, & \text{if } m = 7K(i-1)+7(k-1)+5 \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{v}^{(m)}_{k,l,i,6} = \begin{cases} -1, & \text{if } m = 7K(i-1)+7(k-1)+1 \\ 1, & \text{if } m = 7K(i-1)+7(l-1)+6 \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{v}^{(m)}_{k,l,i,7} = \begin{cases} -1, & \text{if } m = 7K(i-1)+7(k-1)+1 \\ 1, & \text{if } m = 7K(i-1)+7(l-1)+7 \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{v}^{(m)}_{k,l,i,j,8} = \begin{cases} -1, & \text{if } m = 7K(i-1)+7(k-1)+1, \text{for } j = 1,...,J \\ 1, & \text{if } m = 7K(i-1)+7(l-1)+2, \text{ for } j = i \\ 1, & \text{if } m = 7K(j-1)+7(l-1)+4, \text{ for } j \neq i \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{v}^{(m)}_{k,l,i,j,9} = \begin{cases} -1, & \text{if } m = 7K(i-1)+7(k-1)+1, \text{for } j = 1,...,J \\ 1, & \text{if } m = 7K(i-1)+7(l-1)+3, \text{ for } j = i \\ 1, & \text{if } m = 7K(j-1)+7(l-1)+5, \text{ for } j \neq i \\ 0, & \text{otherwise} \end{cases}$$

In this case:

$$\begin{aligned} \alpha_t(\mathbf{z}(t)) \equiv & \sum_{k=1}^{K}\sum_{i=1}^{J} \alpha_{k,i,1}(\mathbf{z}(t))\mathbf{v}_{k,i,1} + \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{J}\sum_{p=2}^{7} \alpha_{k,l,i,p}(\mathbf{z}(t))\mathbf{v}_{k,l,i,p} \\ & + \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{q=8}^{9} \alpha_{k,l,i,j,q}(\mathbf{z}(t))\mathbf{v}_{k,l,i,j,q} \\ \mathbf{B}_t = & \sum_{k=1}^{K}\sum_{i=1}^{J} \alpha_{k,i,1}(\mathbf{z}(t))\mathbf{v}_{k,i,1} \otimes \mathbf{v}_{k,i,1} \\ & + \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{J}\sum_{p=2}^{7} \alpha_{k,l,i,p}(\mathbf{z}(t))\mathbf{v}_{k,l,i,p} \otimes \mathbf{v}_{k,l,i,p} \\ & + \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{q=8}^{9} \alpha_{k,l,i,j,q}(\mathbf{z}(t))\mathbf{v}_{k,l,i,j,q} \otimes \mathbf{v}_{k,l,i,j,q} \end{aligned}$$

Both A_t and B_t are matrices of dimension $7KJ \times 7KJ$. For $k \in H,\ i \in Ser$, let $u = 7K(i-1)+7(k-1)+1$. For $m = 1,...,7KJ$ define:

$$a_t^{(m,u)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,Q,i}(t)}, \quad a_t^{(m,u+1)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,R,i}(t)}, \quad a_t^{(m,u+2)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,U,i}(t)}$$

$$a_t^{(m,u+3)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,A,i}(t)}, \quad a_t^{(m,u+4)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,O,i}(t)}$$

$$a_t^{(m,u+5)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,L,i}(t)}, \quad a_t^{(m,u+6)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,D,i}(t)}$$

5.5.4 Virtual waiting time

Adapting the method set out in [107], I now present a method for calculating the virtual waiting time (VWT) for each service in (5.1). This method differs slightly from [107]. Due to the extension to multiple health states and use of a dynamic server allocation that is dependent on $\mathbf{Z}(t)$, additional assumptions and definitions are required, as detailed below. The method is presented here to give further explicit detail on how this applies in a scenario of multiple services and health states.

Definition 5.5.6 (Virtual waiting time) *For a infinitely patient "virtual customer" arriving to the service and queue at a fixed time τ, $T > \tau \geq 0$, their **virtual waiting time** (VWT) is how long they have to wait until their service begins. This is denoted: $VWT_{k,i}(\tau)$ for each $i \in Ser$ and $k \in H$.*

The method set out in [107] needs to be adapted to calculate the VWT for (5.1). Thus, given the parallel queues and multiple services, the following assumptions are required to calculate the VWT for each $k \in H$ and $i \in Ser$ over the interval $[0, \infty)$:

1. The functions $c_{k,i}(\mathbf{z}(t))$ are continuously differentiable with respect to time;
2. All $\mu_{k,i}(t)$ are continuous;
3. $\delta_{k,R,i}(t), \delta_{k,U,i}(t), \delta_{k,A,i}(t), \delta_{k,O,i}(t)$ and $\theta_{k,i}(t)$ are bounded on compact intervals.

The first assumption places the continuous constraint on all the input parameters when using the dynamic server allocations in section 5.3.1 (thus the second is implicit in 1, unless the dynamic allocations are not modelled). The third assumption ensures that patients who reside in a process state spend a measurable amount of time in it, i.e. if, say, $\delta_{k,U,i}(t) = \infty$, patients spend no time in the process orbit, immediately entering the queue.

To calculate the VWT at time $\tau > 0$, (5.13)-(5.19) are modified. Denoted $\mathbf{Z}^*$, $\mathbf{Z}^*(t) = \mathbf{Z}(t)$ for $\tau > t \geq 0$. Thus, the results of Theorem 5.5.1 and the diffusion

equations still hold in this time period. However, for $t > \tau$, the time after a virtual patient has arrived, the process differs as follows:

1. There are no external arrivals, rejoins, reuses, uses of alternative service or arrivals from other services;
2. Only patients remaining in the queue and service are served after τ;
3. Any patient departing the service and queue process leaves the entire system;
4. There are no health state transitions after τ.

Importantly, the above assumptions simplify the calculation of the VWT in a network, as each health state queue (and service) behaves independently of each other for $t > \tau$. Therefore, the system can be decomposed such that the VWT can be solved for each health state queue. Furthermore, due to the above assumptions for $t > \tau$, I need only focus on the operation of the service orbit.

Applying a scaling limit $\eta > 0$ as in section 5.5.2, the scaled modified process is denoted $\overline{\mathbf{Z}}^{*(\eta)}$. To aid the formulation of the VWT, $\overline{\mathbf{Z}}^{*(\eta)}$ may be defined in terms of a scaled arrival process $\overline{A}_{k,i}^{(\eta)}(t)$ and a scaled departure process $\overline{\Delta}_{k,i}^{(\eta)}(t)$, for $t \geq 0$.

To define these processes, I first introduce new definitions of the service flux term and abandonment flux term for the modified system:

$$\overline{D}_{k,S,i}^{*(\eta)}(t) = \Pi_{k,S,i}\left(\eta \int_0^t \mu_{k,i}(u) \min\left(\overline{Z}_{k,Q,i}^{*(\eta)}(u), C_{k,i}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right) du\right) \Big/ \eta$$
$$\overline{D}_{k,L,i}^{*(\eta)}(t) = \Pi_{k,L,i}\left(\eta \int_0^t \theta_{k,i}(u) \left(\overline{Z}_{k,Q,i}^{*(\eta)}(u) - C_{k,i}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)^+ du\right) \Big/ \eta$$

Notably, $C_{k,i}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)$, is given by the unmodified system since I am calculating the VWT for the unmodified system.

Now, for $\overline{Z}_{k,Q,i}^{*(\eta)}(t) = \overline{A}_{k,i}^{(\eta)}(t) - \overline{\Delta}_{k,i}^{(\eta)}(t)$ and $t > 0$, I define $\overline{A}_{k,i}^{(\eta)}(t)$ to include the patients who are in service at time 0, and $\overline{\Delta}_{k,i}(t)$ to be a continuously differentiable

and non-decreasing function in $[0, \infty)$, as follows:

$$\overline{A}_{k,i}^{(\eta)}(t) = \overline{Z}_{k,Q,i}^{(\eta)}(0) + \frac{\Pi_{\lambda_{k,i}(t)\eta}}{\eta} + \sum_{l=1}^{K} \overline{MS}_{l,R,i}^{(\eta)^{(k)}}(t) + \sum_{l=1}^{K} \overline{MS}_{l,U,i}^{(\eta)^{(k)}}(t) + \sum_{l=1}^{K} \overline{MS}_{l,A,i}^{(\eta)^{(k)}}(t) + \sum_{l=1}^{K} \overline{MS}_{l,O,i}^{(\eta)^{(k)}}(t),\ \tau > t > 0$$

$$\overline{\Delta}_{k,i}^{(\eta)} = \overline{D}_{k,S,i}^{*^{(\eta)}}(t) + \overline{D}_{k,L,i}^{*^{(\eta)}}(t),\ t > 0$$

Noting: $\overline{A}_{k,i}^{(\eta)}(0) = \overline{Z}_{k,Q,i}^{(\eta)}(0), \qquad \overline{A}_{k,i}^{(\eta)}(t) = \overline{A}_{k,i}^{(\eta)}(\tau),\ t \geq \tau, \qquad \overline{\Delta}_{k,i}^{(\eta)}(0) = 0.$

For $k \in H$ and $i \in Ser$, as in [107]:

$$\lim_{\eta \to \infty} \left(\overline{\mathbf{Z}}^{*^{(\eta)}}(t), \overline{A}_{k,i}^{(\eta)}(t), \overline{\Delta}_{k,i}^{(\eta)}(t) \right) = (\mathbf{z}^*(t), A_{k,i}(t), \Delta_{k,i}(t)) \text{ a.s.}$$

which converges uniformly on compact sets of t. As a result, the arrival process, $A_{k,i}(t)$, and service process, $\Delta_{k,i}(t)$, for the modified fluid approximation, $z_{k,Q,i}^*(t) = A_{k,i}(t) - \Delta_{k,i}(t),\ t > 0$, are:

$$A_{k,i}(t) = z_{k,Q,i}^*(0) + \int_0^t \lambda_{k,i}(u) + \sum_{l=1}^{K} s_{l,k,R,i}(t)\delta_{l,R,i}(u)z_{l,R,i}(u) + \sum_{l=1}^{K} s_{l,k,U,i}(t)\delta_{l,U,i}(u)z_{l,U,i}(u) + \sum_{l=1}^{K} s_{l,k,A,i}(t)\delta_{l,A,i}(u)z_{l,A,i}(u) + \sum_{l=1}^{K} s_{l,k,O,i}(t)\delta_{l,O,i}(u)z_{l,O,i}(u)\ du,\ \tau > t > 0$$

$$\Delta_{k,i} = \int_0^t \mu_{k,i}(u) \min\left(z_{k,Q,i}^*(u), c_{k,i}\left(\mathbf{z}(u)\right)\right) + \theta_{k,i}(u)\left(z_{k,Q,i}^*(u) - c_{k,i}\left(\mathbf{z}(u)\right)\right)^+ du, t > 0$$

Noting: $A_{k,i}(0) = z_{k,Q,i}(0), \qquad A_{k,i}(t) = A_{k,i}(\tau),\ t \geq \tau, \qquad \Delta_{k,i}(0) = 0.$

Given the above and following [107], the diffusion approximation is given by:

$$\lim_{\eta\to\infty}\sqrt{\eta}\left(\overline{\mathbf{Z}}^{*(\eta)}(t)-\mathbf{z}^{*}(t),\overline{A}_{k,i}^{(\eta)}(t)-A_{k,i}(t),\overline{\Delta}_{k,i}^{(\eta)}(t)-\Delta_{k,i}(t)\right)\overset{d}{=}\left(\widehat{\mathbf{z}}^{*}(t),\widehat{A}_{k,i}(t),\widehat{\Delta}_{k,i}(t)\right)$$

If the set of time points $\{t\geq 0|\sum_{k=1}^{A}z_{k,Q,i}^{*}(t)=c_{k,i}(\mathbf{z}(t))\}$, has zero measure for all $k\in H$ and $i\in H$, then $\{\widehat{\mathbf{z}}^{*}(t)|t\geq 0\}$ is a Gaussian process for $t\geq\tau$. Therefore, for $t\geq\tau$ the fluid approximation service process is given by:

$$\begin{aligned}z_{k,Q,i}^{*}(t)=z_{k,Q,i}^{*}(\tau)&-\int_{0}^{t}\mu_{k,i}(u)\min\left(z_{k,Q,i}^{*}(u),c_{k,i}\left(\mathbf{z}(u)\right)\right)du\\&-\int_{0}^{t}\theta_{k,i}(u)\left(z_{k,Q,i}^{*}(u)-c_{k,i}\left(\mathbf{z}(u)\right)\right)^{+}du\end{aligned}$$

And the variance of the modified diffusion process, $\mathrm{Var}(\widehat{z}_{k,Q,i}^{*}(t))$, is the solution to:

$$\begin{aligned}\frac{d}{dt}\mathrm{Var}(\widehat{z}_{k,Q,i}^{*}(t))=\;&\theta_{k,i}(t)(z_{k,Q,i}^{*}(t)-c_{k,i}(\mathbf{z}(t)))^{+}\\&+\mu_{k,i}(t)\min(z_{k,Q,i}^{*}(t),c_{k,i}(\mathbf{z}(t)))\\&-2\theta_{k,i}(t)\mathrm{Var}(\widehat{z}_{k,Q,i}^{*}(t))\mathbb{I}_{\{z_{k,Q,i}^{*}(t)>c_{k,i}(\mathbf{z}(t))\}}\\&-\mu_{k,i}(t)\mathrm{Var}(\widehat{z}_{k,Q,i}^{*}(t))\mathbb{I}_{\{z_{k,Q,i}^{*}(t)\leq c_{k,i}(\mathbf{z}(t))\}}\end{aligned}\tag{5.45}$$

where $\mathbb{I}_{\{x\geq 0\}}$ is an identity function such that for $x\geq 0,\mathbb{I}_{\{x\geq 0\}}=1$, otherwise $\mathbb{I}_{\{x\geq 0\}}=0$. Furthermore, it follows that, for $t\geq 0:\widehat{z}_{k,Q,i}^{*}(t)=\widehat{A}_{k,i}(t)-\widehat{\Delta}_{k,i}(t)$.

Having made these extra assumptions and established the modified system, the method now follows that of [107]. To formulate the VWT, define the *potential service initiation* process to be: $\overline{D}_{k,i}^{(\eta)}=\overline{\Delta}_{k,i}^{(\eta)}(t)+C_{k,i}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right)$, such that if $\overline{Z}_{k,Q,i}^{*(\eta)}(t)<C_{k,i}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right)\Rightarrow\overline{A}_{k,i}^{(\eta)}(t)<\overline{D}_{k,i}^{(\eta)}(t)$, giving the service "ahead" of arrivals. Furthermore, by definition, the following limit holds:

$$\lim_{\eta\to\infty}\overline{D}_{k,i}^{(\eta)}(t)=D_{k,i}(t)\qquad\text{a.s.}$$

Convergence is again uniform on compact sets of t [107] and $D_{k,i}(t) = \Delta_{k,i}(t) + c_{k,i}(\mathbf{z}(t)), t \geq 0$. Since $c_{k,i}(\mathbf{z}(t))$ and $\Delta_{k,i}(t)$ are continuously differentiable, $D_{k,i}(t)$ is also continuously differentiable. Additionally, the derivative of $D_{k,i}(t)$ is denoted $d_{k,i}(t)$. A further necessary assumption is that $d_{k,i}(t)$ is strictly positive such that $\lim_{t\to\infty} D_{k,i}(t) > A_{k,i}(\tau)$ (this ensures that the VWT always exists because all patients will eventually be served and is why continuously differentiable parameters and queueing process are required given the dynamic server allocation).

By definition, both $\overline{A}^{(\eta)}_{k,i}(t)$ and $A_{k,i}(t)$ are constant for $t \in [\tau, \infty)$. It is also helpful to define all the processes on the interval $[-T_{k,i}, \infty)$, with $T_{k,i} = c_{k,i}(\mathbf{z}(0))/d_{k,i}(0)$ i.e. whilst no arrivals or departures happen in $[-T_{k,i}, 0)$, the number of servers increases linearly from 0 to $c_{k,i}(\mathbf{z}(0))$ for each $k \in H, i \in Ser$. Thus, for $\widehat{D}_{k,i}(t) = \widehat{\Delta}_{k,i}(t)$:

$$\lim_{\eta\to\infty} \left(\overline{\mathbf{Z}}^{*(\eta)}(t), \overline{A}^{(\eta)}_{k,i}(t), \overline{D}^{(\eta)}_{k,i}(t)\right) = (\mathbf{z}^*(t), A_{k,i}(t), D_{k,i}(t)) \text{ a.s.}$$

$$\lim_{\eta\to\infty} \sqrt{\eta}\left(\overline{\mathbf{Z}}^{*(\eta)}(t) - \mathbf{z}^*(t), \overline{A}^{(\eta)}_{k,i}(t) - A_{k,i}(t), \overline{D}^{(\eta)}_{k,i}(t) - D_{k,i}(t)\right)$$

$$\stackrel{d}{=} \left(\widehat{\mathbf{z}}^*(t), \widehat{A}_{k,i}(t), \widehat{D}_{k,i}(t)\right)$$

Note that $A_{k,i}, D_{k,i}, \widehat{A}_{k,i}, \widehat{D}_{k,i}$ are continuous and $D_{k,i}(-T_{k,i}) = \widehat{D}_{k,i}(-T_{k,i}) = 0$.

Given the stated assumptions, the following processes are well defined and finite with probability 1 for all sufficiently large η [107]. Defining the *first attainment processes*, $\overline{S}^{(\eta)}_{k,i}(t)$ and $S_{k,i}(t)$, for all $t \geq -T_{k,i}$:

$$\overline{S}^{(\eta)}_{k,i}(t) = \inf\left\{s \geq -T_{k,i} : \overline{D}^{(\eta)}_{k,i}(s) > \overline{A}^{(\eta)}_{k,i}(t)\right\}$$

$$S_{k,i}(t) = \inf\left\{s \geq -T_{k,i} : D_{k,i}(s) > A_{k,i}(t)\right\}$$

giving the *attainment waiting time processes*, $\overline{W}^{(\eta)}_{k,i}(t)$ and $W_{k,i}(t)$, as:

$$\overline{W}^{(\eta)}_{k,i}(t) = \overline{S}^{(\eta)}_{k,i}(t) - t, \qquad W_{k,i}(t) = S_{k,i}(t) - t$$

In defining these processes, the scaled VWT at τ for $k \in H$, $i \in Ser$, denoted $\overline{VWT}_{k,i}^{(\eta)}(\tau)$, is calculated by:

$$\overline{VWT}_{k,i}^{(\eta)}(\tau) = \overline{W}_{k,i}^{(\eta)}(\tau)^{+} = \left(\inf\left\{s \geq -T_{k,i} : \overline{D}_{k,i}^{(\eta)}(s) > \overline{A}_{k,i}^{(\eta)}(\tau)\right\} - \tau\right)^{+}$$

It is possible for $\overline{W}_{k,i}^{(\eta)}(\tau)$ and $W_{k,i}(\tau)$ to be negative when $\overline{\mathbf{Z}}^{*(\eta)}(\tau) < C_{k,i}\left(\overline{\mathbf{Z}}^{(\eta)}(\tau)\right)$, hence $\overline{VWT}_{k,i}^{(\eta)}(\tau) = 0$. Thus, if no queue has formed, there is no waiting time. It follows, as in [107], that:

$$\lim_{\eta\to\infty}\left(\overline{\mathbf{Z}}^{*(\eta)}, \overline{A}_{k,i}^{(\eta)}, \overline{D}_{k,i}^{(\eta)}, \overline{W}_{k,i}^{(\eta)}\right) = (\mathbf{z}^{*}, A_{k,i}, D_{k,i}, W_{k,i}) \text{ a.s.}$$

$$\lim_{\eta\to\infty}\sqrt{\eta}\left(\overline{\mathbf{Z}}^{*(\eta)} - \mathbf{z}^{*}, \overline{A}_{k,i}^{(\eta)} - A_{k,i}, \overline{D}_{k,i}^{(\eta)} - D_{k,i}, \overline{W}_{k,i}^{(\eta)} - W_{k,i}\right) \overset{d}{=} \left(\widehat{\mathbf{z}}^{*}, \widehat{A}_{k,i}, \widehat{D}_{k,i}, \widehat{W}_{k,i}\right)$$

and by the theorem and corollary of [120] the diffusion approximation of the virtual waiting time can be calculated as follows:

$$\widehat{W}_{k,i}(t) = \frac{\widehat{A}_{k,i}(t) - \widehat{D}_{k,i}\left(S_{k,i}(t)\right)}{d_{k,i}\left(S_{k,i}(t)\right)}$$

Since the processes $\widehat{A}_{k,i}, \widehat{D}_{k,i}, \widehat{z}_{k,Q,i}^{*}, \widehat{W}_{k,i}$ are continuous with probability 1, they automatically obtain the convergence of finite-dimensional distributions [107].

For the non-trivial case $S_{k,i}(\tau) \geq \tau$ where $z_{k,Q,i}^{*}(\tau) \geq c_{k,i}(\mathbf{z}(\tau))$ if in $[0,\tau]$ the set of points $\{t \mid z_{k,Q,i}^{*}(t) = c_{k,i}(\mathbf{z}(t))\}$ has measure zero, then since $A(S_{k,i}(t)) = A(\tau)$:

$$\begin{aligned}
&\lim_{\eta\to\infty}\overline{W}_{k,i}^{(\eta)}(\tau) = W_{k,i}(\tau) \text{ a.s.}\\
&\lim_{\eta\to\infty}\sqrt{\eta}(\overline{W}^{(\eta)}(\tau) - W_{k,i}(\tau)) \overset{d}{=} \widehat{W}_{k,i}(\tau)\\
&\qquad\qquad = \frac{\widehat{A}_{k,i}(\tau) - \widehat{D}_{k,i}\left(S_{k,i}(\tau)\right)}{d_{k,i}\left(S_{k,i}(\tau)\right)}\\
&\qquad\qquad = \frac{\widehat{A}_{k,i}\left(S_{k,i}(\tau)\right) - \widehat{D}_{k,i}\left(S_{k,i}(\tau)\right)}{d_{k,i}\left(S_{k,i}(\tau)\right)}\\
&\qquad\qquad = \frac{\widehat{z}_{k,Q,i}^{*}(S_{k,i}(\tau))}{d(S_{k,i}(\tau))}
\end{aligned}$$

Hence:

$$\text{Var}[\widehat{W}_{k,i}(\tau)] = \text{Var}\left[\frac{\widehat{z}^*_{k,Q,i}(S_{k,i}(\tau))}{d(S_{k,i}(\tau))}\right] = \frac{\text{Var}\left[\widehat{z}^*_{k,Q,i}(S_{k,i}(\tau))\right]}{d(S_{k,i}(\tau))^2}$$

As shown in [107], this can be solved analytically for each queue. For the non-trivial case of $S_{k,i}(\tau) \geq \tau$ (i.e. the case that $z_{k,Q,i}(\tau) > c_{k,i}(\mathbf{z}(\tau))$):

$$S_{k,i}(\tau) = \min\{t \geq \tau | z_{k,Q,i}(t) = c_{k,i}(\mathbf{z}(t))\}$$

such that, $d(S_{k,i}(\tau))) = c_{k,i}(\mathbf{z}(S_{k,i}(\tau)))\mu_{k,i}(S_{k,i}(\tau)) + \frac{dc_{k,i}(\mathbf{z}(t))}{dt}|_{t=S_{k,i}(\tau))}$. In the next chapter, I solve these equations using MATLAB's built-in ODE solver ode45; however, they may be solved numerically in open source software/software that is readily available to health services.

5.5.5 Production of outcomes

Within a given time period, (5.13)-(5.19) may be adapted to measure the production of outcomes from a service. That is, the number of patients who leave the system at a point in time and are in a given health state. This includes those who leave due to abandonment, $P_{k,L,i}(t)$, and completing service, $P_{k,S,i}(t)$. Over a period of time $[t_s, t_e] \subseteq [0, T]$ the production of patients in health state $k \in H$ from a service $i \in Ser$ is given by:

$$P_{k,S,i}(t) = \sum_{l=1}^{K} MS^{(k)}_{l,S,i}(t_e) - MS^{(k)}_{l,S,i}(t_s) \tag{5.46}$$

$$P_{k,L,i}(t) = \sum_{l=1}^{K} MS^{(k)}_{l,L,i}(t_e) - MS^{(k)}_{l,L,i}(t_s) \tag{5.47}$$

with the analogous fluid approximation of:

$$p_{k,S,i}(t) = \int_{t_s}^{t_e} \sum_{l=1}^{K} s_{l,k,S,i}(u)\mu_{l,i}(u) \min\left(z_{l,Q,i}(u), c_{l,i}(\mathbf{z}(u))\right) du \tag{5.48}$$

$$p_{k,L,i}(t) = \int_{t_s}^{t_e} \sum_{l=1}^{K} s_{l,k,L,i}(u)\theta_{l,i}(u)(z_{l,Q,i}(u) - c_{l,i}(\mathbf{z}(u)))^{+}\, du \tag{5.49}$$

This measure can help to understand how different capacity allocations and changes in time varying systems may affect the output of patients in certain health states from a system, and the system's impact on patients' health.

5.6 Summary and Discussion

Fluid and diffusion approximations for stochastic processes are efficient methods for modelling complex systems of queues that may otherwise be computationally intensive or analytically intractable. I have shown that several flow dynamics may be modelled by these methods, including the sequential use of multiple services, abandonment, rejoin, reuse, health states, and health and time dependent parameters. Furthermore, I have shown that these extensions are mathematically valid.

Whilst the fluid limit is a deterministic approximation of the stochastic system, the variance can be calculated by formulating the diffusion limit. Thus, one can understand both the expected behaviour and variance of the number of patients in each process state and of the virtual waiting time for each queue within the system.

A multifaceted view of system performance can be analysed by combining patient flow and clinical outcomes into a single modelling framework. By using health states the flow of patients with differing resource/service requirements and different capacities to benefit from care. In particular, the model's output is informed by the effect of care, or absence of it, on patient health and the effect of patients with different health care requirements, e.g. service times, on the operation of the system. This may reflect real life where patients of varying health and care needs have markedly different interactions with a health service. This may provide greater insight into the positive and negative clinical effects of a system's process outcomes - providing

a framework for modelling the "flow of outcomes" - discussed further in chapter 6.

Traditionally, parallel queues are inefficient due to the possibility of inactive servers. However, this limitation is overcome by using a dynamic multi-class server allocation since servers are continuously reallocated. Thus, there is no possibility of inactive servers if a queue exists for a service. As a result, the benefits of using multiple queues to represent different health states may be fully utilised. That is, differentiated services can be modelled in order to understand how patients with different levels of health may affect the performance of the system and be used to measure the performance of the system. Furthermore, this priority type allocation can handle the complex flow dynamics of re-entrant patients and may in fact be defined to specifically depend on the process orbits that represent these dynamic.

5.6.1 Limitation

As noted earlier, fluid and diffusion approximations are most accurate for large and heavily loaded systems. There are two reasons for this. Firstly, the approximations are formed by scaling the number of servers and arrivals in the system; thus, by construction, they are more accurate for larger systems. Whilst this may be a limitation, it is also a benefit since the method is scalable and can maintain accuracy and efficiency for larger systems. Secondly, for heavily loaded systems the behaviour of the queues becomes "more deterministic", such that these deterministic approximations hold with greater accuracy. However, this may again limit when and how these approximations may be used.

A second limitation is that, in considering multiple services and several health states, the system can become unwieldy due to the number of input parameters. Furthermore, when time dependence is considered, the implementation of this system can become more complex to code. Whilst the solution to these ODEs is efficient, editing and changing the inputs can be time consuming, especially if several configu-

rations of the system are analysed. One way to overcome this is to use a configurable interface for entering the inputs needed to run these models.

A final limitation, as identified by [107], is that the solution cannot "linger" near $z_{k,Q,i}(t) = C_{k,i}(\mathbf{z}(t))$ when calculating the VWT and its variance. This is because the approximation hinges on the assumption that $\{z_{k,Q,i}(t) = C_{k,i}(\mathbf{z}(t))|t > 0\}$ is of zero measure. As suggested by [107], an alternative, more general formulation of these equations may be used to overcome this, as in [104].

In the next chapter I will explore how these limitations affect the application of these methods within community health care.

5.6.2 Possible avenues for future work

There are several directions in which this work could be extended, adapted or used. Firstly, the concurrent use of multiple services was identified in chapter 2 as a key dynamic of community health care. This would form a useful avenue for future work in understanding how the concurrent uses of services may produce good patient health and process outcomes. One potential direction is to define states that represent the combination of services. For example, a system of two services would be represented by three service states; two pertaining to the use of each single service and one representing the combination.

Secondly, the stochastic system is formulated as a Markovian system with time and health state dependent parameters. Future extensions for this work could involve the relaxation of the Markovian assumptions to form a more generalisable approximation. In addition, it would be insightful to use different parameter definitions. This could include mechanisms for loss that are dependent on the number of patients in different parts of the system, or the introduction of finite waiting space.

Thirdly, the definitions of the capacity allocation $C(\mathbf{Z}(t))$ are illustrative of the types that may be defined by using a fluid and diffusion approach. It would be worth-

while exploring the benefits that the continuous representation of the system may have in producing different definitions that may enable a range of avenues for analysis such as optimisation, capacity allocations and priority queueing. Furthermore, through the combination with patient outcomes novel constraints and objectives may be considered in such analysis and may be used to inform capacity allocations and design referral pathways. For example this could be to maximise positive patients outcomes or minimise adverse flow patterns that lead to poorer outcomes. Similarly, health states have been defined to change at: the completion of service; the point of abandoning the queue; or, upon joining the queue as a rejoin, reuse, alternative service arrival or other service arrival. Hence, it would also be valuable to explore alternative definitions and configurations of this system such as time dependent health transitions that may change whilst a patient resides in a given state.

Fourthly, the methods developed in this chapter are general and may apply to other sectors of health care or industry. They model systems of queues through which a population of entities with different capacities to benefit from service may flow, and whose service times, propensity to abandon, and subsequent use of service, differs. These dynamics may translate into industries such as telecommunications, where health states may be defined as states of satisfaction or opinion.

5.7 Conclusions

In this chapter, I have developed a method for modelling queues of heterogeneous patients whose group (health) may change throughout the service process. These methods contribute to the way in which community services may be modelled, and how patient outcomes may be incorporated into patient flow modelling.

These methods are beneficial in three ways. Firstly, there is a methodological benefit. The approximations are solved as a set of ODEs that are efficient to solve,

even as the system grows large, providing informative performance measures. These include: the number of patients within different health states, process orbits and services in the systems; the virtual waiting time for each service; the variance of each; and, the production of outcomes.

Secondly, complex dynamics may be modelled using these methods, such as: patients reusing a service; future referrals to other services; and, the potential for patients to abandon and potentially rejoin the queue or use another service. Whilst potentially rendering the system analytically intractable or computationally intensive, they can be modelled by producing fluid and diffusion limits.

This leads to the third benefit, that the combination of health states and patient flow provides new avenues for insightful analysis within community care. The methods developed in this chapter highlight how two key perspectives of performance in health care may be united in a single modelling framework. These methods may be used to help understand: how patients use services; the effect of multiple care interactions on patient health; the effect of delayed demand/reuse of services on the operation of the system and on patient health; and, how a dependency between capacity of the system and the future arrival process affects the system.

Moreover, the work presented in this chapter is in line with the findings of this thesis so far. Explicitly, I have produced a time-dependent method that can be numerically solved for modelling patient flow within systems of diverse community services, that may consider the mix of patients who use them (as per chapter 2). Additionally, I contribute to the way in which complex flow dynamics, such as patients reusing services, transitions between multiple services and transitions in patient health may be modelled (as per chapters 3 and 4).

Finally, by extending and combining existing fluid and diffusion approximation methods, I increase the scope for the application and use of these approximations in various settings.

Chapter 6

Application of fluid and diffusion approximations to patient flow in community health care

Given the lack of available data for model validation, in this chapter I present a theoretical understanding of the model. I begin with a discussion of how the methods developed in chapter 5 may apply to community health care, considering how the approximations may be used to represent real world systems.

In addition, I explore the parameter space for these methods to assess when they are most accurate. Starting with a simple steady state model and extending to a multiple health state time dependent scenario, I build up an understanding of when these models are accurate, how the extensions change this and the parameters that are important in determining the accuracy of the method.

I assess the accuracy of the approximations in comparison to simulations, considering the effect of different parameters, empty and non-empty initial conditions, steady state and time varying scenarios, pre-allocation of servers and dynamic multi-class server allocation, and multiple health states with different transition probabilities. Having carried out this analysis, I will discuss how these methods may be used

to model the "flow of outcomes".

The aims of this chapter are to:

1. Discuss how these methods may apply to community health care;
2. Assess the accuracy of the approximations in response to different input parameters and flow dynamics;
3. Develop understanding of the "flow of outcomes" and how it may be modelled.

6.1 Introduction

It is widely established within the literature on fluid and diffusion approximations that they increase in accuracy when modelling systems that are large (many servers) and heavily loaded (when the demand for service persistently exceeds the capability of services to meet it) [102]. This is especially true when abandonment of the queue and rejoins are modelled [105]. Recognising this, there is a need to understand when the methods developed in chapter 5 are accurate. In particular, which input parameters are important for maintaining accuracy, under what conditions the system is heavily loaded and for what size of system accuracy is maintained.

Due to the dynamics considered in chapter 5, the usual definitions of traffic intensity, server utilisation and when systems are heavily loaded, do not hold. In a simple queueing system, *traffic intensity*, denoted a, is a measure of congestion within the system, calculated as the ratio between the mean arrival rate and mean service rate, $a = \lambda/\mu$. In a multi-server system, *service utilisation* is the traffic intensity per server, denoted $\rho = a/c = \frac{\lambda}{c\mu}$.

For a simple stochastic queue, where there is no loss or abandonment, when $\rho < 1$ the average queue length is finite throughout time. Alternatively, when $\rho > 1$

a system is *heavily loaded* such that the queue grows without bound as demand outstrips the service's ability to serve all those who arrive. Notably, this may not be the case if loss is considered.

However, in chapter 5 arrivals to each service consisted of: new patients; patients rejoining the queue or using an alternative service after abandonment; and having finished service, patients reusing a service or using another one. Furthermore, the service process included transitions in health state. As a result, several parameters may be significant in determining whether the queues are heavily loaded in the extended system; thus, determining the accuracy of the approximations in modelling the stochastic system.

I will analyse the accuracy of the approximations by comparing them to a simulation of the stochastic system, and find the pragmatic constraints this places on the input parameters. To do so, I explore several scenarios (some relevant to community health care and some not) in order to test the approximations and identify key limitations. From these findings, I discuss how additional dynamics further effect the use and applicability of these methods.

This chapter makes two main contributions. Firstly, by carrying out a theoretical investigation of the model, the limitations of these approximations in modelling the stochastic system will be understood as I discuss where and how the method may be used for accurate analysis. Secondly, I highlight how these limitations may affect the application of these methods to community health care and the modelling of the "flow of outcomes".

Structure of chapter

In section 6.2, I discuss how the model parameters and dynamics may be interpreted in application to community health care, identifying scenarios for which these models may be used. In section 6.3.1, I briefly describe a simulation model used for

comparison and how the errors between the two models are measured; after which, in section 6.3.2, I conduct steady state analyses of the base system (a single service and a single health state). Here I explore several scenarios and consider how changes in different parameters affect the accuracy of the model. In addition, I introduce time varying behaviour to understand how seasonal changes in the arrival rate of new patients affect the accuracy of the system over time. To conclude this section, I summarise the limitations of the approximations in representing the stochastic system and how this informs the understanding of when the system is heavily loaded. The overall aim of this section is to develop understanding of the most basic model, when it is most accurate and what parameters are important in maintaining accuracy.

Having established the above, in section 6.3.3, I consider a case with two health states. I begin with a steady state analysis to show how health state transitions cause the limitations and parameters that determine the accuracy of the method to differ from those outlined in section 6.3.2. Secondly, I explore a time varying system, introducing a dynamic server allocation for multi-class queues and considering a seasonal spike in arrivals, again to further understand how accuracy may be affected.

In section 6.3.4, I discuss how the work in previous sections informs the understanding of when the complete system - multiple services and transitions between several health states - is accurate in comparison to simulation. Given these findings, the method is illustrated by an application of the fluid and diffusion methods to a larger system. This chapter ends with a discussion of these methods and their use in modelling the "flow of outcomes".

6.2 Application to community health care

First note that conceptually, these methods are appropriate for modelling some aspects of patient flow in community health care due to the dynamics they capture.

Following from chapter 3, these methods include sequential uses of multiple services and the potential for patients to reuse services. Thus, in practice and in the model, the demand for these services is formed of new, previous and current users of care. This effectively creates a higher traffic intensity in the system, increasing the theoretical applicability of these methods (discussed in detail later). In addition, from discussions with community care leads, it was acknowledged that some community services operate with a persistent waiting list throughout the year which may be modelled as a permanent, non-physical queue.

With this in mind, before assessing the accuracy of the approximation method, I now discuss how the parameters and patient flow dynamics within the model may represent different aspects of community health care.

The queues

The methods developed in chapter 5 can be used to model physical and non-physical queues. Considering community health care, a non-physical interpretation is a natural approach to take, such as a waiting list for a service. Whilst in some cases this may have a fixed capacity, I model an infinite waiting space, which is reasonable for many scenarios when patients wait away from the service.

For analytical tractability, I have considered multiple parallel queues each pertaining to the demand of patients in each health state. These queues use servers from a single pool, with patients served on a first come first served (FCFS) basis in each queue. In some real world settings, there may only be a single waiting list made up of patients from all health states; thus, for these scenarios, the use of multiple queues would not be appropriate. However, parallel queues may apply to situations where the waiting list is effectively divided into several lists - such as services made up of several types of health care professional, or where patients with a mixture of morbidities may attend, thus requiring different types of care. This may occur when servers

are reserved/allocated to meet the demand of different groups of patients, or when patients are managed in different groups. In each case, these groups may consist of different health states/health outcomes, different capacities to benefit, different health care needs, or certain morbidities.

Service process

As identified in chapter 3, how patients use services may be markedly different depending on the service and needs of patients. In some instances, a referral may consist of a single use, whilst in others it may represent multiple visits over several days, weeks or months. In this chapter, I consider service to represent the span of a patient's referral - lasting from the start of their first appointment until the end of their final appointment. As a result, referrals may consist of a single appointment or several (as discussed in chapter 3, see Figure 3.4).

Having defined the service process as a continuous, time-inhomogeneous, exponentially distributed process in chapter 5, service time $1/\mu_{k,i}(t)$ represents the average length of referral for patients at time t within health state $k \in H$ for a service $i \in Ser$. Thus, for modelling purposes, the number of appointments that occur does not need to be known, nor the date, time or length of each appointment. Rather, only the average length of referral from start to finish is needed (if required, health states may be defined to reflect information such as the number of appointments).

Using this interpretation, servers may be defined as individual clinicians. However, in modelling the length of patient referrals, a server will be considered busy throughout this period while serving a patient, which is unlikely to be the case in practice. Rather, over the length of a patient's referral, a clinician may treat several patients, managing multiple referrals at a time. To model this, servers may represent the maximum number of patients a service can handle/serve at any one time, given the possible mix of patients. Thus, for example, if a community service consists of

five clinicians who can each manage a maximum of five referrals at any one time, the service is modelled as having 25 servers overall. However, defining the maximum number of referrals a service can theoretically handle is not straightforward; thus, this definition may require a hypothetical/potential capacity.

Abandonment, rejoin and use of alternative services

Abandonment could be used to model patients who leave the queue having died, left the geographical area of service, used another service, or who no longer require care. Considering rejoin, a novel interpretation of abandonment is the use of a health care service outside of the community system e.g. acute services. In particular, abandonment would be considered to occur when the health of a patient, who has waited for a significant length of time, deteriorates such that they require immediate care or have become impatient and sought care elsewhere.

Considering the alternative community service orbit, the interpretation is similar; however, the impact on other community services in the system may be modelled. In this case, the time spent in the alternative service orbit before entering the queue, may be interpreted as: a use of an acute service, the time between leaving the queue and self referring, or the time taken to receive a referral from primary care to an alternative service.

These interpretations may be useful for services that treat long-term conditions because patients often require multiple care interactions over an extended period of time. Likewise, they may be useful for modelling when, due to a long wait, a patient feels they no longer require care as their symptoms subside, yet may require care in the future if the underlying problem still persists. In each case, the use of an acute or primary service may not negate their need for future use.

Furthermore, this interpretation can provide an insightful measure for community health care. As highlighted in chapter 4, a key goal in the provision of community services is to reduce the volume of avoidable acute and primary care demand. Therefore, by modelling a system that possesses high levels of abandonment in this way, flaws in its capacity management and provision may be understood. For example, instances where a delay in care led to increased acute/primary demand may be modelled (only if abandonment is interpreted as a use of care elsewhere). Likewise, the effect of patient behaviour on the operation of these services may be modelled.

However, by interpreting the queue as a waiting list, patients are required to be removed from the list upon using another service, or for them to actively remove themselves and be subsequently re-added as they rejoin - which may be unlikely to occur in real life. Whilst this may not always be true, the use of loss and rejoin provides a helpful measure that is descriptive of community health care. In particular, it can help in understanding how a lack of capacity and long waiting times may influence patient decisions and their health.

Reuse and uses of other services

The reuse orbit and other service orbit may be used to model the future needs and demand of patients who have already used community care. In the case of reuse, this may represent a later episode of care, whilst arrivals from other services could represent a formal referral. In the latter case, the time spent in the orbit may be negligible if patients immediately join the waiting list. However, as before, if this orbit represents a self referral or a referral from primary care between community visits, the time spent in the other service orbit may have a significant value.

Having discussed the possible interpretation of the model, I now begin to explore when the approximations are accurate and how they may be used.

6.3 Exploration of the accuracy of fluid and diffusion approximations for modelling community health care

6.3.1 Description of the simulation model

To evaluate their accuracy, I compare the output of the fluid and diffusion approximations to the equivalent results from a simulation of the stochastic system. Produced in MATLAB, I use discrete event simulation (DES) methods for modelling the stochastic system (see Appendix C for the code). This method is *event* driven with time progressing as events occur. In the system described in section 5.3, an event relates to an arrival of a patient (new, rejoin, reuse, other service, alternative service), the completion of service or an abandonment from the queue. When these events occur, patients may also transition in health state or enter the rejoin/reuse orbits; however, these are instantaneous and part of the above events.

I used DES methods since they provide an intuitive way to model this system given the number of possible events. To this end, I created a single script for modelling several scenarios (steady state, time varying, single or multiple service, single or multiple health states, and all or some of the flow dynamics).

The simulation includes the mechanism of pre-emptive resumption as discussed in section 5.2. Notably, this mechanism may not represent a real life process since servers may remain active until the patient they are serving completes care. However, if it is not feasible for servers to carry on past their "active" period, pre-emptive resumption ensures that all patients complete care once started, and allows for a time varying number of servers. Since the simulation is not designed to be used for single iterations (rather, I use it to find the average behaviour across multiple runs)

this is a reasonable approach to take. For instance, due to stochastic variation, the impact of pre-emptive resumption will vary for each run and will average out over multiple simulations.

Model outputs

Within DES methods it is common to output every event that occurred and when it occurred. However, since the simulation is used to find the average behaviour of the system, this would increase the difficulty in comparing the results. For example, in running the simulation several times, the size of the results matrix would vary, outputting different time markers for each iteration. Instead, I programmed the simulation to output a solution at a given time step, dt, producing an output of length $T/dt + 1$, where T is the total modelled period. This is reasonable since it creates a standardised solution, providing key information throughout the modelling time frame. With careful selection of dt, any loss of detail is insignificant.

Another reason for producing outputs of length $T/dt + 1$ is for easy comparison to the fluid and diffusion approximations. Having used numerical methods for computing these approximations, such as forward Euler and the trapezium rule, the time step is important in determining the accuracy of these solutions and their stability. If the time step is too large, the solution will be less accurate and, in some cases, unstable, see Figure 6.1. By reducing the time step, the stability and accuracy improves; however, this increases the number of calculations required, thus increasing the computation time. This highlights the need to balance accuracy and running time in selecting dt.

By using these numerical methods, solutions of the approximations are calculated at each time step, producing outputs of length $T/dt + 1$. Therefore, formulating the output of the simulation as above helps for comparison between simulations and the numerical solution of the approximations.

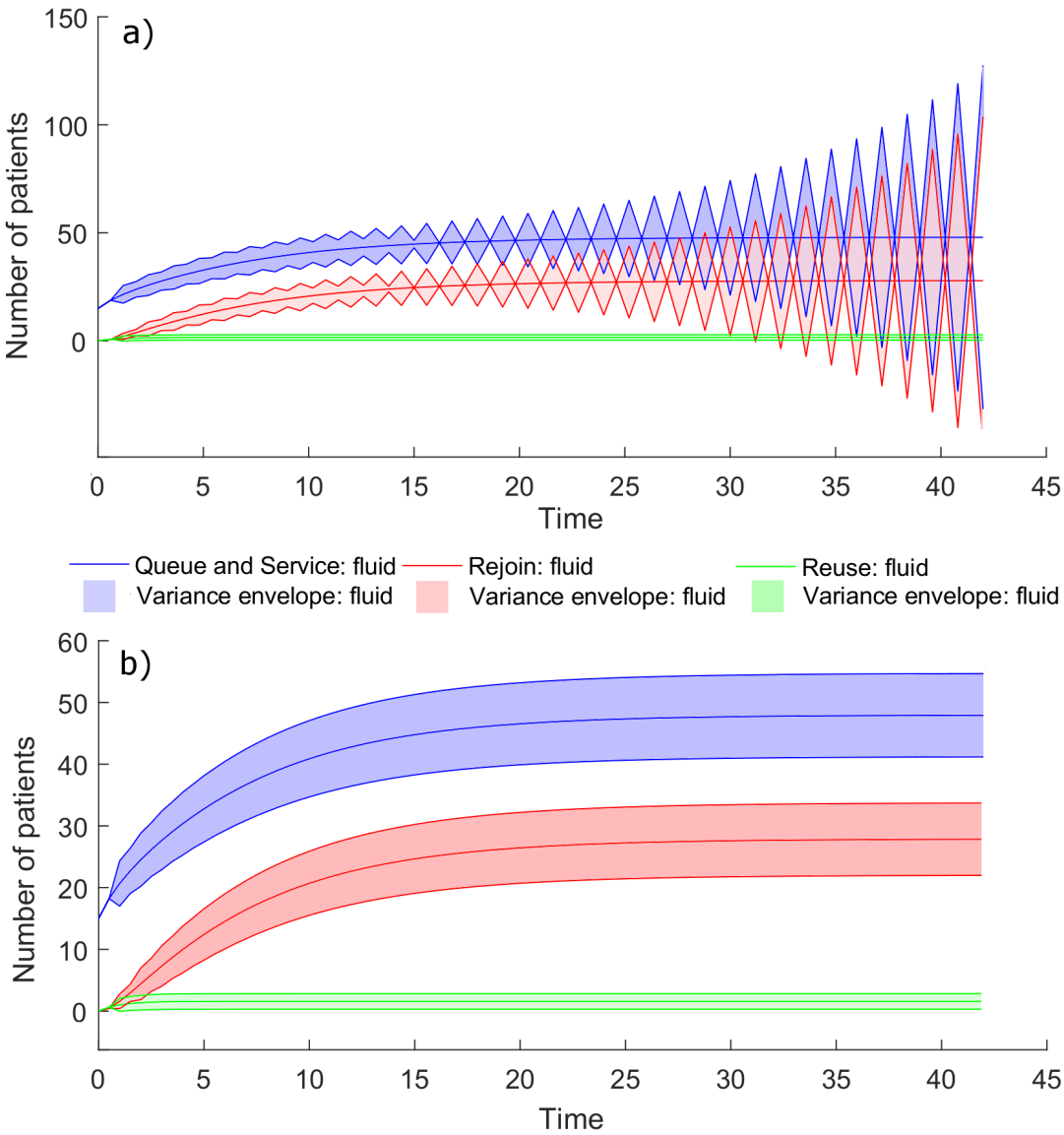

Figure 6.1: (a) Stability issues with forward Euler calculation for the variance when time step is too large, $dt = 0.6$; (b) Improved stability and accuracy of solution for smaller time step, $dt = 0.5$.

Since I have no data or real world scenario to apply the approximation methods to, I take the simulated solution to be "true". Therefore, the aim is to evaluate when and how the fluid and diffusion approximations are faithful to this process. As noted in chapter 5, simulation is one method for modelling these systems when they get large due to the complexity. Thus, I compare two methods that are appropriate for modelling this complex system. Other methods include system dynamic approaches and Markov chains.

Error measurement

To find the average solutions and variance of the simulation, I chose to use 1000 runs for two reasons. Firstly, in comparing the output from 100 (as in [105]), 500, 1000, and 5000 runs (as in [107]), I found that for 1000 runs the stochastic variation is greatly reduced in the averaged solution, providing comparable results, see Figure 6.2 (modelling a single health state, using parameters in Table 6.2, with $q = 0.5$ and $c, \lambda = 20$). Secondly, the number of runs has a significant impact on the running time of the simulation, see Table 6.1.

Number of runs	**100**	**500**	**1000**	**5000**
Total run time (s)	8.78	42.33	84.92	472.68

Table 6.1: Total computation time for the simulation as the number of runs increases

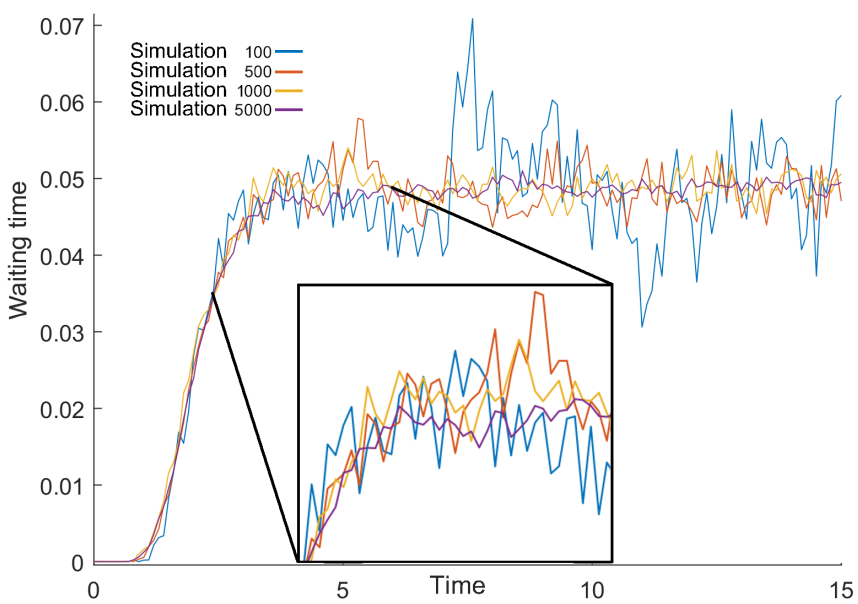

Figure 6.2: Example of increased accuracy in the variance of the simulated waiting time as the number of runs increases

If the simulation was solely used to assess this system, 1000 runs would be reasonable for gaining reliable results. Hence, this analysis will provide a comparison of the approximations to an appropriate method for evaluating this system.

To measure accuracy, I consider the errors between the average number of patients

in each process state, the virtual waiting times (VWT) and the variance of each. For $m \in \{Q, R, U\};\ p, q \in \{0, 1, 2, ..., T/dt\};\ p < q$ such that $t_p = dt \times p$:

$$\text{err}_m(t_q - t_p) = \frac{\sum_{r=p}^{q} \mathbb{E}Z_m(t_r) - z_m(t_r)}{\sum_{r=p}^{q} \mathbb{E}Z_m(t_r)} \tag{6.1}$$

$$\text{verr}_m(t_q - t_p) = \frac{\sum_{r=p}^{q} \text{Var}(Z_m(t_r)) - \text{Var}(z_m(t_r))}{\sum_{r=p}^{q} \text{Var}(Z_m(t_r))} \tag{6.2}$$

$$\text{werr}(t_q - t_p) = \frac{\sum_{r=p}^{q} WT^{\text{Sim}}(t_r) - \text{VWT}(t_r)}{\sum_{r=p}^{q} WT^{\text{Sim}}(t_r)} \tag{6.3}$$

$$\text{wverr}(t_q - t_p) = \frac{\sum_{r=p}^{q} \text{Var}(WT^{\text{Sim}}(t_r)) - \text{Var}(\text{VWT}(t_r))}{\sum_{r=p}^{q} \text{Var}(WT^{\text{Sim}}(t_r))} \tag{6.4}$$

where WT^{Sim} indicates the waiting time gained from the simulation.

6.3.2 Single service and a single health state

Having described a large an complex system in chapter 3, I begin by analysing the approximations in their simplest form. The purpose of this section is to understand the effect that different parameters have on the approximations' accuracy and on the understanding of when the system is "effectively" heavily loaded. Analysis of these approximations in their most basic setting is conducted in order to understand how this may differ as more dynamics and health states are introduced.

The analysis and insights gained add to those published in [105, 107] since: the effect of changes in a range of different parameters is considered, the analysis is conducted over a split time interval, and time dependent behaviour is modelled. Analysing the system over two time intervals shows the accuracy of the model during queue formation and as the system reaches steady state, providing insight into the accuracy of modelling time varying behaviour. This also mitigates the bias that the length of the modelled time period introduces (discussed below).

The parameters used in each scenario are informed by discussion with NELFT staff but do not relate to data. Thus, a hypothetical example is given which may be interpreted for a range of different services e.g. the service time may be interpreted as any length of time, as long as all other times are consistent with its definition.

Steady state analysis

To begin, I conduct a steady state analysis of a single service, single health state model as in [105]. Given that the definition for traffic intensity per server, $\rho = \lambda/c\mu$, does not hold for systems with rejoin and reuse, the authors of [105] identify an effective traffic intensity for this system: $\widetilde{\rho} = \frac{\lambda}{c\mu(1-q)}$, where q is the probability that a patient seeks to reuse the service. This provides a reasonable upper bound for the busyness of servers in the system as it considers that "a proportion of at least $\frac{1}{1-q}$ of ρ will reuse services". In considering this scenario, I will use the notation from [105] where q is the probability that a patient reuses the service after receiving care, and p is the probability that a patient rejoins the queue after abandoning. Rejoins are not included in this calculation since they are captured by the initial λ, and the definition of the traffic intensity does not change for systems with loss.

Given the above, I first investigate the accuracy of these methods as q, the size of λ and the size of c, vary for a system that begins empty and where $\rho = 1$. The parameters used for this comparison are listed in Table 6.2. Whilst in reality this initial condition is unlikely to relate to community services, beginning with an empty system allows for the analysis of the approximations' accuracy when the system moves from being underloaded ($z_Q(t) < c(t)$) to overloaded ($z_Q(t) > c(t)$).

By inspecting the results of the simulation and approximations when $T = 20$, $q = 0.5$ and $c, \lambda = 20$ (Figure 6.3), there are two distinct phases within the solution. This is further highlighted by Figure 6.4 - the error between the two models' outputs for the number of patients in each process state, and Figure 6.5 - the errors in the

Parameter - Value	**Description**
$dt = 0.1$	Time step for the numerical scheme used to solve fluid and diffusion approximations
$T = 15$	Total time period of solution
$\lambda \in \{10, 15, 20\}$	Value range of arrival rates
$c \in \{10, 15, 20\}$	Value range for total number of servers
$\mu = 1$	Service rate
$\theta = 1$	Rate of abandonment
$p = 0.3$	Proportion of patients rejoining after abandonment
$q \in \{0.1, 0.3, 0.5, 0.7\}$	Value range for proportion of patients reusing the service
$\delta_R = 1$	Rate of rejoin after abandonment
$\delta_U = 1$	Rate of reuse after service

Table 6.2: Parameters used to assess the accuracy of the approximations - steady state analyses of the effect of c, λ and q

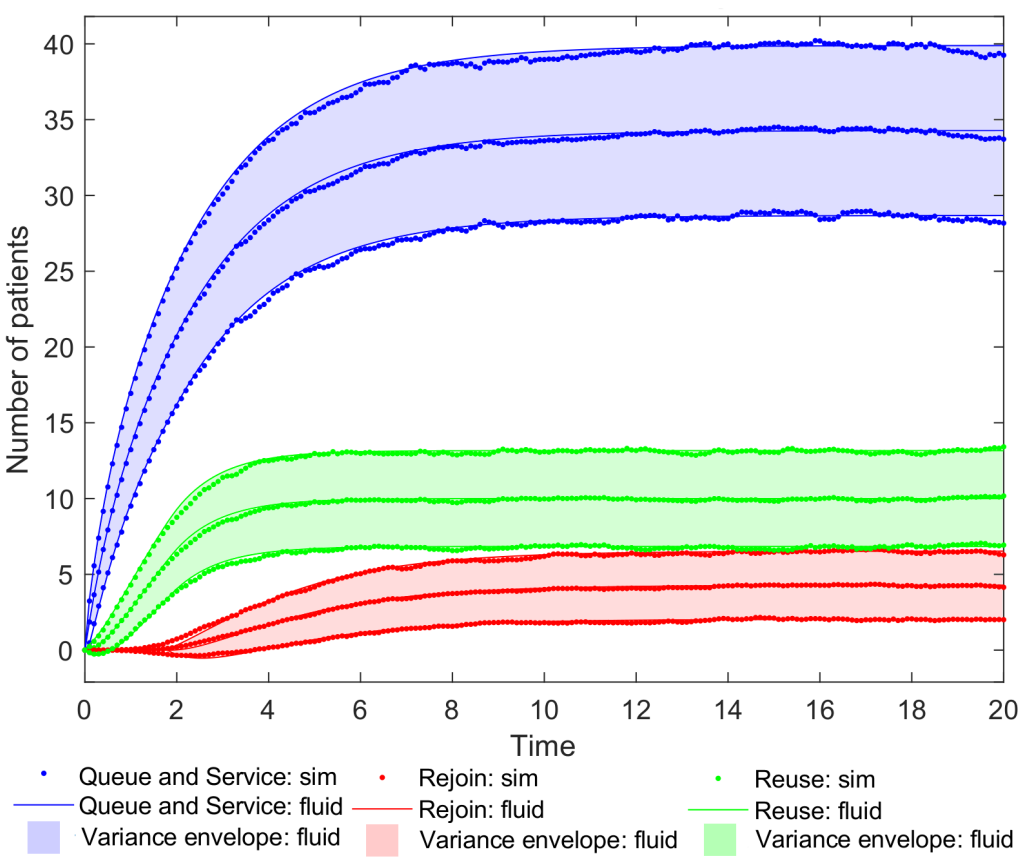

Figure 6.3: Comparison of results from the approximations and the simulation for a system with $\lambda = c = 20, q = 0.5$ and $\widehat{\rho} = 2$

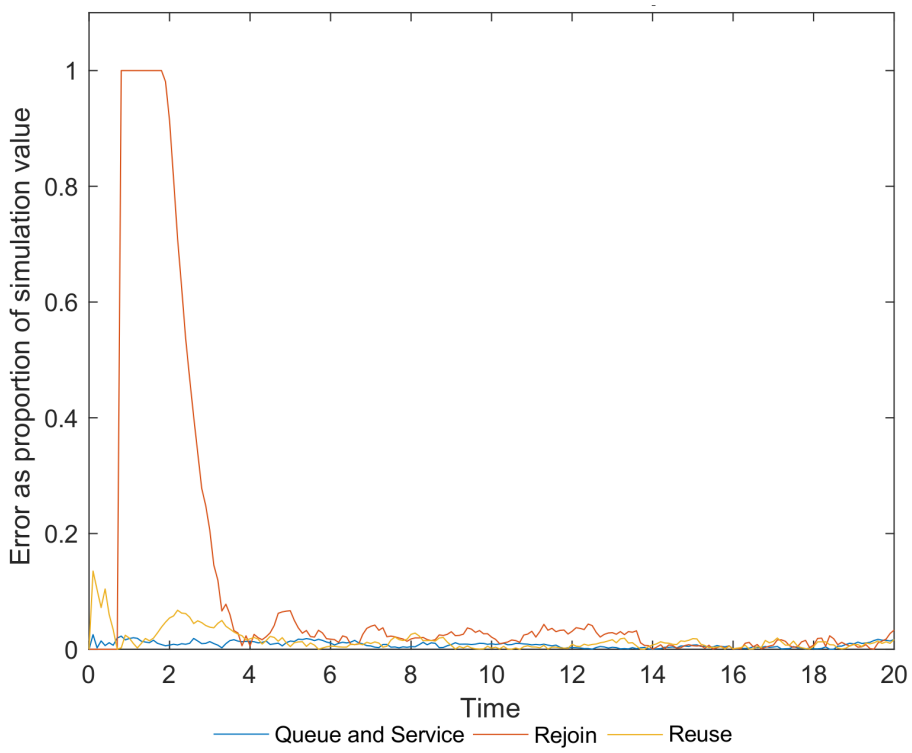

Figure 6.4: Error in the average value of each process state. After time $t = 2.8$ the accuracy of the fluid approximation improves greatly.

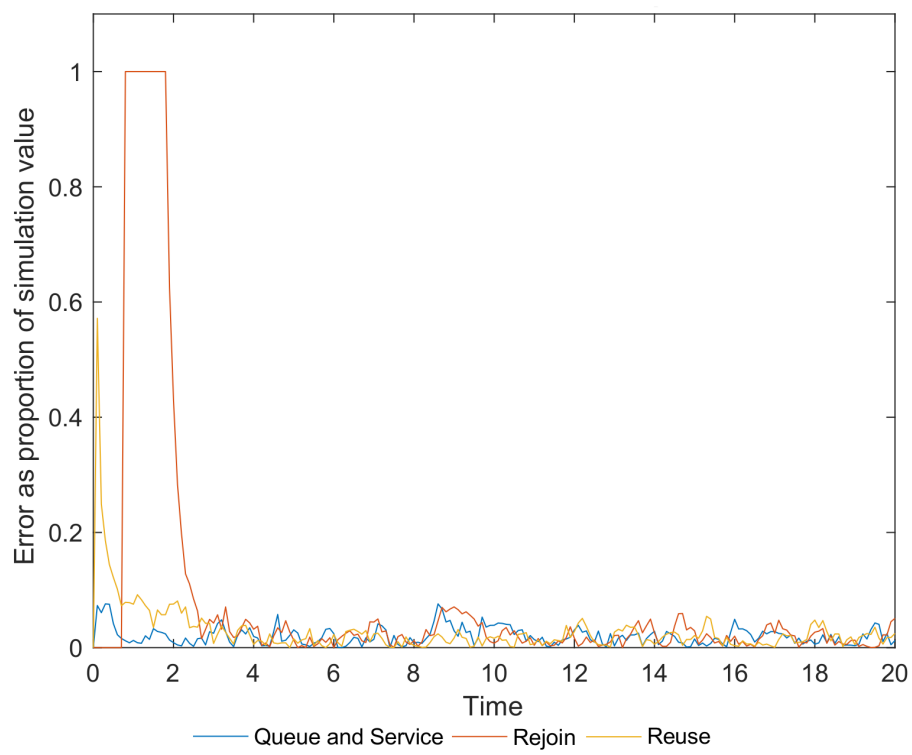

Figure 6.5: Graph of error in the variance of each process state. After time $t = 2.8$ the accuracy of the diffusion approximation improves greatly.

variances given by each model for the number of patients in each process state. Between time $[0, 1.8]$ there is a large error between the approximations and the simulation which diminishes over $[1.8, 2.8]$. Notably, in Figures 6.4 and 6.5 the error for $z_R(t)$ reaches 1, indicating a 100% error. Whilst proportionately large, this is given by $Z_R(t)$ divided by itself since $z_R(t) = 0$ for $0.3 \leq t \leq 1.8$. This error is explained by the system starting underloaded, and the difference between when queues form in the stochastic and deterministic processes.

In a real world system, when servers are free, new arrivals immediately enter service until the system reaches a critical point $z_Q(t) = c(t)$. Subsequent arrivals then form a queue such that $z_Q(t) > c(t)$ from which they may begin to abandon.

Due to random variation in the arrival process for a stochastic system, the existence and size of the queue fluctuates in time such that abandonment may occur throughout the whole time frame. However, within the fluid approximation, this variation does not occur. Instead, there is no queue or loss within the fluid system until the critical point is reached. This delay between the two models causes the initial inaccuracy, after which the approximations may recover and become accurate.

With this in mind, I produce two errors for the system relating to two periods of time. Firstly, I calculate errors as the queue forms in the fluid approximation - for $[0, T_I]$, where $T_I = \max\{t + 1 | z_Q(t) \leq c(t)\}$ - denoted as the "formation error". For the system described in Table 6.2, these errors are shown in Table 6.3. From this I will understand the size of this initial error, the length of time for which this error occurs (T_I) and how the size of the system and $\widetilde{\rho}$ affects both of these. Secondly, I calculate the errors for the remaining time period: $(T_I, T]$, shown in Table 6.4.

Since the errors may diminish as the system reaches steady state, the size of T affects the error measurement if (6.1)-(6.4) are simply calculated over the whole of T. In particular, for larger T the errors will be lower because they are calculated over a large period of time for which the system is more accurate. Thus, informed

by Figures 6.3, 6.4 and 6.5, I consider $T = 15$ to ensure the system has enough time to reach steady state in each situation, whilst diminishing the impact of large time frames on the reported errors. Furthermore, by splitting the time interval and reporting two errors for each scenario, I mitigate the bias of T. Finally, given the modelled parameters, I have chosen T such that $T_I < T$.

It is clear from Tables 6.3 and 6.4 that the approximations are progressively more accurate as the size of the system (number of servers and number of arrivals) and the effective traffic intensity (in relation to the size of q) grow. The method is accurate for the queue and service process (z_Q) and the reuse process (z_U), even during the

		$q = 0.1$		$q = 0.3$		$q = 0.5$		$q = 0.7$	
		$T_I = 3.8$		$T_I = 3.1$		$T_I = 2.8$		$T_I = 2.6$	
c, λ		err	verr	err	verr	err	verr	err	verr
10	z_Q	1.20	2.42	0.76	3.11	0.55	2.42	0.59	3.61
	z_R	95.45	71.96	82.80	56.91	72.15	48.56	62.16	39.22
	z_U	6.09	4.64	6.43	3.70	6.84	7.72	7.53	5.98
15	z_Q	0.79	3.2	0.69	1.52	0.64	2.62	0.59	1.73
	z_R	94.02	66.96	78.40	50.69	65.01	39.91	57.59	32.10
	z_U	4.48	4.66	4.16	5.55	5.69	7.34	4.75	4.80
20	z_Q	0.51	3.29	0.53	2.21	1.54	2.53	0.82	1.62
	z_R	93.33	65.98	72.97	44.19	63.38	43.74	50.96	29.79
	z_U	6.27	7.49	4.06	5.97	2.92	4.04	3.21	3.04
c, λ		werr	wverr	werr	wverr	werr	wverr	werr	wverr
10	z_Q	91.20	75.46	73.71	62.05	61.98	55.91	51.62	50.45
15	z_Q	89.30	73.21	68.31	60.16	52.68	51.84	45.48	45.19
20	z_Q	87.38	74.61	63.11	56.44	49.20	52.25	37.60	44.61

Table 6.3: Errors between the approximations and the simulation during the "formation period" as a percentage of the simulated solution - effect of q, c and λ

		$q = 0.1$ $T_I = 3.8$		$q = 0.3$ $T_I = 3.1$		$q = 0.5$ $T_I = 2.8$		$q = 0.7$ $T_I = 2.6$	
c, λ		err	verr	err	verr	err	verr	err	verr
10	z_Q	1.78	2.02	0.67	1.44	0.61	2.07	0.73	2.20
	z_R	46.41	3.01	10.17	2.24	3.15	2.95	1.79	2.13
	z_U	7.50	5.16	3.53	2.84	1.73	2.51	1.08	1.63
15	z_Q	1.48	2.63	0.47	1.61	0.75	1.51	0.37	1.47
	z_R	39.23	6.21	6.28	2.24	2.17	2.27	1.26	1.95
	z_U	5.16	2.24	1.74	3.11	1.30	2.00	0.69	1.92
20	z_Q	0.93	2.73	0.37	2.05	0.33	1.71	0.37	1.68
	z_R	32.73	7.74	3.23	1.60	1.76	1.67	1.17	2.04
	z_U	5.42	4.02	1.58	1.66	1.08	2.19	1.17	1.78
c, λ		werr	wverr	werr	wverr	werr	wverr	werr	wverr
10	z_Q	48.47	22.90	17.33	3.50	9.60	2.61	7.61	2.52
15	z_Q	40.91	25.17	11.81	3.48	7.69	2.46	4.76	1.89
20	z_Q	35.10	23.67	8.37	3.28	5.54	2.13	4.46	2.16

Table 6.4: Errors between the approximations and the simulation after the "formation period" as a percentage of the simulated solution - effect of q, c and λ

"formation period", often giving errors of less than 5% for the averages and variances.

However, the rejoin orbit has the biggest range in error. In some cases, the approximations recover from the initial inaccuracy, quickly becoming accurate, see Figure 6.3 and Table 6.4 ($q \geq 0.3$). However, Table 6.4 and Figure 6.6 show that for an effective traffic intensity close to 1 the inaccuracy persists, even for large systems.

The same is true for the accuracy of the VWT. Until the queue is formed, the VWT is always 0; therefore, larger errors occur during the formation period, which then diminish once the approximation overcomes the initial delay. In addition, once a queue forms, the variance equations for the VWT change, see (5.47), producing

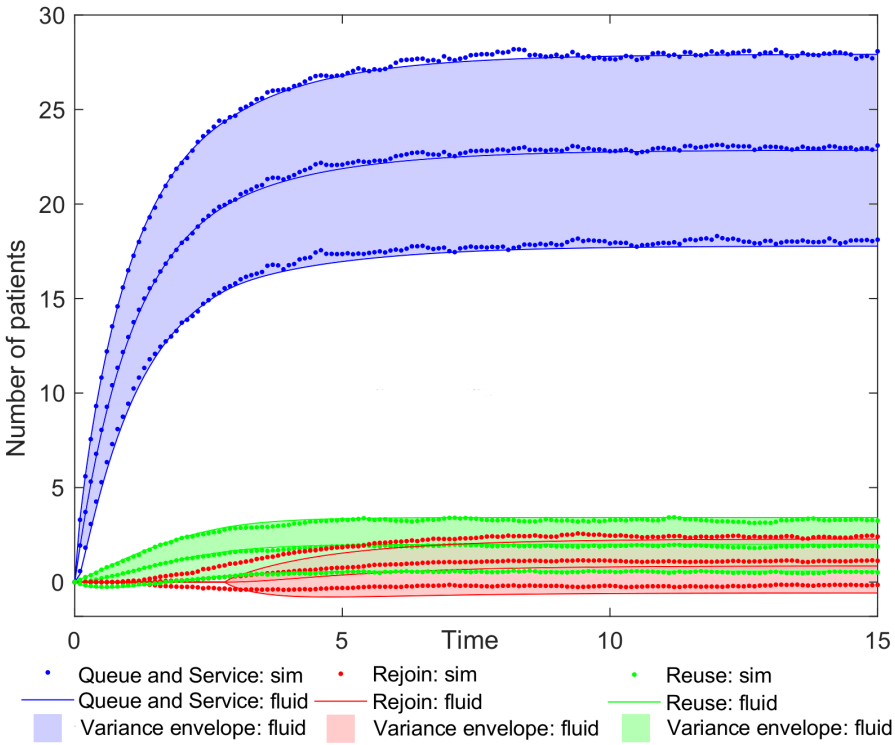

Figure 6.6: Error in the approximation when $\widetilde{\rho} = 1.\dot{1}$

The difference between when loss begins in the simulation compared to the approximation are clearly seen, producing an error that persists throughout the solution.

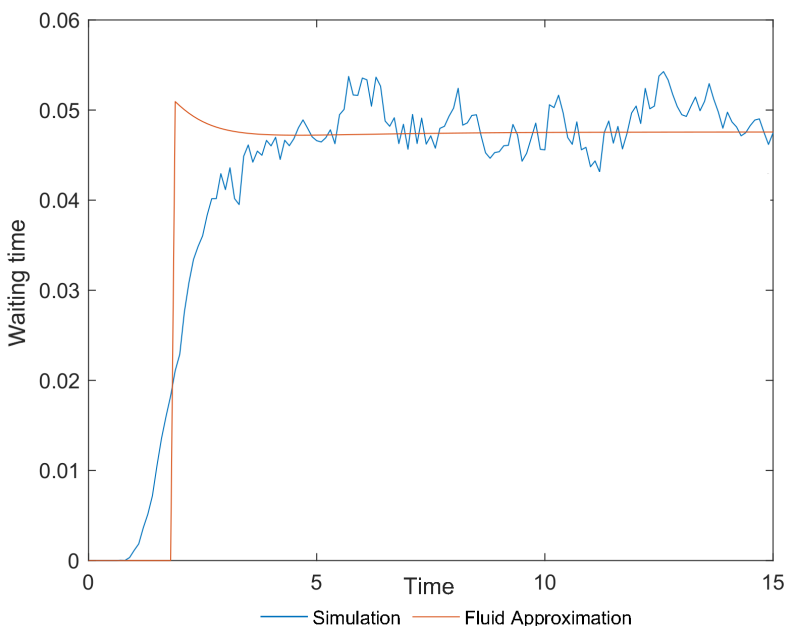

Figure 6.7: Example of an initial spike in the variance of the VWT

spikes in the solution at critical points [107], Figure 6.7. Additionally, for the VWT, the solution is inaccurate when the system is critically loaded for a significant period

of time [107] due to the assumption that $\{z_Q(t) = c(t)|t > 0\}$ is of zero measure.

Importantly, the size of q has a significant effect on the accuracy of the approximation. As q increases, more patients reuse the service, increasing the effective traffic intensity. Thus, in the scenario above, whilst $\rho = 1$, the system operates with a higher effective traffic intensity, increasing the accuracy of the results. Even for the smaller systems, the increased size of q greatly improves the accuracy of results.

From Tables 6.3 and 6.4, the length of the formation period decreases as the effective traffic intensity increases. With a higher effective traffic intensity, queues form quicker, resulting in the approximations becoming accurate at a faster rate. Thus, the error introduced by beginning with an empty system is overcome. Notably, the size of the system had no effect on the length of the formation period.

To further examine the system, I consider the effect of changes in other parameters fixing $c, \lambda = 20$ and $q = 0.3$, and varying θ, δ_R and δ_U as in Table 6.5.

Parameter - Value	**Description**
$dt = 0.1$	Time step for the numerical scheme
$T = 15$	Total time period of solution
$\lambda = 20$	Arrival rates
$c = 20$	Number of servers
$\mu = 1$	Service rate
$p = 0.3$	Proportion of patients rejoining after abandonment
$q = 0.3$	Proportion of patients reusing the service

Table 6.5: Parameters used to assess the accuracy of the approximations - steady state analyses of the effect of θ, δ_R and δ_U

In comparison to the effect of q, μ and the size of the system, these parameters have little effect on the accuracy of the system, Table 6.6. Most effectual is the value of θ, which reduces the error in calculating the VWT after the formation period. Otherwise, each parameter produces a small change in the accuracy of the VWT,

		$t \leq T_I$				$t > T_I$			
Parameters		**err**	**verr**	**werr**	**wverr**	**err**	**verr**	**werr**	**wverr**
$\theta = \frac{1}{2}$	z_Q	1.19	2.80	63.65	52.82	0.74	2.00	5.81	3.54
$\delta_R, \delta_U = 1$	z_R	72.31	48.04	-	-	2.82	2.43	-	-
$T_I = 3.1$	z_U	4.95	6.85	-	-	1.32	2.23	-	-
$\theta = 2$	z_Q	1.86	5.08	59.21	67.26	1.02	3.68	12.42	3.29
$\delta_R, \delta_U = 1$	z_R	68.93	44.11	-	-	4.01	4.67	-	-
$T_I = 3.1$	z_U	2.33	3.90	-	-	2.16	1.82	-	-
$\delta_R = \frac{1}{2}$	z_Q	0.72	2.63	63.55	56.58	0.55	2.09	9.68	3.23
$\theta, \delta_U = 1$	z_R	74.58	42.36	-	-	5.35	2.84	-	-
$T_I = 3.1$	z_U	4.38	4.79	-	-	1.44	2.20	-	-
$\delta_R = 2$	z_Q	0.63	1.54	61.72	60.63	0.40	1.96	8.05	4.02
$\theta, \delta_U = 1$	z_R	68.09	46.66	-	-	3.31	3.06	-	-
$T_I = 3.1$	z_U	3.33	4.12	-	-	1.50	1.73	-	-
$\delta_U = \frac{1}{2}$	z_Q	0.54	2.32	70.43	66.73	0.31	1.54	9.75	3.53
$\theta, \delta_R = 1$	z_R	79.53	53.52	-	-	5.86	2.68	-	-
$T_I = 3.4$	z_U	2.73	5.78	-	-	1.16	2.27	-	-
$\delta_U = 2$	z_Q	1.03	3.91	53.57	56.84	0.39	1.60	9.05	2.43
$\theta, \delta_R = 1$	z_R	65.74	40.11	-	-	2.78	2.62	-	-
$T_I = 2.9$	z_U	4.72	5.92	-	-	1.65	1.72	-	-
$\theta, \delta_R, \delta_U = \frac{1}{2}$	z_Q	1.24	5.97	71.05	64.29	0.97	2.58	7.58	4.72
	z_R	82.47	59.01	-	-	5.84	1.47	-	-
$T_I = 3.4$	z_U	4.36	5.30	-	-	1.29	2.23	-	-
$\theta, \delta_R, \delta_U = 2$	z_Q	1.35	4.86	55.00	60.99	0.55	3.62	13.28	2.75
	z_R	64.60	39.92	-	-	4.44	2.80	-	-
$T_I = 2.9$	z_U	4.39	5.48	-	-	3.09	2.65	-	-

Table 6.6: Error between the approximations and simulation as a percentage of the simulated solution - effect of parameters $\theta, \delta_R, \delta_U$

but overall there is little change for the other outputs. Notably, the length of the formation period reduced as δ_U decreased, highlighting that systems with higher δ_U

become accurate faster. This is expected since patients re-enter the queue sooner, creating a slight "boost" in the total rate arrivals.

By calculating errors during the formation period and after, I conclude that under the right conditions, the error between the two methods is small; thus, the methods are appropriate for steady state analysis. Given the results in this section, for the remainder of the chapter, discussion of the model's accuracy will centre around the effective traffic intensity. This will include analysis of time varying behaviour, when the system is underloaded and overloaded, non-empty initial conditions, and the effect of transitions between multiple health states.

Time varying behaviour

As noted in [106, 107], fluid and diffusion approximations may be used to model time varying systems. Figures 6.8, 6.9 and 6.10 show the results of the approximations as the number of arrivals falls from 20 to 10 at $t = 10$ (this uses the same input parameters as in Table 6.2, except $T = 20$ and $q = 0.7$). Notably, the use of a piecewise continuous arrival rate is allowed for this scenario since the number of servers is constant and continuous. Each output maintains accuracy, closely following the behaviour of the simulation.

However, it should be noted that an inaccuracy may occur when moving from an overloaded to an underloaded phase. To model this behaviour within the approximations, $\widetilde{\rho}$ needs to fall below 1, which may occur due to a fall in arrivals, a fall in reuse or an increase in the service's capability to meet demand (increased c or μ). In Figures 6.11 and 6.12, a similar scenario to Figure 6.8 is modelled but with $q = 0.4$, such that $\widetilde{\rho}(t) = 1.\dot{6}, t \in [0, 10)$ and $\widetilde{\rho}(t) = 0.8\dot{3}, t \in [10, 20]$. This highlights how accuracy may be maintained when passing between overloaded and underloaded phases, as long as the system does not remain critically loaded for too long. Furthermore, in Figure 6.11 a behaviour similar to that which occurs in the

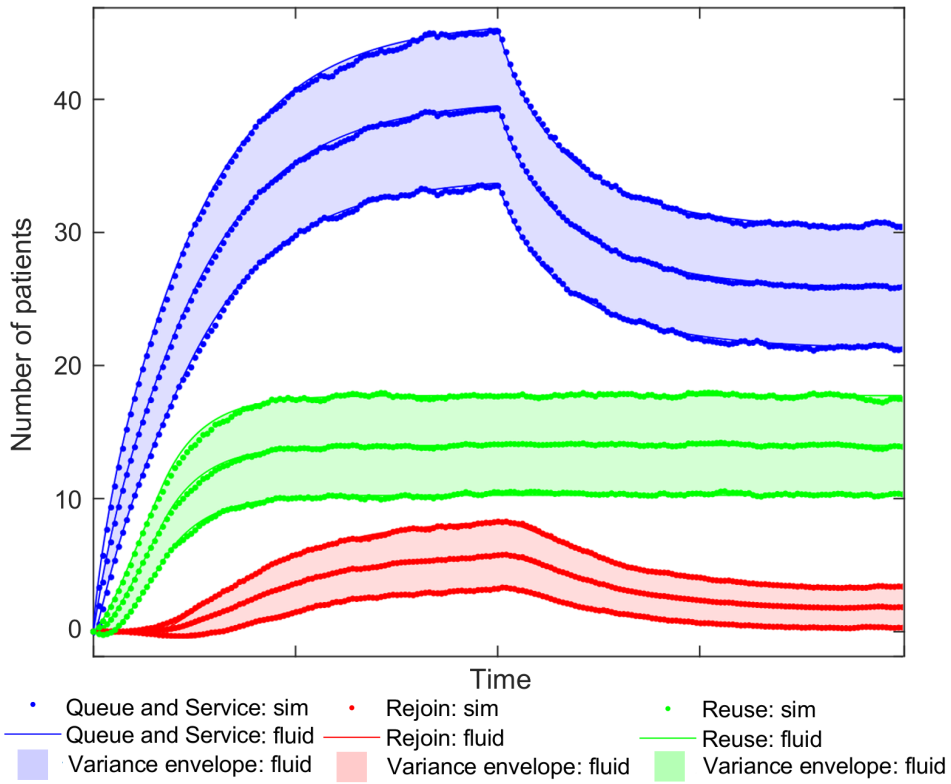

Figure 6.8: Example of the accuracy of the approximations in modelling a fall in the rate of arrival - process states

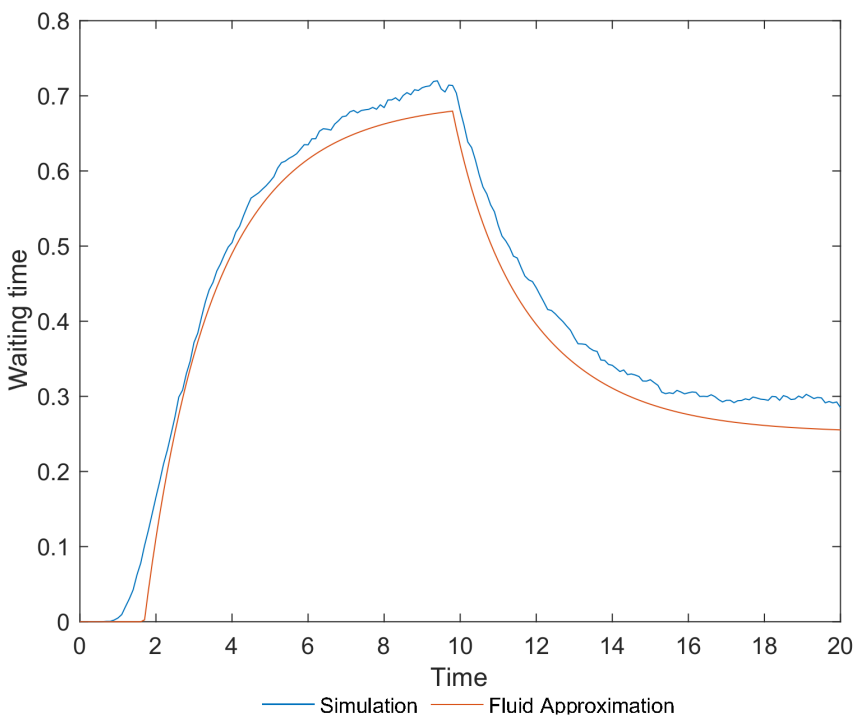

Figure 6.9: Example of the accuracy of the approximations in modelling a fall in the rate of arrival - the VWT

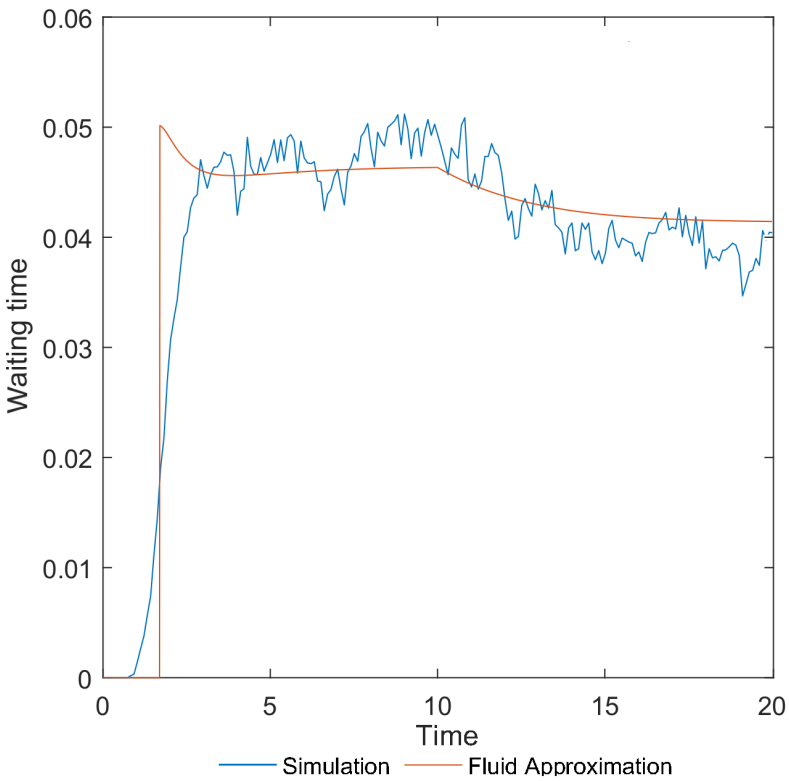

Figure 6.10: Example of the accuracy of the approximations in modelling a fall in the rate of arrival - the variance of the VWT

formation period is seen in the solution to $z_R(t)$ for $12 < t < 16$. I refer to this as a "dispersion" behaviour - where the queue (and thus the number of abandonments and rejoins) diminishes quicker in the fluid approximation, again due to the lack of stochastic variation. In addition, as noted by [107], this error may increase when the difference between these phases are less distinct i.e. the difference between $\widetilde{\rho}$ during the two phases is small.

The above examples highlight that time varying behaviour between overloaded and underloaded phases can be accurately modelled by fluid and diffusion approximations. Importantly, the effective traffic intensity must be significantly greater than 1 in overloaded phases, and significantly lower than 1 in underloaded phases. When this does not occur, the largest errors are seen in the rejoin orbit and the VWT during underloaded phases, and during the formation/dispersion periods.

I now explore the use of the approximations for modelling a spike in demand,

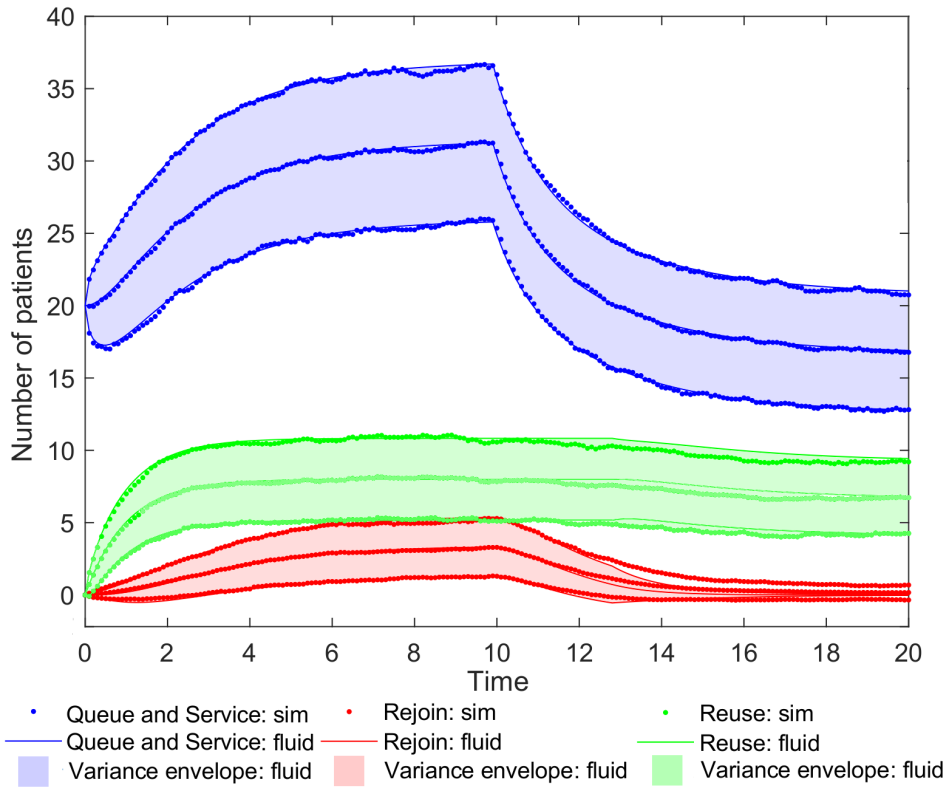

Figure 6.11: Example of the accuracy of the approximations in modelling a system that moves from an overloaded phase to an underloaded phase - process states

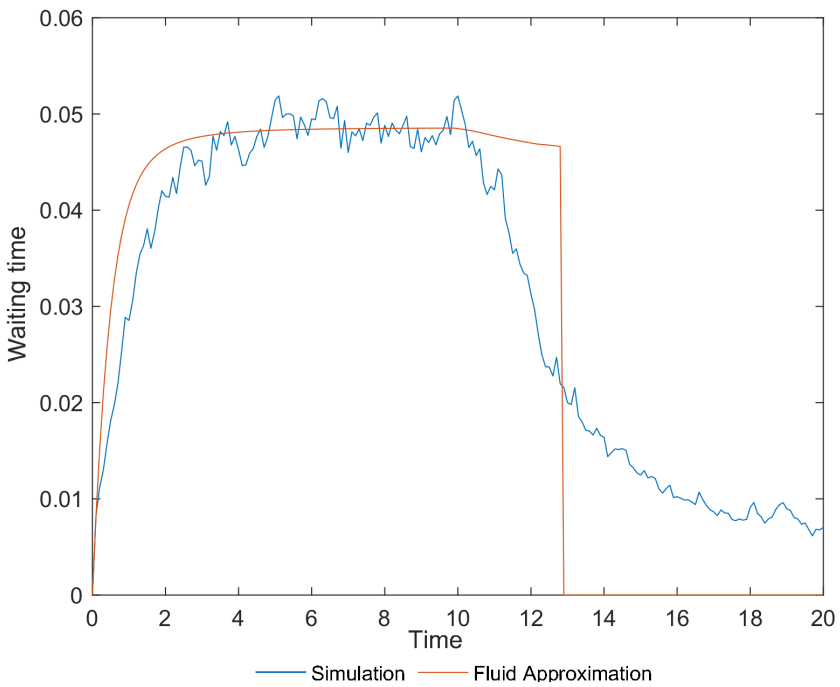

Figure 6.12: Example of the accuracy of the approximations in modelling a system that moves from an overloaded phase to an underloaded phase - VWT

beginning with a non-zero initial conditions $z_Q(0) = 20$ and using the parameters in Table 6.2, setting $q = 0.3, c = 20$ and $T = 15$. The arrival process is:

$$\lambda = \begin{cases} 20, & t \in [0, 6) \cap [8, 15] \\ 40, & t \in [6, 8) \end{cases}$$

Figures 6.13 to 6.17 show that the approximations are accurate throughout the time interval. An error still occurs during the formation period; however, the error is smaller and quickly diminishes. If $z_R(0) > 0$ this initial error greatly diminishes. Furthermore, the solution to the VWT and its variance closely follow the simulation.

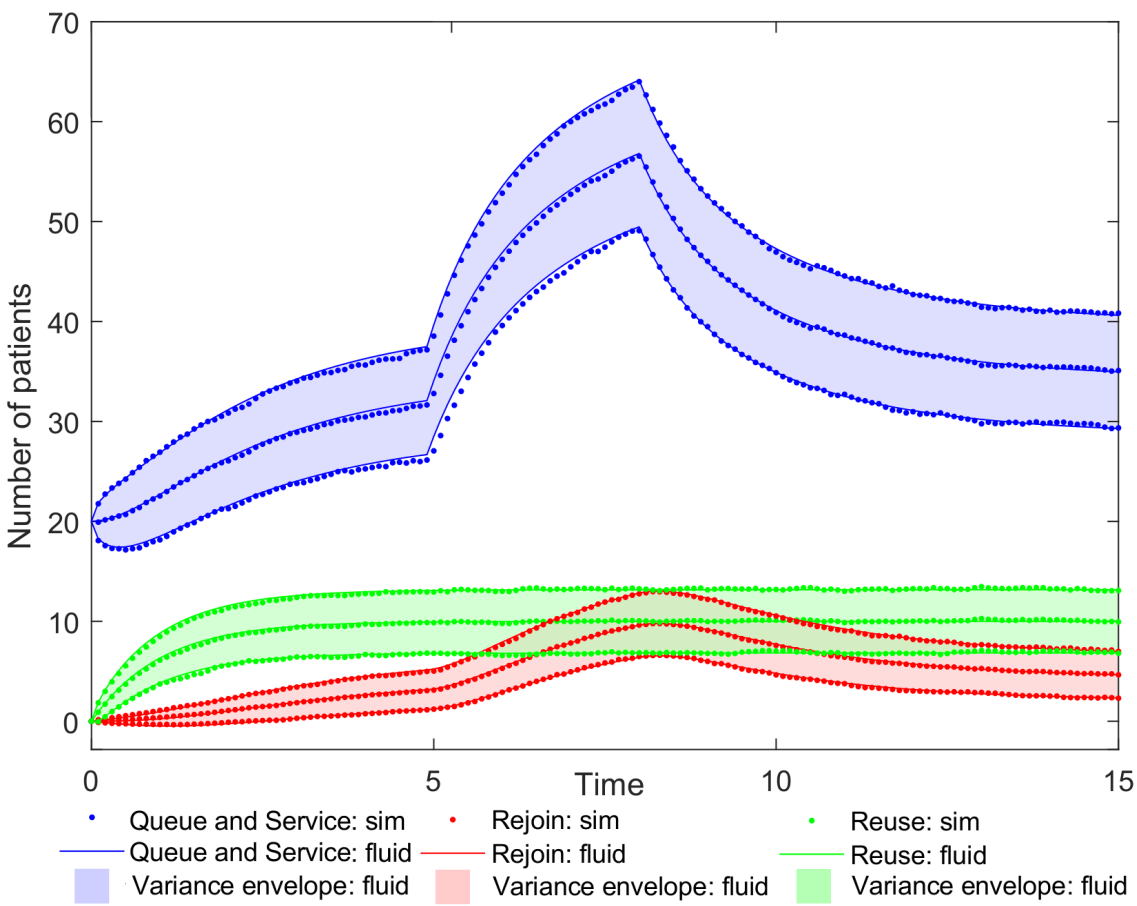

Figure 6.13: Example of a seasonal spike in arrivals - the number of patients in each process state and the variance of each

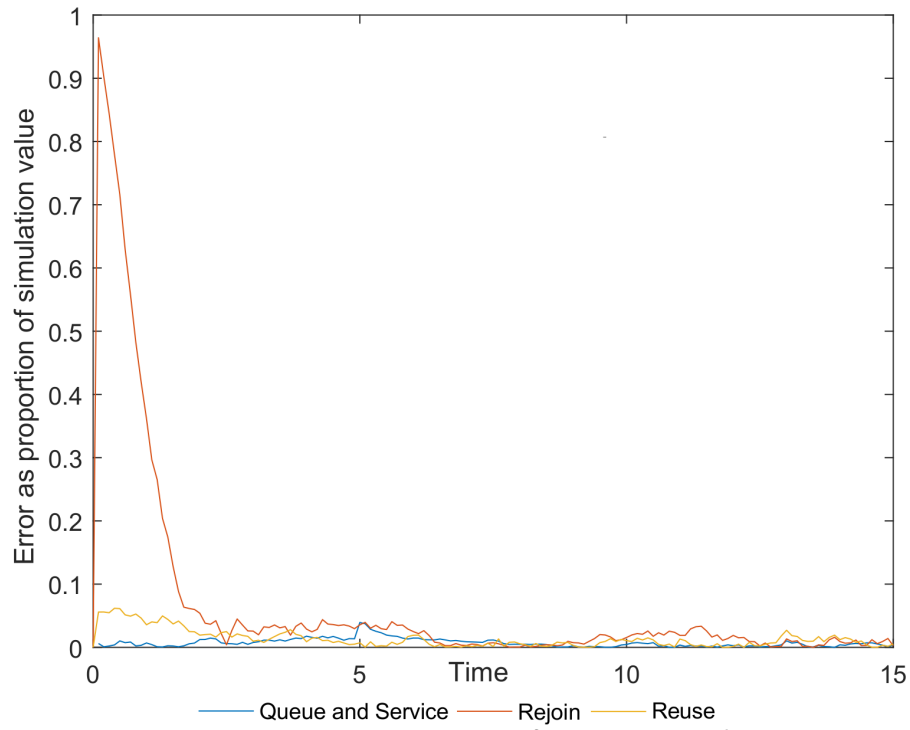

Figure 6.14: Example of a seasonal spike in arrivals - the error in the number of patients in each orbit

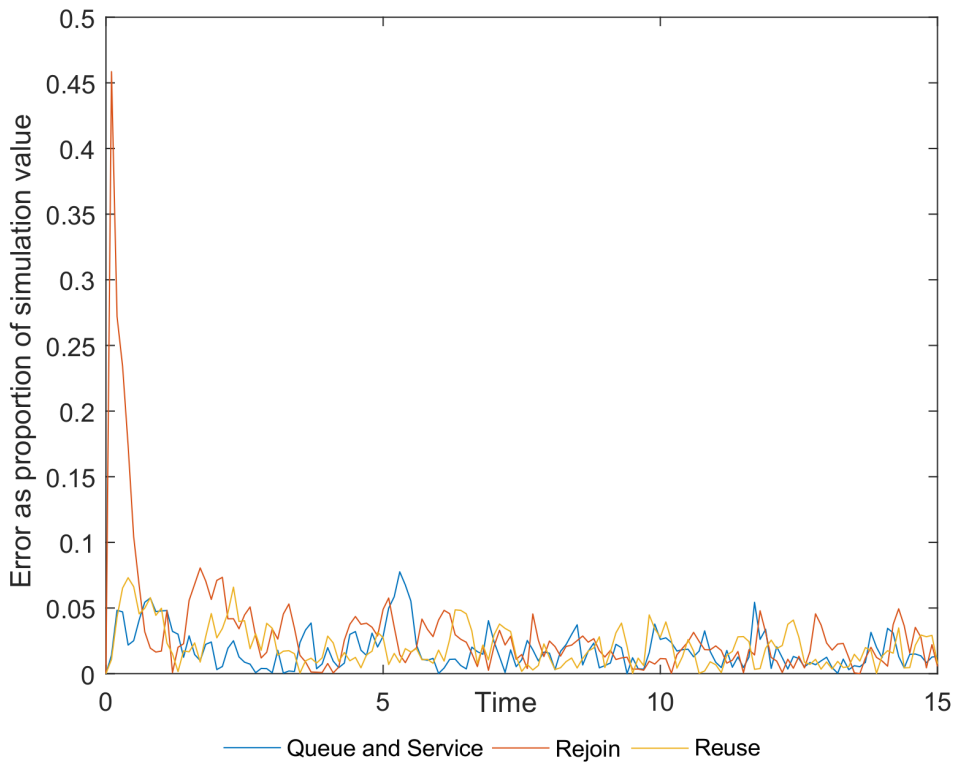

Figure 6.15: Example of a seasonal spike in arrivals - the error in the variance

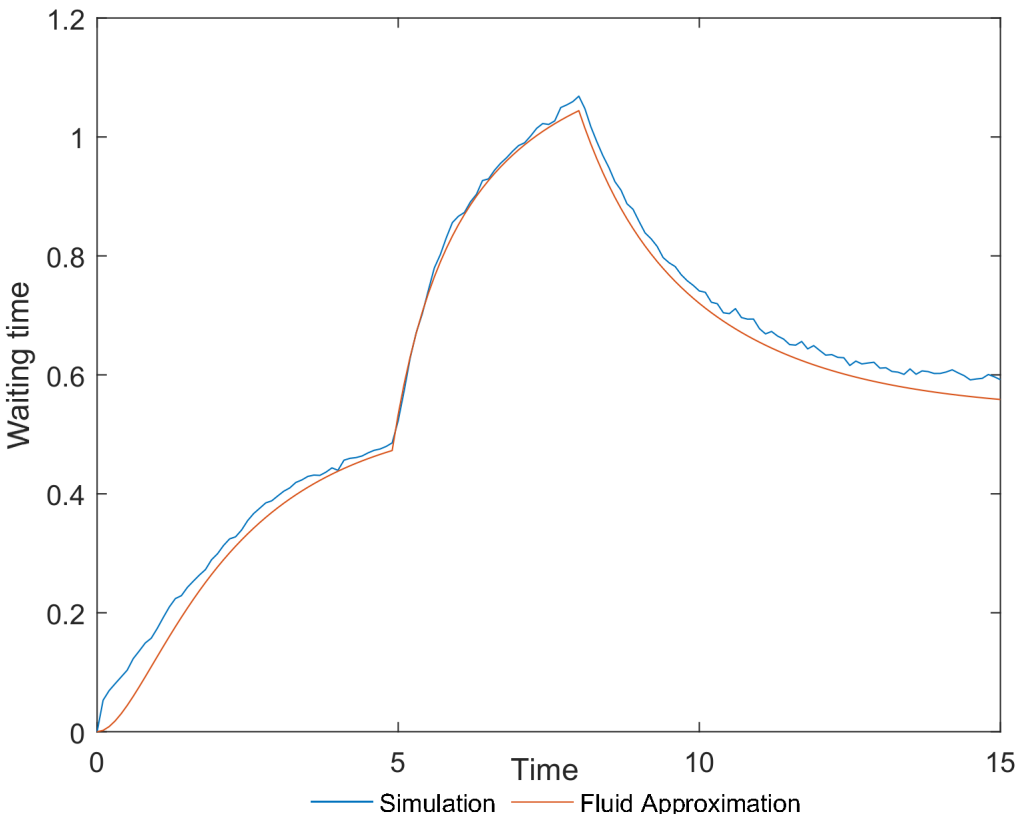

Figure 6.16: Example of a seasonal spike in arrivals - the VWT

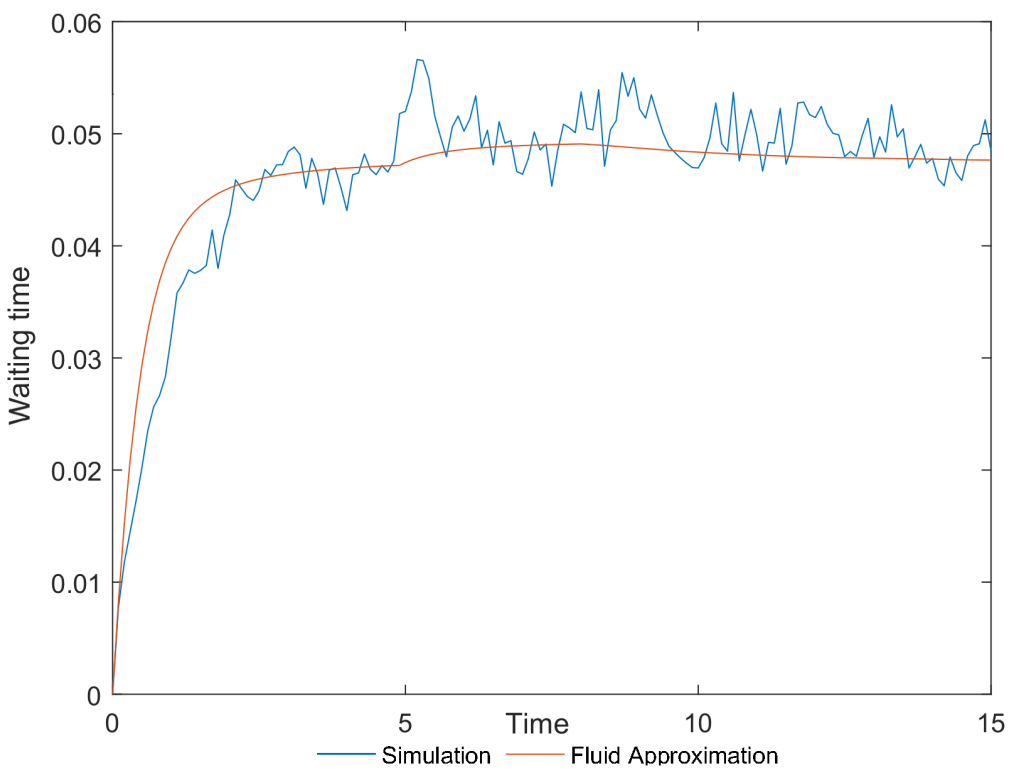

Figure 6.17: Example of a seasonal spike in arrivals - the variance of the VWT

Summary of single service and single health state models

From the above analyses, I note the following:

- Large percentage errors may occur when moving from an underloaded phase to an overloaded phase, caused by a delay between the queue forming and when patients can begin to abandon the queue in the two models - denoted the formation period;
- A similar error is seen when moving from an overloaded phase to an underloaded phase - denoted the dispersion period;
- Accuracy is maintained when moving between phases of underloading and overloading as long as the system does not "linger" in a critically loaded state and $\widetilde{\rho}$ is far from 1;
- Once these delays are overcome, the approximations can be used to accurately model steady state scenarios and time dependent solutions;
- The size of the system (number of servers and arrivals) and the size of the effective traffic intensity $\widetilde{\rho} = \lambda/c\mu(1-q)$ are the most significant parameters in determining the accuracy of the system. Either decreasing θ - the rate at which patients leave the queue - or increasing δ_U - the rate at which patients reusing the service enter the queue - also improves the accuracy of the model;
- The approximation of the number of patients in the reuse orbit, and queue and service orbit, and the variance of these outputs is accurate irrespective of the size of the system, effective traffic intensity and time;
- The VWT approximation and number of rejoin patients are more accurate for higher effective traffic intensities, but may still exhibit large errors within formation and dispersion periods.

I have shown that the approximation methods can be highly accurate for the service and queue process state, and the reuse process state. The majority of the error comes through the rejoin orbit; hence, if this dynamic is negligible, this process state may be dropped to increase accuracy. Nevertheless, when modelling a system with an effective traffic intensity greater than 1, abandonment from the queue is required to maintain a finite average queue length. Alternatively, if abandonment is not considered, there must be periods within the modelled time period when the system is not effectively heavily loaded to inhibit the queue growth [106].

I suggest that these methods are appropriate for modelling community services when significant reuse is anticipated, and demand is regularly high such that the queue persists or seasonal spikes cause excessive demand. I also suggest, in line with [107], that care is taken when evaluating the fluid and diffusion approximations of $z_R(t)$, VWT and their variances when $z_Q(t)$ is close to $c(t)$.

6.3.3 Single service and multiple health states

Building on the previous section, I now explore a simple extension; a system with two health states. I will begin by using a constant and equal allocation of servers across queues. From this I will gain an understanding of how the effective traffic intensity may differ when multiple health states are considered, for both steady state and time varying scenarios.

Steady state analysis

Beginning with a steady state analysis of a system that begins empty, I examine the accuracy of approximations when the system moves from being underloaded to overloaded. Patients may arrive in either a health state $k = 1$ or $k = 2$, and progress through the system according to health state dependent parameters, set out in Table 6.7. In this example, patients from either health state have the same

input parameters, such that $\rho = \lambda/c\mu = 1$ for both queues. By modelling two groups with the same input parameter, this analysis helps to understand the effect that health state transitions have on the accuracy of the approximations.

	Health State	
Parameters	$k = 1$	$k = 2$
μ_k	1	1
θ_k	1	1
λ_k	20	20
c_k	20	20
$p_k = r_{k,L,1,1}$	0.3	0.3
$q_k = r_{k,S,1,1}$	0.3	0.3
$\delta_{k,R}$	1	1
$\delta_{k,F}$	1	1

Table 6.7: Parameters used to assess the accuracy of the approximations - steady state analysis of a single service and two health states

Now to define the health transition matrices with $k = 2$ indicating better health (despite the lack of difference between their flow parameters). Firstly, I assume that receiving service has a potentially beneficial, but not perfect, effect. Thus, patients in $k = 2$ remain in this state after service, whilst those in $k = 1$ are more likely to improve and move to $k = 2$ than stay the same. Secondly, I assume a similar, reverse effect for abandonment - that patients in $k = 1$ remain in this state, yet those in $k = 2$ are more likely to decline in health than stay the same. Thirdly, for those who seek to rejoin, I assume that such patients may use a service outside of the system; hence, there may be an improvement in their health after rejoin. Finally, for those who reuse the service, I assume that their health may change or stay the same after their time in the orbit; however, they are more likely to remain in the state in which they enter. This may represent a delayed benefit of service, or a decline in health post service.

Notably, these matrices highlight the differences in patients' capacities to benefit as the receipt of care, or the absence of it, has a different effect on patients in each health state. The transition matrices for this system are as follows:

$$S_S = \begin{bmatrix} 0.3, 0.7 \\ \\ 0, 1 \end{bmatrix} \; S_L = \begin{bmatrix} 1, 0 \\ \\ 0.6, 0.4 \end{bmatrix} \; S_R = \begin{bmatrix} 0.8, 0.2 \\ \\ 0.5, 0.5 \end{bmatrix} \; S_F = \begin{bmatrix} 0.8, 0.2 \\ \\ 0.2, 0.8 \end{bmatrix}$$

Table 6.8 presents the errors for this system, containing errors for the formation period - $[0, T_I)$, and the error after - $[T_I, 15]$.

Parameters		**err**	**verr**	**werr**	**wverr**
$k = 1$	Z_Q	0.34	2.12	75.32	64.77
	Z_R	63.00	37.00	-	-
$t \leq T_I = 3.4$	Z_U	2.60	3.97	-	-
$k = 1$	Z_Q	0.46	1.45	12.03	6.30
	Z_R	4.24	1.95	-	-
$t > T_I = 3.4$	Z_U	2.67	2.33	-	-
$k = 2$	Z_Q	0.24	1.79	52.56	54.32
	Z_R	65.50	47.85	-	-
$t \leq T_I = 2.9$	Z_U	3.48	4.22	-	-
$k = 2$	Z_Q	0.42	1.30	6.71	1.55
	Z_R	2.05	2.06	-	-
$t > T_I = 2.9$	Z_U	1.41	1.59	-	-

Table 6.8: Error between the approximations and simulation as a percentage of the simulated solution

A clear difference is seen in the accuracy of the fluid and diffusion limits for the two health states. For $k = 2$, after the formation period, the solution is more accurate for both $z_R(t)$ and the VWT, indicating that the queue is more heavily

loaded and that the effective traffic intensity is no longer only dependent on reuse patients and ρ. Moreover, there is a difference between the length of the formation periods for the two health state queues, further highlighting the difference in effective traffic intensity caused by the health state transitions. In particular, a smaller T_I indicates that the system has a higher effective traffic intensity, as in section 6.3.2.

A reason for this is that patients may join the queue for the health state they did not arrive in through either the reuse or rejoin orbits. Previously, rejoining patients were captured by λ in the steady state system; now however, since these patients may join the other queue, this is no longer the case. Hence, these arrivals and the effect of health state transitions are now influential when formulating the effective traffic intensity with multiple health states.

As a final observation, the calculation of the approximations was over 150 times faster than the simulated solution.

Time varying behaviour - dynamic server allocation

Extending this scenario further, I now model a time-varying system that begins non-empty, with all servers busy at $t = 0$ with $z_{k,Q}(0) = 15$ for $k = 1, 2$. Furthermore, the two patient groups now have different flow parameters, see Table 6.9.

Considering a small spike in the arrivals of patients in health state $k = 1$, I analyse the dynamics of queues, using a proportional allocation of servers (equation 5.43). Thus, $c(t) = 30$ for all $t \in [0, 15]$ and $C_{k,i}(\mathbf{Z}(t)) = \left\lfloor \frac{c_i(t) Z_{k,Q,i}(t)}{\sum_{l=1}^{K} Z_{l,Q,i}(t)} \right\rfloor$ for the simulation and $c_{k,i}(\mathbf{z}(t)) = \frac{c_i(t) z_{k,Q,i}(t)}{\sum_{l=1}^{K} z_{l,Q,i}(t)}$ for the approximations. Within the simulation, as in the stochastic process, the number of servers allocated to a queue is updated each time an event occurs that changes the size of $Z_{k,Q}(t)$ i.e. a new arrival, patient completing service, an abandonment from the queue.

Using the same health state transition matrices as before, patients in health state $k = 1$ are now have longer service times, a higher propensity to abandon, a

	Health State	
Parameters	$k=1$	$k=2$
μ_k	1/2	1
θ_k	1	1/2
λ_k	15	15
p_k	0.5	0.3
q_k	0.5	0.3
$\delta_{k,R}$	1	1/2
$\delta_{k,F}$	1	1/2

Table 6.9: Parameters used to assess the accuracy of the approximations - time-varying analysis of a single service and two health states

higher likelihood of rejoin or reuse, and require sequential service sooner. Thus, these patients are more resource intensive since they have longer stays within service and less time between uses.

Since the input parameters are required to be continuous to ensure that $z_Q(t)$ is continuously differentiable, I use a continuous jump in arrivals:

$$\lambda_1(t) = \begin{cases} 15, & t \in [0,4) \cap [7,15] \\ 15 + 15 \times (\sin(\pi(t-4) - \frac{\pi}{2}) + 1), & t \in [4,5) \\ 45, & t \in [5,6) \\ 15 + 15 \times (\sin(\pi(t-6) + \frac{\pi}{2}) + 1), & t \in [6,7) \end{cases}$$

Using this arrival function, I model how parallel queues share servers from a common pool, gaining and losing servers according to the proportion of total demand each queue represents. By one queue "attracting" more servers, they deny the opportunity for the other queue to use those servers.

Figures 6.18 to 6.22 show a comparison of the outputs of the two models, including two new outputs for the system. First, Figure 6.19 shows the dynamic allocation

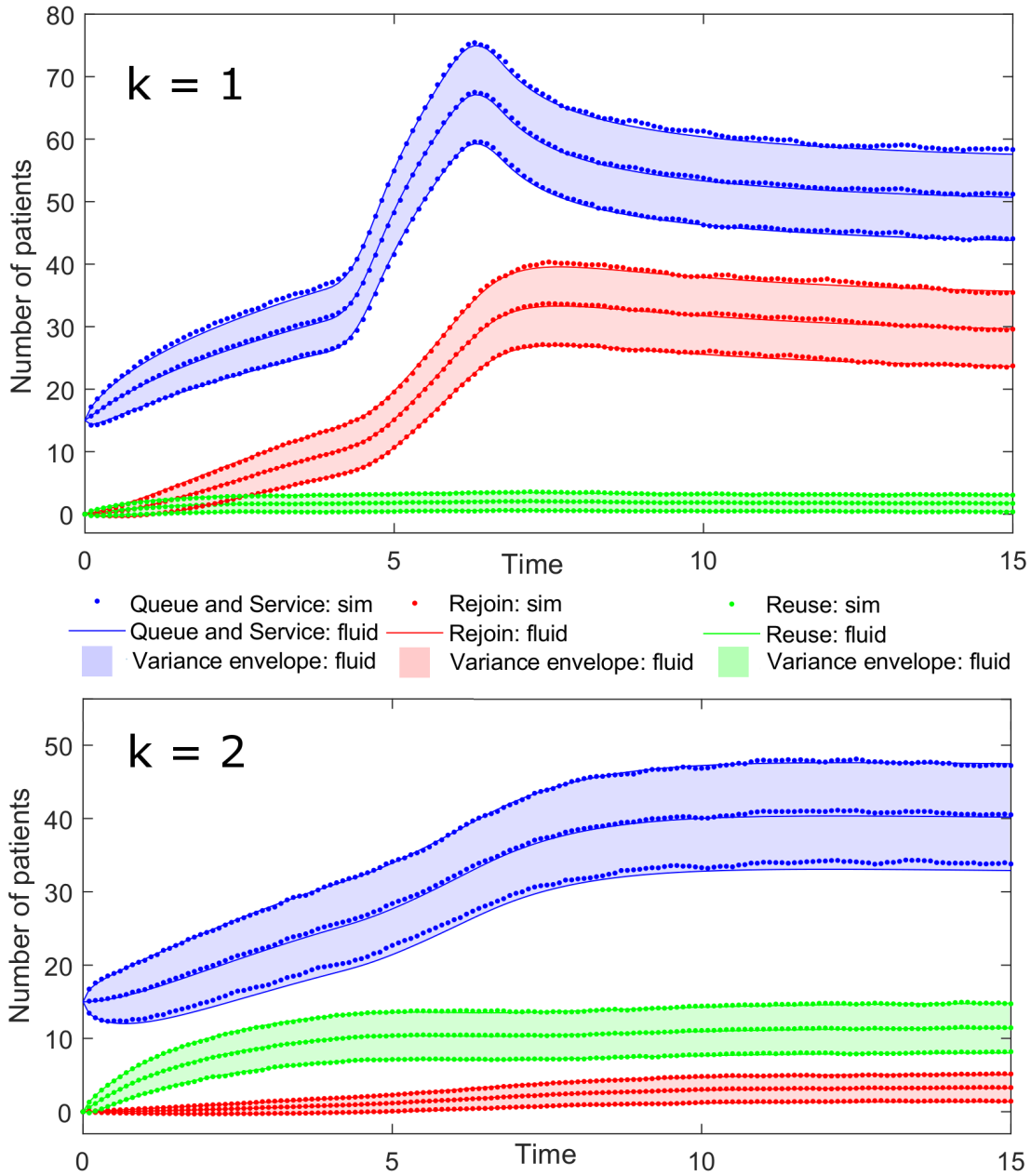

Figure 6.18: Two health state system with dynamic server allocation - number of patients in each process state and their variance

of servers over time as each queue gains servers from/loses servers to the other. Secondly, Figure 6.22 shows the production of outcomes as measured by the rate at which patients in each health state leave the system over time. Overall, the approximations accurately model the system, indicating that it is suitable for this analysis and that the continuous approximation of the capacity allocation holds.

Figure 6.18 shows that the approximations accurately model the number of pa-

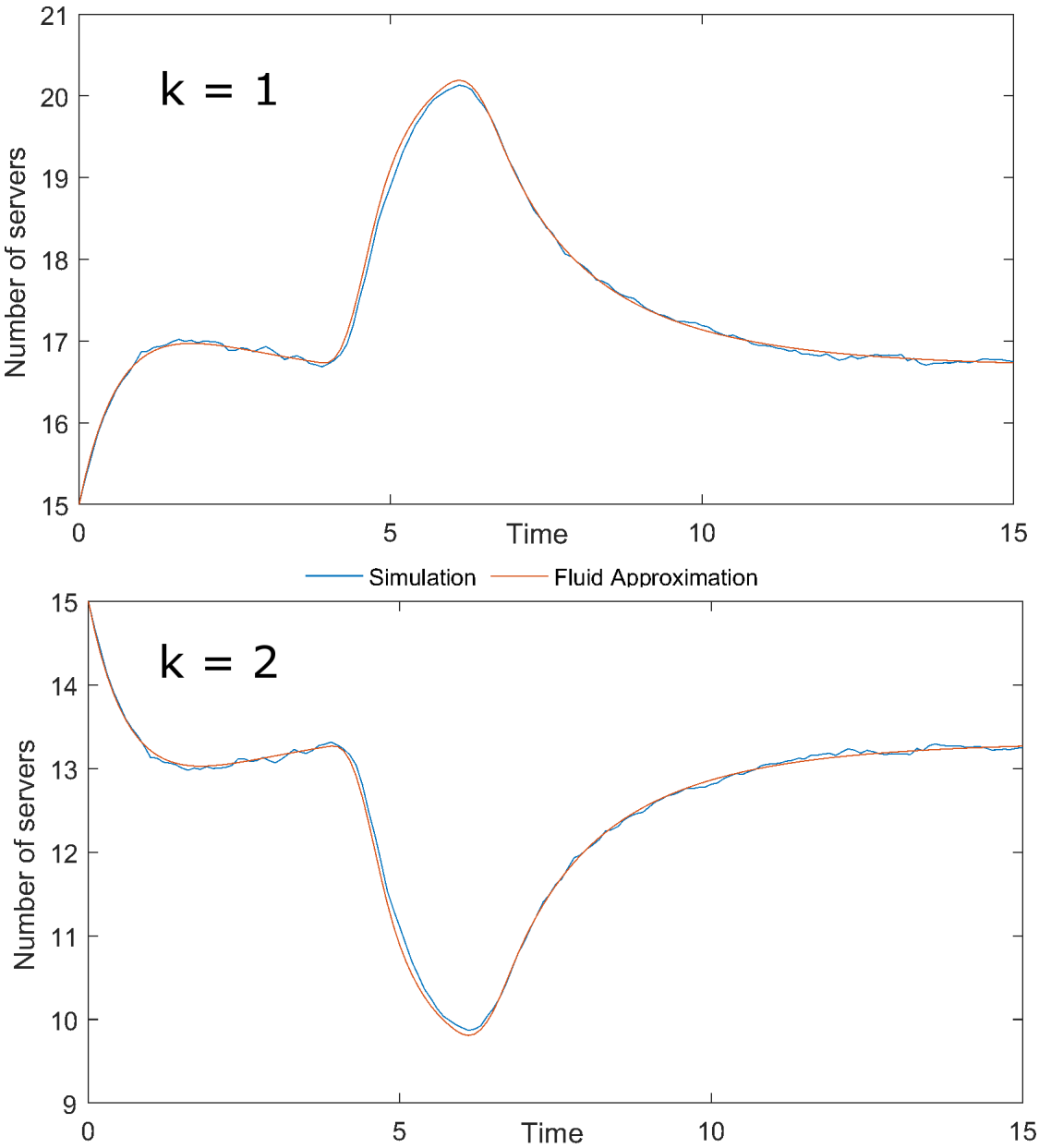

Figure 6.19: Two health state system with dynamic server allocation - number of servers available to each queue over time as they gain/lose servers

tients in each process orbit and the variance of each throughout the modelled time period. Notably, the increased arrivals for the $k = 1$ queue has seemingly little impact on the queue for $k = 2$. There are two reasons for this. Firstly, patients rejoining after abandonment are more likely to be in health state $k = 1$; thus, any limitation on access of $k = 2$ patients results in more rejoins of patients in $k = 1$, rather than a longer $k = 2$ queue. Secondly, the change in servers allocated for $k = 2$

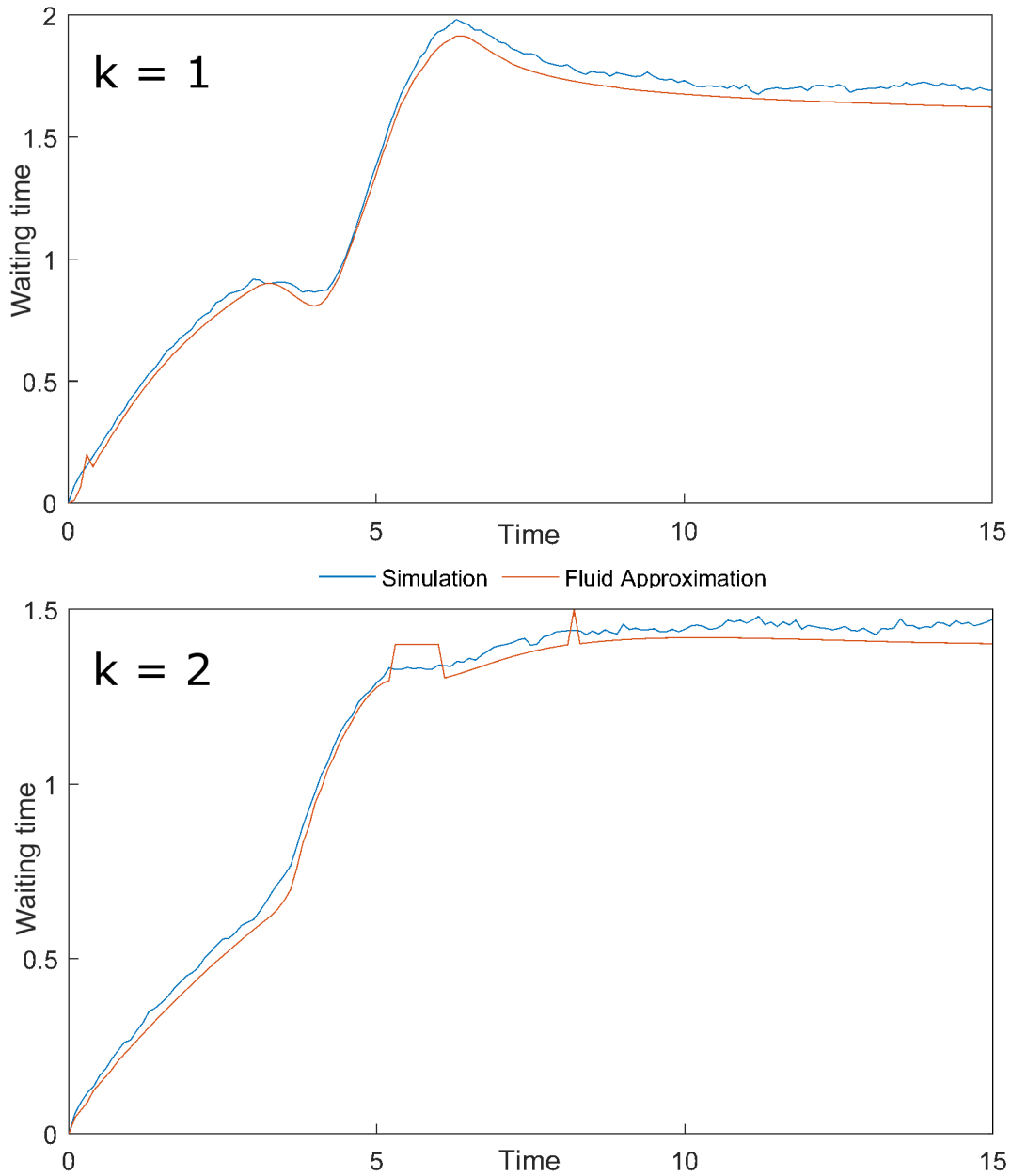

Figure 6.20: Two health state system with dynamic server allocation - the VWT

counteracts any increase in the queue size. If the queue and service were considered separately, a clear difference would be seen i.e. looking at $(z_{k,Q}(t) - c_k(\mathbf{z}(t))^+$ and $\min(z_{k,Q}(t) - c_k(\mathbf{z}(t))$. This is where Figures 6.19 to 6.22 provide more information.

In Figure 6.20, for $k = 2$ the gradient of the VWT increases at $t = 4$, reflecting the loss in available servers to $k = 1$, increased queue lengths and longer waits. Additionally, there is a large increase in the VWT for $k = 1$. Whilst this queue has gained more servers, causing an initial dip, the increase in new arrivals and $k = 1$

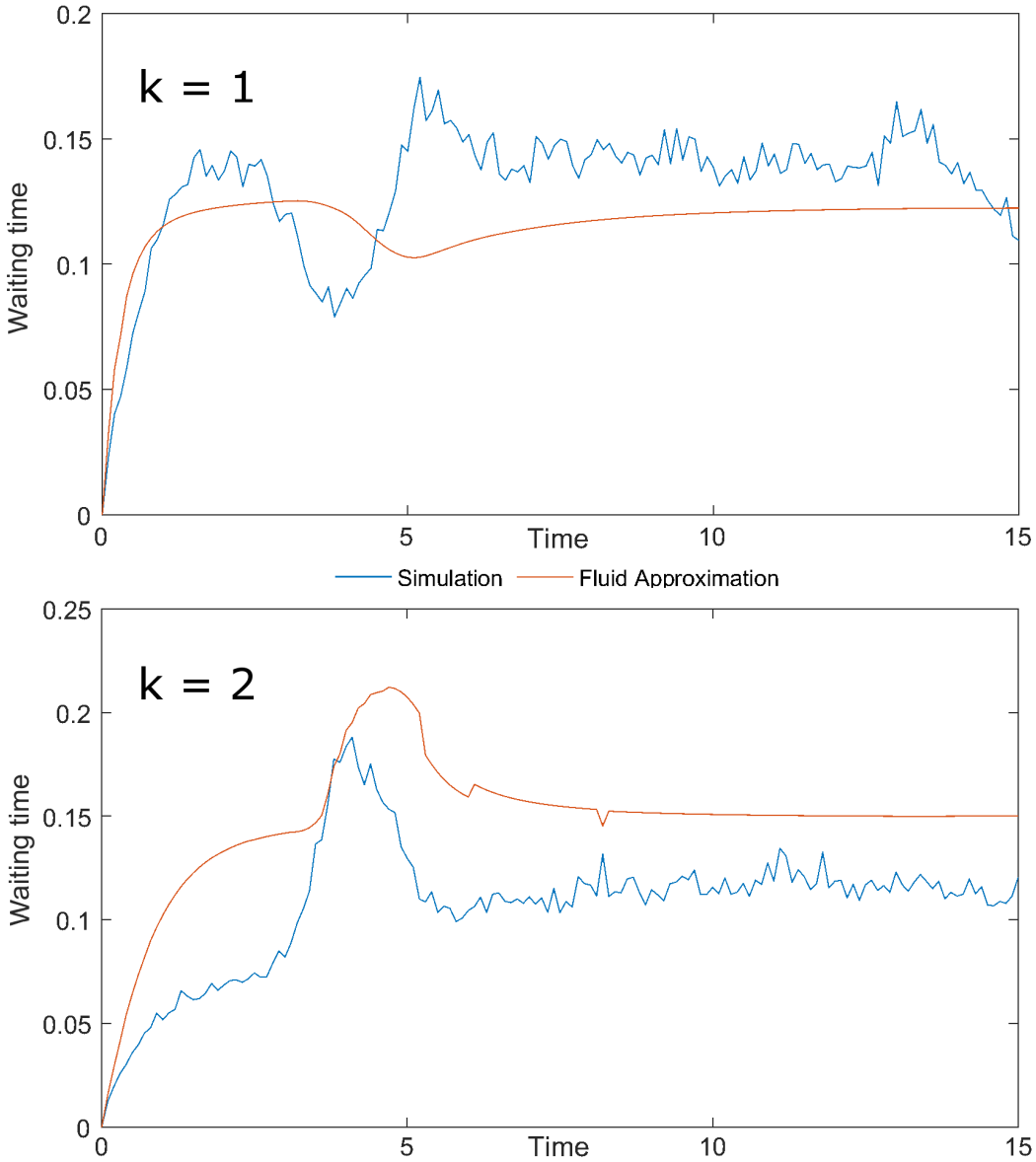

Figure 6.21: Two health state system with dynamic server allocation - variance of the VWT

rejoin patients raises the waiting time. Considering the variance of the VWT, Figure 6.21, the fluid and diffusion approximations match the behaviour but fail to capture the magnitude of the simulated solution. For increased effective traffic intensity and size of system, the results improve; however, it is often inaccurate. One reason is that the variance of the simulated waiting time has the most variability of the system outputs. Combined with the added variability introduced by the dynamic server allocation, this likely produces this inaccuracy.

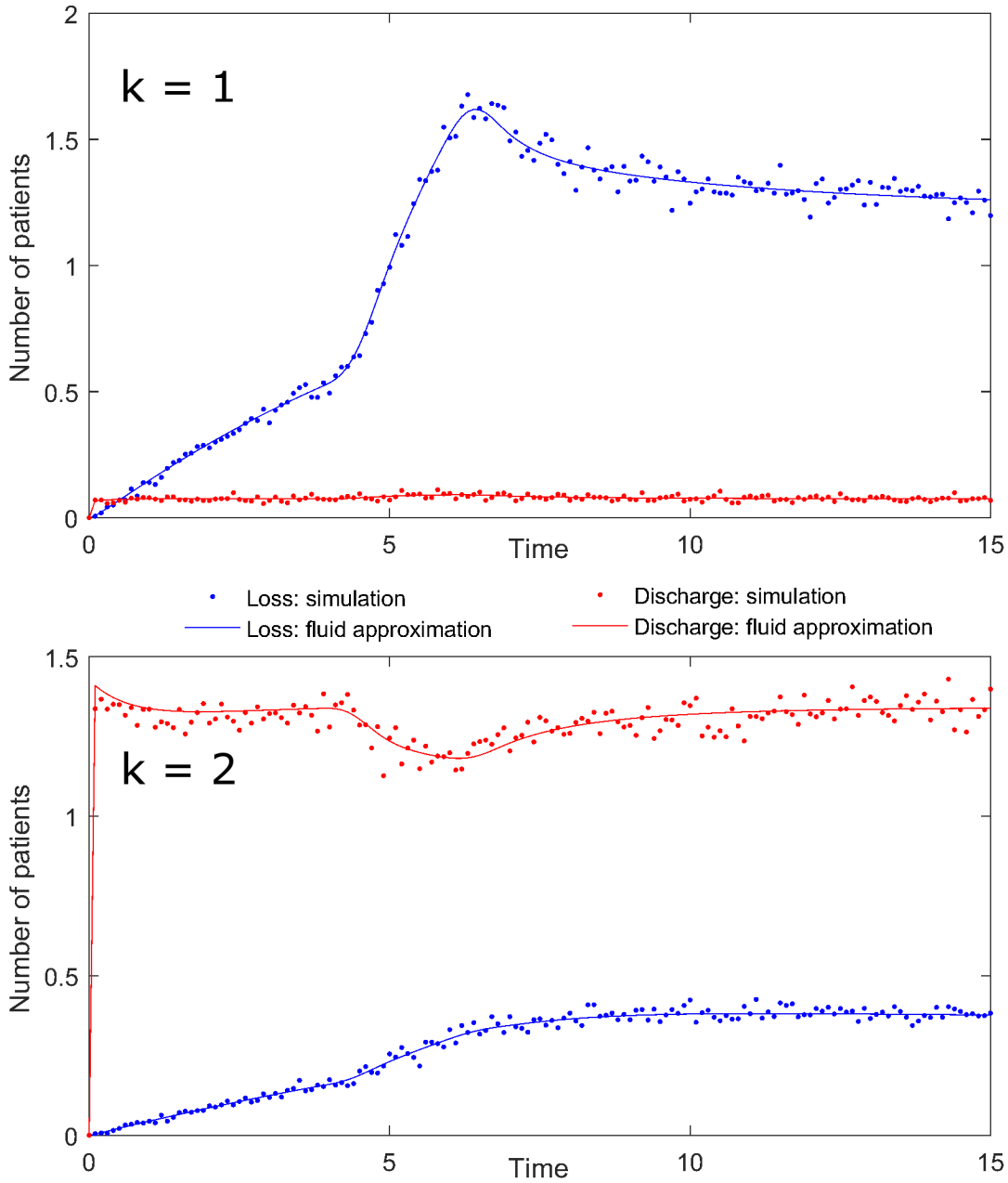

Figure 6.22: Two health state system with dynamic server allocation - the production of outcomes

Alongside the effect on the VWT, the production of outcomes from this system is affected. In Figure 6.22, the number of patients who are lost due to abandonment, and are in the worst health state $k = 1$, greatly increases. Therefore, the impact of the increased arrivals, even with more servers available for treating $k = 1$ patients, results in an increased loss of patients in worse health states - some of who may have transitioned from $k = 2$ at some point. Furthermore, the number of patients

discharged in health state $k = 2$ decreases and the number lost in $k = 2$ increases. This is understandable due to the reduced service of $k = 2$ patients because they are more likely to be in $k = 2$ post service.

This interaction between the two queues and these additional outputs are helpful for modelling and understanding the "flow of outcomes". In particular, how a service produces good and bad outcomes over time in light of patient mix, demand, available/allocated capacity and flow dynamics. This provides a perspective on the quality of service and the performance of the system in relation to process outcomes (such as patient throughput and number of abandonments) and how the differing needs of patients impact the operation of the system.

Finally, to note, in Figures 6.20 and 6.21 there are small jumps in both solutions for $k = 1$ and $k = 2$. Notably, this error is of an order similar to $dt = 0.1$ and lasts for a small amount of time. To investigate further, I compared the results for a scenario with $dt = 0.1, 0.05$ and 0.01 to see how these errors changed, Figure 6.23. As dt decreases, the size of these errors decrease; however, their frequency increases.

There are several possible explanations for this. Firstly, to numerically solve the approximated VWT in MATLAB I used the ODE solver *ode45*, which requires continuous input parameters. However, since the calculation of the VWT relies on the solution to the fluid approximation, i.e. $(c(\mathbf{z}(t)))$, the input parameters were entered over discrete time steps. Thus, at certain points in time, changes in the solution may be small compared to the size of the time step, creating an error in the order of the time step, as seen in the zoomed in panel of Figure 6.23. Secondly, these errors become more frequent as the size of dt reduces because the equations are solved over more time steps, providing more opportunities for the jumps to occur. Thirdly, by using a dynamic server allocation, the fluid approximation becomes non-linear (the combination of equations (5.29)-(5.35) with (5.43) or (5.44)). Hence, solving these systems numerically over discrete time steps may cause these types of error.

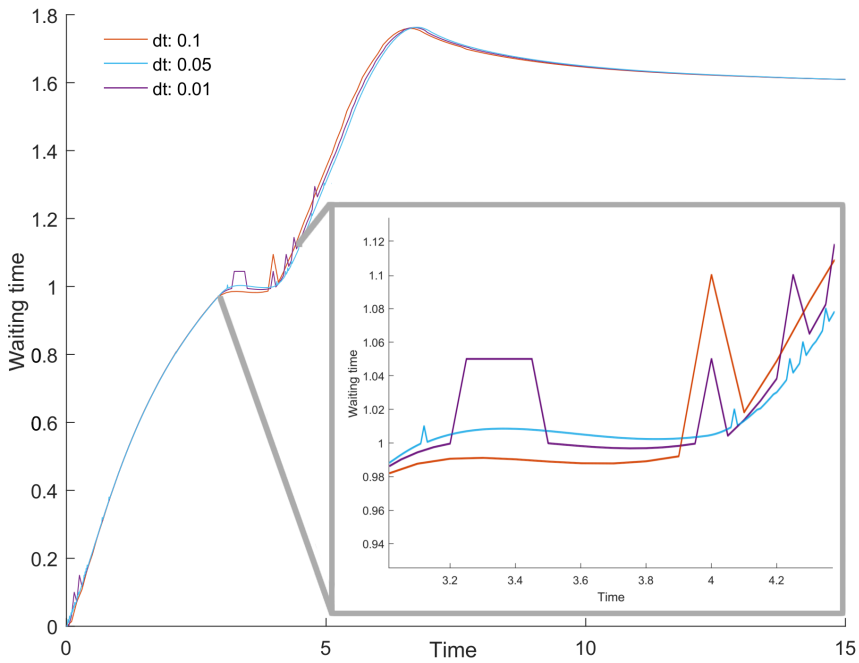

Figure 6.23: Example of small errors in the solution of the VWT when modelling competing queues. These errors are of order dt.

Ultimately, these errors do not have a significant impact on the solution of the VWT or its variance since they last for short time intervals with potentially small magnitude, Figure 6.23. However, in noting these errors, three considerations must be made when choosing the size of dt. Firstly, the size and length of error in comparison to the expected solution e.g. if the VWT equals 1, a time step of $dt = 0.1$ may produce a 10% error that lasts an amount of time in the order of dt. Secondly, the frequency of these errors. Is it better to have fewer larger errors or many smaller errors? Thirdly, the running time, Table 6.10. As the size of dt falls, the time required to solve the numerical scheme increases, creating a trade off between usability and accuracy. Notably, if errors are small, infrequent and short lived, given that this is an approximation of a stochastic system, they may be inconsequential.

dt	0.1	0.05	0.01
Time (s)	6.72	25.50	605.06

Table 6.10: Time taken to solve the fluid approximation for different sizes of dt

Summary of single service and multiple health state models

From this brief exploration, I have shown that by extending to multiple health states, the appropriateness and applicability of the approximations is maintained. Importantly, the transition matrices influence the effective traffic intensity in this system because a patient's health may change throughout their interaction with the system. Thus, each service has arrivals of new patients, reuse patients and rejoins, including those who previously queued within another health state.

This is important for systems where reuse is low for a particular patient group, since these methods may be accurately applied if there is a significant flow of patients arriving from other health states. Therefore, the effective traffic intensity for each $k \in H$ queue, when considering multiple health states, is dependent on the combination of: $\lambda_k(t), c_k(\mathbf{z}(t)), \mu_k(t), \mathbf{S}_{k,m}(t), q_k(t)$ and $p_k(t)$, for all $k \in H$ and $m \in \{S, L, R, U\}$.

6.3.4 Extending to multiple services

The analysis presented above shows when the approximations are accurate and the parameters that determine accuracy in comparison to simulation. Notably these findings still hold when applying both the approximations and the simulation to larger systems since this is equivalent to modelling an amalgamation of these smaller models with additional flow between them. As such, the modelling of larger systems may be implemented through a modular programming of the code which would increase its flexibility and scalability for modelling these scenarios.

By introducing multiple services, the flow dynamics of the other service orbits and alternative service orbits are introduced. This leads to a further change in how the effective traffic intensity is understood, which can be inferred from the model's structure. Notably, this is not significantly different to the previous results; thus, a new comparison is not required as the previous results hold.

In the analysis of multiple services, reuses and rejoins are governed by $r_{k,m,i,i}, m \in \{S, L\}$ respectively. Since patients may use other services after completing service, or use alternative services having abandoned, $r_{k,m,i,j}, m \in \{S, L\}, j \neq i$ may be small for systems of multiple services. However, patients may now arrive from other/alternative services, increasing the number of arrivals to each queue.

Thus, in considering the effective traffic intensity of a service in the network, alongside the parameters previously noted, the values of $r_{k,m,i,j}, m \in \{S, L\}$, for all $k \in H$; $i, j \in Ser$ should also be considered, helping to understand when the approximations are faithful for the multiple service extension. Thus, in such scenarios, the size and value of $\mathbf{S}_{k,m,i}, \mathbf{R}_{k,S,i}, \mathbf{R}_{k,L,i}$, for all $k \in H$, $i \in Ser$ and $m \in \{S, L, R, U, A, O\}$ may combine to increase the model's accuracy.

To illustrate the application to a larger, I now present a fluid and diffusion approximation for a three service and three outcome state system - this system has all the dynamics described in Figure 5.2. The input parameters used to populate this example are shown in Appendix C.6. Service 1 represents an acute prevention service, such as the Community Treatment Team or a service similar to District Nursing. It is modelled to be likely to serve patients in worse health states and represents episodes of care which are typically very short. From service 1 patients are referred into service 2 or 3 depending on their care needs. Service 2 has longer episodes of care and is likely to serve patients in any health state, whilst service 3 has the longest episodes of care and typically serves patients in healthier health states. Service 2 may be interpreted as a short term rehabilitation service and service 3 may represent a service that provides longer term support, such as Nutrition and Dietetics. Furthermore, a patient's health is considered to only improve through service, and may decline in between service.

This scenario highlights how the model may be used to represent a system of diverse care services that each have a different purpose, type of care (indicated by

service rate) and patient mix. Notably, given its small initial condition and arrival rate, the effective traffic intensity for service 3 is significantly increased by the flow from other services. Figure 6.24 shows the number of patients in each process orbit, the variance is not shown to improve the readability of the figure. Figure 6.25 gives the virtual waiting and its variance for each queue, whilst Figure 6.26 shows the dynamic capacity allocation for each health state and service in the system.

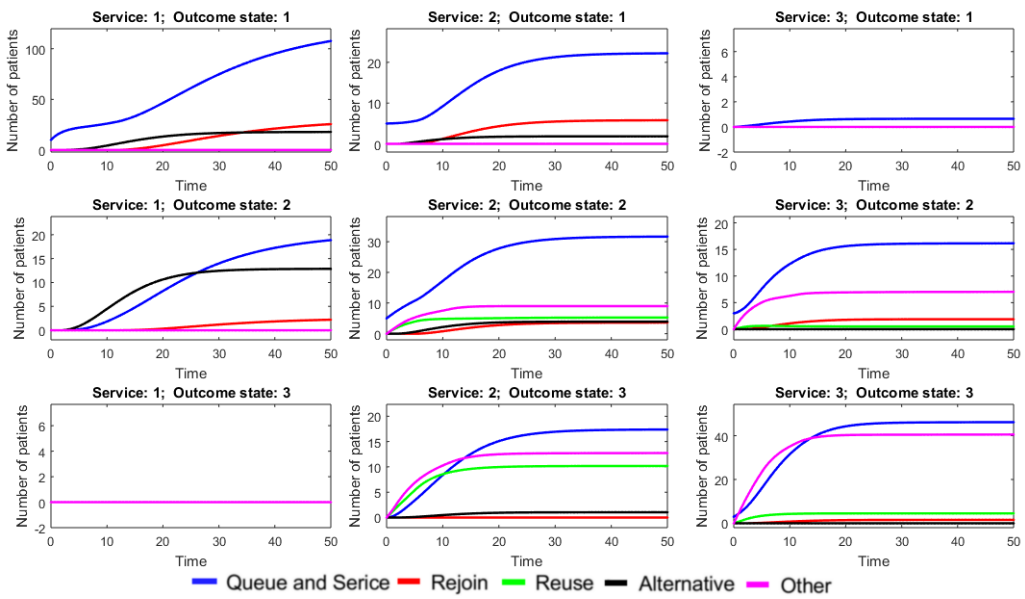

Figure 6.24: Three service and three health state system - number of patients in each process state

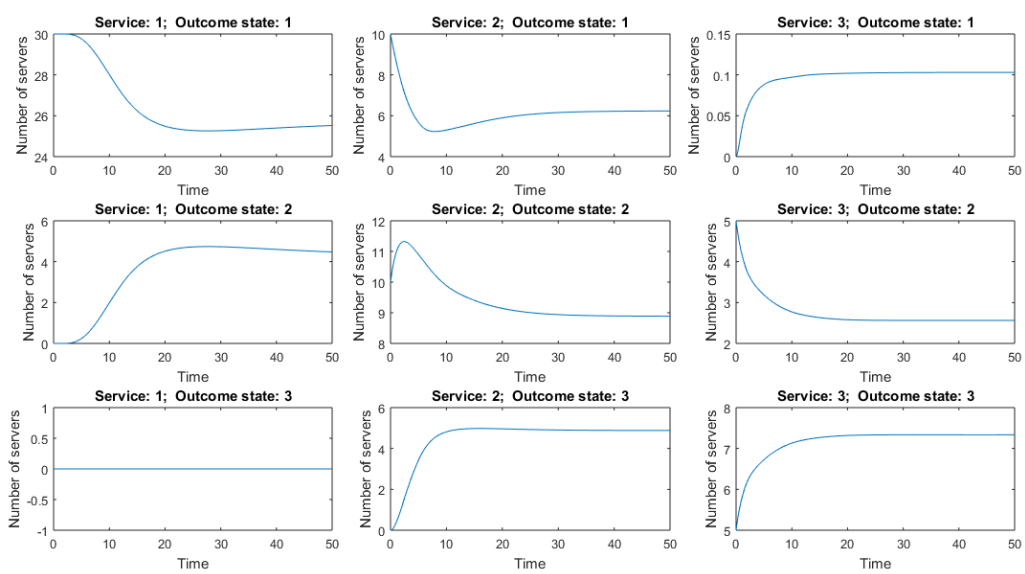

Figure 6.25: Three service and three health state system - dynamic server allocation

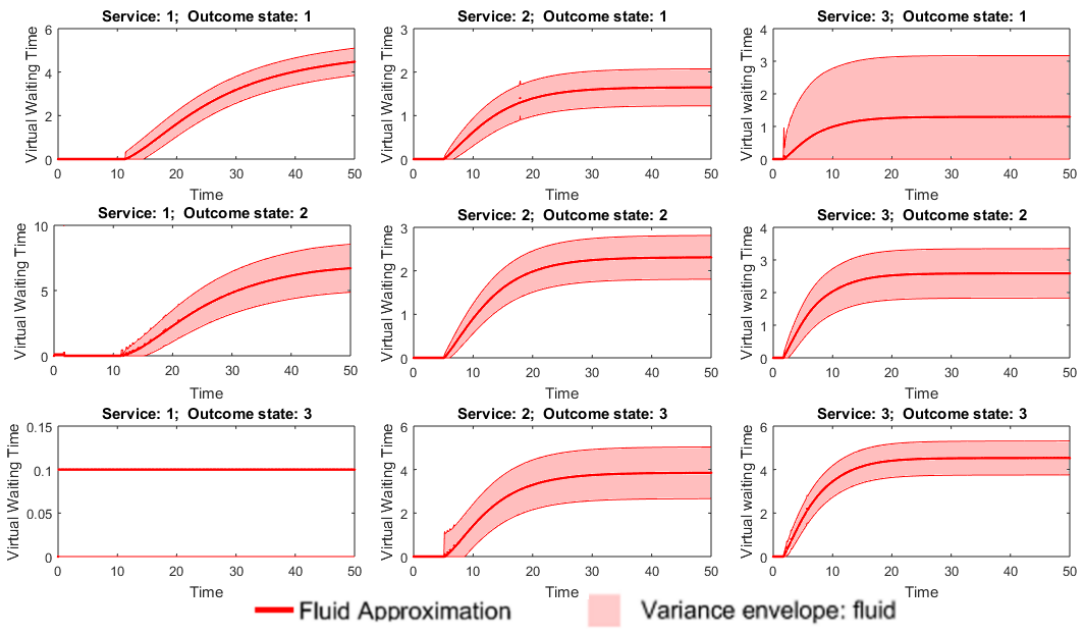

Figure 6.26: Three service and three health state system - the VWT and variance

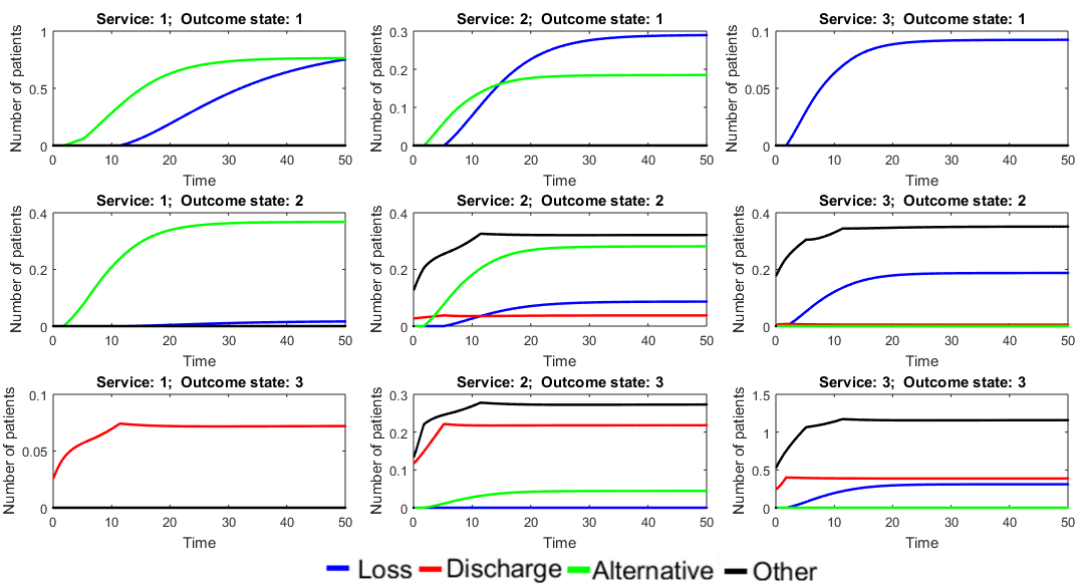

Figure 6.27: Three service and three health state system - the production of outcomes

Figure 6.27 illustrates the benefits of the production output in this scenario. For each service the output of patients in different health states over time is given by the loss and discharge curves. Additional curves correspond to the arrivals to each service of patients in each health state from the alternative service and other service process orbits - highlighting the flow between services. Together, these plots provide greater insight into the flow of patients and outcomes in the system and may be used

to identify negative and positive patterns of flow. For example, whilst service 3 has the highest rate of discharge for patients in the healthiest state, health state 3, there is a significant flow of patients from other services of patients in both health state 2 and 3. Thus, this service does not achieve good outcomes in isolation. Rather, this shows how services combine to produce good outcomes as patients participate in multiple care interactions and use several services.

6.4 Summary and Discussion

I have shown that under the right conditions, the approximations give accurate results and output measures in comparison to a simulation of the stochastic system. For larger systems and higher effective traffic intensity, these methods can be used accurately to model community services, given the parameter space and limitations investigated in this chapter. For example, systems with fewer servers may be accurately modelled when there is a high effective traffic intensity. This may occur when reuse of service, referrals from other services, arrivals from alternative services and the definition of health state transitions combine to increase the referrals.

Furthermore, both steady state and time dependent behaviour may be modelled accurately, increasing the range of scenarios to which these methods apply. Importantly, these methods model several complex dynamics, including multiple services, abandonment and rejoin, reuse, multiple health states, health dependent parameters, time dependence and use a dynamic multi-class server allocation.

When moving between underloaded and overloaded phases there can be periods of larger error due to the discrepancy between the deterministic approximations and stochastic variation of the simulation. In general, these errors are insignificant as they often relate to small values and can be quickly overcome to produce an accurate result (as long as the effective traffic intensity is "far" from 1).

In using parallel queues, I have modelled the clinical impact of care on a population of patients with different health, their different capacities to benefit and the differences in their health care requirements. The impact of these differences on the operation of the services may also be modelled. For example, by introducing competing queues, the effect of different care needs and patient demand on how servers may be dynamically allocated can be analysed, as well as the resulting effect of the allocation on the system's process outcomes e.g. loss, abandonment, throughput.

Discussion of process orbits and future demand

In this model, the total arrivals for each service include new arrivals, rejoins, arrivals from alternative services, reuse and arrivals from other services. Thus, when a system runs with a high utilisation, the total number of arrivals may increase as more patients abandon and rejoin or use alternative services. Similarly, with more servers, more patients complete service, increasing reuse and the arrival of patients from other services.

As a result, there is a dependency between overall patient demand (total arrivals) and system capacity since these orbits act as feedback loops of delayed demand. For example, having arrived, queued and completed service, a patient seeking to reuse the service will wait for a period of time before re-entering the queue. Therefore, more patients in each process state leads to higher future demand. Understanding the effect of these flow dynamics, and how resources may be managed in light of them, is important. Ignoring these orbits may lead to under or over staffing when modelling systems for which these arrivals are significant.

In previous publications, the modelling of these orbits has been most appropriate for systems where arrivals from one of these process states is indistinguishable from new arrivals [105]. However, through the use of health states, patient groups may be defined, for example, to include previous uses of service; thus, overcoming a previous

limitation of these methods.

Discussion of the "flow of outcomes"

By combining patient flow and clinical outcomes into a single modelling framework, a multifaceted view of system performance can be analysed. In particular, the model's output is informed by the effect of care, or absence of it, on patient outcomes and on the operation of services in the system. This provides a greater insight into the positive and negative clinical effects of a system's process outcomes.

In using health state dependent parameters, I modelled differentiated service to understand the flows of patients who have differing resource/service requirements. In combination with transitions between health states, these parameters provide a means for understanding the effect of care on the health of patients who have different capacities to benefit. This reflects real life, when patients of varying health have markedly different interactions with a health care system.

An overarching aim within this thesis has been to develop a method for understanding the "flow of outcomes" - how individual services contribute to the production of good outcomes across services, for patients who may use a range of these services and have multiple care interactions. This has been achieved through the development of these methods in several ways.

For example, the combination of health state transitions and abandonment helps to measure the impact of poor access on patient health and the production of outcomes from a given service. The number of patients who abandon, and the number who seek to rejoin or use an alternative service, are measures of whether patients are able to access and receive adequate care. In combination with health state transitions, this may be used to understand the possible negative impact of poor access. In a system with high abandonment and when rejoin/alternative service use is considered, patients may re-enter the queue in a worse health state than before, requiring

more resource intensive care and increasing the future burden on the health system. Thus, this represents a poor "flow of outcomes".

In addition, the combination of transitions in health state, reuse and uses of other services, helps to understand how multiple service interactions combine to produce good health. For example, how the receipt of care affects a patient's future use of services and the impact on their health. Thus, when a clinical improvement occurs, patients may require fewer interactions, reducing their future demand and the intensity of the care needed, thus producing a positive "flow of outcomes".

The production output is also an insightful measure for the "flow of outcomes", especially in time varying systems. This output gives the rate at which patients leave a service/the system in a given health state and may be used to understand the timely effect of certain server allocations, across the system. Thus, as the allocation of servers and patient demand changes, the wider impact and implications of these allocations may be understood both operationally and clinically.

In addition, "competing" queues present an opportunity for new and interesting analysis. In particular, they can provide insight into how the demand for services and differences in care requirements may affect a time varying allocation of servers and the demand for service. In combination with the production output, the dynamic server allocation may enhance the understanding of the operational and clinical performance of a system. For example, one may analyse how favourably allocating servers to patients in poorer health states affects the process outcomes, patient outcomes and the "flow of outcomes" in a system.

6.4.1 Limitations

There are several limitations of these methods. As discussed, these methods are most accurate for heavily loaded systems, especially when abandonment and rejoin is considered, limiting the scenarios to which these methods may be applied. In

particular, those with an effective traffic intensity close to 1, exhibit large errors for the abandonment and rejoin values, and therefore poorly fit the service and queue, reuse and other service orbits. This limitation may not be too restrictive for some community services, for example, those where reuse/other service use is high within the system.

Furthermore, the size of the system had a clear effect on the accuracy of these methods. For systems with the same effective traffic intensity, the accuracy of the results greatly improved as the system grew in size.

When considering multiple health states and services, the accuracy of the model (and thus the traffic intensity) is affected by: the service rates; the number of rejoins, reuses, alternative service uses and arrivals from other services; and the health state transitions upon exiting each orbit. Therefore, attention must be paid to these parameters when applying these methods. The size of the loss rate, θ, and rate of reuse, δ_U, have a marginal effect on the accuracy of these methods, especially when moving between underloaded and overloaded phases.

The largest errors between the approximations and simulation occur when the system is underloaded, and when the effective traffic intensity is "close" to 1. This is due to the stochastic variation in the queue size, which is not captured by the fluid approximation. Care must therefore be taken when seeking to gain insight for the model near critical points $z_{a,Q,i}(t) = c_{a,i}(\mathbf{z}(t))$. This is especially true for the variance of the VWT (as seen in [107]), where spikes occur in the approximated solution when switching between underloaded and overloaded phases.

When using the dynamic server allocation, errors in the order of dt occur in the solution of the VWT and its variance. This produces a trade off between speed and accuracy, as previously discussed.

Finally, the appropriateness of using these methods to evaluate community health care is not clear. Given the discussion in section 6.2 there is scope to interpret these

methods in the setting of community health care. However, when compared to data the strict Markovian assumptions may limit their accuracy, as well as the limitations of a large heavily loaded system. As a result, these methods may be better used to provide a stylistic representation of the system that helps to understand the dynamics of community health care and the consequences of changes in the system. Alternative methods such as simulation and system dynamics have a greater flexibility in this respect.

6.4.2 Possible avenues for future work

There are several natural directions in which this work could be extended, adapted or used. Firstly, it may be interesting to perform sensitivity analyses and apply dynamical system approaches to the fluid approximation to understand how different parameters affect the behaviour of the solution. This would be in a similar vein to [109]. Such an analysis may help to understand the stochastic system in greater detail, whilst improving understanding of the limitations and dynamics of the approximations.

Secondly, the modelling of the "flow of outcomes", either through these methods or by others, would be worthwhile. In particular, it would be interesting to explore further the benefits of this approach in application to a real world system.

Finally, since the method is computationally efficient, the benefits of the fast calculation may be used in optimisation algorithms and heuristic approaches. This could open up new avenues for informative analysis regarding capacity allocations or different referral policies, in light of patient outcomes, dynamic server allocation and diverse services. In addition, novel constraints can be considered, such as maximising health improvement, or maximising the number of patients in the best health states.

6.5 Conclusions

I have compared fluid and diffusion approximations to equivalent simulated solutions in application to a range of hypothetical scenarios, which feature several different dynamics. By starting with steady state models and working through to time varying behaviour with dynamic server allocation and multiple health states, I have shown how these methods may be used to model community health care and the "flow of outcomes" across multiple services. In particular, I have shown how the effective traffic intensity within different systems is altered by the inclusion of new dynamics, and how this subsequently impacts the accuracy of the approximations.

In developing these methods, I have produced a framework that may be used to model some aspects of community health care, also providing an introduction to the idea of the "flow of outcomes". This combination of health states and patient flow provides new avenues for insightful analysis within community care, highlighting the potential benefits of combining these two key perspectives of service/system performance. They help to understand how patients use services, the effect of multiple care interactions on patient health, and the effect of delayed demand/reuse of services on the operation of a system, in light of patient health.

By implementing a dynamic server allocation, I have produced a method for modelling how the demand of patients with different health, needs and behaviours (represented by health states/health state dependent parameters), combine to influence the number of servers each queue requires in comparison to others. This method is particularly insightful when considering time varying behaviour and the "flow of outcomes".

Chapter 7

Conclusions

At the outset of this project there were two overarching aims. The first was to develop methods for modelling patient flow in community health care, an area of health care that had received little attention in the operational research literature in comparison to acute and primary care. The second was to produce a method that incorporated patient outcomes into patient flow modelling, combining two key perspectives of health care performance.

In fulfilling these two goals, the ambition was to enhance the use of outcome measures within community services by applying a systems view to how they were understood. To do so, I sought to develop the concept of the "flow of outcomes", and methods for modelling it. In achieving this, the aim was to apply these methods to North East London Foundation Trust's (NELFT) services, helping them to better understand their community services.

However, initial conversations with health care professionals and data managers made clear that there was a lack of available data on outcome measurement. To investigate further, I carried out the work detailed in chapter 4, from which I found that there were neither measures for comparing the quality of diverse community services, nor for assessing the impact of multiple services on patient health. Furthermore, the patient data used in chapter 3 was not suitable to validate the model

since it was incomplete and lacked information about the capacity of these services, which was otherwise unavailable. Thus, it was not possible to apply these methods to NELFT services. As a result, the work changed direction towards working with hypothetical scenarios and methods.

The literature review in chapter 2 helped to inform the methods developed in this thesis. The findings of this review indicated that the development of time dependent methods that modelled patient flow within systems of multiple, differing community services, and included a mix of patients whose health could change in response to care would be useful.

This review made a further contribution outside of this thesis. Published in the Health Systems journal [20], this was, to the best of my knowledge, the first literature review to focus on OR methods for modelling patient flow applied to community health care services, and the first to review methods for modelling patient flow and outcomes in combination.

In order to develop a method for modelling patient flow in community health care, the relevant characteristics and dynamics of how patients used these services needed to be understood. To achieve this, I applied methods for visualising patient level data that provided insight into the above, as presented in chapter 3. These methods helped to: understand the vastness and complexity of the system; identify common groups of services (in terms of patient use); understand the levels and types of patient activity (such as reuse, sequential use and concurrent use); and better understand the timing, length and patterns of use in community care.

To my knowledge, this work was the first to use the combination of visualisation method with a specific focus to analyse the referral dynamics of community health care, in particular: patients reusing services, concurrent uses of different services and patterns of subsequent referrals. Whilst applied to a single provider, the methods are generalisable and may be easily applied to other boroughs, trusts and organisations.

They are visually impactful, informative, and simple to create using freely available programs (R and Gephi), increasing their scope for use and application in practice.

Given the data limitations, I developed a theoretical model using fluid and diffusion approximations, which could be used to model complex queue dynamics in a general network of queues. This model featured some of the dynamics identified in chapters 3 and 4, namely: the potential for patients to reuse services and for patients to use different services sequentially; and the potential for patients to abandon the queue, and possibly rejoin later or use an alternate service. Extending current methods, I also incorporated clinical outcomes into these methods in the form of transitions between health states.

This framework can be used to conduct analysis of time varying systems, where parameters are dependent on both time and patient health. By extending and combining existing methods, I produced a framework for calculating: the average number of patients in the system; an estimated waiting time (virtual waiting time); and the variance of each output. This made a contribution to the possible uses and applications of these approximations, the application of a dynamic server allocation for multi-class queues and the potential for modelling the "flow of outcomes".

In chapter 6, the parameter space was explored for these methods to understand whether the introduction of several services and transitions between health states affected the accuracy of the system and when the system was heavily loaded. For systems that are not heavily loaded according to the traditional definition of traffic intensity, the approximations may still produce accurate results since: $p_{k,i}, q_{k,i}, \mathbf{S}_{k,m,i}, \mathbf{R}_{k,n,i}$ for all $k \in H, i \in Ser, m \in \{S, L, R, U, A, O\}$ and $n \in \{s, L\}$ may combine to raise the effective traffic intensity of each service.

In developing these methods I produced a framework that may be used to model some aspects of community health care, and provide an introduction to modelling the "flow of outcomes" and "competing queues". Furthermore, these methods may be

used to understand how: patients use services; the effect of multiple care interactions on patient health; the effect of delayed demand/reuse of services on the operation of the system in light of patient health; and the dependency between the capacity of the system and the future arrival process affect the system.

There are several possible directions for future work. Firstly, there is merit in further exploring the "flow of outcomes". Having developed an illustrative method of the potential benefits, it would be insightful to apply these methods to the large, multi-service real world systems for which they were intended. For example, if there are sets of services or morbidities that have well defined outcome measures, this would be a useful avenue for further research. Likewise it would be beneficial to further explore their benefits and limitations in comparison to other modelling methods, such as system dynamic and Markov chain approaches, when modelling such systems. Furthermore, as noted in chapter 5, it would be beneficial to explore the combination of these methods with optimisation and heuristic approaches given the speed of calculation and ODE representation of the system. In particular, the flexibility in the definition of $C_{k,i}(\mathbf{Z}(t))$ and the inclusion of patent outcomes introduces the implementation of novel constraints and objectives, such as: how best to allocate servers to maximise health improvement and the production of outcomes in a system; or, to minimise the flow of patients through patterns of care that lead to poor outcomes.

Secondly, the mapping work generated a significant amount of interest amongst a range of health care professionals, organisations and researchers. Having presented the visualisation methods at several conferences and to various groups of care managers, future work would be to: distribute these methods further; apply them to other settings, trusts and boroughs; and continue working with NELFT to provide useful and responsive insight into their services. Furthermore, it would be beneficial to explore the network representation further through a deeper analysis, such as cluster analysis, and in combination other visualisation methods. Other methods

such as heat maps and methods which capture the variation of patient use over time would also be helpful.

Thirdly, there is the potential for wider application of the fluid and diffusion methods outside of health care. For example, the method in chapter 5 is a general framework for modelling a system of queues with complex dynamics, through which heterogeneous entities may flow, whose category/class may change throughout in response to, or due to, a lack of service. Furthermore, these groups may be used to model class dependent flow parameters. These dynamics may translate into industries such as telecommunications, where class represents levels of satisfaction, customer opinion or sales.

Bibliography

[1] Munton T, Martin A, Marrero I, Llewellyn A, Gibson K, and Gomersall A, *Evidence: Getting out of hospital?* Health Foundation, 2011.

[2] Ham C, Dixon A, and Brooke B, "Transforming the delivery of health and social care: the case for fundamental change," *King's Fund*, 2012.

[3] Foot C et al., "Managing quality in community health care services," *King's Fund*, 2014.

[4] NHS England, "Five year forward view," 2014.

[5] Sibbald B, McDonald R, and Roland M, "Shifting care from hospitals to the community: a review of the evidence on quality and efficiency," *Journal of Health Services Research & Policy*, vol. 12, no. 2, pp. 110–117, 2007.

[6] Hensher M, "Improving general practitioner access to physiotherapy: a review of the economic evidence," *Health Services Management Research*, vol. 10, no. 4, pp. 225–230, 1997.

[7] Powell J, "Systematic review of outreach clinics in primary care in the UK," *Journal of Health Services Research & Policy*, vol. 7, no. 3, pp. 177–183, 2002.

[8] Whitten P S et al., "Systematic review of cost effectiveness studies of telemedicine interventions," *British Medical Journal*, vol. 324, pp. 1434–1437, 2002.

[9] Côté M J, "Understanding patient flow," *Decision Line*, vol. 31, no. 2, pp. 8–10, 2000.

[10] Utley M, Gallivan S, Pagel C, and Richards D, "Analytical methods for calculating the distribution of the occupancy of each state within a multi-state flow system," *IMA Journal of Management Mathematics*, vol. 20, no. 4, pp. 345–355, 2009.

[11] Cochran J K and Roche K T, "A multi-class queuing network analysis methodology for improving hospital emergency department performance," *Computers & Operations Research*, vol. 36, no. 5, pp. 1497–1512, 2009.

[12] Cavirli T, Veral E, and Rosen H, "Designing appointment scheduling systems for ambulatory care services," *Health Care Management Science*, vol. 9, no. 1, pp. 47–58, 2006.

[13] Zhang Y, Berman O, and Verter V, "Incorporating congestion in preventive healthcare facility network design," *European Journal of Operational Research*, vol. 198, no. 3, pp. 922–935, 2009.

[14] Dshalalow J H, *Advances in Queueing Theory, Methods, and Open Problems*, vol. 4. CRC Press, 1995.

[15] A. S, *Applied probability and queues*, vol. 51. Springer Science & Business Media, 2008.

[16] Depart Of Health, "Our health, our care, our say: a new direction for community," 2006. `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/272238/6737.pdf`, Accessed: 2015 - 08 - 01.

[17] North East London Foundation Trust, "NELFT operational plan 14-16." `https://www.gov.uk/government/uploads/system/uploads/attachment_`

data/file/341093/NELONDON_Operational_Plan_14-16_1_.pdf, Accessed: 2015 - 08 - 01.

[18] Department of Health, *High quality care for all: NHS next stage review final report*. The Stationery Office, 2008.

[19] Liu Y, Avant K C, Aungsuroch Y, Zhang X, and Jiang P, "Patient outcomes in the field of nursing: a concept analysis," *International Journal of Nursing Sciences*, vol. 1, no. 1, pp. 69–74, 2014.

[20] Palmer R, Fulop N J, and Utley M, "A systematic literature review of operational research methods for modelling patient flow and outcomes within community healthcare and other settings," *Health Systems*, pp. 1–21, 2017.

[21] Fakhimi M and Probert J, "Operations research within uk healthcare: a review," *Journal of Enterprise Information Management*, vol. 26, no. 1/2, pp. 21–49, 2013.

[22] Hulshof P J H, Kortbeek N, Boucherie R J, Hans E W, and Bakker P J M, "Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS," *Health Systems*, vol. 1, no. 2, pp. 129–175, 2012.

[23] Gough D, Thomas J, and Oliver S, "Clarifying differences between review designs and methods," *Systematic Reviews*, vol. 1, p. 28, 2012.

[24] Moher D, Liberati A, Tetzlaff J, and Altman D G, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," *Annals of Internal Medicine*, vol. 151, no. 4, pp. 264–269, 2009.

[25] Lane D, Uyeno D, Stark A, Kliewer E, and Gutman G, "Forecasting demand for long-term care services," *Health Services Research*, vol. 20, no. 4, pp. 435–460, 1985.

[26] Lane D, Uyeno D, Stark A, Gutman G, and McCashin B, "Forecasting client transitions in British Columbia's Long-Term Care Program," *Health Services Research*, vol. 22, no. 5, pp. 671–706, 1987.

[27] Koizumi N, Kuno E, and Smith T E, "Modeling patient flows using a queuing network with blocking," *Health Care Management Science*, vol. 8, no. 1, pp. 49–60, 2005.

[28] Kucukyazici B, Verter V, and Mayo N E, "An analytical framework for designing community-based care for chronic diseases," *Production and Operations Management*, vol. 20, no. 3, pp. 474–488, 2011.

[29] Song J, Chen W, and Wang L, "A block queueing network model for control patients flow congestion in urban healthcare system," *International Journal of Services Operations and Informatics*, vol. 7, no. 2-3, pp. 82–95, 2012.

[30] Chao J et al., "The long-term effect of community-based health management on the elderly with type 2 diabetes by the markov modeling," *Archives of Gerontology and Geriatrics*, vol. 59, no. 2, pp. 353–359, 2014.

[31] Bretthauer KM and Côté MJ, "A model for planning resource requirements in health care organizations," *Decision Sciences*, vol. 29, no. 1, pp. 243–270, 1998.

[32] Sterman J D, *Business dynamics: systems thinking and modeling for a complex world*, vol. 19. Irwin/McGraw-Hill Boston, 2000.

[33] Wolstenholme E, "A patient flow perspective of uk health services: exploring the case for new "intermediate care" initiatives," *System Dynamics Review*, vol. 15, no. 3, p. 253, 1999.

[34] Taylor K, Dangerfield B, and Le Grand J, "Simulation analysis of the consequences of shifting the balance of health care: a system dynamics approach," *Journal of Health Services Research & Policy*, vol. 10, no. 4, pp. 196–202, 2005.

[35] Ansah J P et al., "Simulating the impact of long-term care policy on family eldercare hours," *Health Services Research*, vol. 48 2pt2, pp. 773–791, 2013.

[36] Ansah J P et al., "Implications of long-term care capacity response policies for an aging population: a simulation analysis," *Health Policy*, vol. 116, no. 1, pp. 105–113, 2014.

[37] Cepoiu-Martin M and Bischak D P, "Policy choices in dementia care - An exploratory analysis of the Alberta continuing care system (ACCS) using system dynamics," *Journal of Evaluation in Clinical Practice*, 2017.

[38] Xie H, Chaussalet T J, and Millard P H, "A continuous time markov model for the length of stay of elderly people in institutional long-term care," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 168, no. 1, pp. 51–61, 2005.

[39] Xie H, Chaussalet T J, and Millard P H, "A model-based approach to the analysis of patterns of length of stay in institutional long-term care," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 3, pp. 512–518, 2006.

[40] Hare W L, Alimadad A, Dodd H, Ferguson R, and Rutherford A, "A deterministic model of home and community care client counts in British Columbia," *Health Care Management Science*, vol. 12, no. 1, pp. 80–98, 2009.

[41] Pagel C, Richards D A, and Utley M, "A mathematical modelling approach for systems where the servers are almost always busy," *Computational and Mathematical Methods in Medicine*, vol. 2012, 2012.

[42] Garg L, Mcclean S, Barton M, Meenan B J, and Fullerton K, "Intelligent Patient Management and Resource Planning for Complex, Heterogeneous, and Stochastic Healthcare Systems," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 42, no. 6, pp. 1332–1345, 2012.

[43] Deo S, Iravani S, Jiang T, Smilowitz K, and Samuelson S, "Improving health outcomes through better capacity allocation in a community-based chronic care model," *Operations Research*, vol. 61, no. 6, pp. 1277–1294, 2013.

[44] Izady N, "Appointment capacity planning in specialty clinics: A queueing approach," *Operations Research*, vol. 63, no. 4, pp. 916–930, 2015.

[45] Cardoso T, Oliveira M D, Barbosa-Póvoa A, and Nickel S, "Modeling the demand for long-term care services under uncertain information," *Health Care Management Science*, vol. 15, no. 4, pp. 385–412, 2012.

[46] Zhang Y, Puterman M L, Nelson M, and Atkins D, "A simulation optimization approach to long-term care capacity planning," *Operations Research*, vol. 60, no. 2, pp. 249–261, 2012.

[47] Zhang Y and Puterman M L, "Developing an adaptive policy for long-term care capacity planning," *Health Care Management Science*, vol. 16, no. 3, pp. 271–279, 2013.

[48] Clague J E, Reed P G, Barlow J, Rada R, Clarke M, and Edwards R H T, "Improving outpatient clinic efficiency using computer simulation," *International Journal of Health Care Quality Assurance*, vol. 10, no. 5, pp. 197–201, 1997.

[49] Swisher J R and Jacobson S H, "Evaluating the design of a family practice healthcare clinic using discrete-event simulation," *Health Care Management Science*, vol. 5, no. 2, pp. 75–88, 2002.

[50] Matta M E and Patterson S S, "Evaluating multiple performance measures across several dimensions at a multi-facility outpatient center," *Health Care Management Science*, vol. 10, no. 2, pp. 173–194, 2007.

[51] Chand S, Moskowitz H, Norris J B, Shade S, and Willis D R, "Improving patient flow at an outpatient clinic: study of sources of variability and improvement factors," *Health Care Management Science*, vol. 12, no. 3, pp. 325–340, 2009.

[52] Ponis S T, Delis A, Gayialis S P, Kasimatis P, and Tan J, *Applying Discrete Event Simulation (DES) in Healthcare: The Case for Outpatient.* IGI Global, 2014.

[53] Pan C, Zhang D, Kon A W M, Wai C S L, and Ang W B, "Patient flow improvement for an ophthalmic specialist outpatient clinic with aid of discrete event simulation and design of experiment," *Health Care Management Science*, vol. 18, no. 2, pp. 137–155, 2015.

[54] Santibáñez P, Chow V S, French J, Puterman M L, and Tyldesley S, "Reducing patient wait times and improving resource utilization at British Columbia Cancer Agency's ambulatory care unit through simulation," *Health Care Management Science*, vol. 12, no. 4, pp. 392–407, 2009.

[55] Fialho A S, Oliveira M D, and Sá A B, "Using discrete event simulation to compare the performance of family health unit and primary health care centre organizational models in Portugal," *BMC Health Services Research*, vol. 11, p. 275, 2011.

[56] Shi J, Peng Y, and Erdem E, "Simulation analysis on patient visit efficiency of a typical VA primary care clinic with complex characteristics," *Simulation Modelling Practice and Theory*, vol. 47, pp. 165–181, 2014.

[57] Bayer S, Petsoulas C, Cox B, Honeyman A, and Barlow J, "Facilitating stroke care planning through simulation modelling," *Health Informatics Journal*, vol. 16, no. 2, pp. 129–143, 2010.

[58] Patrick J, Nelson K, and Lane D, "A simulation model for capacity planning in community care," *Journal of Simulation*, vol. 9, no. 2, pp. 111–120, 2015.

[59] Qiu Y, Song J, and Liu Z, "A simulation optimisation on the hierarchical health care delivery system patient flow based on multi-fidelity models," *International Journal of Production Research*, vol. 54, no. 21, pp. 6478–6493, 2016.

[60] Zenios S A, "Modeling the transplant waiting list: A queueing model with reneging," *Queueing Systems*, vol. 31, no. 3-4, pp. 239–251, 1999.

[61] Wang Q, "Modeling and analysis of high risk patient queues," *European Journal of Operational Research*, vol. 155, no. 2, pp. 502–515, 2004.

[62] Drekic S, Stanford D A, Woolford D G, and Mcalister V C, "A model for deceased-donor transplant queue waiting times," *Queueing Systems*, vol. 79, no. 1, pp. 87–115, 2015.

[63] Shmueli A, Sprung C L, and Kaplan E H, "Optimizing admissions to an intensive care unit," *Health Care Management Science*, vol. 6, no. 3, pp. 131–136, 2003.

[64] Kim S and Kim S, "Differentiated waiting time management according to patient class in an emergency care center using an open jackson network integrated with pooling and prioritizing," *Annals of Operations Research*, vol. 230, no. 1, pp. 35–55, 2015.

[65] Goddard J and Tavakoli M, "Efficiency and welfare implications of managed public sector hospital waiting lists," *European Journal of Operational Research*, vol. 184, no. 2, pp. 778–792, 2008.

[66] Stanford D A, Lee J M, Chandok N, and McAlister V, "A queuing model to address waiting time inconsistency in solid-organ transplantation," *Operations Research for Health Care*, vol. 3, no. 1, pp. 40–45, 2014.

[67] Diaz R, Behr J, Kumar S, and Britton B, "Modeling chronic disease patient flows diverted from emergency departments to patient-centered medical homes," *IIE Transactions on Healthcare Systems Engineering*, vol. 5, no. 4, pp. 268–285, 2015.

[68] Deo S, Rajaram K, Rath S, Karmarkar U S, and Goetz M B, "Planning for HIV screening, testing, and care at the veterans health administration," *Operations Research*, vol. 63, no. 2, pp. 287–304, 2015.

[69] Liquet B, Timsit J F, and Rondeau V, "Investigating hospital heterogeneity with a multi-state frailty model: application to nosocomial pneumonia disease in intensive care units," *BMC Medical Research Methodology*, vol. 12, no. 79, 2012.

[70] Chan C W, Farias V F, Bambos N, and Escobar G J, "Optimizing intensive care unit discharge decisions with patient readmissions," *Operations Research*, vol. 60, no. 6, pp. 1323–1341, 2012.

[71] Thomsen M S and Nørrevang O, "A model for managing patient booking in a radiotherapy department with differentiated waiting times," *Acta Oncologica*, vol. 48, no. 2, pp. 251–258, 2009.

[72] Li S, Geng N, and Xie X, "Radiation queue: Meeting patient waiting time targets," *IEEE Robotics & Automation Magazine*, vol. 22, no. 2, pp. 51–63, 2015.

[73] Zenios S A, Chertow G M, and Wein L M, "Dynamic allocation of kidneys to

candidates on the transplant waiting list," *Operations Research*, vol. 48, no. 4, pp. 549–569, 2000.

[74] Alagoz O, Maillart L M, Schaefer A J, and Roberts M S, "The optimal timing of living-donor liver transplantation," *Management Science*, vol. 50, no. 10, pp. 1420–1430, 2004.

[75] Gupta D, Natarajan M K, Gafni A, Wang L, Shilton D, Holder D, and Yusuf S, "Capacity planning for cardiac catheterization: a case study," *Health policy*, vol. 82, no. 1, pp. 1–11, 2007.

[76] Yuan Y, Gafni A, Russell J D, and Ludwin D, "Development of a central matching system for the allocation of cadaveric kidneys a simulation of clinical effectiveness versus equity," *Medical Decision Making*, vol. 14, pp. 124–136, 1994.

[77] McLean D R and Jardine A G, "A simulation model to investigate the impact of cardiovascular risk in renal transplantation," in *Transplantation Proceedings*, vol. 37(5), pp. 2135–2143, Elsevier, 2005.

[78] Shechter S M, Bryce C L, Alagoz O, Kreke J E, Stahl J E, Schaefer A J, Angus D C, and Roberts M S, "A clinically based discrete-event simulation of end-stage liver disease and the organ allocation process," *Medical Decision Making*, vol. 25, no. 2, pp. 199–209, 2005.

[79] Panayiotopoulos J C and Vassilacopoulos G, "Simulating hospital emergency departments queuing systems:(GI/G/m(t)):(IHFF/N/inf)," *European Journal of Operational Research*, vol. 18, no. 2, pp. 250–258, 1984.

[80] DeRienzo C M et al., "A discrete event simulation tool to support and predict hospital and clinic staffing," *Health Informatics Journal*, pp. 124–133, 2016.

[81] Van Zon A H and Kommer G J, "Patient flows and optimal health-care resource allocation at the macro-level: a dynamic linear programming approach," *Health Care Management Science*, vol. 2, no. 2, pp. 87–96, 1999.

[82] Cleveland W S and McGill R, "Graphical perception and graphical methods for analyzing scientific data," *Science*, vol. 229, no. 4716, pp. 828–833, 1985.

[83] Plaisant C, Milash B, Rose A, Widoff S, and Shneiderman B, "Lifelines: visualizing personal histories," *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 221–227, 1996.

[84] Plaisant C et al., "LifeLines: using visualization to enhance navigation and analysis of patient records," *Proceedings of the AMIA Symposium*, p. 76, 1998.

[85] Falster M O, Jorm L R, and Leyland A H, "Visualising linked health data to explore health events around preventable hospitalisations in NSW Australia," *BMJ Open*, vol. 6, no. 9, 2016.

[86] West V L, Borland D, and Hammond W, "Innovative information visualization of electronic health record data: a systematic review," *Journal of the American Medical Informatics Association*, vol. 22, no. 2, pp. 330–339, 2015.

[87] Zhang Y, Padman R, and Patel N, "Paving the COWpath: Learning and visualizing clinical pathways from electronic health record data," *Journal of Biomedical Informatics*, vol. 58, pp. 186–197, 2015.

[88] Soulakis N D, Carson M B, Lee Y J, Schneider D H, Skeehan C T, and Scholtens D M, "Visualizing collaborative electronic health record usage for hospitalized patients with heart failure," *Journal of the American Medical Informatics Association*, vol. 22, no. 2, pp. 299–311, 2015.

[89] UCL IT for SLMS, "Data Safe Haven." `https://www.ucl.ac.uk/isd/`

itforslms/services/handling-sens-data/tech-soln, Accessed: 2017-05-19.

[90] Bastian M, Heymann S, and Jacomy M, "Gephi: An Open Source Software for Exploring and Manipulating Networks," in *International AAAI Conference on Weblogs and Social Media*, 2009. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[91] Bojanowski M and Edwards R, "alluvial: R package for creating alluvial diagrams," 2016. R package version: 0.1-2, https://github.com/mbojan/alluvial, Accessed: 2017-09-20.

[92] Bostock M, Rodden K, and Russell K, "sunburstR: 'Htmlwidget' for 'Kerry Rodden' 'd3.js' Sequence Sunburst," 2017. R package version 0.6.5, https://CRAN.R-project.org/package=sunburstR, Accessed: 2017-09-20.

[93] Department of Health, *Hard truths: the journey to putting patients first.* The Stationery Office, 2013.

[94] Francis R, *Report of the Mid Staffordshire NHS Foundation Trust Public Inquiry*, vol. 947. The Stationary Office, 2013.

[95] Berwick D, *A promise to learn–a commitment to act: improving the safety of patients in England.* Department of Health, 2013.

[96] Ham C, Berwick D M, and Dixon J, *Improving quality in the English NHS: a strategy for action.* King's Fund, 2016.

[97] Health Foundation, *Quality improvement made simple.* Health Foundation, 2013.

[98] World Health Organisation, "What is Quality of Care and why is it impor-

tant?," 2017. http://www.who.int/maternal_child_adolescent/topics/quality-of-care/definition/en/, Accessed: 2017 - 08 - 14.

[99] Institute of Medicine, *Crossing the quality chasm: a new health system for the 21st century*. Washington DC: National Academy Press, 1990.

[100] Care Quality Comission and others, "The five key questions we ask," 2016. http://www.cqc.org.uk/what-we-do/how-we-do-our-job/five-key-questions-we-ask, Accessed: 2017 - 08 - 14.

[101] Chen X, Wang L, Ding J, and Thomas N, "Patient flow scheduling and capacity planning in a smart hospital environment," *IEEE Access*, vol. 4, pp. 135–148, 2016.

[102] Remerova M, *Fluid Limit Approximations of Stochastic Networks*. PhD thesis, Vrije Universiteit, 2014.

[103] Hillston J, "Fluid flow approximation of PEPA models," in *Second International Conference on the Quantitative Evaluation of Systems, 2005*, pp. 33–42, IEEE, 2005.

[104] Mandelbaum A, Massey W A, and Reiman M I, "Strong approximations for Markovian service networks," *Queueing Systems*, vol. 30, no. 1, pp. 149–201, 1998.

[105] Ding S, Remerova M, van der Mei R D, and Zwart B, "Fluid approximation of a call center model with redials and reconnects," *Performance Evaluation*, vol. 92, pp. 24–39, 2015.

[106] Yom-Tov G and Mandelbaum A, "Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing," *Manufacturing & Service Operations Management*, vol. 16, no. 2, pp. 283–299, 2014.

[107] Mandelbaum A, Massey W A, Reiman M I, Stolyar A, and Rider B, "Queue lengths and waiting times for multiserver queues with abandonment and retrials," *Telecommunication Systems*, vol. 21, no. 2, pp. 149–171, 2002.

[108] Niyirora J and Zhuang J, "Fluid approximations and control of queues in emergency departments," *European Journal of Operational Research*, vol. 261, no. 3, pp. 1110–1124, 2017.

[109] Pender J, Rand R H, and Wesson E, "Queues with choice via delay differential equations," *International Journal of Bifurcation and Chaos*, vol. 27, 2017.

[110] Federgruen A and Groenevelt H, "M/g/c queueing systems with multiple customer classes: characterization and control of achievable performance under nonpreemptive priority rules," *Management Science*, vol. 34, no. 9, pp. 1121–1138, 1988.

[111] Maglaras C, "Dynamic scheduling in multiclass queueing networks: Stability under discrete-review policies," *Queueing Systems*, vol. 31, no. 3-4, pp. 171–206, 1999.

[112] Ata B, "Dynamic control of a multiclass queue with thin arrival streams," *Operations Research*, vol. 54, no. 5, pp. 876–892, 2006.

[113] Zhang Z G and Tian N, "An analysis of queueing systems with multi-task servers," *European Journal of Operational Research*, vol. 156, no. 2, pp. 375–389, 2004.

[114] Atar R and Mandelbaum A and Reiman M I and others, "Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic," *The Annals of Applied Probability*, vol. 14, no. 3, pp. 1084–1134, 2004.

[115] Pang G, Talreja R, and Whitt W, "Martingale proofs of many-server heavy-traffic limits for Markovian queues," *Probability Surveys*, vol. 4, no. 193–267, p. 7, 2007.

[116] Walck C, *Hand-book on statistical distributions for experimentalists*. 1996.

[117] Skorokhod A V, "Limit theorems for stochastic processes," *Theory of Probability & Its Applications*, vol. 1, no. 3, pp. 261–290, 1956.

[118] Jakubowski A, "The Skorokhod space in functional convergence: a short introduction," in *Abstracts for the Internation Conference Skorokhod Space*, vol. 50, pp. 11–18, 2007.

[119] Ethier S N and Kurtz T G, *Markov processes: characterization and convergence*. John Wiley & Sons, 2009.

[120] Puhalskii A, "On the invariance principle for the first passage time," *Mathematics of Operations Research*, vol. 19, no. 4, pp. 946–954, 1994.

[121] Billingsley P, *Convergence of probability measures*. John Wiley & Sons, 2013.

[122] Reed J and Ward A R, "A diffusion approximation for a generalized jackson network with reneging," in *Proceedings of the 42nd Allerton Conference on Communication, Control, and Computing*, 2004.

Appendix A

Chapter 3

A.1 Details of data cleaning process

Uncleaned dataset: 1,263,914 observations.

1) Keep entries where patients had at least one contact with a service:

- Remove cases where length of stay > 0 and number of contacts/DNA/cancellations $= 0$

2) Edited dates and time. Stored as "datetime" strings in the format MM/DD/YYYY HH:MM:SS I split these for easier computation and data manipulation.

I converted dates to integer values of days since 1 January 1960 - a standard format in STATA.

Date of birth - time dropped due to inaccuracy

Referral Datetime - time retained to chronologically order the data by referral

Discharge Datetime - time was retained to chronologically order the data by discharges

Appointment date:

- Generated "appdate" variable: date of appointment
- Generated "apphour" variable: containing hour of appointment
- Generated "appmin" variable: containing minute of appointment
- Generated "ampm" variable: denoting AM or PM appointment
- "apphour" and "appmin" stored as integers

Cancellation Datetime - time retained for sorting data

3) Removed records where "refdate" was before 1 April 2014, or when a patient had not been discharged by 31 August 2016

Table A.1: Details of the data cleaning process.

4) Referral IDs created.

"refid" was a variable in the original dataset for designating each unique referral that a patient had.

However, this variable was neither chronological nor formed of consecutive integers, hindering data processing.

- Created"ref" variable, an ID of consecutively ordered integers for chronologically identifying unique referrals for each patient.

5) Removed contact data where patients were younger than 65.

- Generated "ageatappointment" variable: the floor of (appointment date - date of birth)
- Removed cases where ageatappointment < 65; 29,579 observations dropped
- 3 cases where refdate $<$ date of birth dropped; 302 observations dropped (3 referrals)

6) "Appid" created, a chronological identifier for unique appointments, for each referral.

7) Created new counts of: contacts, DNAs and cancellations so that the count variables for each only included those between 1 April 2014 and 31 August 2016.

8) Edited source and specialty names for consistent referencing:

- GP: *GP written, GP verbal, General medical practitioner, Out of hours GP Service*
- Care Home: *Care/residential home, Nursing home*
- Acute: *A+E, Ambulance, Hospital admission, Hospital clinical specialty, Hospital consultant, Hospital inpatient service*
- Nutrition and Dietetic Service: *Adult Nutrition & Dietetic Service, Nutrition & Dietetic Service, Acute - Nutrition and Dietetic Service*
- Self referral: *Choose and book*
- Carer/Relative: *Parent, Family*
- Speech and language therapy: *Acute - Speech and language therapy, Speech and Language Therapy - Adult*
- District Nursing Service: *District Nursing Night Service, District Nursing Night Service WF, District Nursing Service WF*
- Health visiting Service: *Health visiting*
- Intermediate Care Service: *Intermediate Care*
- Podiatry service: *Podiatry*
- Prosthetics Service: *Prosthetics*
- Tissue Viability Service: *Tissue Viability*

Table 3.2. (Continued): Details of the data cleaning process.

Appendix B

Chapter 5

B.1 Proof of Theorem 5.5.1

Proof 3 *Re-write the scaled fluid process as follows:*

$$\overline{Z}^{(\eta)}_{k,Q,j}(t) = \overline{Z}^{(\eta)}_{k,Q,j}(0) + G^{(\eta)}_{k,Q,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) + \int_0^t H_{k,Q,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right) du$$

$$\overline{Z}^{(\eta)}_{k,R,j}(t) = \overline{Z}^{(\eta)}_{k,R,j}(0) + G^{(\eta)}_{k,R,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) + \int_0^t H_{k,R,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right) du$$

$$\overline{Z}^{(\eta)}_{k,U,j}(t) = \overline{Z}^{(\eta)}_{k,U,j}(0) + G^{(\eta)}_{k,U,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) + \int_0^t H_{k,U,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right) du$$

$$\overline{Z}^{(\eta)}_{k,A,j}(t) = \overline{Z}^{(\eta)}_{k,A,j}(0) + G^{(\eta)}_{k,A,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) + \int_0^t H_{k,A,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right) du$$

$$\overline{Z}^{(\eta)}_{k,O,j}(t) = \overline{Z}^{(\eta)}_{k,O,j}(0) + G^{(\eta)}_{k,O,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) + \int_0^t H_{k,O,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right) du$$

$$\overline{Z}^{(\eta)}_{k,L,j}(t) = \overline{Z}^{(\eta)}_{k,L,j}(0) + G^{(\eta)}_{k,L,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) + \int_0^t H_{k,L,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right) du$$

$$\overline{Z}^{(\eta)}_{k,D,j}(t) = \overline{Z}^{(\eta)}_{k,D,j}(0) + G^{(\eta)}_{k,D,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) + \int_0^t H_{k,D,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right) du$$

Where, using the definitions (5.20)-(5.27):

$$
\begin{aligned}
G^{(\eta)}_{k,Q,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) :=& \left(\frac{\Pi_{\lambda_{k,j}(t)\eta}}{\eta} - \int_0^t \lambda_{k,j}(u)du\right) \\
&- \left(\overline{D}^{(\eta)}_{k,s,j}(t) - \int_0^t \mu_{k,j}(u)\min\left(\overline{Z}^{(\eta)}_{k,Q,j}(u), C_{k,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)du\right) \\
&- \left(\overline{D}^{(\eta)}_{k,L,j}(t) - \int_0^t \theta_{k,j}(u)\left(\overline{Z}^{(\eta)}_{k,Q,j}(u) - C_{k,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)^+ du\right) \\
&+ \sum_{l=1}^K \left(\overline{MS}^{(\eta)^{(k)}}_{l,R,j}(t) - \int_0^t s_{l,k,R,j}(u)\delta_{l,R,j}(u)\overline{Z}^{(\eta)}_{l,R,j}(u)du\right) \\
&+ \sum_{l=1}^K \left(\overline{MS}^{(\eta)^{(k)}}_{l,U,j}(t) - \int_0^t s_{l,k,U,j}(u)\delta_{l,U,j}(u)\overline{Z}^{(\eta)}_{l,U,j}(u)du\right) \\
&+ \sum_{l=1}^K \left(\overline{MS}^{(\eta)^{(k)}}_{l,A,j}(t) - \int_0^t s_{l,k,A,j}(u)\delta_{l,A,j}(u)\overline{Z}^{(\eta)}_{l,A,j}(u)du\right) \\
&+ \sum_{l=1}^K \left(\overline{MS}^{(\eta)^{(k)}}_{l,O,j}(t) - \int_0^t s_{l,k,O,j}(u)\delta_{l,O,j}(u)\overline{Z}^{(\eta)}_{l,O,j}(u)du\right)
\end{aligned}
$$

$$
\begin{aligned}
G^{(\eta)}_{k,R,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) :=& \overline{MR}^{(j)}_{k,L,j}(t) - \left(\overline{D}^{(\eta)}_{k,R,j}(t) - \int_0^t \delta_{k,R,j}(u)\overline{Z}^{(\eta)}_{k,R,j}(u)du\right) \\
&- \sum_{l=1}^K \int_0^t s_{l,k,L,j}(u)r_{k,L,j,j}(u)\theta_{l,j}(u)\left(\overline{Z}^{(\eta)}_{l,Q,j}(u) - C_{l,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)^+ du \\
G^{(\eta)}_{k,U,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) :=& \overline{MR}^{(j)}_{k,s,j}(t) - \left(\overline{D}^{(\eta)}_{k,U,j}(t) - \int_0^t \delta_{k,U,j}(u)\overline{Z}^{(\eta)}_{k,U,j}(u)du\right) \\
&- \sum_{l=1}^K \int_0^t s_{l,k,S,j}(u)r_{k,S,j,j}(u)\mu_{l,j}(u)\min\left(\overline{Z}^{(\eta)}_{l,Q,j}(u), C_{l,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)du \\
G^{(\eta)}_{k,A,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) :=& \overline{MR}^{(j)}_{k,L,i}(t) - \left(\overline{D}^{(\eta)}_{k,A,j}(t) - \int_0^t \delta_{k,A,j}(u)\overline{Z}^{(\eta)}_{k,A,j}(u)du\right) \\
&- \sum_{l=1}^K \sum_{i=1;i\neq j}^J \int_0^t s_{l,k,L,i}(u)r_{k,L,i,j}(u)\theta_{l,i}(u)\left(\overline{Z}^{(\eta)}_{l,Q,i}(u) - C_{l,i}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)^+ du \\
G^{(\eta)}_{k,O,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) :=& \overline{MR}^{(j)}_{k,s,i}(t) - \left(\overline{D}^{(\eta)}_{k,O,j}(t) - \int_0^t \delta_{k,O,j}(u)\overline{Z}^{(\eta)}_{k,O,j}(u)du\right) \\
&- \sum_{l=1}^K \sum_{i=1;i\neq j}^J \int_0^t s_{l,k,S,i}(u)r_{k,S,i,j}(u)\mu_{l,i}(u)\min\left(\overline{Z}^{(\eta)}_{l,Q,i}(u), C_{l,i}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)du \\
G^{(\eta)}_{k,L,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) :=& \ \overline{MR}^{(\eta)^{(J+1)}}_{k,L,j}(t) - \sum_{l=1}^K \int_0^t r_{k,L,j,J+1}(u)s_{l,k,L,j}(u)\theta_{l,j}(u) \\
&\times \left(\overline{Z}^{(\eta)}_{l,Q,j}(u) - C_{l,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)^+ du
\end{aligned}
$$

$$G^{(\eta)}_{k,D,j}\left(\overline{\mathbf{Z}}^{(\eta)}(t)\right) := \overline{MR}^{(\eta)^{(J+1)}}_{k,s,j}(t) - \sum_{l=1}^{K}\int_0^t r_{k,S,j,J+1}(u)s_{l,k,S,j}(u)\mu_{l,j}(u) \times \min\left(\overline{Z}^{(\eta)}_{l,Q,j}(u), C_{l,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right) du$$

And:

$$\begin{aligned}
\int_0^t H_{k,Q,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\ du :=& \int_0^t \lambda_{k,j}(u) - \mu_{k,j}(u)\min\left(\overline{Z}^{(\eta)}_{k,Q,j}(u), C_{k,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right) \\
&- \theta_{k,j}(u)\left(\overline{Z}^{(\eta)}_{k,Q,j}(u) - C_{k,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)^+ \\
&+ \sum_{l=1}^{K} s_{l,k,R,j}(u)\delta_{l,R,j}(u)\overline{Z}^{(\eta)}_{l,R,j}(u) \\
&+ \sum_{l=1}^{K} s_{l,k,U,j}(u)\delta_{l,U,j}(u)\overline{Z}^{(\eta)}_{l,U,j}(u) \\
&+ \sum_{l=1}^{K} s_{l,k,A,j}(u)\delta_{l,A,j}(u)\overline{Z}^{(\eta)}_{l,A,j}(u) \\
&+ \sum_{l=1}^{K} s_{l,k,O,j}(u)\delta_{l,O,j}(u)\overline{Z}^{(\eta)}_{l,O,j}(u)\ du \\
\int_0^t H_{k,R,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\ du :=& \sum_{l=1}^{K}\int_0^t -\delta_{k,R,j}(u)\overline{Z}^{(\eta)}_{k,R,j}(u) + s_{l,k,L,j}(u)r_{k,L,j,j}(u) \\
&\times \theta_{l,j}(u)\left(\overline{Z}^{(\eta)}_{l,Q,j}(u) - C_{l,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)^+ du \\
\int_0^t H_{k,U,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\ du :=& \sum_{l=1}^{K}\int_0^t -\delta_{k,U,j}(u)\overline{Z}^{(\eta)}_{k,U,j}(u) + s_{l,k,S,j}(u)r_{k,S,j,j}(u) \\
&\times \mu_{l,j}(u)\min\left(\overline{Z}^{(\eta)}_{l,Q,j}(u), C_{l,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)\ du \\
\int_0^t H_{k,A,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\ du :=& \sum_{i=1;i\neq j}^{J}\sum_{l=1}^{K}\int_0^t -\delta_{k,A,j}(u)\overline{Z}_{k,A,j}(u) + s_{l,k,L,i}(u)r_{k,L,i,j}(u) \\
&\times \theta_{l,i}(u)\left(\overline{Z}^{(\eta)}_{l,Q,i}(u) - C_{l,i}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)^+ du \\
\int_0^t H_{k,O,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\ du :=& \sum_{i=1;i\neq j}^{J}\sum_{l=1}^{K}\int_0^t -\delta_{k,O,j}(u)\overline{Z}_{k,O,j}(u) + s_{l,k,S,i}(u)r_{k,S,i,j}(u) \\
&\times \mu_{l,i}(u)\min\left(\overline{Z}^{(\eta)}_{l,Q,i}(u), C_{l,i}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)\ du
\end{aligned}$$

$$\int_0^t H_{k,L,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\, du := \sum_{l=1}^{K}\int_0^t s_{l,k,L,j}(u) r_{k,L,j,J+1}(u) \times \theta_{l,i}(u)\left(\overline{Z}^{(\eta)}_{l,Q,i}(u) - C_{l,i}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)^{+}\, du$$

$$\int_0^t H_{k,D,j}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\, du := \sum_{l=1}^{K}\int_0^t s_{l,k,S,j}(u) r_{k,S,j,J+1}(u) \times \mu_{l,i}(u)\min\left(\overline{Z}^{(\eta)}_{l,Q,i}(u), C_{l,i}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)\, du$$

As shown in [105], from Lemma 5.5.2, we know that the sequence $\left\{\overline{\boldsymbol{Z}}^{(\eta)}(t)\right\}_{\eta=1}^{\infty}$ *is relatively compact with continuous limits. Thus, from the subsequence* $\left\{\overline{\boldsymbol{Z}}^{(\eta_\kappa)}(t)\right\}_{\kappa=1}^{\infty}$ *another subsequence* $\left\{\overline{\boldsymbol{Z}}^{(\eta_{\kappa_\iota})}(t)\right\}_{\iota=1}^{\infty}$ *can be extracted that converges weakly in* $D\Big([0,\infty), \mathbb{R}^{7KJ}\Big)$*, to a continuous process* $\boldsymbol{z}^*(t)$ *- the* ***particular limit*** *of the original sequence* $\left\{\overline{\boldsymbol{Z}}^{(\eta)}(t)\right\}_{\eta=1}^{\infty}$*.*

I now want to show that this gives the unique fluid limit for equations the scaled fluid process. To do this, consider an arbitrary particular limit $\boldsymbol{z}^*(t)$ *for a subsequence* $\left\{\overline{\boldsymbol{Z}}^{(\eta_\kappa)}(t)\right\}_{\kappa=1}^{\infty}$*. By the arbitrariness of* $\boldsymbol{z}^*(t)$*, I will show that this is a unique solution.*

The proof follows that of [105]. Again working in an arbitrary interval, say $[0,T]$ *re-write the system in the form detailed above and rearrange such that, for* $t \in [0,T]$*:*

$$\overline{\boldsymbol{Z}}^{(\eta_\kappa)}(t) - \overline{\boldsymbol{Z}}^{(\eta_\kappa)}(0) - \int_0^t \boldsymbol{H}\left(\overline{\boldsymbol{Z}}^{(\eta_\kappa)}(u)\right) du = \boldsymbol{G}^{(\eta_\kappa)}\left(\overline{\boldsymbol{Z}}^{(\eta_\kappa)}(t)\right)$$

Since $\overline{\boldsymbol{Z}}^{(\eta_\kappa)}(t) \xrightarrow{d} \boldsymbol{z}^*(t)$ *as* $\kappa \to \infty$ *and the limit* $\boldsymbol{z}^*(t)$ *is continuous,*

$$\overline{\boldsymbol{Z}}^{(\eta_\kappa)}(t) - \overline{\boldsymbol{Z}}^{(\eta_\kappa)}(0) - \int_0^t \boldsymbol{H}\left(\overline{\boldsymbol{Z}}^{(\eta_\kappa)}(u)\right) du \xrightarrow{d} \boldsymbol{z}^*(t) - \boldsymbol{z}^*(0) - \int_0^t \boldsymbol{H}(\boldsymbol{z}^*(u))\, du$$

Furthermore, $\boldsymbol{G}^{(\eta_\kappa)}\left(\overline{\boldsymbol{Z}}^{(\eta_\kappa)}(t)\right) \xrightarrow{d} 0$ *since - by (5.36)-(5.42), and the Random time change Theorem in [121] - each term in* $G^{(\eta_\kappa)}_{k,Q,j}(t), G^{(\eta_\kappa)}_{k,R,j}(t), G^{(\eta_\kappa)}_{k,U,j}(t), G^{(\eta_\kappa)}_{k,A,j}(t), G^{(\eta_\kappa)}_{k,O,j}(t)$*,* $G^{(\eta_\kappa)}_{k,L,j}(t)$ *and* $G^{(\eta_\kappa)}_{k,D,j}(t)$ *converge to 0, for all* $k \in Out$ *and* $j \in Ser$*.*

Thus:

$$\overline{\boldsymbol{Z}}^{(\eta_\kappa)}(T) - \overline{\boldsymbol{Z}}^{(\eta_\kappa)}(0) - \int_0^T \boldsymbol{H}\left(\overline{\boldsymbol{Z}}^{(\eta_\kappa)}(u)\right) du \xrightarrow{d} 0$$

It then follows that the particular limit $\boldsymbol{z}^*(t)$ *a.s. satisfies equations (5.13)-(5.19). Also, as noted in [105], the mapping* $\boldsymbol{H}$ *is Lipschitz continuous and by Lemma 1 in [122], equations (5.13)-(5.19) have a unique solution. Therefore, all particular fluid limits are the same, they are the unique solution to (5.13)-(5.19).* ■

Appendix C

Chapter 6

C.1 Code for discrete event simulation of stochastic system

```
function [Z, C] = (c, arrive, serve, leave, dr, df, p, q, dt,T, A, IC, type)

% Discrete Event Simulation of queue with rejoin, reuse and multiple health states
%    exprnd = probabilty distribution of inter-arrival times for pat, service time, time until
     rejoin, and time until reuse

% For each time interval i, such that t_(i-1) <= t < t_(i), and patient group k:
% Inputs:
%  c(i)              = number of available servers
%  arrive(k,i)       = mean arrival rate
%  serve(k,i)        = mean service rate
%  leave(k,i)        = mean abandonment rate
%  dr(k,i), df(k,i) = mean rates of rejoin/reuse
%  p(k,i), q(k,i)    = probability a patient rejoins after abandon/reuses after service
%  dt                = time intervals for output
%  T                 = total simulation time
%  A(:,:,:,j):       = array containing matrices of health state transition, where: j = 1
     post-rejoin; j = 2 post-reuse; j = 3 post-abandon; j = 4 post-service.
%  IC                = initial conditions for the system
%  type              = string indicating which server allocation to use
```

```
% Outputs:
%  ZQ(k),ZR(k),ZF(k) = number in service/queue; in rejoin; in reuse orbits
%  loss(k)           = number of pat lost due to abandonment
%  discharge(k)      = number of pat discharged after service
%  S(k),Q(k)         = number of pat in service/queue

% Set up for simulation
% Number of health states
N = size(A, 1);

% time - vector of time intervals; set initial time t
time = (0:dt:T); t = time(1); i = 1;

% Initital conditions
ZQ = IC(:,1); ZR = IC(:,2); ZF = IC(:,3); loss = zeros(N, 1); discharge = zeros(N, 1);

% S,Q - Variables for tracking number of patients in service/queue
S = zeros(N, 1); Q = zeros(N, 1);

% Queue - variable for storing number in queue at each time step for calculating VWT
Queue = zeros(N,length(t));

% Define Z to store system state information
% Z(:, 1, :) - Number in queue/service
% Z(:, 2, :) - Number in rejoin
% Z(:, 3, :) - Number in reuse
% Z(:, 4, :) - Number discharged
% Z(:, 5, :) - Number lost
% Z(:, 6, :) - Time interval
% Z(:, 7, :) - VWT
Z = zeros(N, 7, length(time));

% C - variable for assigning servers across each outcome state
C = zeros(N, length(time)); i = 1;

% Assign servers according to type of server allocation
% Even split over each queue
if strcmp(type,'even')
  C(:,i) = round(c(i) ./ N);
```

```
  % Algorithm for assigning servers if not divisible into integers
  while sum(C(:,i)) > c(i)
    k = randi([1 N],1);
    C(k,i) = C(k,i) - 1;
  end
  while sum(C(:,i)) < c(i)
    k = randi([1 N],1);
    C(k,i) = C(k,i) + 1;
  end

% Continuous allocation of servers based on proportion of patient in queue and service for each
    health state
elseif strcmp(type,'continuous')
  C(:,i) = round(ZQ(:) .* c(i) ./ sum(ZQ(:)));

  % Algorithm for assigning servers if not divisible into integers
  while sum(C(:,i)) > c(i)
    k = randi([1 N],1);
    C(k,i) = C(k,i) - 1;
  end
  while sum(C(:,i)) < c(i)
    k = randi([1 N],1);
    C(k,i) = C(k,i) + 1;
  end

% Continuous allocation of servers based on proportion of patient in queue and service for each
    health state, weighted by service time
elseif strcmp(type,'byserve')
  C(:,i) = round(ZQ(:) .* c(i) .* serve(:,i) ./ sum(ZQ(:) .* serve(:,i)));

  % Algorithm for assigning servers if not divisible into integers
  while sum(C(:,i)) > c(i)
    k = randi([1 N],1);
    C(k,i) = C(k,i) - 1;
  end
  while sum(C(:,i)) < c(i)
    k = randi([1 N],1);
    C(k,i) = C(k,i) + 1;
  end
end
```

```
% Begin simulation
% Create patient matrix for storing information
pat = inf(1,5);

% Populate patient matrix with all possible new arrivals in time frame [t_0,T]
for j = 1:N

  % Set service and queue counts according to IC
  S(j) = min(IC(j,1),C(j,1)); Q(j) = max(IC(j,1) - C(j,1),0);

  % Use interarrival rate of patients in each health state, j, for time t_0
  % i.e. exprnd(1./arrive(j,1))
  pat(i,:) = [t + exprnd(1./arrive(j,1)), 0, j, inf, inf];
  i = i + 1;
  while pat(i-1,1) < max(time)
    P = find(time > pat(i-1,1),1,'first');
    pat(i,:) = [pat(i-1,1) + exprnd(1./arrive(j,P)), 0, j, inf, inf];
    i = i + 1;
  end

  % If the system starts with more patients in ZQ(j) than servers C(j,1)
  if IC(j,1) > C(j,1)
    k = C(j,1);

    % Populate all servers with patients
    while k >= 1
      pat(i,:) = [exprnd(1./serve(j,1)), 3, j, inf, inf];
      i = i + 1; k = k - 1;
    end

    % Place remaining patients in the queue
    k = IC(j,1) - C(j,2);
    while k >= 1
      pat(i,:) = [exprnd(1./leave(j,1)), 4, j, k, inf];
      i = i + 1; k = k - 1;
    end

  % If the system starts with fewer patients in ZQ(j) than servers C(j,1)
  else
    k = IC(j,1);
```

```
    % Place all patients into service
    while k >= 1
      pat(i,:) = [exprnd(1./serve(j,1)), 3, j, inf, inf];
      i = i + 1; k = k -1;
    end
  end

  % If the system starts with patients in ZR(j) assign a rejoin time
  if IC(j,2) > 0
    k = IC(j,);
    while k >= 1
      pat(i,:) = [exprnd(1./dr(j,1)), 1, j, inf, inf];
      i = i + 1; k = k - 1;
    end
  end

  % If the system starts with patients in ZF(j) assign a reuse time
  if IC(j,3) > 0
    k = IC(j,3);
    while k >= 1
      pat(i,:) = [exprnd(1./df(j,1)), 2, j, inf, inf];
      i = i + 1; k = k - 1;
    end
  end
end

% PAT - variable for storing pat information at each time step for calculating VWT
PAT(:,:,:) = zeros(size(pat,1),5,length(t));

% Loop through time steps - updating variables accordingly
%   Time intervals [t_i-1, t_i]
for i = 2:(length(time))

  % Assign capacity in next time interval
  C(:,i) = C(:,i-1);

  % Update capacity allocation if type = 'even'
  if strcmp(type,'even')
    C(:,i) = round(c(i) ./ N);
```

```
  % Algorithm for assigning servers if not divisible into integers
  while sum(C(:,i)) > c(i)
    k = randi([1 N],1);
    C(k,i) = C(k,i) - 1;
  end
  while sum(C(:,i)) < c(i)
    k = randi([1 N],1);
    C(k,i) = C(k,i) + 1;
  end
end

t = min(pat(:,1));
next = find(pat(:,1) == t,1,'first');

% Update parameters when time interval is crossed
while t < time(i)

  % Event: new patient joins service
  if pat(next, 2) ==  0 && S(pat(next, 3)) < C(pat(next, 3),i)
    % Time to complete service
    pat(next,1) = t + exprnd(1./serve(pat(next,3),i));

    % 3: In service
    pat(next,2) = 3;

    % Update number in queue/service and number in service
    ZQ(pat(next, 3)) = ZQ(pat(next, 3)) + 1;  S(pat(next, 3))  = S(pat(next, 3))  + 1;

  % Event: new patient joins queue
  elseif pat(next, 2) ==  0 && S(pat(next, 3)) >= C(pat(next, 3),i)
    % Time until abandonment
    pat(next,1) = t + exprnd(1./leave(pat(next,3),i));

    % 4: In queue
    pat(next,2) = 4;

    % Position in queue
    pat(next,4) = Q(pat(next, 3)) + 1;

    % Update number in queue/service and number in queue
    ZQ(pat(next, 3)) = ZQ(pat(next, 3)) + 1;  Q(pat(next, 3)) = Q(pat(next, 3)) + 1;
```

```
% Event: rejoin patient joins service
elseif pat(next, 2) ==  1 && S(pat(next, 3)) < C(pat(next, 3),i)
  % Update number in rejoin orbit
  ZR(pat(next, 3)) = ZR(pat(next, 3)) - 1;

  % Assign new health state
  pat(next,3) = find(mnrnd(1,A(pat(next, 3),:,1)) == 1);

  % Time until complete service
  pat(next,1) = t + exprnd(1./serve(pat(next,3),i));

  % 3: In service
  pat(next,2) = 3;

  % Update number in queue/service and number in service
  ZQ(pat(next, 3)) = ZQ(pat(next, 3)) + 1;  S(pat(next, 3)) = S(pat(next, 3)) + 1;

% Event: rejoin patient joins queue
elseif pat(next, 2) ==  1 && S(pat(next, 3)) >= C(pat(next, 3),i)
  % Update number in rejoin orbit
  ZR(pat(next, 3)) = ZR(pat(next, 3)) - 1;

  % Assign new health state
  pat(next,3) = find(mnrnd(1,A(pat(next, 3),:,1)) == 1);

  % Time until abandonment
  pat(next,1) = t + exprnd(1./leave(pat(next,3),i));

  % 4: In queue
  pat(next,2) = 4;

  % Position in queue
  pat(next,4) = Q(pat(next, 3)) + 1;

  % Update number in queue/service and number in queue
  ZQ(pat(next, 3)) = ZQ(pat(next, 3)) + 1;  Q(pat(next, 3)) = Q(pat(next, 3)) + 1;

% Event: reuse patient joins service
elseif pat(next, 2) ==  2 && S(pat(next, 3)) < C(pat(next, 3),i)
  % Update number in reuse
  ZF(pat(next, 3)) = ZF(pat(next, 3)) - 1;
```

```
  % Assign new health state
  pat(next,3) = find(mnrnd(1,A(pat(next, 3),:,2)) == 1);

  % Time until complete service
  pat(next,1) = t + exprnd(1./serve(pat(next,3),i));

  % 3: In service
  pat(next,2) = 3;

  % Update number in queue/service and number in service
  ZQ(pat(next, 3)) = ZQ(pat(next, 3)) + 1;  S(pat(next, 3)) = S(pat(next, 3)) + 1;

% Event: reuse patient joins queue
elseif pat(next, 2) ==  2 && S(pat(next, 3)) >= C(pat(next, 3),i)
  % Update number in reuse orbit
  ZF(pat(next, 3)) = ZF(pat(next, 3)) - 1;

  % Assign new health state
  pat(next,3) = find(mnrnd(1,A(pat(next, 3),:,2)) == 1);

  % Time until abandonment
  pat(next,1) = t + exprnd(1./leave(pat(next,3),i));

  % 4: In queue
  pat(next,2) = 4;

  % Position in queue
  pat(next,4) = Q(pat(next, 3)) + 1;

  % Update number in queue/service and number in queue
  ZQ(pat(next, 3)) = ZQ(pat(next, 3)) + 1;  Q(pat(next, 3)) = Q(pat(next, 3)) + 1;

% Event: patient completes service
elseif pat(next, 2) ==  3
  % Update number in queue/service and number in service
  ZQ(pat(next, 3)) = ZQ(pat(next, 3)) - 1;  S(pat(next, 3))  = S(pat(next, 3))  - 1;

  % If there are free servers and patients waiting to resume
  if isempty(pat(pat(:,3) == pat(next, 3) & pat(:,2) == 5)) == 0 && S(pat(next, 3)) <
 C(pat(next, 3),i)
```

```
    % Select first patient in resume queue
    row = find(pat(:,3) == pat(next, 3) & pat(:,2) == 5, 1,'first');

    % Time completed service
    pat(row,1)  = t + pat(row,5);

    % 3: In service
    pat(row,2)  = 3;

    % inf: not in resume
    pat(row,5)  = inf;

    % Update number in service
    S(pat(next, 3)) = S(pat(next, 3)) + 1;

  % If there are pat in the queue and servers are available
  elseif Q(pat(next, 3)) > 0
    % Select first patient in queue
    row = find(pat(:,2) == 4 & pat(:,3) == pat(next, 3) & pat(:,4) == 1);

    % Time completed service
    pat(row,1) = t + exprnd(1./serve(pat(next, 3),i));

    % 3: In service
    pat(row,2) = 3;

    % inf: not in queue
    pat(row,4) = inf;

    % Adjust positions of patients in queue
    pat(pat(:,2) == 4 & pat(:,3) == pat(next, 3) & pat(:,4) > 1,4) = pat(pat(:,2) == 4 &
    pat(:,3) == pat(next, 3) & pat(:,4) > 1,4) - 1;

    % Update number in queue and number in service
    Q(pat(next, 3)) = Q(pat(next, 3)) - 1;  S(pat(next, 3)) = S(pat(next, 3)) + 1;
  end

  % Assign new health state
  pat(next,3) = find(mnrnd(1,A(pat(next, 3),:,3)) == 1);

  % Bernoulli trial - patient seek to reuse service; 2: yes
```

```
    pat(next,2) = binornd(1,q(pat(next,3),i)) * 2;

    if pat(next,2) == 2
      % Time until reenter queue
      pat(next,1) = t + exprnd(1./df(pat(next,3),i));

      % Update number in reuse orbit
      ZF(pat(next,3)) = ZF(pat(next,3)) + 1;

    else
      % Remove patient from system
      pat(next,1) = inf;  pat(next,2) = inf;

      % Update number discharged
      discharge(pat(next,3)) = discharge(pat(next,3)) + 1;
    end

  % Event: Patient abandons queue
  elseif pat(next, 2) == 4
    % Update number in queue/service and number in queue
    ZQ(pat(next, 3)) = ZQ(pat(next, 3)) - 1;  Q(pat(next, 3)) = Q(pat(next, 3)) - 1;

    % Adjust all queue positions of patients behind abandoning patient
    pat(pat(:,2) == 4 & pat(:,3) == pat(next, 3) & pat(:,4) > pat(next,4),4) = pat(pat(:,2) == 4 &
    pat(:,3) == pat(next, 3) & pat(:,4) > pat(next,4),4) - 1;

    % Assign new health state
    pat(next,3) = find(mnrnd(1,A(pat(next, 3),:,4)) == 1);

    % Bernoulli trial - patient rejoin queue; 1: yes
    pat(next,2) = binornd(1,p(pat(next,3),i));
    pat(next,4) = inf;

    if pat(next,2) == 1
      % Time until reenter queue
      pat(next,1) = t + exprnd(1./dr(pat(next,3),i));

      % Update number in rejoin orbit
      ZR(pat(next,3)) = ZR(pat(next,3)) + 1;
```

```
    else
      % Remove patient from system
      pat(next,1) = inf;  pat(next,2) = inf;

      % Update number lost
      loss(pat(next,3)) = loss(pat(next,3)) + 1;
    end
  end

% Continuous allocation of servers based on proportion of patient in queue and service for each
 health state
if strcmp(type,'continuous')
  % Algorithm for assigning servers if not divisible into integers
  C(:,i) = round(ZQ(:) .* c(i) ./ sum(ZQ(:)));
  while sum(C(:,i)) > c(i)
    k = randi([1 N],1);
    C(k,i) = C(k,i) - 1;
  end
  while sum(C(:,i)) < c(i)
    k = randi([1 N],1);
    C(k,i) = C(k,i) + 1;
  end

% Continuous allocation of servers based on proportion of patient in queue and service for each
 health state, weighted by service time
elseif strcmp(type,'byserve')
  C(:,i) = round(ZQ(:) .* c(i) .* serve(:,i) ./...
  sum(ZQ(:) .* serve(:,i)));
  while sum(C(:,i)) > c(i)
    k = randi([1 N],1);
    C(k,i) = C(k,i) - 1;
  end
  while sum(C(:,i)) < c(i)
    k = randi([1 N],1);
    C(k,i) = C(k,i) + 1;
  end
end

% Preemptive resumption: If the number of servers falls below the number in service
if sum(S(:) > C(:,i)) > 0
  bigger = find(S(:) > C(:,i));
```

```
    for l = 1:length(bigger)
      % Set health state indicator
      k = bigger(l);

      % Find pat who are in service and in health state k
      row = find(pat(:,2) == 3 & pat(:,3) == k);

      % Remove (S(k) - C(k,i)) pat from service
      % Calculate remaining time in service
      pat(row(C(k,i) + 1:end), 5) = pat(row(C(k,i) + 1:end), 1) - t;
      pat(row(C(k,i) + 1:end), 1) = inf;
      pat(row(C(k,i) + 1:end), 2) = 5;

      % Update number in service
      S(k) = C(k,i);
    end
  end

  % Check: are servers available and patients waiting to resume service?
  % Loop for outcome groups in system
  for k = 1:N
    % If there are pat in infinite buffer resume space and servers are available
    while isempty(pat(pat(:,3) == k & pat(:,2) == 5)) == 0 && S(k) < C(k,i)
      % Select first patient in resume queue
      row = find(pat(:,3) == k & pat(:,2) == 5, 1,'first');
      pat(row,1)  = t + pat(row,5);
      pat(row,2)  = 3;
      pat(row,5)  = inf;

      % Update number in service
      S(k) = S(k) + 1;
    end

    % If there are pat in the queue and servers are available
    while Q(k) > 0 && S(k) < C(k,i)
      % Select first patient in queue
      row = find(pat(:,2) == 4 & pat(:,3) == k & pat(:,4) == 1, 1,'first');
      pat(row,1) = t + exprnd(1./serve(k,i));
      pat(row,2) = 3;
      pat(row,4) = inf;
```

```
        % Adjust all patient queue places
        pat(pat(:,2) == 4 & pat(:,3) == k & pat(:,4) > 1,4) = pat(pat(:,2) == 4 & pat(:,3) == k
        & pat(:,4) > 1,4) - 1;

        % Update number in queue and number in service
        Q(k) = Q(k) - 1; S(k) = S(k) + 1;
      end
    end

    % Update time to correspond with next event
    t = min(pat(:,1));

    % Identify patient corresponding to next event
    next = find(pat(:,1) == t, 1, 'first');
  end

  % Update output vector and variables for calculating VWT (PAT, Queue)
  Z(:,:,i) = [ZQ(:), ZR(:), ZF(:), loss(:), discharge(:), [time(i);time(i)], [0;0]];
  PAT(:,:,i) = pat(:,:);
  Queue(:,i) = Q;
end

% Calculate the VWT for each time interval and health state
for i = 1:length(time)
  for l = 1:N
    if Queue(l,i) > 0 && C(l,i) > 0
      VWT = SimVWT((0:dt:4*T), i, PAT(PAT(:,3,i) == l,:,i),...
      C(l,i), Queue(l,i), [C(l,:),C(l,end) * ones(1, 3 * length(time))],...
      [serve(l,:),serve(l,end) * ones(1, 3 * length(time))]);
      Z(l,7,i) = VWT - time(i);
    end
  end
end
```

C.2 Code for discrete event simulation of virtual waiting time

```
function [VWT] = SimVWT(time, j, pat, S, Q, C, serve)
% Function for calculating the simulated VWT
% Inputs:
%   time  - total modelled time frame
%   j     - time index indicating point in modelled time VWT is found for
%   pat   - matrix of patient information for j-th time interval
%   S     - number of patients in service in j-th time interval
%   Q     - number of patients in queue in j-th time interval
%   C     - number of servers available in j-th time interval
%   serve - vector of service rate for the modelled time period

% Outputs:
%   VWT - waiting time from simulated system

% Select patient arriving closest to time(j)
hold = find(pat(:,1) >= time(j) & pat(:,2) == 0,1,'first');

% Ensure that this patient cannot abandon the queue, set queue length
pat(hold,1) = inf;
pat(hold,2) = 4;
pat(hold,4) = Q + 1; Q = Q + 1;

% Remove all patients who arrive after time(j)
pat(pat(:,1) > time(j) & pat(:,2) == 0,:) = inf;
pat(pat(:,2) == 4 & pat(:,4) > Q,1) = inf;
pat(pat(:,2) == 1,:) = inf;
pat(pat(:,2) == 2,:) = inf;

% Update pat variable to include only those in the queue or service
pat = pat(pat(:,1) < inf | (pat(:,4) > 0 & pat(:,4) < inf),:);

% Loop through time steps - updating variables accordingly
% Time intervals [t_i-1, t_i]
for i = j:(length(time))
```

```
% Preemptive resumption: If the number of servers falls below the number in service
if S > C(i)

  % Find all patients in service
  row = find(pat(:,2) == 3);

  % Remove (S - C) pat from service
  % Calculate remaining time in service
  pat(row(C(i) + 1:end), 5) = pat(row(C(i) + 1:end), 1) - t;
  pat(row(C(i) + 1:end), 1) = inf;
  pat(row(C(i) + 1:end), 2) = 5;

  % Update number in service
  S = C(i);
end

% Check: are servers available and patients waiting to resume service?
% If there are pat in infinite buffer resume space and
% servers are available, place them in service
while isempty(pat(pat(:,2) == 5)) == 0 && S < C(i)

  % Select first patient in resume queue place in service
  row = find(pat(:,2) == 5, 1,'first');

  pat(row,1)  = t + pat(row,5);
  pat(row,2)  = 3;
  pat(row,5)  = inf;

  % Update number in service
  S = S + 1;
end

% If there are pat in the queue and servers are available
while Q > 0 && S < C(i)

  % Select first patient in queue
  row = find(pat(:,2) == 4 & pat(:,4) == 1, 1,'first');

  pat(row,1) = t + exprnd(1./serve(i));
  pat(row,2) = 3;
  pat(row,4) = inf;
```

```
  % Adjust all patient queue places
  pat(pat(:,2) == 4 & pat(:,4) > 1,4) = pat(pat(:,2) == 4 & pat(:,4) > 1,4) - 1;

  % Update number in queue and number in service
  Q = Q - 1; S = S + 1;
end

% If queue is empty, output VWT
if Q == 0
  VWT = t;
  break;
end

% Update time to correspond with next event
t = min(pat(:,1));

% Identify patient corresponding to next event
next = find(pat(:,1) == t, 1, 'first');

while t < time(i) && Q > 0
  % Event: patient completes service
  if pat(next, 2) ==  3

    % Update number in service
    S = S - 1;

    % If there are free servers and patients waiting to resume
    if isempty(pat(pat(:,2) == 5)) == 0 && S < C(i)

      % Select first patient in resume queue
      row = find(pat(:,2) == 5, 1,'first');

      % Time completed service
      pat(row,1)  = t + pat(row,5);

      % 3: In service
      pat(row,2)  = 3;

      % inf: not in resume
      pat(row,5)  = inf;
```

```
    % Update number in service
    S = S + 1;

  % If there are pat in the queue and servers are available
  elseif Q > 0

    % Select first patient in queue
    row = find(pat(:,2) == 4 & pat(:,4) == 1);

    % Time completed service
    pat(row,1) = t + exprnd(1./serve(i));

    % 3: In service
    pat(row,2) = 3;

    % inf: not in queue
    pat(row,4) = inf;

    % Adjust positions of patients in queue
    pat(pat(:,2) == 4 & pat(:,4) > 1,4) = pat(pat(:,2) == 4 & pat(:,4) > 1,4) - 1;

    % Update number in queue and number in service
    Q = Q - 1; S = S + 1;
  end

  % Remove patient from system
  pat(next,:) = inf;

% Event: Patient abandons queue
elseif pat(next, 2) == 4
  Q  = Q - 1;

  % Adjust all queue positions of patients behind abandoning patient
  pat(pat(:,2) == 4 & pat(:,4) > pat(next,4),4) = pat(pat(:,2) == 4 & pat(:,4) >
  pat(next,4),4) - 1;

  % Remove patient from the system
  pat(next,1) = inf; pat(next,2) = inf; pat(next,4) = inf;
end
```

```
    % If queue is empty, output VWT
    if Q == 0
      VWT = t;
      break;
    end

    % Update time to correspond with next event
    t = min(pat(:,1));

    % Identify patient corresponding to next event
    next = find(pat(:,1) == t, 1, 'first');
  end
end
end
```

C.3 Code for the fluid and diffusion approximation

```
function [Z, ZP, ZM, C, V, dcdt] = FluidComp(c, arrive, serve, leave, dr, df, p, q, dt, T, S, IC,
    type)
% Function for calculating fluid approximation of a stochastic queueing system
% Inputs:
%   c       - number of servers
%   arrive  - rate of new arrival
%   serve   - rate of service
%   leave   - rate of abandonment
%   dr      - rate of rejoin
%   df      - rate of reuse
%   p       - probability of rejoin
%   q       - probability of reuse
%   dt      - time step
%   T       - total length of time
%   S       - matrices of health state transition
%   IC      - initial conditions
%   type    - sever allocation used
```

```
% Outputs:
%   Z       - number of patients in each state
%   ZP      - upper variance envelope of Z
%   ZM      - lower variance envelope of Z
%   C       - capacity allocation from model
%   V       - variance of Z
%   dcdt    - vector of derivative for queue and service states (for VWT)

%Using trapezoidal rule  in MatLab, solving iteratively:
% Calculate number of intervals for iteration
NT = T/dt;

% A - number of outcome states
A = size(S,1);

% VAR - variance matrix; ; V - variance output;
VAR = zeros(5*A, 5*A, NT + 1);  V = zeros(A,5,NT+1);

% Set up error vector  used to stop iteration -
err = 1 / 1000000;

% Z - number of patients in each state; ZP/ZM - variance envelopes for Z
Z(:,:,:) = zeros(A,5,NT+1);  ZP  = zeros(A,5,NT + 1);  ZM  = zeros(A,5,NT + 1);

% dcdt - matrix of derivatives
dcdt = zeros(A,NT+1);

% Y - variable for iteration
Y(:,:,:)  = zeros(A,5,NT+1);

% C - capacity allocation variable
C(:,:)    = zeros(A,NT+1);

% alpha - functions for producing diffusion equations
alpha     = zeros (A, A, NT+1, 6);

% AT, B, BT - matrices for solving diffusion equations
AT  = zeros(5*A, 5*A, NT + 1);  B   = zeros(5*A, 5*A, NT + 1, 9);  BT  = zeros(5*A, 5*A, NT + 1);

% Set initial condition
```

```
Z(:,1:3,1) = IC; ZP(:,1:3,1) = IC; ZM(:,1:3,1) = IC;

if strcmp(type,'even')
  C(:,1) = round(c(1) ./ A);
end
% System starts empty  Z(a,i) where a is the patient group, i is the time step
if sum(Z(:,1,1)) == 0
  error('Must have non-zero initial conditions for the queue and service.');
else
  if strcmp(type,'continuous')
    C(:,1) = c(1) .* Z(:,1,1) ./ sum(Z(:,1,1));
  elseif strcmp(type,'byserve')
    C(:,1) = c(1) .* Z(:,1,1) .* serve(:,1) ./ sum(Z(:,1,1) .* serve(:,1));
  end
end

% Calculate fluid approximation over multiple time steps
for i = 2:NT+1
  % Y is a holding variable
  Y(:,:,i) = Z(:,:,i-1);

  % Allocation of servers based on previous number of patients in queue, or expected arrivals, or
     equal allocation to each patient
  if strcmp(type,'continuous')
    C(:,i) = c(i) .* Z(:,1,i) ./ sum(Z(:,1,i));
  elseif strcmp(type,'byserve')
    C(:,i) = c(i) .* Z(:,1,i) .* serve(:,i) ./ sum(Z(:,1,i) .* serve(:,i));
  elseif strcmp(type,'even')
    C(:,i) = round(c(i) ./ A);
  end

  ZR = S(:,:,1)' * (dr(:,i) .* (Z(:,2,i-1) + Y(:,2,i))) * dt/2;
  ZF = S(:,:,2)' * (df(:,i) .* (Z(:,3,i-1) + Y(:,3,i))) * dt/2;
  ZS = q(:,i) .* S(:,:,3)' * (serve(:,i) .* (min(Z(:,1,i-1),C(:,i-1)) + min(Y(:,1,i),C(:,i)))) *
     dt/2;
  ZA = p(:,i) .* S(:,:,4)' * (leave(:,i) .* (max(Z(:,1,i-1) - C(:,i-1),0) + max(Y(:,1,i) -
     C(:,i),0))) * dt/2;
  ZD = (1-q(:,i)) .* S(:,:,3)' * (serve(:,i) .* (min(Z(:,1,i-1),C(:,i-1)) + min(Y(:,1,i),C(:,i)))) *
     dt/2;
  ZL = (1-p(:,i)) .* S(:,:,4)' * (leave(:,i) .* (max(Z(:,1,i-1) - C(:,i-1),0) + max(Y(:,1,i) -
     C(:,i),0))) * dt/2;
```

```
Z(:,1,i) = Z(:,1,i-1) + arrive(:,i) * dt + ZF + ZR - serve(:,i) .* (min(Z(:,1,i-1), C(:,i-1)) +
   min(Y(:,1,i), C(:,i))) * dt/2 - leave(:,i) .* (max(Z(:,1,i-1) - C(:,i-1), 0) + max(Y(:,1,i) -
   C(:,i), 0)) * dt/2;
Z(:,2,i) = Z(:,2,i-1) - dr(:,i) .* (Y(:,2,i) + Z(:,2,i-1)) * dt/2 + ZA;
Z(:,3,i) = Z(:,3,i-1) - df(:,i) .* (Y(:,3,i) + Z(:,3,i-1)) * dt/2 + ZS;
Z(:,4,i) = Z(:,4,i-1) + ZL;
Z(:,5,i) = Z(:,5,i-1) + ZD;

% Continue iteration until negligible benefit in continuing
while sum(sum(Z(:,:,i) - Y(:,:,i) > err)) > 0
  % Y is a holding variable
  Y(:,:,i) = Z(:,:,i);

  % Allocation of servers
  if strcmp(type,'continuous')
    C(:,i) = c(i) .* Z(:,1,i) ./ sum(Z(:,1,i));
  elseif strcmp(type,'byserve')
    C(:,i) = c(i) .* Z(:,1,i) .* serve(:,i) ./ sum(Z(:,1,i) .* serve(:,i));
  elseif strcmp(type,'even')
    C(:,i) = round(c(i) ./ A);
  end

  ZR = S(:,:,1)' * (dr(:,i) .* (Z(:,2,i-1) + Y(:,2,i))) * dt/2;
  ZF = S(:,:,2)' * (df(:,i) .* (Z(:,3,i-1) + Y(:,3,i))) * dt/2;
  ZS = q(:,i)  .* S(:,:,3)' * (serve(:,i) .* (min(Z(:,1,i-1),C(:,i-1)) + min(Y(:,1,i),C(:,i)))) *
   dt/2;
  ZA = p(:,i)  .* S(:,:,4)' * (leave(:,i) .* (max(Z(:,1,i-1) - C(:,i-1),0) + max(Y(:,1,i) -
   C(:,i),0))) * dt/2;
  ZD = (1-q(:,i)) .* S(:,:,3)' * (serve(:,i) .* (min(Z(:,1,i-1),C(:,i-1)) + min(Y(:,1,i),C(:,i))))
   * dt/2;
  ZL = (1-p(:,i)) .* S(:,:,4)' * (leave(:,i) .* (max(Z(:,1,i-1) - C(:,i-1),0) + max(Y(:,1,i) -
   C(:,i),0))) * dt/2;

  Z(:,1,i) = Z(:,1,i-1) + arrive(:,i) * dt + ZF + ZR - serve(:,i) .* (min(Z(:,1,i-1),C(:,i-1))  +
   min(Y(:,1,i),C(:,i))) * dt/2 - leave(:,i) .* (max(Z(:,1,i-1) - C(:,i-1),0) + max(Y(:,1,i) -
   C(:,i),0)) * dt/2;
  Z(:,2,i) = Z(:,2,i-1) - dr(:,i) .* (Y(:,2,i) + Z(:,2,i-1)) * dt/2 + ZA;
  Z(:,3,i) = Z(:,3,i-1) - df(:,i) .* (Y(:,3,i) + Z(:,3,i-1)) * dt/2 + ZS;
  Z(:,4,i) = Z(:,4,i-1) + ZL;
  Z(:,5,i) = Z(:,5,i-1) + ZD;
```

```
end

dcdt(:,i) = arrive(:,i) + S(:,:,2)' * (df(:,i) .* Z(:,3,i)) + S(:,:,1)' * (dr(:,i) .* Z(:,2,i)) -
   serve(:,i) .* min(Z(:,1,i), C(:,i)) - leave(:,i) .* max(Z(:,1,i) - C(:,i), 0);

% Calculate variance of system
% Define values of alpha over each time interval
alpha(:,:,i,1) = dr(:,i)     .* (S(:,:,1) .* Z(:,2,i));
alpha(:,:,i,2) = df(:,i)     .* (S(:,:,2) .* Z(:,3,i));
alpha(:,:,i,3) = p(:,i)'     .* (S(:,:,4) .* leave(:,i) .* (max(Z(:,1,i) - C(:,i),0)));
alpha(:,:,i,4) = (1-p(:,i))' .* (S(:,:,4) .* leave(:,i) .* (max(Z(:,1,i) - C(:,i),0)));
alpha(:,:,i,5) = q(:,i)'     .* (S(:,:,3) .* serve(:,i) .* (min(Z(:,1,i), C(:,i))));
alpha(:,:,i,6) = (1-q(:,i))' .* (S(:,:,3) .* serve(:,i) .* (min(Z(:,1,i), C(:,i))));

% Set initial value for while loop
a  = 1;

% Build matrices for variance calculation
while a <= A
  b = 1; k = 5 * a - 4;
  while b <= A
    l = 5 * b - 4;
    if a == b
      if Z(a,1,i) <= C(a,i)
        AT(k,k,i)   = - serve(a,i);
        AT(k+2,k,i) = q(a,i) * serve(a,i) * S(a,a,3);
        AT(k+4,k,i) = (1-q(a,i)) * serve(a,i) * S(a,a,3) ;
      else
        AT(k,k,i)   = - leave(a,i) * (1-c(i).*(sum(Z(:,1,i))-Z(a,1,i))/sum(Z(:,1,i))^2) -
   serve(a,i) * c(i) .* ((sum(Z(:,1,i))-Z(a,1,i))/sum(Z(:,1,i))^2);
        AT(k+1,k,i) = p(a,i) * leave(a,i) * S(a,a,4) * (1 - c(i) .*
   (sum(Z(:,1,i))-Z(a,1,i))/sum(Z(:,1,i))^2) + p(a,i) .* (sum(leave(:,i) .* S(:,a,4) .* c(i) .*
   (Z(:,1,i)/sum(Z(:,1,i))^2)) - leave(a,i) * S(a,a,4) * c(i) .* (Z(a,1,i)/sum(Z(:,1,i))^2));
        AT(k+2,k,i) = q(a,i) * serve(a,i) * S(a,a,3) * (c(i) .*
   (sum(Z(:,1,i))-Z(a,1,i))/sum(Z(:,1,i))^2) + q(a,i) .* (sum(serve(:,i) .* S(:,a,3) .* c(i) .*
   (-Z(:,1,i)/sum(Z(:,1,i))^2)) - serve(a,i) * S(a,a,3) * c(i) .* (-Z(a,1,i)/sum(Z(:,1,i))^2));
        AT(k+3,k,i) = (1-p(a,i)) * leave(a,i) * S(a,a,4) * (1 - c(i) .*
   (sum(Z(:,1,i))-Z(a,1,i))/sum(Z(:,1,i))^2) + (1-p(a,i)) * (sum(leave(:,i) .* S(:,a,4) .* c(i)
   .*(Z(:,1,i)/sum(Z(:,1,i))^2)) - leave(a,i) * S(a,a,4) * c(i) .* (Z(a,1,i)/sum(Z(:,1,i))^2));
        AT(k+4,k,i) = (1-q(a,i)) * serve(a,i) * S(a,a,3) * c(i) .*
   ((sum(Z(:,1,i))-Z(a,1,i))/sum(Z(:,1,i))^2) + (1-q(a,i)) * (sum(serve(:,i) .* S(:,a,3) .* c(i)
```

```
.*(-Z(:,1,i)/sum(Z(:,1,i))^2)) - serve(a,i) * S(a,a,3) * c(i) .* (-Z(a,1,i)/sum(Z(:,1,i))^2));
   end

   AT(k,k+1,i) = S(a,a,1) * dr(a,i);
   AT(k,k+2,i) = S(a,a,2) * df(a,i);

   AT(k+1,k+1,i) = - dr(a,i);
   AT(k+2,k+2,i) = - df(a,i);
 else
   AT(k,l+1,i) = S(b,a,1) .* dr(b,i);
   AT(k,l+2,i) = S(b,a,2) .* df(b,i);

   if Z(b,1,i) <= C(b,i)
     AT(k+2,l,i) = q(a,i) * serve(b,i) * S(b,a,3);
     AT(k+4,l,i) = (1-q(a,i)) * serve(b,i) * S(b,a,3);
   else
     AT(k,l,i)   = serve(a,i) * c(i).* Z(a,1,i)/sum(Z(:,1,i))^2 - leave(a,i) * c(i).*
Z(a,1,i)/sum(Z(:,1,i))^2;
     AT(k+1,l,i) = p(a,i) * leave(b,i) * S(b,a,4) * (1 -c(i).*
(sum(Z(:,1,i))-Z(b,1,i))/sum(Z(:,1,i))^2) + p(a,i) * (sum(leave(:,i) .* S(:,a,4) .* c(i) .*
(Z(:,1,i)/sum(Z(:,1,i))^2)) - sum(leave(b,i) .* S(b,a,4) .*  c(i).*
(Z(b,1,i)/sum(Z(:,1,i))^2)));
     AT(k+2,l,i) = q(a,i) * serve(b,i) * S(b,a,3) * c(i).*
((sum(Z(:,1,i))-Z(b,1,i))/sum(Z(:,1,i))^2) + q(a,i) * (sum(serve(:,i) .* S(:,a,3) .* c(i) .*
(-Z(:,1,i)/sum(Z(:,1,i))^2)) - serve(b,i) .* S(b,a,3) .* c(i) .* (-Z(b,1,i)/sum(Z(:,1,i))^2));
     AT(k+3,l,i) = (1-p(a,i)) * leave(b,i) * S(b,a,4) * c(i).* (1
-(sum(Z(:,1,i))-Z(b,1,i))/sum(Z(:,1,i))^2) + (1-p(a,i)) * (sum(leave(:,i) .* S(:,a,4) .* c(i)
.* (Z(:,1,i)/sum(Z(:,1,i))^2)) - sum(leave(b,i) .* S(b,a,4) .* c(i) .*
(Z(b,1,i)/sum(Z(:,1,i))^2)));
     AT(k+4,k,i) = (1-q(a,i)) * serve(b,i) * S(b,a,3) * c(i).*
((sum(Z(:,1,i))-Z(b,1,i))/sum(Z(:,1,i))^2) + (1-q(a,i)) * (sum(serve(:,i) .* S(:,a,3) .* c(i)
.* (-Z(:,1,i)/sum(Z(:,1,i))^2)) - serve(b,i) .* S(b,a,3) .* c(i) .*
(-Z(b,1,i)/sum(Z(:,1,i))^2));
   end
 end

 B(k,k,i,1) = arrive(a,i);

 B(l,l,i,2)              =  alpha(a,b,i,1) + B(l,l,i,2);
 B(5*a - 3, 5*a - 3,i,2) =  alpha(a,b,i,1) + B(5*a - 3, 5*a - 3,i,2);
 B(l, 5*a - 3,i,2)       = -alpha(a,b,i,1) + B(l, 5*a - 3,i,2);
```

```
    B(5*a - 3, l,i,2)       = -alpha(a,b,i,1) + B(5*a - 3, l,i,2);

    B(l,l,i,3)              =  alpha(a,b,i,2) + B(l,l,i,3);
    B(5*a - 2, 5*a - 2,i,3) =  alpha(a,b,i,2) + B(5*a - 2, 5*a - 2,i,3);
    B(l, 5*a - 2,i,3)       = -alpha(a,b,i,2) + B(l, 5*a - 2,i,3);
    B(5*a - 2, l,i,3)       = -alpha(a,b,i,2) + B(5*a - 2, l,i,3);

    B(k,k,i,4)              =  alpha(a,b,i,3) + B(k,k,i,4);
    B(5*b - 3, 5*b - 3,i,4) =  alpha(a,b,i,3) + B(5*b - 3, 5*b - 3, i,4);
    B(k, 5*b - 3,i,4)       = -alpha(a,b,i,3) + B(k, 5*b - 3,i,4);
    B(5*b - 3, k,i,4)       = -alpha(a,b,i,3) + B(5*b - 3, k,i,4);

    B(k,k,i,5)              =  alpha(a,b,i,4) + B(k,k,i,5);
    B(5*b - 1, 5*b - 1,i,5) =  alpha(a,b,i,4) + B(5*b - 1,5*b - 1,i,5);
    B(5*b - 1, k,i,5)       = -alpha(a,b,i,4) + B(5*b - 1,k,i,5);
    B(k, 5*b - 1,i,5)       = -alpha(a,b,i,4) + B(k,5*b - 1,i,5);

    B(k,k,i,6)              =  alpha(a,b,i,5) + B(k,k,i,6);
    B(5*b - 2, 5*b - 2,i,6) =  alpha(a,b,i,5) + B(5*b - 2, 5*b - 2,i,6);
    B(k, 5*b - 2,i,6)       = -alpha(a,b,i,5) + B(k, 5*b - 2,i,6);
    B(5*b - 2, k,i,6)       = -alpha(a,b,i,5) + B(5*b - 2, k,i,6);

    B(k,k,i,7)              =  alpha(a,b,i,6) + B(k,k,i,7);
    B(5*b, 5*b,i,7)         =  alpha(a,b,i,6) + B(5*b, 5*b,i,7);
    B(5*b, k,i,7)           = -alpha(a,b,i,6) + B(5*b, k,i,7);
    B(k, 5*b,i,7)           = -alpha(a,b,i,6) + B(k, 5*b,i,7);

    b = b + 1;
  end
  a = a + 1;
end

BT(:,:,i) =  B(:,:,i,1) + B(:,:,i,2) + B(:,:,i,3) + B(:,:,i,4) + B(:,:,i,5) + B(:,:,i,6) +
   B(:,:,i,7);

% Forward Euler to solve variance equations
VAR(:,:,i) = VAR(:,:,i-1) + dt * (VAR(:,:,i-1) * AT(:,:,i-1)' + AT(:,:,i-1) * VAR(:,:,i-1) +
   BT(:,:,i-1));
hold = diag(VAR(:,:,i))';

for k = 1:A
```

```
    V(k,:,i) = sqrt(diag(VAR(5 * (k - 1) + 1 : 5 * k, 5 * (k - 1) + 1 : 5 * k,i)))';
    ZP(k,:,i) = Z(k,:,i) + sqrt(hold(5 * (k - 1) + 1 : 5 * k));
    ZM(k,:,i) = Z(k,:,i) - sqrt(hold(5 * (k - 1) + 1 : 5 * k));
  end
end
end
```

C.4 Code for fluid and diffusion approximation of virtual waiting time

```
function [VWT, VARV] = FluidVWT(F, serve, leave, C, time,  v, dcdt, a, Z, c)
% Function for calculating fluid approximation of a stochastic queueing system
% Inputs:
%   F     - output from the approximation script
%   serve - vector of service rates for modelled time period
%   leave - vector of abandonment rates for modelled time period
%   C     - capacity allocation
%   time  - modelled time period
%   v     - variance of queueing process
%   dcdt  - vector of derivatives for queue and service orbit
%   a     - health state of queue
%   Z     - fluid approximation of queue and service
%   c     - total servers available for a service
% Outputs:
%   VWT   - virtual waiting time approximation
%   VARV  - variance of virtual waiting time approximation

% Set initial conditions
y = [F,v];

% Set criteria for ending the ode solver: y(1) - C(i) = 0
function [value,isterminal,direction] = event_function(t,y)
  value = y(1) - C(i);  % when value = 0, an event is triggered
  isterminal = 1; % terminate after the first event
  direction  = 0; % get all the zeros
end
```

```
for i = 2:length(C)
  % Set IC for each interation due to piecewise approximation of continuous functions
  tspan = [time(i-1) time(i)];
  IC = y(end,1:2)';
  y = IC;

  % Set the ode solver
  odefun = @(t,y) [- leave(i) * y(1) + (leave(i) - serve(i)) * C(i);...
  - 2 * leave(i) * y(2) + leave(i) * (y(1) - C(i) + serve(i) * C(i)];

  % create an options variable
  options = odeset('Events',@event_function);

  % Solve the system of ODEs
  [t,y] = ode45(odefun, tspan, IC, options);
  if y(end,1) - C(i) <= 0.00001
    VWT  = t(end);
    VARV = y(end,2)/((C(i) * serve(i) + ((c(i) * (dcdt(a,i) * sum(Z(:,1,i)) - Z(a,1,i) *
     sum(dcdt(:,i))))/(sum(Z(:,1,i))^2)))^2);
    break
  end
  end
end
```

C.5 Code for implementing comparison between the models

```
% Define parameters for simulation and fluid approximation
T = 15; dt =  round(0.1,1); t = (0:dt:T); NT = length(t); IC = [[15,0,0];[15,0,0]]; sims = 1000;

c = 30 * ones(1, NT);
arrive(1,:) = 15 * ones(1, NT);     arrive(2,:) = 15 * ones(1, NT);
serve(1,:) = 1/2 * ones(1, NT);     serve(2,:) = 1 * ones(1, NT);
leave(1,:) = 1 * ones(1, NT);       leave(2,:) = 1/2 * ones(1, NT);
p(1,:) = 0.7 * ones(1, NT);         p(2,:) = 0.3 * ones(1, NT);
q(1,:) = 0.7 * ones(1, NT);         q(2,:) = 0.3 * ones(1, NT);
dr(1,:) = 1 * ones(1, NT);          dr(2,:) = 1/2 * ones(1, NT);
```

```
df(1,:) = 1 * ones(1, NT);          df(2,:) = 1/2 * ones(1, NT);
type = 'continuous';

% Health state transitions
S(:,:,1) = [0.8, 0.2; 0.5, 0.5]; % rejoin
S(:,:,2) = [0.8, 0.2; 0.2, 0.8]; % reuse
S(:,:,3) = [0.3, 0.7; 0, 1];     % service
S(:,:,4) = [1, 0; 0.6, 0.4];     % abandon

A = size(S,1);
%Simulation loop
Y = zeros(A, 7, NT, sims);
CapSim = zeros(A, NT, sims);

for U = 1:sims
  [Y(:,:,:,U), CapSim(:,:,U)] = SimMult(c, round(arrive), serve, leave, dr, df, p, q, dt, T, S, IC,
     type);
end
M  = sum(Y,4)/sims; V = sqrt(var(Y,1,4)); MP = M + V; MM = M - V;

% Fluid and diffusion approximation
VWT = zeros(A,3,NT);

% Solve fluid and diffusion model
[F, FP, FM, Cap, Var, dcdt] = FluidComp(c, arrive, serve, leave, dr, df, p, q, dt, T, S, IC, type);

% Calcualte approximation of the VWT and it's varaince
for i = 2:length(t)
  for a = 1:A
    if F(a,1,i) >= Cap(a,i)
      [vwt, varv] = FluidVWT(F(a,1,i), [serve(a,i:end), serve(a,end)*ones(1,3*NT)], [leave(a,i:end),
     leave(a,end)*ones(1,3*NT)], [Cap(a,i:end),   Cap(a,end)*ones(1,3*NT)],
     (t(i):dt:4*T),(Var(a,1,i)^2), dcdt, a, F, c);
      VWT(a,1,i) = vwt - t(i);
      VWT(a,2,i) = VWT(a,1,i) + sqrt(varv);
      VWT(a,3,i) = VWT(a,1,i) - sqrt(varv);
    end
  end
end

% Calculate error during/after warm up period for each queue
```

```
for a = 1:A
  WU = round(max(t(F(a,2,:) == 0)),1);
  if isempty(WU) == 0
    under = (t < WU); over  = (t >= WU);

    errunder   = sign(M(a,1:3,under) - F(a,1:3,under)) .* (M(a,1:3,under) - F(a,1:3,under));
    total.errunder = sum(errunder,3)./sum(M(a,1:3,under),3) * 100;

    errover   = sign(M(a,1:3,over) - F(a,1:3,over)) .* (M(a,1:3,over) - F(a,1:3,over));
    total.errover = sum(errover,3)./sum(M(a,1:3,over),3) * 100;
    verrunder   = sign(V(a,1:3,under) - Var(a,1:3,under)).*(V(a,1:3,under) - Var(a,1:3,under));
    total.verrunder = sum(verrunder,3)./sum(V(a,1:3,under),3) * 100;

    verrover   = sign(V(a,1:3,over) - Var(a,1:3,over)).*(V(a,1:3,over) - Var(a,1:3,over));
    total.verrover = sum(verrover,3)./sum(V(a,1:3,over),3) * 100;

    werrunder   = sign(M(a,7,under) - VWT(a,1,under)).*(M(a,7,under) - VWT(a,1,under));
    total.werrunder = sum(werrunder,3)./sum(M(a,7,under),3) * 100;

    werrover   = sign(M(a,7,over) - VWT(a,1,over)).*(M(a,7,over) - VWT(a,1,over));
    total.werrover = sum(werrover,3)./sum(M(a,7,over),3) * 100;

    vwerrunder   = sign(V(a,7,under) - VWT(a,2,under) + VWT(a,1,under)).*(V(a,7,under) -
     VWT(a,2,under) + VWT(a,1,under));
    total.vwerrunder = sum(vwerrunder,3)./sum(V(a,7,under),3) * 100;

    vwerrover   = sign(V(a,7,over) - VWT(a,2,over) + VWT(a,1,over)).*(V(a,7,over) - VWT(a,2,over) +
     VWT(a,1,over));
    total.vwerrover = sum(vwerrover,3)./sum(V(a,7,over),3) * 100;
  end

  % Calculate pointwise errors
  keep = M(a,1:3,:);
  err = sign(keep - F(a,1:3,:)) .* (keep - F(a,1:3,:));
  TotalERR = sum(err,3)./sum(keep,3) * 100;

  PointERR = err ./ keep ;
  PointERR(isnan(PointERR)) = 0;

  keepv = V(a,1:3,:);
  Verr = sign(keepv - Var(a,1:3,:)).*(keepv - Var(a,1:3,:));
```

```
TotalVERR = sum(Verr,3)./sum(keepv,3) * 100;

PointVERR = Verr ./ keepv ;
PointVERR(isnan(PointVERR)) = 0;

keepw = M(a,7,:);
Werr = sign(keepw - VWT(a,1,:)).*(keepw - VWT(a,1,:));
TotalWERR = sum(Werr,3)./sum(keepw,3)*100;

PointWERR = Werr ./ keepw ;
PointWERR(isnan(PointWERR)) = 0;

figure(1)
subplot(A,1,a);
varNames1 = {'Queue and Serice: sim','Queue and Serice: fluid',...
'Rejoin: sim','Rejoin: fluid','Reuse: sim','Reuse: fluid'};
plot(squeeze(M(a,6,:)),squeeze(M (a,1,:)),'.b',t,squeeze(F (a,1,:)),'b',...
squeeze(M(a,6,:)),squeeze(M (a,2,:)),'.r',t,squeeze(F (a,2,:)),'r',...
squeeze(M(a,6,:)),squeeze(M (a,3,:)),'.g',t,squeeze(F (a,3,:)),'g',...
squeeze(M(a,6,:)),squeeze(MP(a,1,:)),'.b',t,squeeze(FP(a,1,:)),'b',...
squeeze(M(a,6,:)),squeeze(MM(a,1,:)),'.b',t,squeeze(FM(a,1,:)),'b',...
squeeze(M(a,6,:)),squeeze(MP(a,2,:)),'.r',t,squeeze(FP(a,2,:)),'r',...
squeeze(M(a,6,:)),squeeze(MM(a,2,:)),'.r',t,squeeze(FM(a,2,:)),'r',...
squeeze(M(a,6,:)),squeeze(MP(a,3,:)),'.g',t,squeeze(FP(a,3,:)),'g',...
squeeze(M(a,6,:)),squeeze(MM(a,3,:)),'.g',t,squeeze(FM(a,3,:)),'g');
ylim = ([0 max(max(max(MP(a,1:3,:))),max(max(FP(a,1:3,:))))]);
legend(varNames1);
xlabel('Time')
ylabel('Number of patients');

figure(2)
subplot(A,1,a);
varNames2 = {'Queue and Service','Rejoin','Reuse'};
plot(squeeze(M(a,6,:)),squeeze(PointERR));
legend(varNames2);
xlabel('Time')
ylabel('Error as proportion of simulation value');

figure(3)
subplot(A,1,a);
varNames3 = {'Queue and Service','Rejoin','Reuse'};
```

```
  plot(squeeze(M(a,6,:)),squeeze(PointVERR));
  legend(varNames3);
  xlabel('Time')
  ylabel('Error as proportion of simulation value');

  figure(4)
  subplot(A,1,a);
  varNames4 = {'Simulation','Fluid Approximation'};
  plot(squeeze(M(a,6,:)), squeeze(V(a,7,:)).^2, t, squeeze((VWT(a,2,:)-VWT(a,1,:)).^2)); %
  legend(varNames4);
  xlabel('Time')
  ylabel('Waiting time');
  title(['Variance in the virtual waiting timein health state', num2str(a)]);

  figure(5)
  subplot(A,1,a);
  varNames5 = {'Simulation','Fluid Approximation'};
  plot(squeeze(M(a,6,:)), squeeze(M(a,7,:)), t, squeeze(VWT(a,1,:))); %
  legend(varNames5);
  xlabel('Time')
  ylabel('Waiting time');
  title(['Virtual waiting time for patients in health state ', num2str(a)]);

  figure(6)
  subplot(A,1,a);
  varNames1 = {'Loss: simulation', 'Loss: fluid approximation', 'Discharge: simulation', 'Discharge:
     fluid approximation'};
  plot(squeeze(M(a,6,:)),squeeze(M (a,4,:)),'.b',t,squeeze(F (a,4,:)),'b',...
  squeeze(M(a,6,:)),squeeze(M (a,5,:)),'.r',t,squeeze(F (a,5,:)),'r');
  legend(varNames1);
  xlabel('Time')
  ylabel('Number of patients');

  figure(7)
  subplot(A,1,a);
  plot(squeeze(M(a,6,:)), sum(CapSim(a,:,:),3)'/sims, t, Cap(a,:));
  legend(varNames5);
  xlabel('Time')
  ylabel('Number of servers');
end
```

C.6 Parameters used in the fluid and diffusion approximation of section 6.3.4

Health state transition matrices, $i \in \{1, 2, 3\}$:

$$S_{S,1} = \begin{bmatrix} 0, 0.4, 0.6 \\ 0, 0.5, 0.5 \\ 0, \ 0, \ 1 \end{bmatrix} \; S_{S,2} = \begin{bmatrix} 0, 0.4, 0.6 \\ 0, 0.2, 0.8 \\ 0, \ 0, \ 1 \end{bmatrix} \; S_{S,3} = \begin{bmatrix} 0, \ 0, \ 1 \\ 0, 0.2, 0.8 \\ 0, \ 0, \ 1 \end{bmatrix}$$

$$S_{L,1} = \begin{bmatrix} 1, \ 0, \ 0 \\ 0.6, 0.4, 0 \\ 0, \ 1, \ 0 \end{bmatrix} \; S_{L,2} = \begin{bmatrix} 1, \ 0, \ 0 \\ 0.6, 0.4, 0 \\ 0.6, 0.4, 0 \end{bmatrix} \; S_{L,3} = \begin{bmatrix} 1, \ 0, \ 0 \\ 0.6, 0.4, 0 \\ 0, 0.6, 0.4 \end{bmatrix}$$

$$S_{R,i} = \begin{bmatrix} 1, \ 0, \ 0 \\ 0.2, 0.8, 0 \\ 0, 0.2, 0.8 \end{bmatrix} \; S_{U,i} = \begin{bmatrix} 1, \ 0, \ 0 \\ 0.1, 0.9, 0 \\ 0, 0.1, 0.9 \end{bmatrix} \; S_{A,i} = \begin{bmatrix} 1, \ 0, \ 0 \\ 0.2, 0.8, 0 \\ 0, 0.2, 0.8 \end{bmatrix} \; S_{O,i} = \begin{bmatrix} 1, \ 0, \ 0 \\ 0.1, 0.9, 0 \\ 0, 0.1, 0.9 \end{bmatrix}$$

Service routing matrices:

$$R_{1,S} = \begin{bmatrix} 0, 0.9, 0 \\ 0, 0.6, 0.4 \\ 0, 0.2, 0.7 \end{bmatrix} \; R_{2,S} = \begin{bmatrix} 0, 0.9, 0 \\ 0, 0.5, 0.4 \\ 0, 0.2, 0.6 \end{bmatrix} \; R_{3,S} = \begin{bmatrix} 0, 0.9, 0 \\ 0, 0.4, 0.5 \\ 0, 0.2, 0.2 \end{bmatrix}$$

$$R_{1,L} = \begin{bmatrix} 0.6, 0, 0 \\ 0.8, 0, 0 \\ 0.8, 0, 0 \end{bmatrix} \; R_{2,L} = \begin{bmatrix} 0.4, 0.4, 0 \\ 0.2, 0.6, 0 \\ 0, 0.8, 0 \end{bmatrix} \; R_{3,L} = \begin{bmatrix} 0.4, 0.4, 0 \\ 0.3, 0.3, 0.2 \\ 0, 0.1, 0.2 \end{bmatrix}$$

	Service		
Parameters	$i=1$	$i=2$	$i=3$
$C(0)$	30	20	10
$\lambda_{k=1}$	10	0	0
$\lambda_{k=2}$	5	5	0
$\lambda_{k=3}$	0	3	3
$\mu_{k=1}$	2/14	4/14	2/14
$\mu_{k=2}$	1	10/14	6/14
$\mu_{k=3}$	2	1	8/14
$\theta_{k=1}$	3/14	2/14	1/14
$\theta_{k=2}$	7/14	5/14	3/14
$\theta_{k=3}$	1	7/14	4/14
$d_{k=1,F}$	3/14	2/14	1/14
$d_{k=2,F}$	7/14	5/14	3/14
$d_{k=3,F}$	1	7/14	4/14
$d_{k=1,O}$	3/14	2/14	1/14
$d_{k=2,O}$	7/14	5/14	3/14
$d_{k=3,O}$	1	7/14	4/14
$d_{k=1,R}$	6/14	4/14	2/14
$d_{k=2,R}$	1	10/14	6/14
$d_{k=3,R}$	2	1	8/14
$d_{k=1,A}$	6/14	4/14	2/14
$d_{k=2,A}$	1	10/14	6/14
$d_{k=3,A}$	2	1	8/14

Table C.1: Parameters used in the fluid and diffusion approximation of section 6.3.4