Modeling and Mitigating Bottlenecks in Healthcare Systems for Older Adults

Tim Rens de Boer

Modeling and Mitigating Bottlenecks in Healthcare Systems for Older Adults

Tim Rens de Boer

Modeling and Mitigating Bottlenecks in Healthcare Systems for Older Adults

Tim Rens de Boer

ISBN: 978-94-6536-148-2
DOI: https://doi.org/10.5463/thesis.1658

Typeset by LaTeX.
Printed by: Ipskamp Printing
Cover design by: Salim Salmi

VRIJE UNIVERSITEIT

Modeling and Mitigating Bottlenecks in Healthcare Systems for Older Adults

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
volgens besluit van de decaan
van de Faculteit der Bètawetenschappen
in het openbaar te verdedigen
op woensdag 17 juni 2026 om 13.45 uur
in de universiteit

door

Tim Rens de Boer

geboren te Zaanstad

promotoren:	prof.dr. R.D. van der Mei prof.dr. S. Bhulai
copromotoren:	dr. R. Bekker prof.dr. B.M. Buurman-van Es
promotiecommissie:	prof.dr. G.M. Koole dr. D. Moeke prof.dr. G.G. van Merode prof.dr. P. Harper dr. J.L. MacNeil-Vroomen

Voorwoord

Dit proefschrift is tot stand gekomen dankzij de steun en inzet van velen. Ook degenen die niet direct bij het onderzoek betrokken waren, hebben hieraan een waardevolle bijdrage geleverd. Mocht ik iemand niet bij naam noemen, dan is dat omdat het er eenvoudigweg te veel zijn om op te noemen.

Laat ik beginnen met R^3: Rob van der Mei, René Bekker en Rebekka Arntzen. Met een beetje creativiteit via mijn tweede naam waren we op donderdagochtend R^4. Rob, bedankt voor je vertrouwen en expertise. Je hebt me laten zien hoe belangrijk het is om wiskundige ideeën niet alleen begrijpelijk, maar ook boeiend en toegankelijk te maken voor een breed publiek. Daar ben ik je zeer dankbaar voor, net als voor de sturing die je het onderzoek gaf op de momenten waarop dat nodig was. René, bedankt voor het introduceren van wachtrijtheorie. Vanaf het begin liet je mij al zien dat wachtrijtheorie niet saaie stof hoeft te zijn. Jouw enthousiasme voor moeilijke problemen werkten vaak aanstekelijk. Rebekka wil ik graag bedanken voor haar gezelligheid en steun, eerst als begeleider tijdens mijn stage, daarna als projectmaatje. Dankzij de kennis die jij al vóór mijn start bij Dolce Vita had opgebouwd, kon ik meteen goed beginnen. Ik kon met jou alles bespreken, van wiskunde in de zorg tot alles wat er bij een PhD komt kijken. Dank aan jullie alle drie voor de fijne samenwerking en meetings waarin inhoud en gezelligheid altijd samenkwamen.

Ik heb, naast R^3 ook veel samengewerkt met Oscar Smeekes: mijn vraagbaak over alles zorg gerelateerd. Samen hebben we drie mooie onderzoeken gedaan. Tijdens onze samenwerking was er ook altijd ruimte om andere onderwerpen te bespreken, zoals: vakanties, ADE, weekendplannen, en nog veel meer.

Natuurlijk wil ik ook de rest van de Dolce Vita-projectgroep bedanken. Zonder Sandjai weet ik niet of ik bij de VU terecht was gekomen. Zijn enthousiasme

tijdens de open dag zorgde ervoor dat ik een studie aan de VU ging doen. Bianca wil ik graag bedanken voor het vertrouwen en de kennis. Wiskundige ideeën konden zo vertaald worden naar resultaten waar de zorg wat aan had. Hanna wil ik graag bedanken voor haar scherpe blik. Na meetings met jou en Oscar was vaak duidelijk wat de volgende stappen waren. Robert wil ik graag bedanken voor zijn brede visie en het feit dat hij verder keek dan alleen de wetenschappelijke bijdrage. Ismail, Eline en Sofie, de nieuwe lichting PhD-studenten binnen Dolce Vita, ik wil jullie bedanken voor jullie frisse blik en nieuwe energie. Ik kijk uit naar de nieuwe onderzoeken waar jullie mee komen. Katja, Joost en Louisa, in de zes maanden die jullie bij Dolce Vita doorbrachten, hebben jullie een grote indruk achtergelaten. Gezelligheid, goede discussies en leerzame momenten stonden daarin centraal.

Alle onderzoeken in dit proefschrift had ik niet kunnen doen zonder een fijne werkplek en geweldige collega's: de Stochastics-groep onder leiding van Bert Zwart bij het Centrum Wiskunde & Informatica (CWI). CWI heb ik altijd ervaren als een plek waar hard gewerkt wordt, maar waar ook genoeg ruimte was voor andere dingen: tafeltennissen, koffiegesprekken over van alles en nog wat, borrels, pub quizzen en vele teamuitjes. Op CWI deelde ik een kantoor met Jesse. Waar ik kies voor een vakantie op het strand en Jesse voor eentje in de bergen, bleken we verrassend snel overeenkomsten te hebben in boeken, films en nog veel meer. Ons kantoor was gelukkig groot genoeg voor nog meer gezelschap. Zo heb ik het mogen delen met Rens, María, Barbara en Emma tijdens het schrijven van hun masterscriptie. Eén voor één zorgden jullie voor nieuwe energie en de nodige dosis humor op kantoor. Als er op ons kantoor te hard werd gewerkt, dan had ik gelukkig nog erg gezellige buren: Joris, Arwin en Ruurd. Deze drie heren waren altijd in voor een discussie, filosofisch of wiskundig van aard. Joris, Voor anderen leek het misschien alsof we elkaar constant plaagden, maar wij weten dat dit onze band alleen maar sterker heeft gemaakt. Arwin bood altijd een luisterend oor. Bij hem kon ik met eventuele machine learning vragen terecht, maar ook voor een potje kaarten. Ruurd wil ik bedanken voor zijn humor. Hoewel we bijna tien jaar geleden onze studie tegelijkertijd begonnen, hebben we elkaar de afgelopen 4 jaar pas beter leren kennen. Onze Brusseltrip met liveblog zal ik altijd blijven herinneren en hopelijk volgen er in de toekomst nog veel pubquizzen. Verderop de gang zaten de rest van mijn collega's, waaronder: Guus, Salim, Etienne, Hilde, Robin, Berend, Britt, Elisabeth, Jan, Joris, Moos, Eda, Simon, Martijn en waarschijnlijk mis ik er nog veel meer. Met allen heb ik fijne herinneringen, zoals: mijn tijd bij 113 met Guus en Salim, een eerste congres-bezoek met Etienne en Britt, escape rooms met de hele groep, pubquizzen (die we gewonnen hebben) en nog veel meer. Jullie hebben de lat erg hoog gelegd voor toekomstige collega's.

De A&O-groep, onder leiding van Ger Koole aan de VU, bleef naast CWI gelukkig een vertrouwde werkplek. Ik wil deze fijne groep collega's bedanken met name voor de interessante seminars, waardoor ik op de hoogte bleef van de onderzoeken binnen en buiten onze afdeling. Graag zou ik Erica, Jasper, Jeroen, Leon, Yasmin, Mathijs, Rik, Ritsaart, Corné en Hanneke bedanken voor de gezellige sfeer, gastvrijheid, boekenclubs en spelavonden. Wees gerust, ik zal niet meer

iemands vaste mok of bureau inpikken.

Mijn proefschrift bevat ook een onderzoek uitgevoerd in samenwerking met 113 Zelfmoordpreventie. Ik wil hier graag al mijn oud-collega's van 113 bedanken. Saskia en Renske: bedankt voor jullie expertise en de fijne samenwerking. Het project was een top begin van mijn PhD-tijd.

I would like to express my gratitude to the members of the doctoral committee for reviewing my thesis, and I look forward to the discussions during the defense.

Niet alleen collega's zijn belangrijk geweest tijdens mijn PhD-tijd. Misschien wel net zo belangrijk was de ontspanning met vrienden. Ik wil Joris, Marc, Melissa en Ali bedanken, de vaste spellengroep. Onze spelavonden waren vaak nodig om te ontsnappen aan mijn onderzoek. Deze avonden zijn vaak gevuld met veel gelach, fanatieke potjes en goeie gesprekken. Ik ken jullie al sinds de middelbare school, en het gemak waarmee we toen met elkaar konden lachen is er nog steeds. Joris en Marc, het is niet voor niets dat ik jullie ook heb gevraagd als getuigen, ik weet dat ik altijd bij jullie terecht kan. Melissa, bedankt voor de gezelligheid en de mooie reizen naar Brussel en Parijs samen met Manon. Ali, op de middelbare school wist ik al: als je wilt lachen, moet je bij jou zijn en jaren later is dat nog steeds zo, of het nu om imitaties of andere verhalen gaat. Ik kijk ernaar uit om met jullie nog veel spellen te spelen, hopelijk met meer overwinningen dan verliezen, maar waarschijnlijk niet.

Ook wil ik mijn familie bedanken: mijn ouders, zussen, zwagers, al mijn neefjes en nichtjes en mijn hondje Saar. Mama, bedankt voor je vertrouwen en je hulp. Je kon me misschien niet helpen met de wiskundige formules, maar als dat had gekund, had je vast direct met pen en papier klaargestaan. Ik herinner me nog goed hoe mijn vader met me meeging naar de eerste informatieavonden over studiekeuzes, en hoe er thuis vaak een warme maaltijd voor me klaarstond als ik laat thuiskwam. Zonder jullie steun en hulp waar nodig was dit proefschrift er niet geweest. Mijn zussen Lisanne en Charlotte en mijn zwagers Robin en Jace wil ik graag bedanken. Mijn zussen stonden altijd voor me klaar, of het nu ging om hulp bij boekverslagen of zorgen dat ik veilig thuiskwam na een avond uit. Daar ben ik ze dankbaar voor als jongste broertje. Hoewel ze nog nieuwe toevoegingen aan de familie zijn en dit misschien nog niet helemaal kunnen lezen, wil ik mijn neefjes en nichtjes Suze, Wout, Joah, Lotte en Fynn bedanken. Ik hoop straks de titel Dr. te mogen dragen, maar mijn neefjes en nichtjes hebben mij al een andere titel gegeven: oom Tim. Een titel om trots op te zijn. Ook wil ik de rest van mijn familie en schoonfamilie bedanken voor hun interesse, steun en gezelligheid. Of het nu ging om feestdagen, verjaardagen bij Ingrid, Jeroen, Ilse, Stef of Sander, een bezoek aan Anja en Rob op Terschelling, een dagje Bladel met de familie Antenbrink, of andere mooie herinneringen, jullie aanwezigheid maakte het altijd bijzonder.

Tot slot wil ik Manon bedanken, ik had dit avontuur met niemand anders wil-

len aangaan dan met jou. Mijn vriendin, verloofde en toekomstige vrouw: je bent altijd mijn steun en toeverlaat geweest. Dankzij jou komen we letterlijk verder: eerst Brussel en nu in Limburg. Al negen jaar brengen we bijna elke dag samen door, en nog steeds kan ik met je lachen, plezier maken, serieus discussiëren en eigenlijk alles delen. Dit proefschrift had ik niet zonder jou kunnen schrijven.

Contents

List of Abbreviations

(I)LP	(Integer) Linear Programming
(S)ARIMA	(Seasonal) AutoRegressive Integrated Moving Average
AIC	Akaike Information Criterion
AI	Artificial Intelligence
ANOVA	Analysis of Variance
ASB	After-Service Blocking
BSB	Before-Service Blocking
Card	Cardiology hospital department
CBM	Condition-Based Management
CIZ	Care Needs Assessment Centre
DTMC	Discrete-Time Markov Chain
ED	Emergency Department
FCFS	First-Come-First-Served
Gastro	Gastroenterology hospital department
GP	General Practitioner
GR	Geriatric Revalidation
HC	Home Care
HH	Household Help
HL	Helpline
HOME	At home without any professional care

HOS	Hospitalization
Int Med	Internal Medicine hospital department
LSTM	Long Short-Term Memory
LTC	Long-Term Care
Lung	Lung diseases hospital department
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MDP	Markov Decision Process
ML	Machine Learning
MSE	Mean Squared Error
Neur	Neurology hospital department
NHH	Nursing Home care at Home
NH	Nursing Home
NN	Neural Network
NO	No home care
Ortho	Orthopedics hospital department
OR	Odds Ratio
Pall	Palliative
PC	Personal Care
PG	Psycho-Geriatric
PM	Process Mining
POMDP	Partially Observable Markov Decision Process
Rehab	Rehabilitation
RNN	Recurrent Neural Network
RSB	Repetitive Service Blocking
SD	System Dynamics
SES-WOA	Socioeconomic status score developed by [96]
SMDP	Semi-Markov Decision Process
SOM	Somatic
STC	Short-Term residential Care
Std Dev	Standard Deviation
STROBE	STrengthening the Reporting of OBservational studies in Epidemiology

Surg	Surgery hospital department
TBM	Time-Based Management
Uro	Urology hospital department

1 Introduction

In many countries, healthcare systems face great challenges, particularly in the domain of elderly care. One key driver is the rising gray pressure [86], which refers to the ratio of people aged 65 and older to those between 20 and 65 years. This increase is due to multiple reasons, most importantly a decrease in birth rates and an increase in life expectancy [220]. The aging population creates a growing demand for elderly care [142], such as Home Care (HC) [164] or Long-Term Care (LTC) in an institution [242]. At the same time, the cost of elderly care increases [125] and the healthcare system must deal with a shortage of available personnel [94]. All of these developments put considerable pressure on the sustainability, availability and effectiveness of the healthcare system.

Many countries around the world are currently experiencing an increase in the number of older adults [72, 185]. In the United States of America, the number of older adults is expected to increase to 72 million by 2030, accounting for 20% of the total population [244]. In 2050, 25% of the population in Europe and North America is expected to be 65 years or older [220]. Figure 1.1 illustrates that the old-age dependency ratio (i.e., the ratio of older adults to the working age population), along with life expectancy, has been on the rise over the past 75 years across the USA, Europe, and the Netherlands [169].

In the last two decades, various regions have experienced an increase in health expenditure per capita. The Netherlands witnessed an increase from $2,466 in 2000 to $7,179 in 2021 (see Figure 1.2). With advancing age, individuals place greater demands on the healthcare system, necessitating either more frequent or more costly care [117]. This observation is also confirmed by data from the Netherlands, where geriatric rehabilitation care expenditure has seen an increase in the last five years [213], rising from €1.2 billion in 2021 to an expected €1.5 billion in 2025. There is also a well-documented shortage in healthcare workforce [69, 147], also visible in the number of new and open healthcare vacancies in the Netherlands over the years (see Figure 1.3). This together, results in a healthcare system that is overloaded.

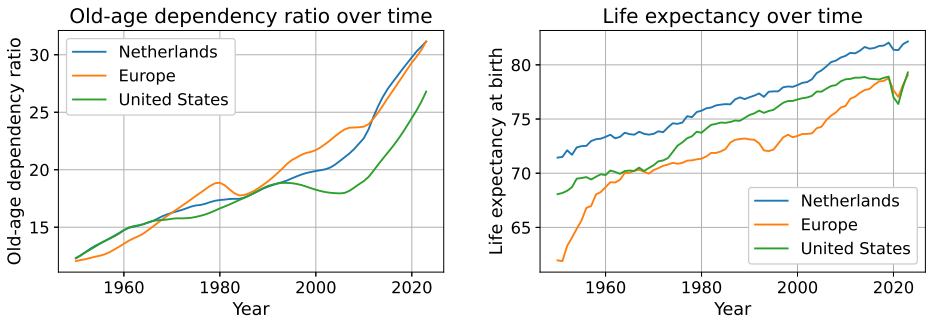


Figure 1.1: Trends in old-age dependency ratio, defined as the number of adults aged 65 years and older per 100 individuals of working age (15–64), and life expectancy at birth for the Netherlands, Europe, and the United States. Data from [169].

The challenges outlined above represent only a portion of the issues facing the healthcare system. Additional examples include fragmentation of care [35], the prevalence of multi-morbidity [187], and variations in regional costs [255]. In a positive sense, there are also promising strategies to improve the healthcare system or relieve pressure on it. These include preventive care [154], proactive care [173], and digital care [102]. Other strategies involve shifting care from the institutional setting to the home setting [252] and using Artificial Intelligence (AI) or technology support to reduce the workload of nurses [129]. These interventions are not a one-size-fits-all remedy. Their effectiveness depends on the local context, patient preferences, cost structures, and regulatory environments, and therefore, a careful context-specific evaluation is required.

In this thesis, the system of elderly care in the Netherlands is examined using two fundamentally different perspectives: a *flow-based* and a *capacity-based* perspective. The perspectives examine the system from the demand side of care by

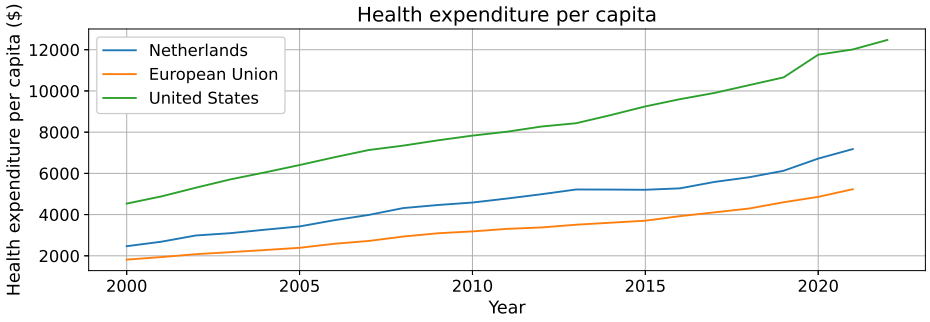


Figure 1.2: Health expenditure per capita for the Netherlands, the European Union and the United States over time. Data from [169].

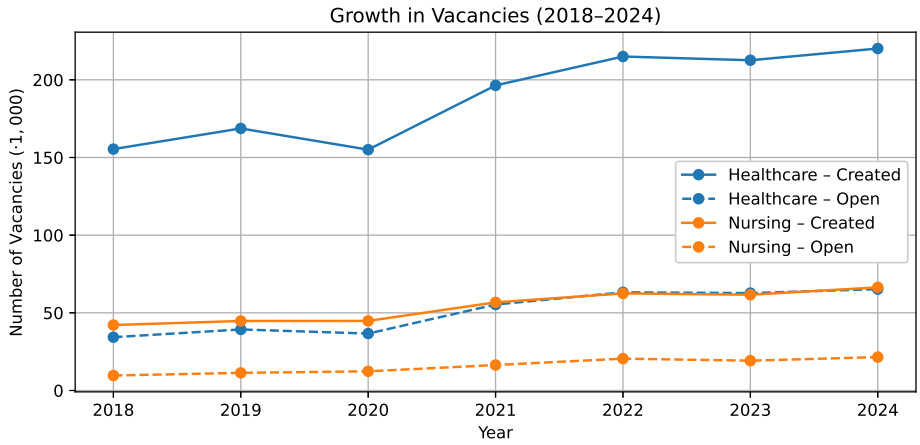


Figure 1.3: Trends in vacancies in the Netherlands for healthcare and nursing (2018-2024). Solid lines represent newly created vacancies, while dashed lines represent open vacancies [34].

asking, "How do older adults present themselves to the care system?" and from the supply side of care by questioning, "How does the limited capacity introduce issues in the healthcare system, and how do we avoid or mitigate them?"

1.1 DOLCE VITA

The thesis was developed as part of the NWO project *Data-Driven Optimization for a Vital Elderly Care System in the Netherlands* (*DOLCE VITA*), a collaborative project between mathematicians and elderly care professionals. The project aims to use queueing methods and data science to aid the elderly care system. The organizations collaborating in this project are: Centrum Wiskunde & Informatica (CWI), Amsterdam UMC, Vrije Universiteit Amsterdam (VU), and Sigra [198], the overarching organization representing healthcare providers in the Amsterdam region. In addition to these core partners, the project is supported by a broad user group that brings together a wide range of healthcare stakeholders. The user group consists of 23 different healthcare organizations, including healthcare insurers, several major care organizations, and General Practitioner (GP) organizations, active in the Amsterdam region.

DOLCE VITA was motivated by the Sigra report *Krakende Ketens* [197], detailing various bottlenecks in the healthcare system. These bottlenecks range from critical to minor and could be caused by numerous reasons, such as: older adults having a small social network, opening hours of various care institutions, capacity-related issues and inaccurate, incomplete, or no access to the right patient information. These bottlenecks were found mainly at the transition points between different

forms of care, where there was a mismatch between demand and supply and little or no coordination. An overview of these bottlenecks is visualized in Figure 1.4.

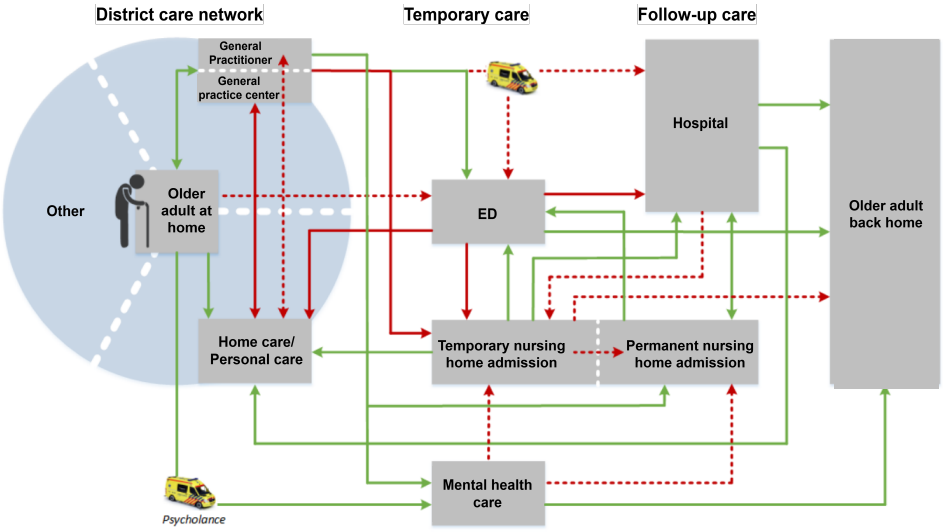


Figure 1.4: Overview of bottlenecks found in the elderly care system in the Amsterdam region, from [197]. Green line: no significant bottleneck; red dotted line: area that requires attention; red solid line: major bottleneck.

An illustration of the effect of these bottlenecks can be found in Figure 1.5. Here, we see a frail older adult living at home with HC, possibly waiting for a Nursing Home (NH) admission. Then a crisis occurs, the patient falls and has to go to the Emergency Department (ED). The ED is overcrowded, complicating the health-care professional's ability to assess the individual's health, thus hindering precise care. Surgery is required and this person is admitted to the hospital. The patient may have worsened in health and has now to be admitted to the NH immediately. The NH has no capacity available and the patient is required to wait in the hospital and block a hospital bed. Eventually, capacity becomes available and the patient can be admitted to the NH. It is still possible that this person is admitted to a NH that is not preferred by the patient. These bottlenecks have various causes; the question of when to admit a person to the NH and the difficulty in assessing the patient's health in the ED are examples of bottlenecks concerning the flow and demand side of care, while *bed-blocking* in the hospital due to limited capacity at the NH is a key example of capacity and supply side of the care system. These types of bottlenecks can lead to unnecessary waiting times, additional transfers, and avoidable care, resulting in higher costs and more stress for the patients and healthcare professionals. It should be noted that such problems do not occur in each case, but when they do, their impact can be substantial.

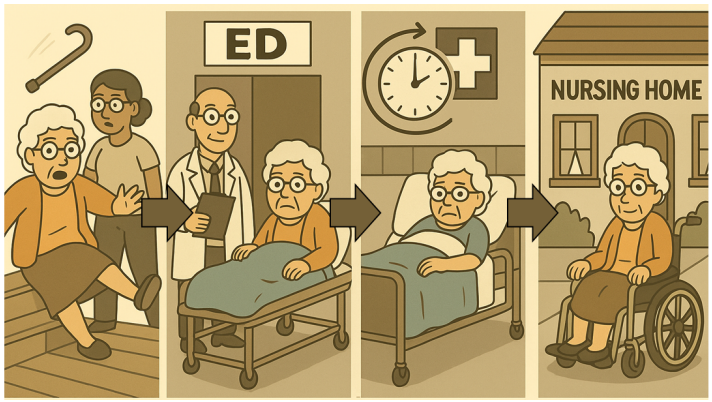


Figure 1.5: Care pathway of an older adult from a fall at home to the emergency department, surgery, and eventual nursing home care.

The goal of this thesis is to help prevent and reduce such bottlenecks. To this end, it focuses on the two complementary sides of the healthcare system: the *flow side* and the *capacity side*. Together, these perspectives provide a comprehensive view of how demand and supply imbalances arise and how they can be mitigated.

1.2 Care System for Older Adults

The healthcare system consists of different forms of healthcare, each interconnected and often mutually dependent. For example, a hospital discharge may only be allowed if followed by NH admission or if HC is present at the patient's home. Each care form has its own target population, capacity, length-of-stay distribution, opening hours, arrival patterns, costs, and goals. All forms of care collectively create a service network that older adults navigate, leading to either recovery or decline. Each form of care is spread across different institutions and locations that collaborate at multiple organizational levels.

Figure 1.6 shows a stylized view of the healthcare system of the Netherlands. This system consists of several key forms of care, including various types of HC, hospital care across different departments, geriatric rehabilitation, short-term residential care, LTC, and other types of care. Below, a brief overview of these care forms is provided.

Home Care (HC)
HC is an umbrella term for all care provided at home. This could be Household Help (HH) to support a patient's prolonged living at home, or Personal Care (PC), where a nurse provides medical care and treatments at home to improve or restore health. A person can also receive both HH and PC. In addition, Nursing Home

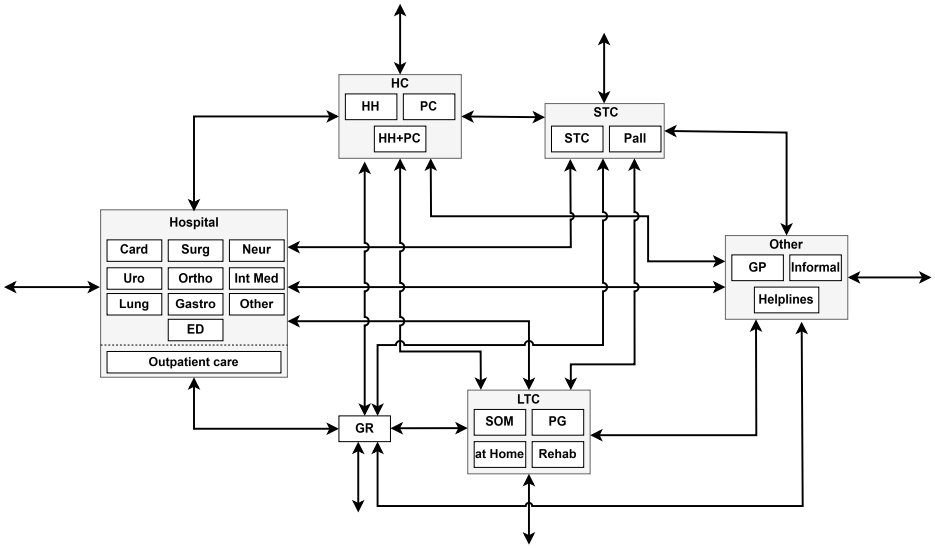


Figure 1.6: Stylized view of the elderly care system. Arrows between care forms indicate possible patient flows, while arrows leading from or to empty space represent arrivals to or departures from the system, i.e., individuals who previously did not require care or who will no longer require care.

care at Home (NHH), also known as Long-Term Care at home (LTC at home), is an option. This is often more intensive care than just PC and often precedes a nursing home admission. All these HC types vary in how they are financed, by whom they are provided, in intensity, and for how long it is expected to be received [256].

Hospital Care

Hospital care is a crucial part of the older adult care system, even though hospitals are not specifically targeted at older adults. Older adults are disproportionately represented in unplanned admissions, with around 300,000 people aged 65+ hospitalized each year between 2017 and 2021 [232]. We distinguish between ED visits and hospitalization, which can occur in a wide range of hospital departments. These include, but are not limited to, cardiology (Card), surgery (Surg), neurology (Neur), urology (Uro), orthopedics (Ortho), internal medicine (Int Med), lung diseases (Lung), and gastroenterology (Gastro). With respect to hospital care, this thesis primarily concentrates on inpatient services delivered within hospitals. Nonetheless, hospitals also deliver a range of outpatient services, including specialist consultations, day treatments, diagnostic procedures, and rehabilitation. Following hospitalization or outpatient treatment, many older adults require further care, including follow-up visits, home care, or admission to long-term care facilities, to support recovery and prevent readmissions.

Geriatric Revalidation (GR)
GR focuses on rehabilitation following hospitalization, for example, after a stroke or an operation. The goal of this type of care is to allow older adults to return home. In 2019, around 53,000 older adults in the Netherlands received GR, mainly following fractures and strokes [256].

Short-Term (Residential) Care (STC)
STC is temporary residential care for individuals unable to return home, but who are expected to eventually return. It covers high- or low-complex care needs. A distinction is often made between non-palliative and palliative STC. STC was introduced as a separate form of care in 2017, with about 30,000 recipients in that year, growing to nearly 38,000 in 2020 [256].

Long-Term Care (LTC)
LTC is care provided to older adults who are expected to require this care permanently. Before receiving LTC, a person is required to have an indication of the type and intensity of care, provided by CIZ [21], where each type and intensity corresponds with a specific profile. We distinguish between LTC in a nursing home and at home, which could also be seen as HC. The LTC in nursing homes is classified into three different types: Somatic (SOM), Psycho-geriatric (PG), and Rehabilitation (Rehab). SOM is intended for older adults with primarily physical health needs, such as mobility limitations, chronic diseases, or complex medical conditions requiring daily assistance. PG is provided for older adults with cognitive impairments. Rehab offers support aimed at restoring functional independence after events such as surgery, stroke, or other acute conditions. In 2017, approximately 223,000 older adults received LTC, most of whom lived in facilities due to dementia-related problems [256].

Other types of care
Finally, there are many other types of care that do not fit the categorizations made above, such as informal care and the GP. Another example is helplines; while not a physical form of care, helplines are often seen as the entry point to the care system. Examples of helplines are the emergency number (112), GP phone, 113 Suicide Prevention [170], and the listen line [59], with each helpline having a specific expertise. These helplines can help people obtain information and receive the correct type of care. In recent years, helplines have also played an important role in managing the shortage in workforce, ensuring that callers can be efficiently directed to the right form of care [143].

Together, these types of care and support systems create a complex and interdependent network of services. Understanding their interactions and transitions is crucial to understanding and improving the network and the healthcare system as a whole.

1.3 Overview of the Thesis

This thesis consists of two parts: (i) *flow-based* and (ii) *capacity-based* research. Figure 1.7 shows which chapters are connected to which part of the healthcare system. Chapters 2-4, which are referred to as the chapters with a *flow-based* perspective and shown in light blue in the figure, focus on understanding *where*, *how*, and *when* healthcare care demand arises. Chapters 5-7, throughout referred to as the chapters with a *capacity-based* perspective and shown in light yellow in the figure, focus on how capacity influences the system and how capacity-related problems can be mitigated by implementing a control policy or altering the capacity. The dotted areas in the figure indicate the focus areas of each chapter. Chapters 2, 5, and 6 address the healthcare system as a whole, while Chapter 3 focuses on the connection between home care and hospital care. Chapter 4 examines the flow to nursing homes, and Chapter 7 studies the helplines. Note that while outpatient care is included in the figure, this type of care is not included in this thesis.

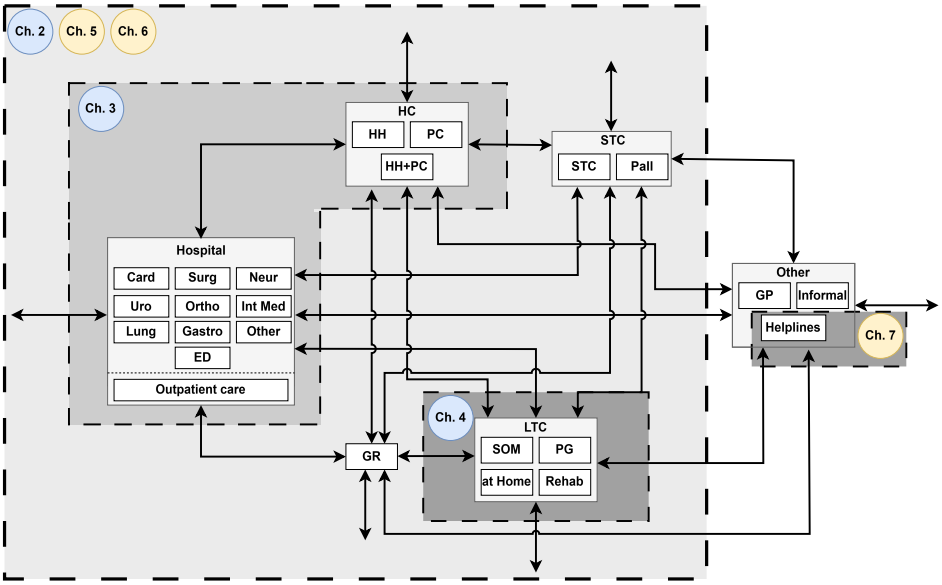


Figure 1.7: Elderly care system visualized, with the included healthcare forms of each chapter highlighted. The light-blue circles correspond to the flow-based chapters, while the light-yellow circles indicate the capacity-based chapters.

This thesis starts with the chapters concerning research from a *flow-based* perspective. Figure 1.7 shows how these chapters are related, but note that these only examine the flow from and to healthcare forms and not the capacity or aspects related to the length of stay.

The goal of Chapter 2 is to gain an understanding of the routes older adults take through the healthcare system. Understanding these longitudinal patterns of healthcare utilization among older adults is essential in designing and improving care coordination across settings. This chapter applies process mining techniques to healthcare usage data, focusing on adults 65 years of age or older. This data is obtained from Statistics Netherlands [139]. The most common care pathways are visualized for various subpopulations, based on age and drug use. The results provide valuable insights into improving healthcare, by showing which care forms are most utilized and which group has the most complex care needs. The study is based on [52].

Chapter 3 aims to examine how different forms of HC are associated with adverse outcomes among older adults. Figure 1.7 shows that this research is primarily concerned with hospital care and HC. However, flows from the hospital and HC to other forms of care are also examined in this chapter. Recognizing the different types of HC can reveal subgroups with varying risk levels, which helps in directing targeted interventions. This is done by examining data that contain the usage of HC in older adults and data on adverse outcomes, such as ED visits, acute hospitalizations, NH admissions and death. The results show that all forms of HC are linked to substantially higher risks than no care, with PC recipients often facing more than twice the risk of hospitalization. Shifts between different types of HC occur often, usually moving towards more intensive care or resulting in death. These findings demonstrate that by differentiating among HC types, we can identify subpopulations with distinct risk profiles. Notably, PC recipients make up a substantial group with significant prevention potential. These results highlight the necessity of tailored interventions to reduce avoidable acute care usage and improve outcomes for vulnerable older adults. This chapter is based on [199, 200].

Chapter 4 explores how the timing of NH admissions can be optimized using a Markov Decision Process (MDP) framework. The health deterioration (and improvements) of older adults is modeled as a Discrete-Time Markov Chain (DTMC), where older adults change in health status, where worse health statuses correspond with a higher risk of a crisis. This DTMC is combined with realistic costs for HC, NH admissions, checks and crises. The model provides an optimal policy that specifies both when to admit patients to an NH and when to monitor their health status. The results indicate that, given current cost-estimates, condition-based monitoring is preferable, with more frequent check-ups for patients in poorer health, and that delaying admission until the final health state can reduce both NH and overall costs. These findings highlight the trade-offs between crisis prevention, monitoring, and institutional care, and demonstrate how data-driven decision models can support more efficient and individualized admission strategies. This chapter is based on [55].

Next, we continue with the chapters concerning research from a *capacity-based* perspective. Chapters 5 and 6 focus on the complete healthcare system at a

macro-level, both introducing a methodology to understand or reduce the effect of bed-blocking, induced by capacity restrictions. Chapter 5 starts by studying tandem queueing systems with zero buffers and After-Service Blocking (ASB), also known as bed-blocking in healthcare settings. Queueing networks with ASB are notably difficult to analyze, as the inclusion of ASB means that no closed-form formulas are available. To address this, this chapter introduces several heuristics that approximate the effects of ASB on system performance. These new heuristics are shown to provide more accurate and computationally efficient estimates than existing methods, making them suitable for use in optimization and real-time decision making. The results demonstrate how such tools can support better capacity allocation and resource management, particularly in settings where budgets and system capacity are constrained. This chapter is based on [53].

Chapter 6 aims to introduce a control mechanism to avoid *deadlocks*. Deadlocks are situations in which patients are directly or indirectly blocked by themselves. These situations are prone to happen in queueing networks with finite buffers, ASB and feedback loops. Such models are also relevant in many other application areas, where bottlenecks and flow disruptions occur frequently. This chapter studies both the *detection* and *resolution* of deadlocks, as a resolution mechanism is essential to prevent the system from becoming indefinitely stuck in a permanent deadlock. The effectiveness of our detection and resolution approach is demonstrated with System Dynamics (SD) and discrete-time simulation models, showing that frequent resolution reduces rejections and blocked servers, especially in systems with limited capacity. A case study on the care system for older adults in Amsterdam further highlights the practical impact, showing that coordinated deadlock resolution can reduce patient rejections by up to 10% and wrong-bed days by up to 20% over a ten-year horizon, thereby improving patient flow without requiring capacity expansion. This chapter is based on [54].

In Chapter 7, the focus shifts to helplines, aiming to determine required staff levels to ensure timely support. Using data from the Dutch 113 Suicide Prevention helpline [170], this chapter combines forecasting and queueing models to improve operational efficiency. A detailed simulation model is developed and validated using trace-driven simulation, capturing features such as triage, abandonment, and multi-phase service durations. Furthermore, daily and hourly call and chat volumes are forecasted using models including (S)ARIMA, linear regression, and LSTM, showing that (S)ARIMA performs best short-term, while linear regression suits longer horizons. Counselor feedback is integrated to link perceived workload to objective measures. Together, these tools support dynamic staffing to reduce delays and improve counselor well-being. The chapter is based on [56, 57].

1.4 Publications of the Author

- de Boer, T. R., Arntzen, R. J., Bekker, R., Buurman, B. M., Willems, H. C., & van der Mei, R. D. (2025). Process mining on national healthcare data for the discovery of patient journeys of older adults. *Journal of the American Medical Directors Association*, 26(1), 105333. [52]
- Smeekes, O. S., de Boer, T. R., van der Mei, R. D., Buurman, B. M., & Willems, H. C. (2024). Differentiating between home care types to identify older adults at risk of adverse health outcomes in the community. *Journal of the American Medical Directors Association*, 25(11), 105257. [199]
- Smeekes, O. S., de Boer, T. R., van der Mei, R. D., Buurman, B. M., & Willems, H. C. (2024). Receiving home care forms and the risk for emergency department visits in community-dwelling Dutch older adults, a retrospective cohort study using national data. *BMC Public Health*, 24(1), 1792. [200]
- de Boer, T. R., Bekker, R., van der Mei, R. D., & van den Nieuwenhof, J. M. (2025). Too soon or too late? A Markov decision process approach for preventive nursing home admission. *Submitted for publication.* [55]
- de Boer, T. R., Bekker, R., & van der Mei, R. D. (2024). After-service blocking in tandem queues. *In Proceedings of Valuetools* (Milan, Italy, December 2024). [53]
- de Boer, T. R., Bekker, R., & van der Mei, R. D. (2025). Deadlock detection and resolution in queueing networks with blocking. *Submitted for publication.* [54]
- de Boer, T. R., Mérelle, S., Bhulai, S., Gilissen, R., & van der Mei, R. D. (2023). Forecasting call and chat volumes at online helplines for mental health. *BMC Public Health*, 23(1), 984. [56]
- de Boer, T. R., Mérelle, S., Bhulai, S., & van der Mei, R. D. (2022). Performance modeling for call centers providing online mental health support. *International Journal on Advances in Life Sciences*, 14(3&4), 120. [57]
- de Boer, T. R., Mérelle, S., Bhulai, S., & van der Mei, R. D. (2022). A call center model for online mental health support. *In Proceedings of International Conference on Prediction Solutions for Technical and Societal Systems* (Nice, France, July 2022). [51]

Part I

Flow-Based Perspective

2

Patient Journeys of Older Adults in the Healthcare System

Based on:
de Boer, T. R., Arntzen, R. J., Bekker, R., Buurman, B. M., Willems, H. C., & van der Mei, R. D. (2025). Process mining on national healthcare data for the discovery of patient journeys of older adults. *Journal of the American Medical Directors Association, 26*(1), 105333.

Abstract

It is crucial to comprehend the longitudinal patterns of healthcare utilization among older adults for improving patient experiences and strengthening coordination across various care settings. This chapter explores the most common care pathways among older adults in the Netherlands by applying process mining techniques to national healthcare data from 2017 to 2019. The analysis focused on all residents aged 65 and older as of January 1, 2017, using healthcare declarations from multiple care domains over the three-year period, including only journeys with clearly defined starting points. After extensive preprocessing and merging, a unified event log was created for process mining, and patients were categorized by age and drug use to identify differences in care trajectories. The resulting visualizations illustrated how care types were connected and how changes in one affected others. Among the 3,177,203 individuals aged 65 and older, 44% experienced one or more patient journeys, totaling 2,469,663 journeys. Most were short and simple: 95% of journeys for those aged 65+ and 90% for those 85+ involved four or fewer care transitions. Chronic care forms, such as home care, personal care, and institutional long-term care, accounted for the majority of time spent in care. The results show that while most older adults follow relatively straightforward care paths involving emergency department visits and hospitalizations, a smaller, older subgroup requires more complex and prolonged care. Although this group is smaller, reducing the number of transitions they experience could lead to significant improvements in the healthcare system.

2.1 Introduction

The demographic challenges discussed in Chapter 1 raise the urgency for a better understanding of the healthcare usage of older adults. To make any improvements in the system, first an understanding of how the system is currently utilized is required. Such insights can be obtained from Process Mining (PM), a method that uses advanced algorithms to analyze large quantities of data. These analyses are presented as visualizations of patient journey patterns, which show which healthcare forms are mainly utilized and in which order. This helps us to understand how patients flow through the healthcare system. PM enables us to anticipate on coming challenges, such as the aging population. Thus far, PM in healthcare has been used on healthcare datasets with a smaller scope, such as a hospital [141, 251]. However, PM has not been applied to healthcare in a system-wide manner. For that, a larger complete dataset is required. Statistics Netherlands collects data on care declarations of all inhabitants of the Netherlands [139], which enables us to map the complete care system.

In this chapter, we address the following two research questions: (a) What are the most frequently occurring patient journeys within the healthcare system of older adults? (b) Is there a difference in healthcare usage between various age groups and groups with different cumulative drug use? The hypothesis is that older adults have long and intricate patient journeys, since most research often reports either Long-Term Care (LTC) pathways [105] or depict the older adults as a frail population with recurrent healthcare usage [42]. It is expected that the length and complexity of these journeys will only increase as the age or drug use of the sub-populations increases, since age [10] and drug use [182] are both shown to correlate with an increase in healthcare usage and frailty, and drug use is often seen as proxy for frailty.

The remainder of this chapter is organized as follows. Section 2.2 outlines the study's design, methodology, and stratification criteria. Section 2.3 presents the findings, which include demographic data and details of patient journeys. Section 2.4 provides a discussion of the study, comparing it with previous research, and highlighting its strengths, limitations, and suggestions for future research. Lastly, Section 2.5 offers a conclusion to this chapter.

2.2 Methods

This section describes the data sources, preprocessing steps, and analytical methods used to study healthcare utilization trajectories of older adults in the Netherlands. We outline the study design, the construction of patient journeys, the application of process mining techniques, and the stratification of the population by age and polypharmacy.

2.2.1 Study design

For this study, we used data that contain the usage of healthcare in older adults in the Netherlands from 2017 to 2019. Throughout, we use the following definition of a *patient journey*: A sequence of transitions between different forms of care, allowing for a maximum gap of four weeks without professional healthcare. Healthcare usage is defined as older adults utilizing professional healthcare, ranging from home care to hospital care. This four-week limit is based on research indicating that healthcare events within this timeframe are often connected [234]. The dataset utilized in this study originates from multiple healthcare datasets provided by Statistics Netherlands (CBS) [139][1]. Permission was obtained to access the microdata via a secured portal.

The dataset includes the following forms of care: Household Help (HH), Personal Home Care (PC), LTC at home or in an institution (NHH or NH), Geriatric Rehabilitation (GR), Short-Term Residential Care (distinction made between non-palliative and palliative, STC and STC-pall. respectively), Emergency Department visits (ED) and Hospitalizations (HOS). No care at home (HOME) is used to define that a person used no declared care. Since the LTC reforms of 2015 [246], several types of long-term and home care are paid using different budgets. The municipality is responsible for HH, this includes household help, personal support, and low-intensity personal care. PC comprises of personal care based on medical diagnoses, this is provided by registered nurses and paid by health insurance companies under the Healthcare Insurance act. GR encompasses specialized interventions tailored to address the multifaceted needs of frail older adults [77] and is also paid by Healthcare Insurance Act. LTC can either be provided at home (NHH) or in a nursing home (NH), this care includes all-day-long care by a healthcare professional and is paid by the Long Term Care Act. Several healthcare activities fall beyond the scope of this research, for example the GP, this choice was made due to constraints in data availability and the need for clarity in large-scale analysis. Data limitations restricted access to comprehensive information for all older adults across every healthcare activity, and incorporating all activities would introduce numerous shorter paths relevant on a small scale but not aligned with the large-scale focus of this study. Preprocessing of the datasets was necessary to create a single event file suitable for PM [28].

The dataset comprises data from 3,177,203 individuals aged 65 or older on January 1, 2017. Data from 2017 to 2019 was selected to reflect the healthcare system, due to anomalies related to COVID-19 in 2020 [152] and the healthcare reform in 2015 [246]. Each row contains: an activity, start date and end date. Rows combined are seen as patient journeys, an example of such a patient journey could be the following: a patient has ED, HOS and returns home with PC. This journey then counts 3 care transitions and forms. The duration of this journey is then the length from start date of ED until the end date of PC. In addition, people could

[1]The results are based on calculations by Centrum Wiskunde & Informatica using non-public microdata from Statistics Netherlands.

have multiple forms of care at the same time. Periods with multiple care forms were handled by removing overlap based on a ranking in which the most intense care form remains. The ranking is based on intensity in the following order: Death, ED, HOS, GR, STC, NH, NHH, PC and HH.

The data may contain care journeys that have begun before 2017. We decided to include only those journeys where we can be sure there was a minimum of four weeks without care prior to the first recorded healthcare event. As a result, only care journeys starting after February 1, 2017, are taken into account. It is certain that these journeys are not continuations from journeys that have started before 2017 and therefore, are certain we capture the beginning of the healthcare pathways. These new journeys are characterized by their number of care transitions (i.e. change of care form utilized by the patient), the used healthcare types, and the time spent per journey.

2.2.2 Process mining

The aim of PM [228] is what is called *process discovery* [229], i.e. discovery of processes or paths in a system. This is done by analyzing log files using the R-library named bupaR [28]. We first generate a table containing the number of persons and patient journeys of each sub-population, after which we generate a table with the most frequently occurring patient journeys of each sub-population. The journeys are then examined further, as we take a closer look at the distribution of the complexity of these journeys using a histogram and the time length using Lorenz curves. Finally, process maps are generated, which can provide a picture of the complexity of the taken care journeys of each group. These process maps are set to have a coverage of 0.8, meaning that 80% of the journeys are visualized. Increasing the coverage to 1.0 often creates maps containing a superfluous amount of flows, so-called 'spaghetti maps' [230], which are too detailed to be informative.

2.2.3 Age and polypharmacy

Finally, the population can be split based on age and drug use. Specifically, the age ranges 65-74, 75-84 and 85+ are distinguished, where the age of the person on January 1, 2017, is used. Similarly, a distinction is made based on drug use of a person in 2017, as drug use can be seen as a proxy for frailty [182]. The population is divided into people who use 0-4 distinct types of drugs, 5-9, or 10+. These sub-groups enable comparisons between different demographic and drug profiles, with sample sizes of 100,000 persons per sub-population facilitating meaningful comparisons. Comparisons will be done between the 65+ and the 85+ population in this chapter, results for other sub-groups are provided in Appendix 2.A.

2.3 Results

This section presents the main findings of this chapter, starting with a description of the study population, followed by an overview of the most frequent care journeys,

their characteristics, and the corresponding process maps.

2.3.1 Population data

Table 2.1 provides an overview of the population used in this analysis. The majority of the population is relatively young: 58% is aged between 65 and 74, and has little drug use, only 21% has a drug use of ten or more drugs. The fraction of persons with a care journey increases with age and drug use, but quite surprisingly, the older sub-population of 85+ consists of fewer patients with a care journey (44% compared to 52% in the patient group 75-84 years of age). This can be explained by the fact that many older adults aged 85 or older were already in a care journey in January 2017. In this chapter, these journeys are omitted, which can be seen in the last column of Table 2.1, here it is also visible that the 85+ subgroup has relatively more omitted journeys, compared to the 75-84 subgroup. However, it should be noted that the excluded 612,470 journeys do not correspond to 612,470 excluded persons, but only with 412,261 excluded persons, as people with excluded journeys could also have an included journey. We also see that in terms of mean journey length the sub-population of 85+ differs most from the complete population, with an average of 2.45 care transitions compared to 2.06 overall. This difference also is clear when comparing journeys with a similar starting time, which can be seen in Table 2.A.1 in Appendix 2.A.

Group	All	Age			Drug use		
		65–74	75–84	85+	0–4	5–9	10+
Number of persons (%)	3,177,203 (100%)	1,816,172 (58%)	969,814 (31%)	357,172 (11%)	1,381,582 (44%)	1,122,587 (35%)	673,034 (21%)
Number of patient journeys[2]	2,469,663	1,241,416	936,401	273,261	605,638	948,280	915,745
Persons with a journey (%)	44	39	52	44	30	49	64
Mean number of journeys per person	1.78	1.73	1.86	1.75	1.48	1.72	2.14
Mean journey length (in care transitions)	2.06	1.89	2.18	2.45	1.81	1.96	2.33
Number of excluded patient journeys	612,470	150,243	236,985	223,212	160,857	189,503	262,110

Table 2.1: Population size as of January 1, 2017.

[2]Journeys are included only if their starting point is clearly identifiable, meaning that only journeys not present in January 2017 are considered. Individuals may have both excluded and included journeys; frailer sub-populations tend to have a relatively higher proportion of excluded journeys.

2.3.2 Frequently occurring journeys

Figure 2.1 illustrates the ten most frequent patient journeys for the 65+ and 85+ populations, respectively. In the supplementary material, the ten most frequently patient journeys of all (sub-)populations are shown in Tables 2.A.2-2.A.8. In both the figures and tables, many of these journeys involve less than five care transitions. The most frequently occurring journey for each (sub-)population is in all cases a journey containing only one healthcare form, often the journey consisting of an ED visit and afterwards returning home without professional healthcare. This can be seen in Figure 2.1a, where we see that this specific journey occurs 24,350 times in a sample of 100,000 persons. However, for the 85+ sub-population, the most frequently occurring journey is patients having PC and returning to home without care, also visible in Figure 2.1b, which occurs 12,008 times in our sample of 100,000 persons. The importance of ED and the hospital is evident in the top three journeys of the complete 65+ population, as they can be described by only the hospital and ED. This can also be seen in other top ten journeys of other sub-populations.

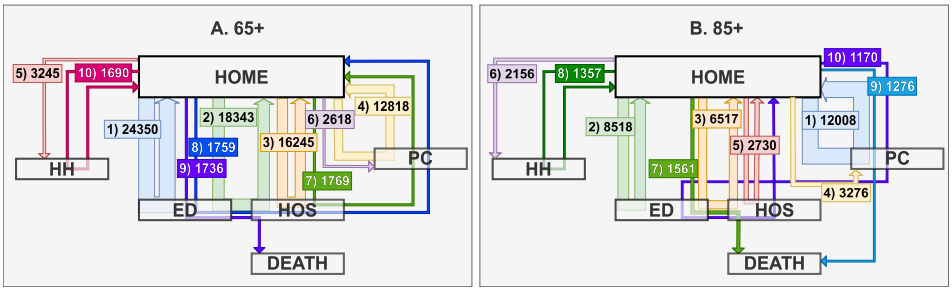


Figure 2.1: The ten most frequently occurring patient journeys for the (A) 65+ and (B) 85+ population (taken from a sample of 100,000 persons). It shows how often the specific journeys occur within the sample.

2.3.3 Care transitions and time distribution

We observe that by far most care usage of older adults is simple and short when looking at the distribution of the number of care transitions per journey, see Figure 2.2a (see Figure 2.A.1 for the distribution of all sub-populations). Both the complete population and the older sub-population exhibit a right-tailed distribution, with 59% of journeys for the 65+ population and 53% for the 85+ population involving only one care form. Additionally, 95% of journeys for the 65+ population and 90% for the 85+ population have four or fewer care transitions. Figure 2.2b visualizes the amount of time spent in the healthcare system using a Lorenz curve (see Figure 2.A.2 for the Lorenz curves of all sub-populations). In the 65+ group the 90% shortest journeys account for only 27% of the total time. Hence, the 10% of the longest journeys account for 73% of the total time spent in the system, these 10% longest journeys consist mostly of care forms aimed at chronic care, such as HH, PC, LTC, nine of the ten most frequently occurring patient journeys of this

10% consist solely of HH, PC, LTC. Among those aged 85 and over, the disparity in time is less pronounced.

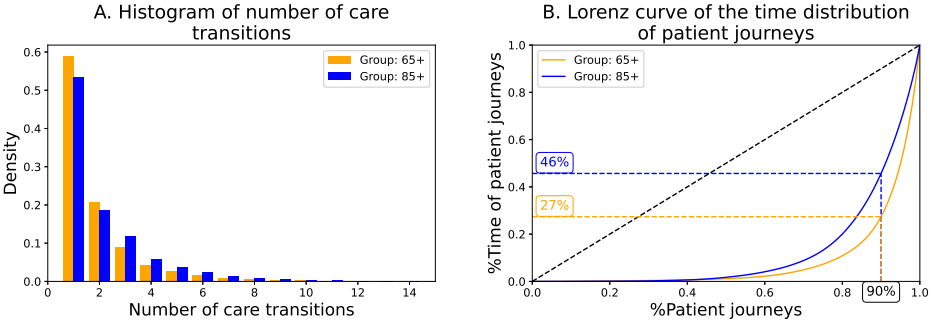


Figure 2.2: (A) Exploring variations in care transition frequency distribution: a comparative analysis between age groups 65+ and 85+ and (B) a Lorenz curve of patient journey time allocation between age groups 65+ and 85+. The curve illustrates the cumulative distribution of time spent across patient journeys, highlighting differences in healthcare utilization patterns among older adults.

2.3.4 Process maps

Figure 2.3 presents process maps for individuals aged 65 or older, while Figure 2.4 displays maps for those aged 85 or older, the maps for the other sub-populations can be found in Figures 2.A.3-2.A.7. These maps offer a complete view of healthcare usage patterns within these age groups. It shows that while the majority of both groups have a short and straightforward healthcare journey (as can be seen in Figures 2.1a and 2.1b). However, by comparing the most common journeys that together provide 80% coverage of the 65+ population and that of the 85+ population, it can be seen that the older sub-population has a more complex path (see Figures 2.3 and 2.4). Notably, the process map for the 85+ subgroup, with the same coverage (0.8), exhibits more transitions and a less structured layout compared to the map for the 65+ population, indicating that healthcare usage among the older age group is less predictable and cannot be easily categorized into distinct pathways. This disparity is visually depicted by the 'lasagna process' appearance of the map for the 65+ population, contrasting with the 'spaghetti process' observed in the map for individuals aged 85 or older.

2.4 Discussion

This is the first study to identify patient journeys over a large-scale healthcare system, using PM. It was found that most frequently occurring patient journeys of older adults contain either one or two and often not more than four care transitions. Similar to the simplicity it was found that most journeys span a relatively

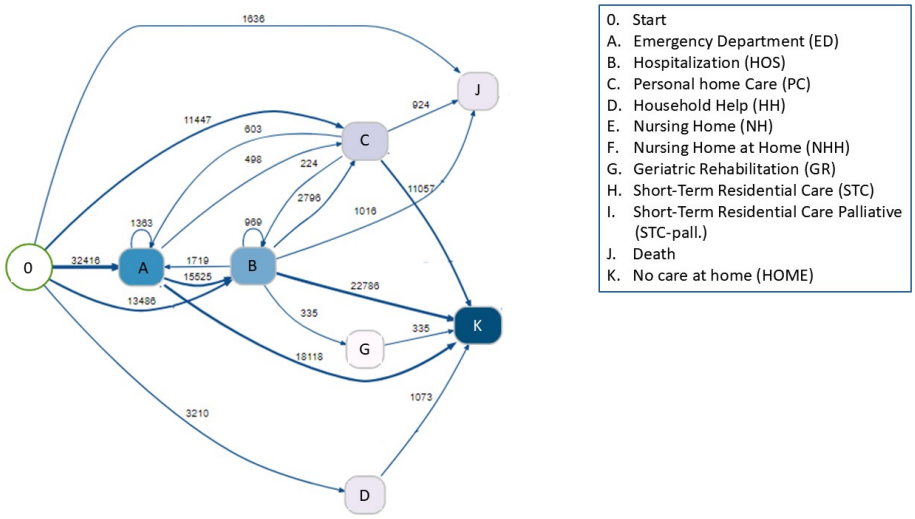


Figure 2.3: Process map of 80% of the most frequently occurring journeys of the 65+ population (taken from a sample of 100,000 persons), the number above the lines show how often that transition occurs in the sample, given an 80% coverage.

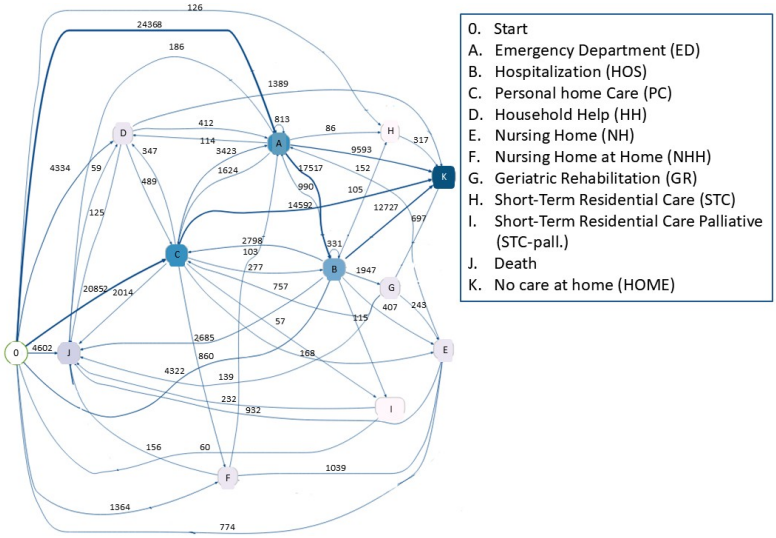


Figure 2.4: Process map of 80% of the most frequently occurring journeys of the 85+ population (taken from a sample of 100,000 persons), the number above the lines show how often that transition occurs in the sample, given an 80% coverage.

short amount of time, visible in our Lorenz curve. This confirms the previous observation that most older adults have simple and short healthcare journeys. Only a small part of the population covers more than half of the total time spent on

all journeys, implying that this group requires more coordination in the system. Noteworthy is that this inequality diminishes for the 85+ population, this can be contributed to the fact that healthcare journeys that have started before February 1, 2017, are not included, and this sub-population often utilizes LTC, either at home or in an institution.

Another interesting observation is the consistent presence of ED visits or hospitalizations in nearly all frequently occurring journeys across all (sub)populations. This underscores the important role of the ED and hospital as entry points into the healthcare system for older adults. Therefore, proper healthcare at the ED is crucial and strategies to avoid ED visits should be studied. Additionally, it was observed that hospitalizations often followed ED visits, indicating the acute nature of these hospital admissions. Streamlining this transition is crucial given the prevalence of ED in healthcare journeys, as inefficiencies in ED transfers can limit its effectiveness and lead to more care utilization later in the patient journey [19, 30].

Notably, the process maps of the older sub-population 85+ depicted a more complex map than that of the population 65+, suggesting a greater diversity of healthcare journeys among older adults aged 85 and above. As the population of the Netherlands is aging more in the future [168], the strain on the system will rise further. More specifically, it is expected that the number of older adults aged 80 or above also increases over the coming years, for which we show that this group is harder to capture using a few transitions. This makes it more difficult to predict, anticipate and coordinate what journey an older adult aged 85 or older will embark on in the system. Simplifying these journeys could minimize efficiency lost in possible transfers, as transfers can be the source of a higher care demand, inefficiencies or dangers [192]. Integrated care for this population might be essential to limit the number of transfers [99]. Additionally, integrated care can enhance continuity of care, reduce redundancies, and improve overall patient outcomes by ensuring that all aspects of an older adult's health needs are managed cohesively, which is especially important for 85+. For this sub-population four out of the ten most frequently occurring journeys include PC, providing policy makers with a clear entryway into how to target this part of the population.

2.4.1 Strengths and limitations

This research distinguishes itself from previous research into patient journeys, by providing a scope on the whole healthcare system, while previous research often focus on either a part of the healthcare system, e.g. the hospital, and only a small sample [87, 95, 222]. Previous research on patient journeys of healthcare usage is often focused on designing the patient journey [17, 214], while this chapter focuses on first understanding what the patient journeys of older adults are, since the system should first be understood before adaptions can be made.

The authors acknowledge that this research has also some limitations and therefore the results should be used with caution. An example of a limitation is the fact that

it is centered around the Dutch healthcare system. However, given the similarities in population demographics and healthcare structures between the Netherlands and countries in Northern Europe [76] and Canada [112], the findings are also applicable to these regions. In contrast, the markedly different healthcare systems in the USA [221] and Asia [254] present challenges for direct application of these results, necessitating further investigation to determine their relevance in these contexts. Another limitation is that the analysis only includes patient journeys that begin after January 2017, which may exclude the longest and most intense patient journeys that started earlier. Additionally, we are limited to data from 2017 to 2019, potentially missing longer patient journeys that extend beyond this period. Furthermore, while the analysis differentiates between age and drug use groups, it does not account for other important characteristics that could provide a more nuanced understanding of patient journeys. This limitation suggests that future research could benefit from incorporating a broader range of variables.

2.4.2 Future work

Some research opportunities fell beyond the scope of this chapter and could be topics of future research. Previous research done by the NZa [150] that focuses on the healthcare usage 100 days preceding death or nursing home admittance shows that these journeys are often different and visualize a different part of the population. This research approach is promising and could be examined further, especially given the importance of ED visits and hospitalizations. Another opportunity for follow-up research is to include other groups of the population and a longer time window. Future research could also focus on different factors such as patient comorbidity or geographical region, here the comparison between different regions with different types of healthcare is especially interesting, for example, how does the proximity of ED [18] affect the patient journeys from patients. Another interesting aspect could be the inclusion of overlap, in this chapter the authors chose to remove overlap of different care forms. However, overlap of care forms could signal that the provided care from one care form is insufficient and this could characterize a specific sub-population.

2.5 Conclusion and Implications

In conclusion, this study employs PM on a dataset from CBS to unravel the complexity of patient journeys of older adults in the Netherlands, addressing critical challenges in healthcare of older adults. The research aims to provide insights that can inform policymakers for more effective healthcare decisions, providing an overview of which and how different forms of care are used by the older adult's population of the Netherlands.

In summary, our study highlights how most older adults tend to have straightforward healthcare needs, often involving short visits to the emergency department or brief hospital stays. However, a smaller group among them requires more complex

and prolonged care. This suggests that when designing or adapting new healthcare policies, focusing solely on the complex cases might overlook the majority who rely on simpler healthcare services. Therefore, it is important for policymakers to consider the broader picture, including routine emergency and hospital visits, to ensure that any changes made benefit the majority of older adults accessing healthcare. However, the 85+ population has a longer and more complex healthcare pathway. This implies that reducing the number of care transitions or journey length for only this group, might result in a larger effect on the system, due to the interconnectivity of various forms of healthcare that this group requires.

2.A Appendix

This appendix provides additional analyses and visualizations that complement the main results presented in Chapter 2. It includes detailed tables on journey lengths and the most frequently occurring patient journeys for various sub-populations, as well as extended figures illustrating care transitions and process maps for different age and drug use groups.

2.A.1 Length of journeys with a similar starting time

Table 2.A.1 shows a comparison of mean journey length per sub-population of journeys starting in a specific month.

Group		**Mean journey length of journeys starting in 2017:**					
		Feb	**Mar**	**Apr**	**May**	**Jun**	**Jul**
All		2.18	2.14	2.19	2.17	2.16	2.17
Age	**65-74**	1.90	1.89	1.95	1.92	1.93	1.94
	75-84	2.35	2.30	2.32	2.34	2.31	2.31
	85+	2.85	2.75	2.77	2.72	2.67	2.70
Drug use	**0-4**	1.75	1.67	1.66	1.66	1.65	1.66
	5-9	1.97	1.91	1.93	1.93	1.92	1.94
	10+	2.55	2.52	2.59	2.54	2.54	2.53

Table 2.A.1: Mean journey length of journeys starting in specific months.

2.A.2 Frequently occurring patient journeys

The ten most frequently occurring journeys of each sub-group are given in Tables 2.A.2-2.A.8.

Journey	Group: All				
	1st Activity	2nd Activity	3rd Activity	4th Activity	Number of journeys
1st	ED	HOME			24,350
2nd	ED	Hos	HOME		18,343
3rd	Hos	HOME			16,245
4th	PC	HOME			12,818
5th	HH				3,245
6th	PC				2,618
7th	Hos	PC	HOME		1,769
8th	ED	Hos	PC	HOME	1,759
9th	ED	Hos	Death		1,736
10th	HH	HOME			1,690

Table 2.A.2: The ten most frequently occurring patient journeys for 65+ population (taken from a sample of 100,000 persons).

Journey	Group: Aged 65-74				
	1st Activity	2nd Activity	3rd Activity	4th Activity	Number of journeys
1st	ED	HOME			17,862
2nd	Hos	HOME			12,460
3rd	ED	Hos	HOME		9,342
4th	PC	HOME			5,202
5th	HH				1,539
6th	ED	ED	HOME		1,085
7th	Hos	PC	HOME		1,040
8th	PC				869
9th	HH	HOME			830
10th	ED	Hos	PC	HOME	715

Table 2.A.3: The ten most frequently occurring patient journeys for population aged 65–74 (taken from a sample of 100,000 persons).

Journey	Group: Aged 75–84				
	1st Activity	2nd Activity	3rd Activity	4th Activity	Number of journeys
1st	ED	HOME			16,805
2nd	PC	HOME			13,543
3rd	ED	Hos	HOME		10,695
4th	HH				3,361
5th	PC				2,825
6th	HH	HOME			1,512
7th	Hos	PC	HOME		1,482
8th	ED	Hos	PC	HOME	1,446
9th	ED	Hos	Death		1,098
10th	PC	ED	Hos	HOME	939

Table 2.A.4: The ten most frequently occurring patient journeys for population aged 75–84 (taken from a sample of 100,000 persons).

Journey	Group: Aged 85+				
	1st Activity	2nd Activity	3rd Activity	4th Activity	Number of journeys
1st	PC	HOME			12,008
2nd	ED	HOME			8,518
3rd	ED	Hos	HOME		6,517
4th	PC				3,276
5th	Hos	HOME			2,730
6th	HH				2,156
7th	ED	Hos	Death		1,561
8th	HH	HOME			1,357
9th	PC	Death			1,276
10th	PC	ED	Hos	HOME	1,170

Table 2.A.5: The ten most frequently occurring patient journeys for population aged 85+ (taken from a sample of 100,000 persons).

Journey	Group: Drug use 0–4				
	1st Activity	2nd Activity	3rd Activity	4th Activity	Number of journeys
1st	ED	HOME			10,870
2nd	Hos	HOME			6,447
3rd	PC	HOME			5,067
4th	ED	Hos	HOME		4,780
5th	HH				1,278
6th	PC				1,048
7th	Hos	PC	HOME		577
8th	HH	HOME			567
9th	ED	Hos	PC	HOME	553
10th	ED	ED	HOME		480

Table 2.A.6: The ten most frequently occurring patient journeys for population with medicine use between 0 and 4 (taken from a sample of 100,000 persons).

Journey	Group: Drug use 5–9				
	1st Activity	2nd Activity	3rd Activity	4th Activity	Number of journeys
1st	ED	HOME			18,315
2nd	Hos	HOME			11,792
3rd	PC	HOME			10,059
4th	ED	Hos	HOME		9,682
5th	HH				2,608
6th	PC				2,070
7th	Hos	PC	HOME		1,284
8th	HH	HOME			1,197
9th	ED	ED	HOME		1,009
10th	ED	Hos	PC	HOME	994

Table 2.A.7: The ten most frequently occurring patient journeys for population with medicine use between 5 and 9 (taken from a sample of 100,000 persons).

Journey	Group: Drug use 10+				
	1st Activity	2nd Activity	3rd Activity	4th Activity	Number of journeys
1st	ED	HOME			24,165
2nd	ED	Hos	HOME		18,240
3rd	Hos	HOME			16,406
4th	PC	HOME			12,955
5th	HH				3,288
6th	PC				2,633
7th	Hos	PC	HOME		1,822
8th	HH	HOME			1,790
9th	ED	Hos	Death		1,777
10th	ED	ED	HOME		1,728

Table 2.A.8: The ten most frequently occurring patient journeys for population with medicine use 10+ (taken from a sample of 100,000 persons).

2.A.3 Care transitions and time distribution

For the distribution of care transitions and the Lorenz curve for the time spent in the system figures with comparisons of all sub-groups were also made and can be found in Figures 2.A.1 and 2.A.2.

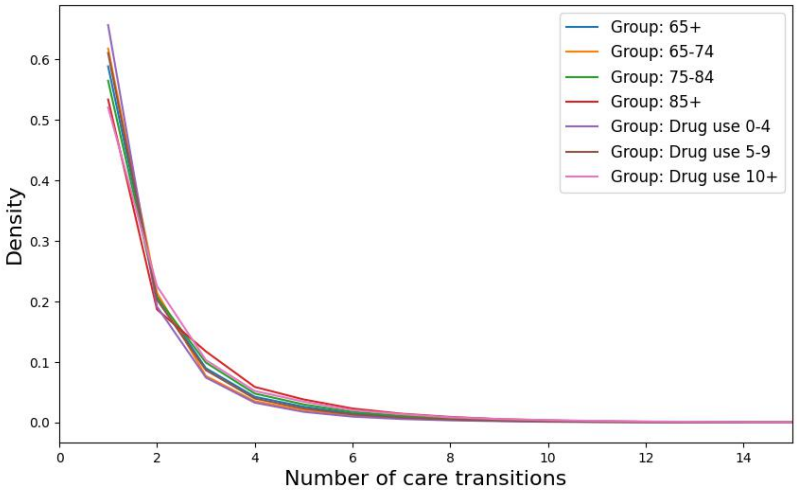


Figure 2.A.1: Exploring variations in care transition frequency distribution: a comparative analysis between (sub-)populations.

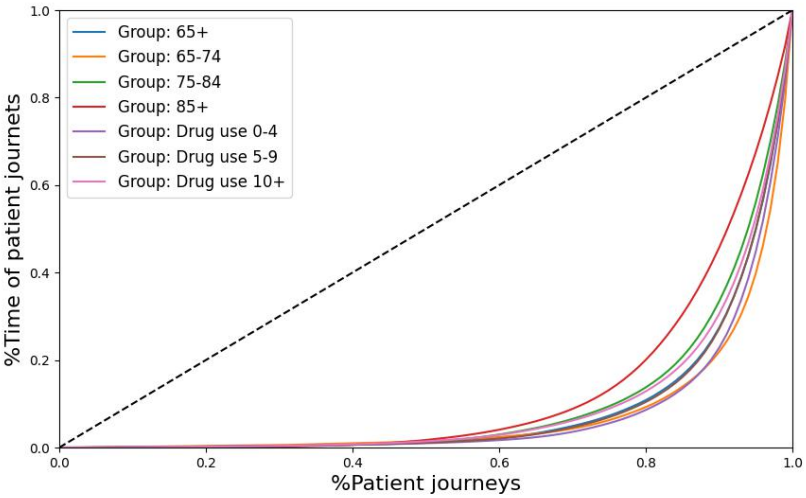


Figure 2.A.2: Interpreting the Lorenz curve: a comparative analysis of patient journey time allocation between (sub-)populations. The curve illustrates the cumulative distribution of time spent across patient journeys, highlighting differences in healthcare utilization patterns among older adults.

2.A.4 Process maps

The remaining process maps of other sub-groups can also be found in Figures 2.A.3 till 2.A.7.

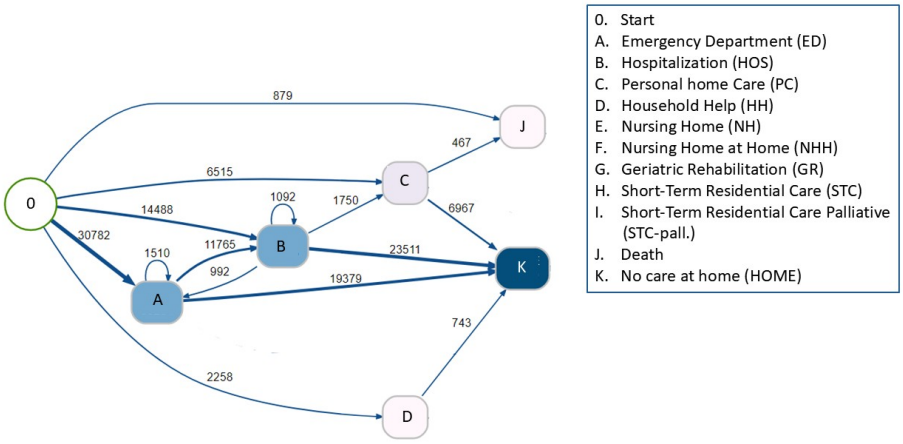


Figure 2.A.3: Process map of 80% of the most frequently occurring journeys of the 65-74 population (taken from a sample of 100,000 persons), the number above the lines show how often that transition occurs in the sample, given an 80% coverage.

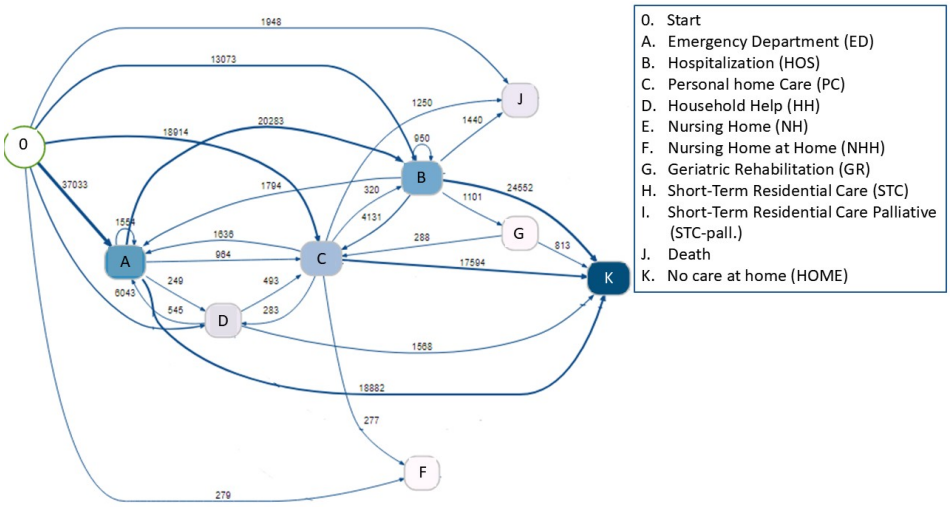


Figure 2.A.4: Process map of 80% of the most frequently occurring journeys of the 75-84 population (taken from a sample of 100,000 persons), the number above the lines show how often that transition occurs in the sample, given an 80% coverage.

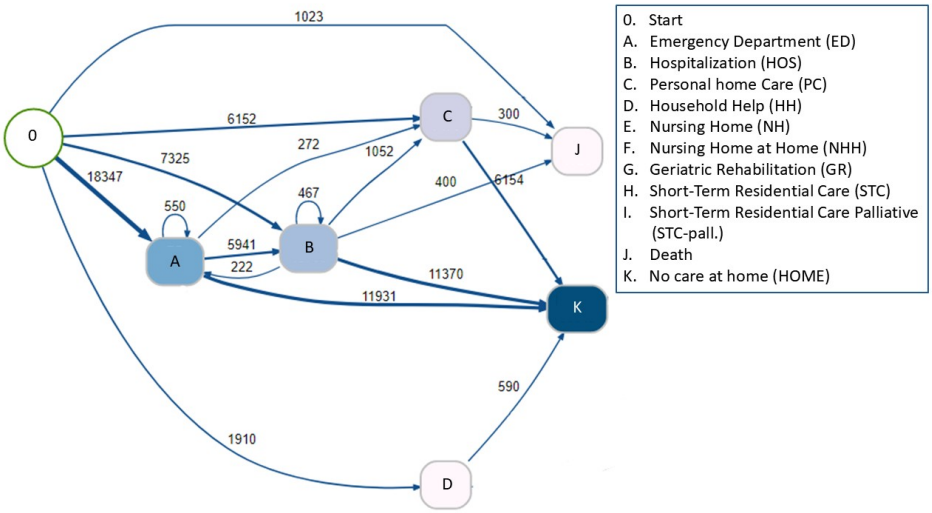


Figure 2.A.5: Process map of 80% of the most frequently occurring journeys of the 65+ population with medicine use between 0 and 4 (taken from a sample of 100,000 persons), the number above the lines show how often that transition occurs in the sample, given an 80% coverage.

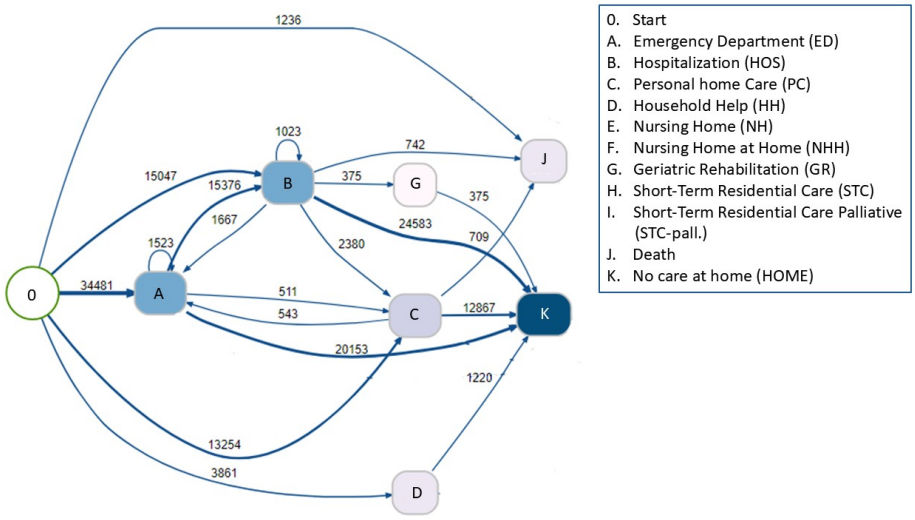


Figure 2.A.6: Process map of 80% of the most frequently occurring journeys of the 65+ population with medicine use between 5 and 9 (taken from a sample of 100,000 persons), the number above the lines show how often that transition occurs in the sample, given an 80% coverage.

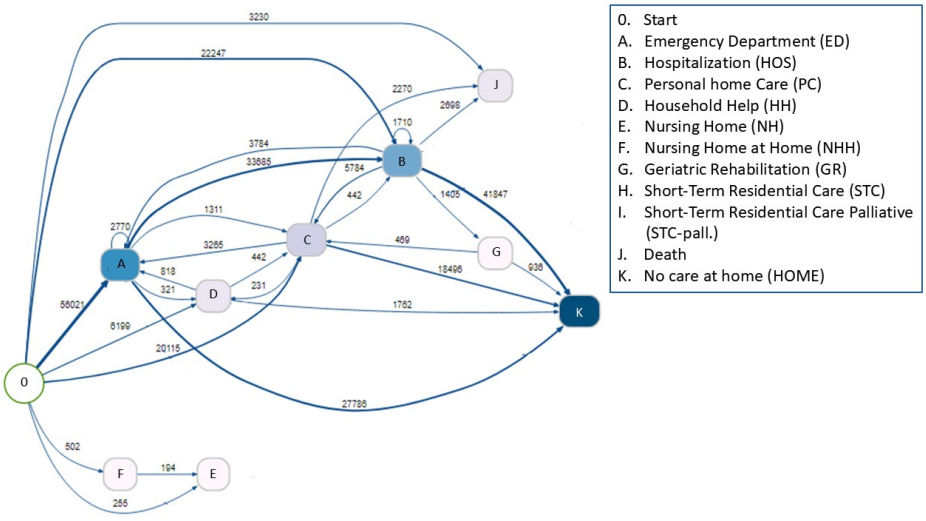


Figure 2.A.7: Process map of 80% of the most frequently occurring journeys of the 65+ population with medicine use higher than or equal to 10 (taken from a sample of 100,000 persons), the number above the lines show how often that transition occurs in the sample, given an 80% coverage.

3

Differentiating Home Care Types and Their Association with Adverse Health Outcomes in Older Adults

Based on:
Smeekes, O. S., de Boer, T. R., van der Mei, R. D., Buurman, B. M., & Willems, H. C. (2024). Differentiating between home care types to identify older adults at risk of adverse health outcomes in the community. *Journal of the American Medical Directors Association, 25*(11), 105257.

Smeekes, O. S., de Boer, T. R., van der Mei, R. D., Buurman, B. M., & Willems, H. C. (2024). Receiving home care forms and the risk for emergency department visits in community-dwelling Dutch older adults, a retrospective cohort study using national data. *BMC Public Health, 24*(1), 1792.

Abstract

Older adults receiving home care are at increased risk of emergency department visits, acute hospitalization, institutionalization and death compared with community-dwelling older adults not receiving home care. These risks may be partly explained by greater comorbidity burdens and reduced functional autonomy among home care recipients. As different forms of home care are provided according to varying health and autonomy profiles, distinguishing between these forms may reveal sub-populations with different risk levels and guide targeted interventions. We conducted a retrospective cohort study using 2019 Dutch national claims data on individuals aged 65+ years. Participants were categorized as receiving no home care, household help, personal care, household help combined with personal care, or nursing home care at home. Outcomes included ED visits, acute hospitalization, recurrent admissions, institutionalization, death, and changes in home care type. Logistic regression models were applied to compare risks between groups. All home care groups showed higher risks of ED visits and adverse outcomes compared with the no home care group, with risk increases often exceeding two-fold for personal care recipients. In all groups, shifts in home care were common, often involving more intensive care or mortality. Differentiating between types of home care identifies sub-populations with distinct risks for ED visits and other adverse outcomes. Personal care recipients represent a large group with considerable prevention potential. These findings highlight the need for tailored interventions to mitigate avoidable acute care use and improve outcomes in vulnerable older adults

3.1 Introduction

As the population ages with increasing comorbidity, more older adults are presenting at the Emergency Department (ED) [3, 142, 175]. In the Netherlands, 33% of ED visits are by older adults [142], and this is expected to increase further, placing a burden on patients, society, and the health system as a whole [3, 218]. Older peoples' ED visits differ from younger people because these visits are more often a result of comorbidities and geriatric syndromes, require more time, diagnostics and hospital admittances [3]. ED's are not designed to properly address these care needs, which leads to further health deterioration, leaves more care needs unmet, creates a need for more intense long-term care, and increases healthcare costs [23, 63, 76, 98, 166, 225, 226]. ED visits are part of the adverse outcomes we examined in this chapter, other adverse outcomes are acute hospitalization, recurrent admissions, institutionalization, and death. These are important adverse health outcomes in older adults [153, 225, 226]. Such adverse outcomes present challenges to many countries on a population level because they increase the need for care in societies with rapidly aging populations and limited healthcare personnel [63, 166].

Multiple risk factors have been associated with the adverse health outcomes in older adults, such as geriatric syndromes, multi-morbidity, low socioeconomic status, and living alone [31, 33, 70, 71, 107]. Therefore, finding ways to identify older adult populations at risk for adverse outcomes, is essential to develop targeted interventions. Where detailed information on comorbidities, geriatric syndromes and functional autonomy helps identify older adults at the individual level, it is difficult to obtain this information on a population level. The use of home care reflects the formal support needed to address these comorbidities, geriatric syndromes and loss of functional autonomy in the community and its data is often more easily accessible on a population level. Older adults receiving home care are particularly prone to these risk factors [31, 83], but few studies have investigated the effect of different types of home care on the risk of adverse health outcomes.

National data from the Dutch healthcare authority in January 2016 showed that the probability of older adults who received any type of home care visiting the ED was 30% compared with 9-25% in older adults not receiving home care and 17% in adults receiving nursing home care, either at home or in a nursing home [142]. A prospective cohort study followed 47 newly admitted Canadian home care recipients for 12 months and counted the number of ED visits during their home care [201]. The rate of ED visits was 3.3 per 1,000 home care days and Kaplan–Meier curves predicted that, if all recipients were followed for 5.3 months, then 50% would visit the ED. A prospective three-year follow-up study in 1,300 older Finnish adults who presented with urgent geriatric health problems showed that home care was independently associated with a more than two-fold increased risk for institutionalization after correction for age, gender, living alone, falls, cognitive status, number of medications, contact with health services, and a selection of diseases [189]. In addition, a national register-based cohort study showed that older Danish adults who received home care were at increased risk of readmission

and mortality within 30 days of a short ED admission (< 48h) [66]. They also showed that the number of minutes of home care received was associated with an increased risk of readmission and mortality. For example, patients receiving more than 120 minutes of home care per week had twice the risk of readmission and four times the risk of dying within 30 days.

Furthermore, in the USA, two retrospective cohort studies of Medicare [241] beneficiaries, one involving 18,555 patients discharged from surgical medical units and the other involving 95,531 patients discharged after hospitalization for heart failure, found higher hazards for hospitalization, readmissions, and death within one to three months after discharge in those receiving home healthcare than in those not receiving home care [196, 204]. Taken together, these studies in acute care, post-acute care, and community cohorts show that older adults receiving home care are at risk of ED visits, acute hospitalization, recurrent admissions, institutionalization, and death. However, it is unknown how these adverse outcomes are affected by the types of home care remains unknown. Differentiating between the type of home care received may help us identify older adults in the community who are at risk of developing these adverse outcomes and to target preventive interventions accordingly.

Since the 2015 long-term care reforms in the Netherlands, the home care system has evolved to offer various types of support tailored to individual needs, enabling older adults to maintain independent living. Funded by government budgets through social taxes and mandatory healthcare premiums, the system aims to provide support to all in need. Household Help (HH) involves non-healthcare professionals performing domestic tasks, and is covered under the Social Support Act, managed by municipalities, and requires a minor contribution from the recipient [164]. Personal Care (PC) is provided by qualified nurses for medical diagnoses and functional limitations, and is funded by health insurance companies under the Healthcare Insurance Act [243]. HH and PC can be combined if needed. Long-term or nursing-home care can also be provided at home under the Long-Term Care Act, and requires an income-dependent contribution from the recipient [242]. This comprehensive care includes up to 24/7 assistance by qualified nurses or nurse assistants for severe medical, psychological, or functional limitations expected to be present until death. When care needs exceed the maximum possible care that can be provided at home, older adults are institutionalized. Most Dutch older adults are community-dwelling and live in a private household without home care. A minority receives home care so they can stay at home [217].

This study aimed to compare the risk of adverse health outcomes in older adults receiving different types of home care with those not receiving home care in a national Dutch cohort. It also aimed to determine how quickly these outcomes occur to find whether there is a window for prevention and how often older adults change their care form.

The remainder of this chapter is structured as follows. Section 3.2 details the

methodologies used, including the study design, participant selection, outcome measures, and the statistical methods applied. Section 3.3 outlines the findings, detailing population characteristics, differences in the risk of adverse outcomes across groups, and the frequency of transitions between home care types. Section 3.4 explores the chapter's contributions, emphasizing its strengths and limitations, while also proposing potential avenues for future research. Lastly, Section 3.5 presents the chapter's conclusions.

3.2 Methods

This section describes the study design, data sources, population selection, outcome definitions, and statistical analyses used in this study.

3.2.1 Study design and participants

We conducted this retrospective cohort study using national aggregated claims data from 2019. We selected data from 2019 because this was after the long-term care reforms were introduced in the Netherlands in 2015 and before the COVID-19 pandemic. We recovered data on demographics, home care status, acute hospitalization, and institutionalization from Statistics Netherlands [139], which covers the entire Dutch population. Dates of death were obtained from the death registry (see Appendices 3.A and 3.B). Ethics approval was not required because the data were anonymized and Statistics Netherlands provides guidelines for use and output control. This study was developed using the STROBE guidelines [237].

We included community-dwelling Dutch adults who were 65 years or older on January 1, 2019. We excluded individuals who were institutionalized because, despite their higher comorbidity and functional decline burden compared to community-dwelling peers, they have a lower risk of emergency department visits. This is because nursing homes often manage care needs internally and do not refer to the ED [142, 248]. Because their (emergency) care needs are met differently, we excluded them from comparison. We classified included individuals into the following five groups based on their home care status: no home care (NO), HH, PC, HH+PC, and nursing home care at home (NHH). In order to be classified in a home care group, individuals had to receive at least one day of home care in January 2019, according to earlier analyzes by the Dutch healthcare authority [142]. A description of these types is included in Tables 3.1 and 3.2.

An ED visit is defined as any registered contact with an ED resulting in assessment or treatment, regardless of subsequent hospital admission. We defined acute hospitalization according to the Dutch health authority, including a clinical nursing day following within 24h of an acute presentation to any department in a hospital [20]. We defined institutionalization as admittance to a nursing home under the Long-Term Care Act. We used the Statistics Netherlands definitions of income class and socioeconomic status [8]. Income class referred to a percentage

of the social minimum household income of the household income registry and socioeconomic status referred to a composite score (SES score of the SES-WOA registry) of education, income, and employment. The score ranges from −2 to 1, with a lower score corresponding to lower income, education, and/or employment. The reference SES-WOA score is 0 and represents the mean score of all Dutch adults [96].

Household help (HH)	Act	Social Support Act [164]
	Purpose	To ensure that individuals who have difficulty managing daily household tasks can continue to live independently in their own homes.
	Types of service	Assistance with cleaning, laundry, and other household tasks
	Provided by	Non–healthcare personnel
	Time	Varies depending on what is needed, from once to a few times a week; mostly a few hours a week.
	Eligibility	Available to all Dutch inhabitants who can demonstrate that they are unable to perform household tasks independently due to age, illness, disability, or other limitations, and who have an insufficient support system to assist them with these tasks.
	Responsibility	Coordinated by municipal governments responsible for assessing care needs, arranging services, and monitoring quality and effectiveness.
	Procedure	Individuals can apply for household help from municipalities. A municipal consultant assesses the application, conducts a home visit, and advises the municipality to what extent care is needed.
	Funding	Social taxes
	Contribution	Approximately 20 Euros per month
Personal care (PC)	Act	Healthcare Insurance Act [243]
	Purpose	To provide medical care and treatments at home needed to maintain, improve, or restore health.
	Types of service	Medical nursing care, such as wound care, injections, and other medical procedures. Helping with activities of daily living.
	Provided by	Qualified nurse (often referred to as district nurse)
	Time	Varies depending on what is needed, from once per week to a few times a day; mostly a few hours per week.
	Eligibility	Available to all Dutch inhabitants needing medical care for medical diagnoses and functional limitations.
	Responsibility	Organized by healthcare insurance companies in collaboration with healthcare providers, managing insurance coverage, reimbursing healthcare costs, and contracting with providers.
	Procedure	Individuals can apply for personal care from district nursing providers. A district nurse assesses the application, conducts a home visit, and indicates to what extent care is needed.
	Funding	Social taxes and healthcare insurance premiums
	Contribution	None

Table 3.1: Description of home care types with their combinations and definitions (part 1).

Household help combined with personal care		See household help and personal care (HH+PC)
Nursing home care at home (NHH)	Act	Long-Term Care Act [242]
	Purpose	To provide intensive care or supervision, often for a long period or the rest of the recipient's life.
	Types of service	Medical nursing care and helping with disabilities more complex and intensive in nature than those needing personal care.
	Provided by	Qualified nurse (often referred to as district nurse) or nurse assistant
	Time	Varies depending on what is needed, from once per day to 24h care; mostly a few hours a day.
	Eligibility	Available to all Dutch inhabitants with an expected need for around-the-clock care or supervision for the remainder of an individual's life due to medical, psychological, or functional limitations.
	Responsibility	Organized by care offices, which are part of healthcare insurance companies, managing assessment, coordination, quality assessment and financial management.
	Procedure	Individuals need to submit a formal application to the Care Needs Assessment Centre (CIZ), an independent governing body operating under Dutch government supervision. In the application they need to explain how they meet the eligibility criteria along with their medical file for examination by CIZ. Once approved, care offices, overseen by government and linked to health insurance companies, facilitate the necessary care.
	Funding	Social taxes
	Contribution	Income depended, maximized at approximately 800 euros per month

Table 3.2: Description of home care types with their combinations and definitions (part 2).

3.2.2 Outcomes

Primary outcomes were the number and rate of individuals per group that visited the ED, experiencing acute hospitalization, recurrent admissions, institutionalization, or death in 2019. We defined an ED visit in the same way as the Dutch health authority, including an acute presentation to any department in a hospital [20]. ED visits were scored as the number and rate of individuals who visited the ED once, more than once, and four or more times at the end of follow-up. Four or more visits were studied, because it is the most common used definition for frequent ED use and is to be associated with comorbidity burden [137, 236]. The number and rate of individuals per group that received home care, were institutionalized, or died are also presented and these outcomes were recorded from February to December 2019 because of the classification method. Any changes in home care

status or institutionalization were scored in the following month to avoid double registration in one group. Lastly, the time to the adverse event, starting on January 1, 2019, for all adverse events except institutionalization, which was measured from February 1, 2019, is examined.

3.2.3 Measurements

The age, gender, demographics, and socioeconomic status of all groups were determined on January 1, 2019, as previously described [200]. Drug use was reported in an ATC-4 registration file that recorded the number of drugs prescribed per person in 2019.

3.2.4 Statistical analysis

To compare baseline differences between all groups, we used Chi-square tests for dichotomous variables. For continuous variables, we first used t-tests to compare the NO group with each home care group separately. Then, we used one-way ANOVA to compare the different home care groups. Logistic regression was used to investigate the association between home care types and all primary outcomes after a one-year follow-up (11-month follow-up for institutionalization). Two separate models were employed in the analyses. In model one, the multivariate logistic regression was adjusted for age and gender. In model two, regression was adjusted for age, gender, and death. We used only model one to analyze death to avoid double correction because model two already adjusted for it. We limited the adjustments to correcting only for differences in age, gender, and death. Correcting for our other measures, which we believe reflect parts of the underlying mechanisms that put older adults receiving home care at risk for ED visits, would result in over-correction [3, 83, 162]. If we were to adjust for drug use, it would partially mitigate the important effect of comorbidities on the risk of ED visits among older adults receiving home care.

To determine changes in status during follow-up in each group, the number and rate of individuals who received home care, were institutionalized, or who died were calculated monthly and visualized graphically. For the time till events, cumulative incidence curves were used to analyze the time to acute hospitalization and time to institutionalization whereas Kaplan–Meier curves were used to visualize time to death. The incidence of these outcomes over time was visually assessed. This method was chosen because it only provided information on the potential window for prevention alongside the primary outcome if there was a substantial deviation of the curve during the follow-up period. No cumulative incidence curves were made for recurrent admissions because this would have entailed plotting the time to the second ED visit, which was not the aim of the study. Older adults were allocated to groups based on their home care status in January 2019, whereas the time to institutionalization was reported from February 1, 2019, onward. This one-month difference between institutionalization and the other outcomes influenced the results.

3.3 Results

This section presents the results of the analyses on population characteristics, adverse health outcomes, and changes in home care use over time. The findings are structured to first describe the study population, followed by the associations between home care types and adverse health outcomes, and finally the temporal dynamics of home care status.

3.3.1 Population characteristics and differences in adverse health outcomes per group

In total, 3,174,953 community-dwelling persons in the Dutch national population were aged 65 years and older at baseline. Of these, 2,758,093 (86.9%) persons were included in the NO group, 131,260 (4.1%) in the HH group, 154,462 (4.9%) in the PC, 96,526 (3.0%) in the HH+PC group, and 34,612 (1.1%) in the NHH group.

Tables 3.3 and 3.4 report the baseline characteristics of the population as well as the number and rate of each adverse event per home care group. Adults in the home care groups were significantly older than those in the NO group ($p< 0.001$). They were also more often female, living alone, had a lower SES-WOA score, had a lower income, and used more drugs ($p< 0.001$). Older adults in the HH or HH+PC group were more often female (74.4%), lived alone more often (73.7%), had a lower SES-WOA score (−0.6), and used more drugs than older adults in the other groups ($p< 0.001$).

	NO	HH	PC	HH+PC	NHH	P value
n	2,758,093	131,260	154,462	96,526	34,612	
Age, mean (Std Dev)	73.1 (6.2)	78.9 (7.4)	81.1 (7.7)	82.3 (7.7)	83.2 (7.8)	< .001
Female, %	51.0%	74.4%	58.2%	75.1%	64.5%	< .001
Lives alone, %	26.8%	73.7%	48.2%	77.5%	42.6%	< .001
SES-WOA, mean (Std Dev)	−0.1 (0.5)	−0.6 (0.5)	−0.3 (0.5)	−0.6 (0.5)	−0.3 (0.5)	< .001
Income, mean (Std Dev)	238.0 (128.0)	143.9 (54.2)	188.4 (97.7)	143.2 (57.0)	178.2 (92.6)	< .001
Medication, mean (Std Dev)	5.5 (4.2)	8.8 (4.7)	9.9 (5.0)	10.9 (5.0)	9.0 (4.8)	< .001

Table 3.3: Demographic and socioeconomic characteristics by home care type.

	NO	HH	PC	HH+PC	NHH	P value
Number of individuals that visited the ED, n (%),	381,523 (13.8%)	31,431 (24.0%)	53,531 (34.7%)	33,866 (35.1%)	10,902 (31.5%)	< .001
1 ED visit, n (%)	254,135 (9.2%)	19,040 (14.5%)	29,744 (19.3%)	18,614 (19.3%)	6,384 (18.4%)	< .001
>1 ED visit, n (%)	127,388 (4.6%)	12,391 (9.4%)	23,787 (15.4%)	15,252 (15.8%)	4,518 (13.1%)	< .001
$\geq$ 4 ED visits, n (%)	21,732 (0.8%)	2,307 (1.8%)	5,159 (3.3%)	3,200 (3.3%)	765 (2.2%)	< .001
Acute hospitalization, n (%)	176,681 (6.4%)	17,726 (13.5%)	35,681 (23.1%)	22,798 (23.6%)	6,903 (19.9%)	< .001
Recurrent admissions[1], n (%)	56,655 (2.1%)	6,246 (4.8%)	13,754 (8.9%)	8,866 (9.2%)	2,414 (7.0%)	< .001
Institutionalized, n (%)	10,117 (0.4%)	2,001 (1.5%)	11,592 (7.5%)	8,541 (8.9%)	12,007 (34.7%)	< .001
Died, n (%)	51,583 (1.9%)	6,199 (4.7%)	25,141 (16.3%)	13,158 (13.6%)	7,206 (20.8%)	< .001

Table 3.4: Health outcomes by home care type.

All home care groups had a higher rate of ED visits than the NO group did after one year. The PC and HH+PC groups had the highest incidences of ED visits and the results between these two groups were similar. The NHH group had a higher incidence than the HH group, but a lower incidence than the PC groups. The NO group had the lowest rate of individuals who visited the ED (13.8%), but contributed most to the total number of individuals visiting the ED (381,523) because of its larger size. All home care groups had a relatively higher rate of more than one ED visit, and especially four or more visits, compared to the NO group. These rates were highest in the PC groups. The number of persons who experienced an adverse health outcome was significantly higher in the home care groups than in the NO group ($p < 0.001$). The rates of acute hospitalization and recurrent admissions were highest in the HH+PC and the PC group. The rate of adverse outcomes was lowest in the NO group, even though the number of persons experiencing acute hospitalization (176,681) and recurrent admissions (56,655) was highest in this group because of the large group size. The rates of institutionalization and death were highest in the NHH group, and there was more variation in the rate of institutionalization between the groups than in the other adverse events.

3.3.2 The association between home care types and adverse health outcomes

Table 3.5 shows the results of the logistic regression analysis. All home care groups had a significantly higher risk of all adverse events than the NO group. The HH and NHH groups have the highest risk for ED visits. The risk of acute hospitalization

[1]Recurrent admissions were defined as more than one acute hospitalization.

(OR 2.60, 95% 2.55 – 2.64) and recurrent admissions (OR 2.60, 95% 2.55 – 2.65) were highest in the HH+PC group. The risk of institutionalization (OR 63.22, 95% 60.94 – 65.58) and the risk of dying (OR 7.59, 95% CI 7.35 – 7.85) were highest in the nursing-home care at home group.

Group/Model	OR ED visit (95% CI)	OR acute hospitalization (95% CI)	OR recurrent admissions (95% CI)	OR institutionalization (95% CI)	OR death (95% CI)
NO					
Model 1	Ref	Ref	Ref	Ref	Ref
Model 2	Ref	Ref	Ref	Ref	-
HH					
Model 1	1.67 (1.63 – 1.67)	1.88 (1.85 – 1.92)	2.03 (1.97 – 2.09)	2.60 (2.46 – 2.75)	1.96 (1.90 – 2.02)
Model 2	1.63 (1.60 – 1.65)	1.80 (1.77 – 1.83)	1.80 (1.77 – 1.84)	2.55 (2.42 – 2.70)	-
PC					
Model 1	2.56 (2.51 – 2.61)	3.12 (3.07 – 3.17)	3.41 (3.33 – 3.49)	11.49 (11.12 – 11.88)	6.11 (5.99 – 6.23)
Model 2	2.08 (2.05 – 2.10)	2.35 (2.31 – 2.38)	2.34 (2.31 – 2.38)	10.30 (9.96 – 10.65)	-
HH+PC					
Model 1	2.57 (2.51 – 2.61)	3.24 (3.18 – 3.30)	3.64 (3.54 – 3.74)	12.80 (12.34 – 13.28)	5.00 (4.87 – 5.12)
Model 2	2.20 (2.17 – 2.24)	2.60 (2.55 – 2.64)	11.59 (11.17 – 12.03)	11.59 (11.17 – 12.03)	-
NHH					
Model 1	2.11 (2.08 – 2.16)	2.40 (2.33 – 2.48)	2.46 (2.35 – 2.59)	70.20 (67.71 – 72.79)	7.59 (7.35 – 7.85)
Model 2	1.55 (1.51 – 1.60)	1.56 (1.52 – 1.60)	1.55 (1.50 – 1.60)	63.22 (60.94 – 65.58)	-

Table 3.5: Adjusted odds ratios for acute hospitalization, recurrent admissions, institutionalization, and death per home care group. Recurrent admissions were defined as more than one acute hospitalization. Model one was corrected for age and gender. Model two was corrected for age, gender, and death.

3.3.3 How quickly adverse health outcomes occur per group over time

Incidences of acute hospitalization, institutionalization, and death increased gradually in most groups during follow-up. In contrast, the incidence of institutionalization increased substantially in the nursing-home care at home group in the first six months before stabilizing to a more gradual rate (Figure 3.1). All other curves are presented in Appendix 3.C.

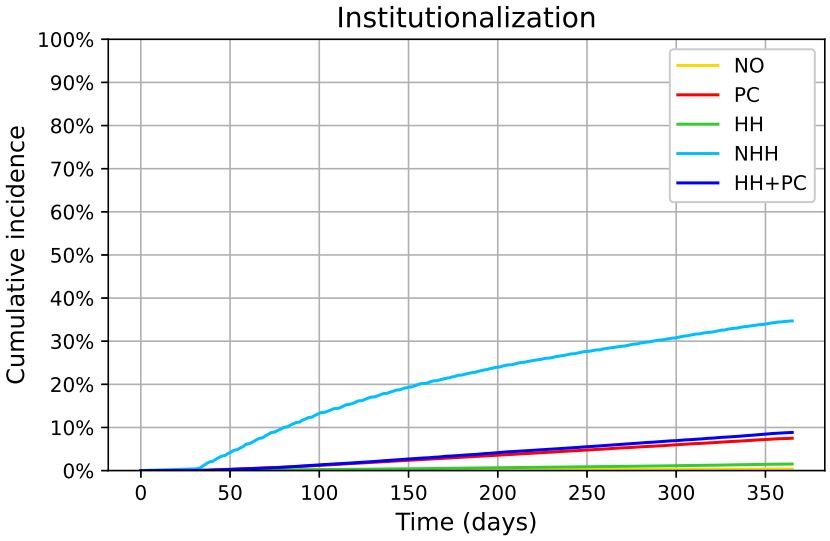


Figure 3.1: Cumulative incidence curve of institutionalization per home care type. Participants were allocated to groups based on their home care status in January 2019, whereas the time to institutionalization was measured from February 2019. This one-month difference influenced the end results. Recurrent admissions were not included because the time to the second admission would have had to be included, which was not the aim of the outcome.

3.3.4 The dynamics of home care during follow-up

Figure 3.2 describes the monthly results for home care, institutionalization, and death from February to December 2019. In the NO group, 6.4% (n = 175,950) changed in home care status (including 1.6% to HH, 2.2% to PC, 0.4% to HH+PC, 0.2% to NHH, 0.3% to institutionalization and 1.9% died). In the HH group, 22.9% (n = 30,053) changed in home care status (including 4.4% to NO, 0.6% to PC, 11.3% to HH+PC, 1.1% to NHH, 1.2% to institutionalization and 4.7% died). In the PC group, 51.9% (n = 80,115) changed in home care status (including 16.8% to NO, 1.7% to HH, 7.0% to HH+PC, 5.4% to NHH, 5.7% to institutionalization and 16.3% died). In the HH+PC group, home care status changed in 36.3% (n = 35,026) of individuals (including 1.1% to NO, 7.3% to HH, 2.5% to PC, 6.1% to NHH, 6.8% to institutionalization and 13.6% died). In the NHH group, home care status changed in 53.1% (n = 18,394) of individuals (3.3% to NO, 0.8% to HH, 1.8% to PC, 1.6% to HH+PC, 26.5% to institutionalization and 20.8% died). It seems unlikely that individuals of the NHH group who changed to no claimed home care (n = 1,158) did not require any home care because they had a previous 24h nursing home care at home indication. A possible explanation could be that these individuals changed to privately funded home care.

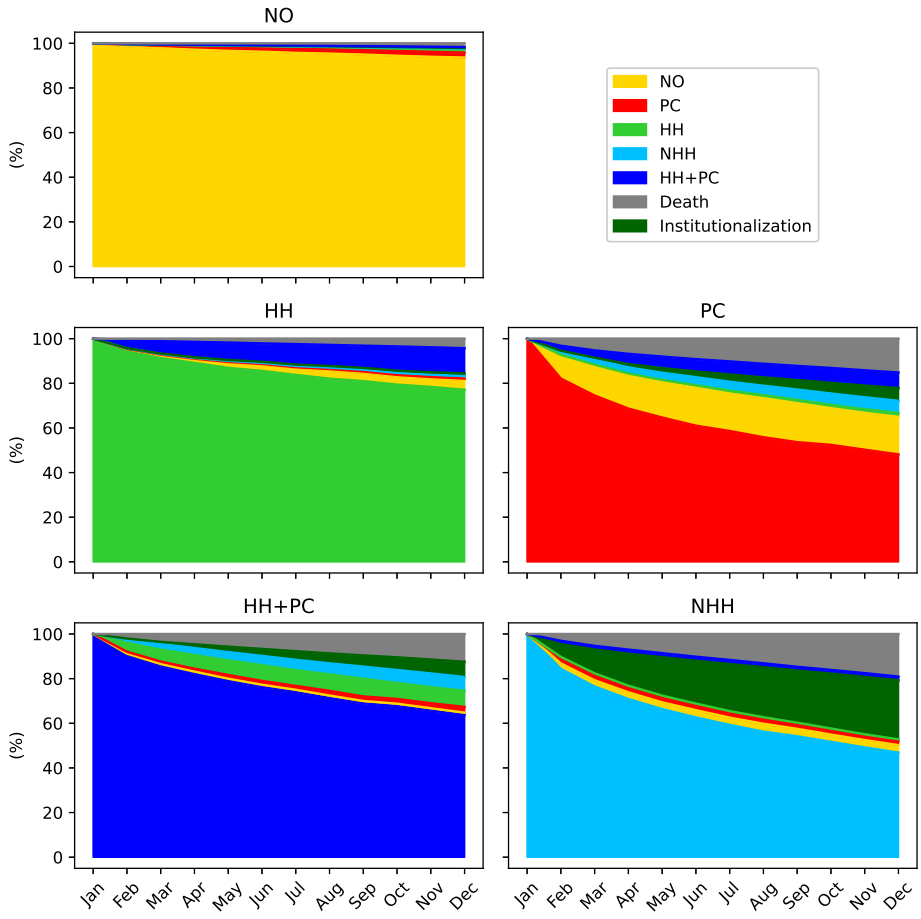


Figure 3.2: The dynamics of home care, institutionalization, and death in 2019 in each group. The vertical axis describes the percentage of the cohort that remains unchanged home care status from January-December 2019.

3.4 Discussion

This national study shows that older adults receiving home care have a higher risk of visiting the ED, acute hospitalization, recurrent admissions, institutionalization, and death than those not receiving home care. Compared with individuals who do not receive home care, the risk of these adverse health outcomes also differed depending on the type of home care received. The risk of ED visits increases to more than two-fold for those receiving nurse-based, medically and functionally indicated PC. The risk for acute hospitalization and recurrent admissions was approximately two-fold higher in those receiving HH and three-fold higher in those receiving PC than for those receiving no home care. The chance of institutional-

ization was higher in those receiving PC (7.5 – 8.9%; 12-fold increase) and NHH (34.5%; 60-fold increase). The chance of death was also two- to seven-fold higher in groups receiving home care. These adverse outcomes increased gradually during follow-up, except for institutionalization in the nursing-home care at home group, which increased substantially in the first six months. Furthermore, home care status changed considerably over the one-year follow-up period, mostly to a more intense form of home care, institutionalization, or death.

Earlier studies have compared the risks of adverse health outcomes between older adults who receive home care and those who do not [66, 142, 189, 196, 201, 204, 248, 250]. In line with these studies, we found higher risks for ED visits, acute hospitalization, recurrent admissions, institutionalization, and death in older adults receiving home care. The difference in ED visit risk between those not receiving home care and those who did was greatest for more frequent visits. This supports the hypothesis that chronic comorbidities play a key role in the ED visit risk for home care recipients [137, 236]. Singular visits often result from an acute care need that is effectively addressed, while recurrent visits are more often due to chronic conditions that are insufficiently addressed [137, 236]. The increases we observed in the risk of acute hospitalization and recurrent admissions were similar to those found in earlier studies [66, 201, 248]. These observations could be due to higher frailty in home care recipients. We base this hypothesis on the following arguments. First, the indication for home care is based on an individual's ability to live independently, so reflects a person's functionality, which is one of the domains used to diagnose frailty [164, 242, 243]. Compared with individuals receiving no home care, home care recipients are probably frailer, varying from mildly frail (limited functional decline) in individuals receiving HH, intermediately frail (more distinct functional decline) in individuals receiving PC, to severely frail (obvious functional restrictions) in individuals receiving NHH. Second, frailty has been identified as an important driver for ED visits and acute hospitalization in recipients of PC [22, 151]. However, our results showed a lower ED risk in the NHH group, which is probably the most frail group. We have two explanations for this. First, recipients of NHH in the Netherlands are often admitted to nursing homes instead of EDs when acute care demands arise, because they require 24h care or nursing home care [203]. Second, as demonstrated in Figure 3.2, a large portion of this group was institutionalized during follow-up. Further research is needed to test this hypothesis.

We found that most older Dutch adults who visited the ED live independently without any formal support. These results are in line with those of previous reports on older adults visiting the ED [3, 83, 142, 151, 157, 162, 175, 218]. Older adults living independently without any formal support are less frail, present less frequently, have fewer geriatric syndromes, and fewer unstable comorbidities, so present more often with acute complaints [151, 248]. These differences may require different preventive approaches. Therefore, to fully address the increasing volume of older adults at the ED, future research should focus on how to prevent ED visits in older adults not receiving home care.

Our study revealed a greater increase in the risk of institutionalization (three- to sixty-fold versus two- to three-fold in Salminen et al. [189]) and death (adjusted five- to eight-fold versus unadjusted two- to five-fold in Elkjaer et al. [66]). These discrepancies can partly be explained by a lower incidence of institutionalization and death in those not receiving home care in our cohort compared with cohorts published in other studies (0.4% incidence of institutionalization in our cohort compared with 5% reported in [189] and a 1.6% incidence of death in our cohort compared with 10% reported in [66]). The odds of adverse events were calculated by comparison with the NO group, so the low rate we observed in our comparison group increased the risk in the home care groups substantially. Higher incidences were observed in the comparison groups of [189] and [66] because these groups comprised older adults seeking emergency care, who have acute illnesses and therefore a higher chance of adverse outcomes than individuals not receiving home care [3, 83]. These observations indicate that differentiating in the type of home care is potentially more effective at identifying older adults at risk for institutionalization or death in the community than in those seeking emergency care.

It is already known that sicker and more cognitively or physically impaired older adults will need more home care and may ultimately end up with acute problems or require 24h nursing home care, especially when home care is insufficient because of nursing staff shortages and deficient funding [16, 23, 31, 58, 63, 98, 101, 124, 166, 225, 226]. However, our findings have helped identify subgroups that are most at risk of adverse outcomes and that would benefit greatly from prevention strategies. This knowledge is vital for implementing intervention strategies effectively on a population level.

One high-risk subgroup is older adults receiving PC with or without HH, who constitute 8% of the older community-dwelling Dutch population. This subgroup faces high risks for all adverse outcomes, particularly for acute hospitalization and recurrent admissions. These outcomes cause functional decline for the patient and extensive healthcare costs for the system [40, 225, 226]. Acute care demands of those receiving PC have already been reported, and frailty, comorbidity, and severity of acute disease have been identified as important contributing factors [83, 113, 114, 238]. Comorbidity and frailty also increase the need for more intensive home care and nursing home care [31]. Furthermore, previous studies have shown that when registered nurses are involved in care, there may be more referrals to the ED, for example, as the nurses can identify concerns and will refer whereas a housekeeper might not [122, 142, 144, 201]. There is a great potential for prevention in older adults receiving PC, because they are often visited daily by trained nurses, and their acute care demands have been shown to originate gradually over days to weeks. Furthermore, earlier reports have indicated that these nurses often refer patients to the ED because they have insufficient time, resources, and back up from primary care physicians or intermediate care facilities to address developing demands effectively [113, 122, 138, 144]. Moreover, multiple studies have demonstrated that intensifying home care can prevent these referrals and adverse

health outcomes by providing nurses with the support they need [50, 124, 144, 212, 238, 250].

We also showed that older adults receiving NHH face very high risks for death and institutionalization within a year, with transitions to a nursing home often occurring within six months. These results seem logical given the severe and permanent impairments that qualify individuals for this most intensive form of home care, raising questions about whether these adverse outcomes can be prevented at such an advanced stage. Furthermore, earlier reports show that acute care demands in this group are often met by admission to nursing homes rather than referral to a hospital emergency department [203]. This suggests that aggravating care needs in this group are better addressed in nursing homes than in hospitals or at home. Instead of focusing on prevention strategies, it may be more beneficial to concentrate on optimizing the quality of life and care for this group through advanced care planning, whether in a home or nursing home setting. Taken together, our findings suggest that prevention strategies should prioritize older adults receiving PC, and that advanced care planning could be more beneficial to those receiving NHH.

3.4.1 Strengths and limitations

A strength of this research is the size of the study sample, which provides a national perspective. There are also some limitations. First, we did not have accurate data on the exact daily or weekly duration of home care received within each home care group, so we did not include this in our analysis. Second, relevant data on comorbidities and geriatric syndromes were unavailable in the Statistics Netherlands database, which limits our ability to compare the underlying risk factors for adverse outcomes between groups. However, the impairments caused by these comorbidities and geriatric syndromes are reflected in the eligibility criteria for the different types of home care, as described in Tables 3.1 and 3.2. Informal care was also not included in the data, and therefore not included in this chapter. Third, we limited our correction in the regression analyses to age, gender, and death to prevent overcorrecting. However, other measurements could have had a confounding effect. Fourth, this study focused on a Dutch national sample receiving home healthcare within the Netherlands. Home care organization varies widely across countries because of differences in healthcare systems, government policies, and cultural norms [76, 112, 221, 254]. In (Northern) Europe and Canada, home care is available to all residents and is funded publicly through social insurance or taxes, while in the US, the emphasis is on private care. Whether care can be covered by public funds depends on the type of care needed and the individual's income. In Western societies, care is often organized with formal services, whereas in Asia, care is traditionally centered around family and community – although this has recently evolved towards more formal care. Despite these differences, the challenges in providing sufficient home care across the US, Canada, Europe, and Asia are similar, and include increasing demand and healthcare costs, personnel shortages, and effective coordination of care [76, 112, 144, 166, 221, 254]. We have made ef-

forts to thoroughly characterize the sample and the Dutch home healthcare system to assess the generalizability of our findings. Finally, there are inherent limitations to this type of study: for example, group risk may not always reflect individual risk, and causal factors may vary, and these limitations must be considered when developing interventions.

3.4.2 Practical use and future research

The type of home care older adults receive can be used by government and municipal policymakers to identify older adults at risk of adverse health outcomes on a population level. Changes in home care were observed in all groups during the follow-up period, and most of these changes involved more intense care, institutionalization, or death. Our findings in individuals receiving PC illustrate the value of a monthly follow-up. Older adults who transitioned to the NO group typically did so within 2 months, while those who did not recover in the first few months gradually needed more intense home care, or were institutionalized or died. Thereby, we identified two different subgroups. We hypothesize that the subgroup that recovers quickly has a temporary need for support and is temporarily frail, while the other group gradually needs more intense care because of increasing frailty. This hypothesis is endorsed by earlier studies [31, 36, 97, 101, 126, 145, 235].

Further research is needed to optimize prevention strategies for older adults with profiles similar to those in our PC group. Multiple studies have shown that sufficient home care can prevent an increase in home care demand [50, 138, 144, 212, 250] and that system integration, caring case management, and relational continuity are essential [45, 138, 144]. To optimize the quality of PC, research should determine whether addressing factors such as frailty, comorbidity, and the development of acute disease can reduce adverse events in older adults receiving PC [83, 113, 114]. The timely deployment of healthcare personnel can be optimized by early detection of health deterioration through prediction models and through collection of real-life data on activity via robots or domotics. Additionally, interventions can be assessed by integrating home care into simulation models of healthcare systems, which traditionally focus on acute care but can also be used to assess the effects of long-term care. Using the type of home care to predict adverse outcomes in these models is attractive because this information is simple, easy to access, and has minimal data requirements, making it suitable for regions with limited population data.

3.5 Conclusion and Implications

This study shows that the type of home care can identify sub-populations at increased risk of ED visits, acute hospitalization, recurrent admissions, institutionalization, and death compared with community-dwelling older adults that do not receive home care. Older adults receiving nurse-based, medically and functionally

indicated PC are most at risk. They face a more than twofold risk compared to their peers not receiving home care, in particular for more frequent visits. These older adults are visited daily by professionals, so there is potential for targeted interventions in the future. This requires further investigation into the causal factors contributing to the risk of ED visits among home care recipients, particularly comorbidities, geriatric syndromes, and the quality of both primary and home care. Furthermore, when assessing the risk of visiting the ED in older adults, one should take into account that home care changes considerably over a one-year period, mostly to more intense care, institutionalization, or death. Moreover, older adults not receiving home care contribute most to the national volume of ED visits by older adults, despite their lower ED visit risk and may need a different preventive approach.

Older adults receiving nurse-based, medically and functionally indicated PC are particularly at risk of acute hospitalization and recurrent admissions, which can lead to further health deterioration and an increased need for care. These adults form a substantial population in the community and have skilled nurses visiting them frequently, so there is potential for prevention. However, further research is needed to determine the causal factors, such as comorbidity, developing frailty, acute disease, and insufficient care, and to investigate whether targeting these factors improves prevention. The type of home care can be used to help policymakers target prevention and model its effect on national and regional scales.

3.A Appendix: Used Registries

The following databases of CBS were used for this chapter (description of data included in these registries is in Dutch):

- GBAHUISHOUDENBUS
  https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/gbahuishoudensbus-huishoudenskenmerken
- MEDICIJNTAB
  https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/medicijntab-geneesmiddelen-op-atc-code--4--
- INHATAB
  https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/inhatab-inkomen-van-huishoudens
- KOPPELPERSOONHUISHOUDEN
- GBAOVERLIJDENTAB
  https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/gbaoverlijdentab-datum-van-overlijden-van-personen-ingeschreven-in-het-gba
- SESWOA
  https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/seswoa-sociaaleconomische-statusscores-huishoudens
- WLZINTAB
  https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/wlzzintab-personen-metgebruik-wlz-zorg-in-natura
- GEBWMOTAB
  https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/gebwmotab-personen-met-wmo-maatwerkvoorzieningen
- ZVWWVPTAB
  https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/zvwwvptab-personen-met-gebruik-van-zvw-wijkverpleging
- GBAPERSOONKTAB
  https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/gbapersoonktab-persoonskenmerken-beperkt-in-de-brp

- MSZZorgactiviteitenVEKTTAB
  `https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/mszzorgactiviteitenvekttab-zorgactiviteiten-diagnose`

3.B Appendix: Used Codes

- Household help (WMO codes): 006,007,100,101,102,103,104,105,107,711
- Personal care (ZVW codes): 1, 2, 3 ,4
- Nursing home care at home (WLZ codes): 2 and 3
- Acute hospitalization codes: clinical nursing day within 24hours after 19015 (ED admittance) or 19016 (acute admittance outside ED)
- Institutionalization (WLZ codes): 1, 11 and 12.

3.C Appendix: Additional Figures

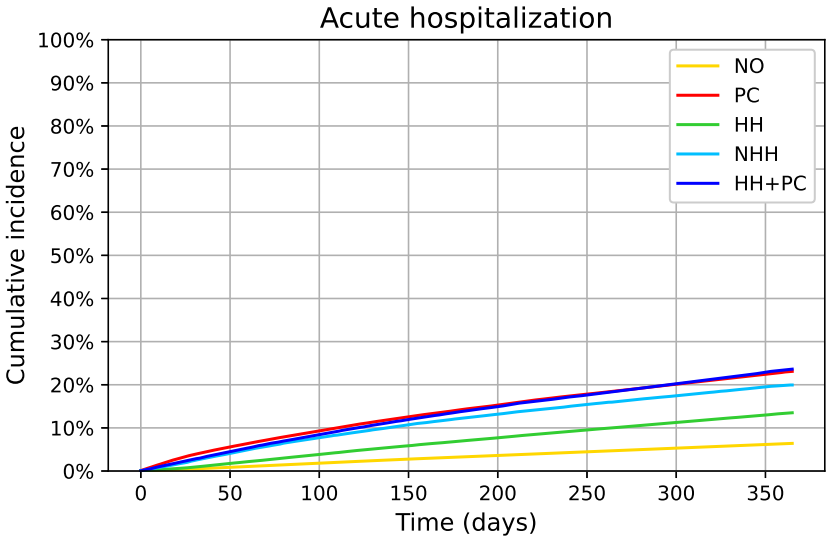


Figure 3.C.1: Cumulative incidence curves of acute hospitalization per home care form. Recurrent admissions were not included because the time to the second admission would have had to be included, which was not the aim of the outcome.

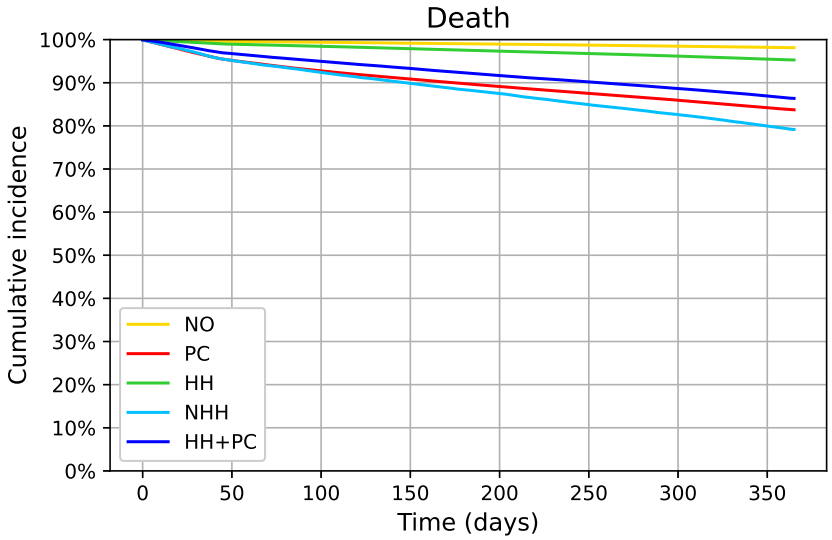


Figure 3.C.2: Kaplan–Meier curves of death per home care form.

4

Too Soon or Too Late? A Markov Decision Process Approach for Preventive Nursing Home Admission

Based on:
de Boer, T. R., Bekker, R., van der Mei, R. D., & van den Nieuwenhof, J. M. (2025). Too soon or too late? A Markov decision process approach for preventive nursing home Admission. *Submitted for publication.*

Abstract

This chapter presents a Markov decision process framework for determining the optimal timing of nursing home admissions in elderly care with monitoring of health status. By modeling health deterioration as a discrete-time Markov chain and incorporating realistic healthcare costs, we derive an optimal policy that minimizes expected costs. The policy recommends condition-based monitoring, with more frequent check-ups for patients with poorer health. The analysis shows that, given current cost estimates, the optimal policy postpones nursing home admission until the poorest health state, thereby minimizing both nursing home and overall costs. We evaluated the Markov decision process under various scenarios, including changes in nursing home costs, crisis costs, health state granularity, and admission delays. The model is shown to be robust in all settings, highlighting key trade-offs between crisis prevention, patient monitoring, and institutional care. These insights can support data-driven, individualized admission decisions that improve efficiency and patient outcomes.

4.1 Introduction

There are various methods to reduce the demand for healthcare for older adults, including e-health options [176], fall prevention strategies [219], and exercise programs [32]. Understanding and implementing effective demand-reduction strategies is therefore crucial to ensure that healthcare systems remain sustainable and capable of meeting the needs of an aging population. Long-term care (LTC) represents one of the most constrained areas of elderly care, particularly in nursing homes (NHs). LTC primarily serves older adults, with around 40% of users 80 years or older [149]. It is provided in two main forms: *home-based* care and *institutional* care, such as that provided in an NH. Individuals with lower care intensity needs are generally supported at home. However, once the home environment becomes inadequate, admission to an NH may be required. The NH facilities provide continuous support for individuals who can no longer live independently [249]. With NH admissions typically being permanent, NH beds have become a scarce resource. This combination of high demand and constrained resources makes efficient admission policies critical to ensuring that those with the greatest need receive timely care while preventing unnecessary strain on the system.

Previously, the focus of research on NH admissions was primarily on strategies to increase facility capacity [46] or shorten wait times [7]. Another commonly proposed strategy is to postpone admission [65], enabling older adults to live independently [155] or with the support of home care services [216]. This approach is often considered cost-effective [38] and living at home is generally preferred by individuals [186]. However, studies indicate that older adults living at home with home care are at increased risk for negative outcomes [200], such as visits to the Emergency Department (ED) in crisis situations. Such crises often lead to an accelerated deterioration [80, 156], necessitating additional care for the elderly, and an admission to the NH. An estimated 50% of NH admissions occur following such crisis situations [106]. These admissions are typically *reactive* rather than *proactive*, addressing emergencies instead of preventing them. Crises often involve an acute decline in health or the breakdown of informal caregiving arrangements, leading to substantial financial costs and significant psychosocial burdens for older adults and their families.

Some crisis situations could be avoided by preventive admission to the NH, as people admitted to NHs have a lower risk of adverse events, as shown in Chapter 3. Therefore, it remains critical to formulate an admission strategy that selects patients for admission, with the aim of balancing cost reduction and the mitigation of crisis situations while maximizing the time spent living at home. Consequently, this approach could lower NH occupancy rates and further decrease waiting times.

4.1.1 The admission dilemma

Developing an admission policy for NHs requires balancing clinical need, uncertainty, monitoring costs, and limited capacity. Determining the optimal timing for

admission is challenging: Patients typically favor remaining at home as it often enhances their quality of life; however, postponing admission for too long increases the risk of crises and negatively impacts long-term health. The health status of older adults plays a crucial role in determining their risk of experiencing a crisis. By categorizing individuals into specific health classes, admission policies can be tailored to prevent unnecessary costs [111, 120, 252] and avoid last-minute crisis-driven placements.

To illustrate the dilemma of balancing maximizing the time spent at home and minimizing the number of crises and costs, consider the following stylized example. We follow the trajectory of an older adult who starts relatively healthy and whose probability of experiencing a crisis is low. Over time, their condition slowly deteriorates and the probability of a crisis increases. We consider three different admission scenarios; see Figure 4.1 and below:

1. **Late Admission**: An older adult with prolonged residence at home experiences a crisis, which can cause poorer health and increased costs, leading to admission to NH.

2. **Early Admission**: NH admission occurs well before the risk of a crisis increases significantly. While the crisis is avoided, the older adult spends more time than necessary in the NH, potentially reducing their well-being and incurring avoidable care costs.

3. **Timely Admission**: The older adult is admitted as the risk of a crisis becomes substantial. Despite a shorter stay at home, the crisis is effectively averted, thus better balancing the trade-off between crisis prevention and a longer period at home.

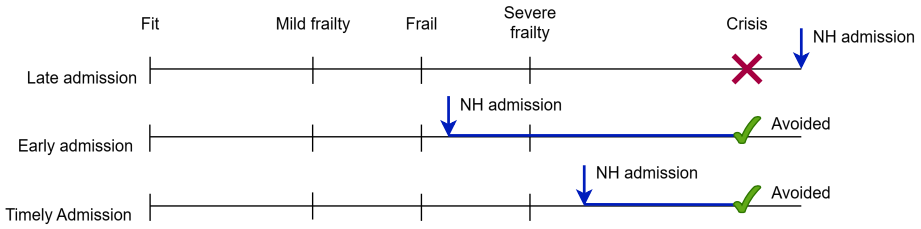


Figure 4.1: Timeline of three different scenarios of NH admission: late, early, and timely admission.

Another challenge is the uncertainty in the progression of health: an individual's precise condition is often unknown, requiring periodic assessments to make informed decisions. However, these *checks* come with additional effort and costs, which means that their frequency must be carefully balanced against the value of the information they provide. Lastly, the limited capacity of NHs encourages

delaying admissions. As these facilities often operate near full capacity and admissions are typically for a prolonged period, policies must ensure that capacity is reserved for those most in need while sustaining the system.

4.1.2 Contributions and outline

This chapter presents three main contributions in the context of NH admissions. First, we present a novel application of a Markov Decision Process (MDP) to model health progression among older adults and the associated decision-making around NH placement. We derive an optimal policy based on realistic health transition rates and cost estimates. Second, we highlight the critical role of condition-based monitoring in reducing uncertainty. By dynamically adjusting the frequency of health check-ups based on an individual's current health state, the model supports early detection of decline and better timing of interventions, preventing costly crises while avoiding premature institutionalization. Third, we demonstrate the flexibility of the MDP approach by systematically varying key parameters, such as crisis costs, NH costs, the granularity of the classification of health status, and the presence of admission delays. These sensitivity analyses reveal how the MDP and its policy can adapt to different economic and operational constraints, providing valuable information for policy makers and care providers.

The remainder of this chapter is organized as follows. Section 4.2 reviews the relevant literature on health modeling, decision-making under uncertainty, and NH admission. Section 4.3 introduces the health model and Section 4.4 presents the MDP formulation to optimize admission and check-up strategies. Section 4.5 reports the results in various realistic settings. Finally, Section 4.6 discusses implications, limitations, and directions for future research.

4.2 Literature

This section reviews the literature in three related areas: (i) modeling the health status of older adults, (ii) maintenance problems and uncertainty in health status, and (iii) research on admission policies of NH.

The health status of older adults

A growing body of research models the health status of older adults using Markov-based frameworks or, more generally, uses discrete health states to represent individual well-being. Dolan [64] introduced a framework for evaluating different health states, incorporating the time spent in each state to reflect how patients perceive their state. Although primarily focused on utility evaluation rather than transitions or interventions, this work highlights the importance of distinguishing discrete health conditions in care planning. More closely aligned with our approach is Kaushik et al. [119], which used a Markov chain to model health state transitions based on sensor data from living environments of older adults. Their

study demonstrated the feasibility of probabilistic models in tracking health dynamics over time, even in the presence of noisy or incomplete data.

Frailty has become a central concept in the modeling of health among older adults, frequently serving as a proxy of vulnerability to adverse outcomes such as hospitalization, falls, or death. Standardized frailty indices and phenotypes have been developed to aggregate multidimensional characteristics, ranging from biological to sociodemographic factors, into a single measure of health risk [37, 79, 158, 193]. Hoogendijk et al. [101] demonstrated the predictive value of frailty for future health deterioration, underscoring its relevance for anticipatory care. However, a key limitation is that most frailty measures are updated infrequently, limiting their applicability to real-time monitoring or dynamic decision-making. More recent work has explored alternative models, such as hidden Markov models and transition-based health prediction tools, to better capture the temporal evolution of dependency or health status [116, 239]. These approaches support more responsive and personalized care strategies taking into account the stochastic and time-varying nature of health trajectories in older populations.

Status uncertainty and maintenance problems

There are clear conceptual similarities between the health deterioration of older adults and the degradation of the system in maintenance problems. In both domains, interventions are required to prevent costly failures, while the objective is to determine the optimal timing of interventions to minimize expected costs and adverse outcomes. As preventive maintenance problems are well studied, we limit our discussion to the critical features. Preventive maintenance strategies are often classified as time-based (TBM) or condition-based (CBM). In TBM, maintenance is performed at fixed intervals, whereas CBM triggers interventions based on observed system state or sensor data [135]. Recent studies confirm that CBM generally achieves lower costs and higher availability than TBM, in particular when system lifetimes are neither perfectly predictable nor entirely random [4], as CBM better balances the trade-off between maintenance and unexpected failure.

The application of CBM requires an optimal scheduling of inspections and repairs in an environment of stochastic system degradation. In the literature, such policies are typically derived using stochastic control models, especially MDPs or Semi-Markov Decision Processes (SMDP). For example, Chen and Trivedi formulated an SMDP for CBM that jointly optimizes the inspection rate and maintenance policy [39]. They show that when the deterioration rates are identical between stages, the optimal SMDP policy reduces to a dynamic threshold rule, with the threshold depending on the inspection rate. Liao et al. [132] propose a reliability-centered sequential preventive maintenance model for repairable deteriorating systems. Unlike traditional fixed-interval strategies, their approach schedules maintenance actions based on continuously monitored system reliability. Liu et al. [136] extend this perspective by incorporating age- and state-dependent operating costs that increase even when the system remains functional. Their res-

ults show that accounting for such types of operating costs leads to more cautious and cost-effective maintenance strategies than traditional models that ignore them.

To address uncertainty in system state information, recent research has increasingly focused on dynamic inspection scheduling using MDPs and partially observable MDPs (POMDPs). Assis and Marques [9] propose a method that adapts inspection intervals based on safe and unsafe time windows, reducing costs by anticipating risk. Noori et al. [160] develop a two-stage stochastic model that links inspection timing to maintenance planning under demand uncertainty. In situations where system states cannot be directly observed, POMDPs offer significant advantages, as demonstrated by Papakonstantinou and Shinozuka [172] and Guo and Liang [88]. Policies may feature information-gathering actions that are more cost-effective over time.

Admission policies of NH

Timely admission to a NH is critical to preventing acute hospitalizations and managing patient flow in elderly care [14]. Given workforce shortages and limited capacity, expanding LTC capacities is not feasible in the short term, making better admission strategies and managing waiting lists essential. Several studies have explored prioritization and allocation policies. Burkell et al. [29] compared First-Come-First-Served (FCFS) and needs-based admissions, showing that the latter significantly reshapes placement decisions. Meiland et al. [146] found that urgency-based prioritization benefits high-risk patients without disadvantaging lower-risk individuals. Preference-based models, such as that of Arntzen et al. [6], have reduced abandonment rates and improved waiting times by incorporating patient preferences. Hallal [91] used a Markov model to simulate patient flow through home care, hospitals, and NHs, identifying policies that reduce bottlenecks. Those studies frequently neglect the importance of a patient's preference for remaining at home, resulting in policies that focus on only reducing waiting time.

This chapter introduces an MDP framework that adapts concepts from *preventive maintenance* to the context of NH admissions. We model an older adult's health as a Discrete-Time Markov Chain (DTMC), interpret an acute crisis as a system failure, and treat check-ups as state inspections that reveal uncertain system conditions. Admitting an individual to an NH can be viewed as a preventive repair action, but with the fundamental difference that, in care for older adults, the system effectively terminates after such an intervention. Unlike prior work on patient flow, prioritization, or maintenance scheduling, our approach integrates these elements into a single model that optimizes when to inspect and when to admit, balancing the costs of early placement against the risk of sudden health deterioration. The result is a personalized, cost-sensitive policy for a timely and efficient admission to LTC.

4.3 Modeling of Patient Health Status

We employ a DTMC to model health decline in older adults. This model, referred to as the health model $\mathbb{H}$, is used in Section 4.4 to determine the optimal timing for NH admission. The model $\mathbb{H}$ consists of N health states and one absorbing *crises state* (C), where each state represents a different set of health conditions. Hence, the state space is $S_{\mathbb{H}} = \{1, \ldots, N, C\}$, and the transition matrix $\mathbf{P}_{\mathbb{H}}$ is defined as

$$\mathbf{P}_{\mathbb{H}} = \begin{bmatrix} \mathbf{Q} & \mathbf{S} \\ \mathbf{0} & 1 \end{bmatrix}, \tag{4.1}$$

$$\text{with } \mathbf{Q} = \begin{bmatrix} P_{\mathbb{H}}(1,1) & \cdots & P_{\mathbb{H}}(1,N) \\ \vdots & \ddots & \vdots \\ P_{\mathbb{H}}(N,1) & \cdots & P_{\mathbb{H}}(N,N) \end{bmatrix} \text{ and } \mathbf{S} = \begin{bmatrix} P_{\mathbb{H}}(1,C) \\ \vdots \\ P_{\mathbb{H}}(N,C) \end{bmatrix}.$$

Here, the matrix $\mathbf{Q}$ corresponds to the transitions for transient states and $\mathbf{S}$ is the exit vector due to a crisis. In the model, states with higher indices are associated with a declining health condition, such that the risk of a crisis $P_{\mathbb{H}}(i, C)$ increases with the corresponding state i. In the experiments, we will additionally assume that the time step is sufficiently small such that only transitions to neighboring states or the absorbing state are possible, i.e., $P_{\mathbb{H}}(i, j) = 0$ for $i, j \in \{1, \ldots, N\}$ and $|i - j| > 1$; this assumption is made for convenience and is not required by the model. The health condition in which a patient enters the system at time 0 is described by the initial distribution, denoted by the vector $\boldsymbol{\beta} = [\beta_1 \ \cdots \ \beta_N \ \beta_C]$, with $\beta_i \in [0, 1]$ for $i \in \{1, \ldots, N, C\}$ and $\sum_{i=1}^{N} \beta_i + \beta_C = 1$. This vector captures the extent of health information accessible once a patient becomes visible within the health system, where we allow patients to arrive in state C, (cf. [52]). The health model is visualized in Figure 4.2.

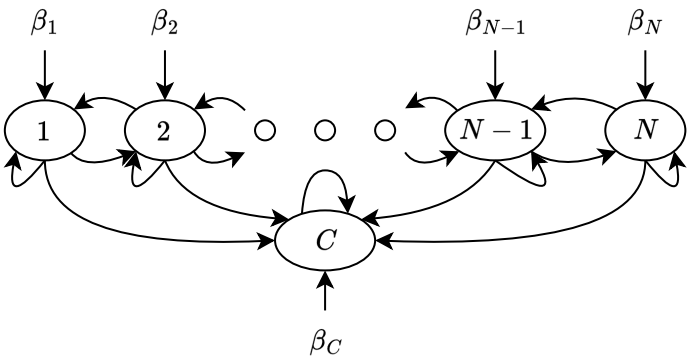


Figure 4.2: Health model $\mathbb{H}$ with initial distribution $\boldsymbol{\beta} = [\beta_1 \ \cdots \ \beta_N \ \beta_C]$ and only transitions to neighboring states or the crisis state.

Clearly, the evolution of the health state during n consecutive time steps follows from the n-step transition matrix $\mathbf{P}_{\mathbb{H}}^{(n)} = \mathbf{P}_{\mathbb{H}}^{n}$, which yields the n-step transition

probabilities $P_{\mathbb{H}}^{(n)}(i,j)$ for $i,j \in S_{\mathbb{H}}$. As the health model has a finite state space and a single absorbing state, each patient will ultimately enter state C if there is no intervention. The expected time to enter the crisis state follows from standard results (see e.g. [184, Section 4.6]) and is given by the fundamental matrix $\mathbf{F} = (\mathbf{I} - \mathbf{Q})^{-1}$. In particular, the (i,j)-th entry of $\mathbf{F}$ is interpreted as the expected number of times that the DTMC visits the transient state j when starting in i. Consequently, the expected time until absorption is given by the column vector $\mathbf{t} = \mathbf{F}\mathbf{1}$, with $\mathbf{1}$ a column vector of ones. Hence, incorporating the initial distribution and combining the above yields the expected time until absorption:

$$\begin{bmatrix}\beta_1 & \cdots & \beta_N\end{bmatrix} \mathbf{t} = \begin{bmatrix}\beta_1 & \cdots & \beta_N\end{bmatrix} (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{1}. \tag{4.2}$$

4.4 Markov Decision Process of Nursing Home Admission

We introduce an MDP, denoted $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, to optimize admission decisions, balancing the costs of early admission against the risks of delayed admission. The MDP is an extension of the health model $\mathbb{H}$. Here, state $s_t = (i,j)$ indicates that the patient's last observed health status was i for $i = \{1,\dots,N\}$, and j time steps have passed since that observation. The state space is

$$\mathcal{S} = \{(i,j) \mid i \in \{1,\dots,N\},\ j \in \{0,\dots,T\}\} \cup \{C,\ L\},$$

with L an absorbing state that represents LTC in a NH and T the maximum time between health assessments.

We define $\mathcal{A} = \{\mathit{wait},\ \mathit{check},\ \mathit{admit}\}$ as the set of potential actions. The action *check* is feasible in all states and corresponds to a health assessment, after which the health status information is up to date; the corresponding transition probabilities are governed by the health model $\mathbb{H}$. The action *admit* corresponds to an immediate admission to a NH, i.e., results in a transition to state L. An admission is only feasible immediately after *check*, so that the health status is known at admission, similar to its real-life counterpart. State L is absorbing, meaning that once a patient is admitted to the NH, the process ends and no further actions are taken. Finally, the action *wait* indicates that a patient will not be admitted and there will be no health check. Regardless of the action taken, there is always a risk that the patient experiences a crisis and transitions to state C. In our model, a crisis represents a severe situation such that the only subsequent action is to *admit* a patient to a NH. A less severe crisis situation may be incorporated in the direct costs. The complete model $\mathcal{M}$ is visualized in Figure 4.3. The transition probabilities and direct costs of each action are explained below.

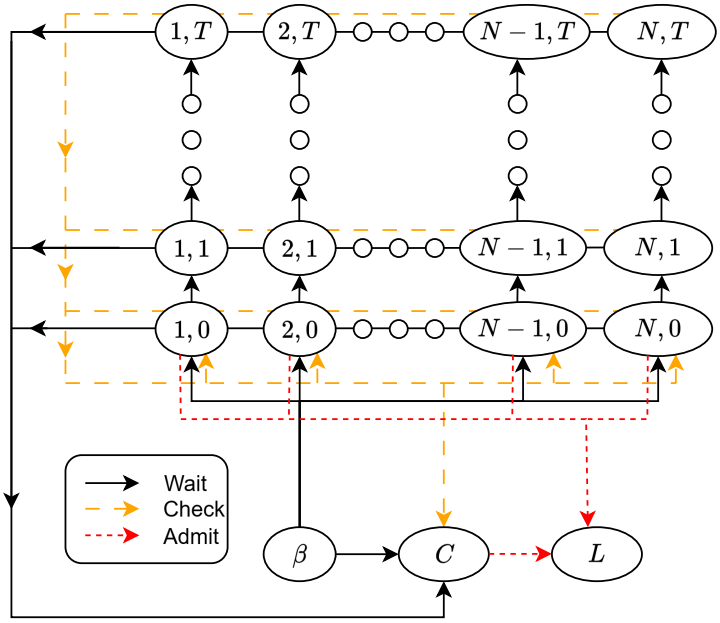


Figure 4.3: MDP for preventive nursing home admission with actions wait, check, and admit, and transitions to crisis or nursing home states.

Action: wait The action *wait* corresponds to the absence of the action check or admit. This action is feasible in states (i, j) for $i \in \{1, \ldots, N\}, j \in \{0, \ldots, T-1\}$. From state (i, j) there are two possible transitions: the process moves to state $(i, j+1)$, or the patient experiences a crisis and moves to state C. The transition probabilities are obtained from the matrix of n-step probabilities of the health model $\mathbb{H}$ ($\mathbf{P}_{\mathbb{H}}^{(n)}$) as follows:

$$\begin{aligned}\mathbb{P}(s_{t+1} = C | s_t = (i,j), a_t = wait) &= \frac{\mathbb{P}(s_{t+1} = C \cap s_t \neq C | s_{t-j} = (i,0))}{\mathbb{P}(s_t \neq C | s_{t-j} = (i,0))} \\ &= \frac{P_{\mathbb{H}}^{(j+1)}(i,C) - P_{\mathbb{H}}^{(j)}(i,C)}{1 - P_{\mathbb{H}}^{(j)}(i,C)}, \end{aligned} \tag{4.3}$$

which holds because if $s_t = (i, j)$, the patient did not experience a crisis up to time t. Obviously, the probability of moving to state $(i, j+1)$ is given by $\mathbb{P}(s_{t+1} = (i, j+1) | s_t = (i, j), a_t = wait) = 1 - \mathbb{P}(s_{t+1} = C | s_t = (i, j), a_t = wait)$.

Action: check The action *check* corresponds to a health professional meeting the patient to perform a health assessment. After a check, the actual health state is revealed and the process moves to $(k, 0)$, for $k \in \{1, \ldots, N\}$ or to state C in case of a crisis. Note that $|k - i| \leq j + 1$ in case only transitions to neighboring

transient health states are feasible. The probability of a crisis is similar to (4.3):

$$\mathbb{P}(s_{t+1} = C | s_t = (i,j), a_t = check) = \frac{P_{\mathbb{H}}^{(j+1)}(i,C) - P_{\mathbb{H}}^{(j)}(i,C)}{1 - P_{\mathbb{H}}^{(j)}(i,C)}.$$

Utilizing the n-step transition probabilities of the health model, we have, for $k \in \{1,\dots,N\}$,

$$\begin{aligned}\mathbb{P}(s_{t+1} = (k,0) | s_t = (i,j), a_t = check) &= \frac{\mathbb{P}(s_{t+1} = (k,0) \cap s_t \neq C | s_{t-j} = (i,0))}{\mathbb{P}(s_t \neq C | s_{t-j} = (i,0))} \\ &= \frac{P_{\mathbb{H}}^{(j+1)}(i,k)}{1 - P_{\mathbb{H}}^{(j)}(i,C)}.\end{aligned}$$

Action: admit The action *admit* represents the decision to admit a patient to a NH. This action can only be performed in states C and $(i,0)$, for $i \in \{1,\dots,N\}$. Admission happens instantaneously, preventing any risk of a patient crisis during this transition. The transition probability is trivial, for $l \in \{(1,0),\dots,(N,0),C\}$,

$$\mathbb{P}(s_{t+1} = L | s_t = l, a_t = admit) = 1.$$

Costs We consider the following four cost components in the MDP: (i) *home-care cost* ($\mathrm{K_{HC}}(i)$), (ii) *nursing-home cost* ($\mathrm{K_L}(i)$), (iii) *checking cost* ($\mathrm{K_{check}}$), and (iv) *crisis cost* ($\mathrm{K_C}$). Cost (i) represents the cost associated with older adults living at home, where they generally receive professional care. The cost of home care per time step $\mathrm{K_{HC}}(i)$ will increase in i, since worse health typically requires more care. Cost (ii) is incurred only when the action admit is taken; it reflects the expected cost for the remainder of the patient's stay in the NH, which again depends on the patient's health status, because healthier patients tend to stay longer. Cost (iii) is the cost of a healthcare professional, such as the general practitioner, performing a health assessment to update the patient's health status. Finally, cost (iv) captures the consequences of a crisis, such as an ED visit, hospitalization, or other acute events. Currently, only monetary costs are included; however, a utility component could be incorporated for each cost to reflect patient preferences, for example, a desire to stay at home or avoid a crisis. Thus, the costs of action a in state s leading to state s' are given by

$$R(s,a,s') = \begin{cases} \mathrm{K_C} & \text{if } a = \text{wait and } s = (i,j),\ s' = C, \\ \mathrm{K_{HC}}(i) & \text{if } a = \text{wait and } s = (i,j),\ s' = (i,j+1), \\ \mathrm{K_C} + \mathrm{K_{check}} & \text{if } a = \text{check and } s = (i,j),\ s' = C, \\ \mathrm{K_{HC}}(i) + \mathrm{K_{check}} & \text{if } a = \text{check and } s = (i,j),\ s' = (k,0), \\ \mathrm{K_L}(N) & \text{if } a = \text{admit and } s = C,\ s' = L, \\ \mathrm{K_L}(i) & \text{if } a = \text{admit and } s = (i,0),\ s' = L. \end{cases}$$

Value Function and Optimal Policy Let the total expected cumulative cost under policy π be $R^\pi = \mathbb{E}\left[\sum_{t=0}^{\infty} R(s_t, \pi(s_t), s_{t+1})\right]$. The goal of the MDP is to determine an optimal policy π^* that minimizes the expected cumulative cost incurred over time. To do this, we define the value function $V(s)$ as the minimal expected cost starting from state s and following the optimal policy thereafter. The Bellman equation for the value function is given by:

$$V(s) = \min_{a \in \mathcal{A}(s)} \sum_{s' \in \mathcal{S}} P(s' \mid s, a) \left[R(s, a, s') + V(s')\right], \tag{4.4}$$

where $\mathcal{A}(s)$ the set of feasible actions in state s. We have $V(L) = 0$, since no further actions are taken once a patient is admitted to a NH.

To compute the value function and optimal policy, we apply *value iteration*, an iterative dynamic programming method. Starting from an arbitrary initial value function $V_0(s)$ (e.g., all zeros), the value function is updated iteratively using:

$$V_{t+1}(s) = \min_{a \in \mathcal{A}(s)} \sum_{s' \in \mathcal{S}} P(s' \mid s, a) \left[R(s, a, s') + V_t(s')\right],$$

until convergence, i.e., until

$$\max_{s \in \mathcal{S}} |V_{t+1}(s) - V_t(s)| < \varepsilon$$

for a small threshold $\varepsilon > 0$. Once convergence is achieved, the optimal policy is obtained by:

$$\pi^*(s) = \arg\min_{a \in \mathcal{A}(s)} \sum_{s' \in \mathcal{S}} P(s' \mid s, a) \left[R(s, a, s') + V(s')\right].$$

Hence, for each state s, the policy $\pi^*(s)$ takes the action that minimizes the expected cumulative cost, effectively balancing the risks and costs of premature admission, delayed intervention, and health deterioration.

For any policy π, the two main performance measures are the total expected cost R^π and the probability of an avoidable crisis $P^\pi(C)$, given by $P^\pi(C) = \sum_{t=1}^{\infty} \mathbb{P}^\pi(s_t = C)$, with $\mathbb{P}^\pi$ denoting the probability measure induced by policy π. Note that we omit patients entering the system in a crisis $\mathbb{P}(s_0 = C) = \beta_C$, as they could not have been avoided. The two performance measures follow directly from the Markov chain associated with policy π. Specifically, the Markov chain has state space $\mathcal{S}$ in which L is an absorbing state. The one-step transition matrix $\mathbf{P}^\pi$, depending on policy π, has dimension $(N \cdot (T+1)+2) \times (N \cdot (T+1)+2)$ and can be partitioned as

$$\mathbf{P}^\pi = \begin{bmatrix} \mathbf{Q}^\pi & \mathbf{S}^\pi \\ \mathbf{0} & 1 \end{bmatrix},$$

with $\mathbf{Q}^\pi$ describing the transitions among the transient states and $\mathbf{S}^\pi$ the exit vector. Now, the *fundamental matrix* $\mathbf{F}^\pi = (\mathbf{I} - \mathbf{Q}^\pi)^{-1}$ gives the expected number

of visits to each transient state before absorption. For the expected cost, let $r^\pi(s)$ be the expected immediate cost incurred when taking the action prescribed by policy π in transient state $s \in \mathcal{S} \setminus L$:

$$r^\pi(s) = \sum_{s' \in \mathcal{S}} \mathbb{P}(s' \mid s, a = \pi(s))\, R(s, a, s'),$$

with corresponding vector $\mathbf{r}^\pi = (r^\pi(s))_{s \in \mathcal{S} \setminus L}$. Define the vector $\mathbf{g}$ by

$$\mathbf{g} = \mathbf{F}^\pi \mathbf{r}^\pi = (\mathbf{I} - \mathbf{Q}^\pi)^{-1} \mathbf{r}^\pi.$$

The s-th element of $\mathbf{g}$ gives the expected total cost when starting in transient state s. Using the initial distribution, the total expected cost is given by: $R^\pi = \sum_{i=1}^N \beta_i g_{(i,0)} + \beta_C g_C$. For the risk of crisis, observe that state C is visited exactly once if a crisis event occurs, whereas state C is not visited otherwise. Hence, from a current state s, the risk of a crisis corresponds to the (s, C)-th element of $\mathbf{F}^\pi$. Incorporating the initial distribution, yields

$$P^\pi(C) = \frac{1}{1 - \beta_C} \sum_{i=1}^N \beta_i \mathbf{F}^\pi_{(i,0),C}.$$

In some scenarios, there may be a desired upper bound for the risk of a crisis. Suppose that the constrained optimization problem is $\min_\pi R^\pi$ s.t. $P^\pi(C) \leq \alpha_C$, for some target risk α_C. Its Lagrangian relaxation with multiplier λ is

$$\begin{aligned} &\min_\pi \; R^\pi + \lambda\,(P^\pi(C) - \alpha_C) \\ &= \min_\pi \; \mathbb{E}\left[\sum_{t=0}^\infty R(s_t, \pi(s_t), s_{t+1}) + \lambda \mathbb{1}_{\{s_{t+1} = C\}}\right] - \lambda \alpha_C. \end{aligned} \tag{4.5}$$

Thus, by appropriately choosing the cost parameters, the objective of the MDP can naturally incorporate constraints.

Remark 4.1. Observe that the optimal policy also specifies the timing of the next check immediately after an assessment, allowing an appointment to be scheduled in advance, provided that no crisis occurs in the meantime. An alternative MDP formulation embeds the process at assessment epochs; the state space is then reduced to $\{1, \ldots, N, C, L\}$, while the action space is extended to include the scheduling of the next check.

Remark 4.2. Our MDP differs fundamentally from typical formulations in preventive maintenance. Preventive maintenance models usually involve component replacement, whereas our model includes an absorbing state. As a result, the value function in preventive maintenance models generally increases with the state of the system, since greater degradation leads to higher future costs. In contrast, in our model the value function $V(i, j)$ decreases with i, since worse health states are associated with lower expected costs over the remaining lifetime.

4.5 Results

This section presents results obtained using a realistic health model. In Section 4.5.1, we use real-world data to calibrate the health model $\mathbb{H}$, and the cost parameters are discussed in Section 4.5.2. The optimal policy can be found in Section 4.5.3, where it is compared to some benchmarks reflecting current practice. More specifically, performance is assessed based on (i) the total expected cost R^π, and (ii) fraction of patients who suffered a crisis before NH admission that could have been avoided $P^\pi(C)$. In Section 4.5.4 we study the impact of the cost parameters, whereas Section 4.5.5 investigates the impact of a maximum crisis fraction on the total cost. Section 4.5.6 considers the impact of the number of health classes. We study the impact of delay between the admission decision and the actual NH admission, during which patients may still deteriorate or experience a crisis in Section 4.5.7.

4.5.1 Calibrating the health model

To calibrate the health model $\mathbb{H}$ with real-world data, we used non-public microdata from Statistics Netherlands (CBS) [139], with which we created healthcare trajectories of older adults in Chapter 2. We extracted survival curves that represent the time between the initiation of healthcare services and the onset of a crisis event. We used data from 2017-2019. Due to the relatively short time span, censoring of survival lengths plays a prominent role, which we addressed using Kaplan-Meier estimates [118]. Our goal is to determine the parameters of a parsimonious health model $\mathbb{H}$ that fits the empirical survival curve using a weekly time step.

To define the parameter structure of the health model, we impose the following simplifying assumptions. We choose $N = 20$, as larger values make classification overly complex for health professionals, while smaller values result in a poorer fit and were considered impractical by practitioners. The risk of a crisis is assumed to increase exponentially with the health state:

$$P_{\mathbb{H}}(i, C) = p_C \cdot g_C^i,$$

for some $p_C < 0.001$ reflecting a very small risk of crisis in a healthy state and $g_C > 1$ the exponential growth factor reflecting the accelerating risk. The probability of a health improvement decreases, while the probability that the health deteriorates increases, yielding

$$P_{\mathbb{H}}(i, i-1) = p_u - g_u \frac{i}{N}, \qquad P_{\mathbb{H}}(i, i+1) = p_d + g_d \frac{i}{N},$$

for some $p_u, p_d \in [0, 1]$ and $g_u \in [0, p_u]$, $g_d \in [0, p_d]$. Allowing only jumps to neighboring transient states, we have $P_{\mathbb{H}}(i, i) = 1 - P_{\mathbb{H}}(i, i-1) - P_{\mathbb{H}}(i, i+1) - P_{\mathbb{H}}(i, C)$, which should be non-negative. Finally, for the initial distribution, we have the probability β_C of entering a crisis immediately upon arrival. The initial

distribution for the transient states is assumed to decrease linearly with the health state, giving, for $i = 1, \ldots, N$,

$$\beta_i = i \cdot \frac{1 - \beta_C}{\sum_{j=1}^{N} j} = \frac{2i(1 - \beta_C)}{N(N+1)}.$$

The set of parameters $\theta = (\beta_C, p_C, g_C, p_d, g_d, p_u, g_u)$ is determined by random search [5] while evaluating the quality of the fit by

$$\sum_{t=1}^{\infty} \left(S_{\mathbb{H}}(t;\theta) - S_{\mathrm{KM}}(t)\right)^2, \tag{4.6}$$

with $S_{\mathbb{H}}(t;\theta)$ the survival probability of the fitted DTMC with parameter set θ and $S_{\mathrm{KM}}(t)$ the survival probability based on empirical data and the Kaplan-Meier estimator. Several low-error fits were obtained, demonstrating that the health model is flexible in various settings. The parameter values listed in Table 4.1 both achieve low error and align with the intuition of healthcare professionals. The survival curves of the data and the health model, as presented in the left panel of Figure 4.4, also show that the health model fits the data well. To obtain additional insight in the time to crisis for various health states, the right panel of Figure 4.4 shows the expected absorption times per health state. Lower health states are clearly associated with longer expected times until crisis, consistent with a progressive deterioration model.

Table 4.1: Fitted health model parameters.

Parameter	Value
β_C	0.1052
p_C	0.001
g_C	1.22
p_{d}	0.0011
g_{d}	0.2109
p_{u}	0.0428
g_{u}	0.0129

Table 4.2: Cost settings for the model.

Parameter	Value
$\mathrm{K}_{\mathrm{HC}}^{-}$	115.96
$\mathrm{K}_{\mathrm{HC}}^{+}$	322.32
$\mathrm{K}_{\mathrm{L}}^{-}$	192,383.10
$\mathrm{K}_{\mathrm{L}}^{+}$	337,974.70
$\mathrm{K}_{\mathrm{check}}$	50
K_{C}	11,700

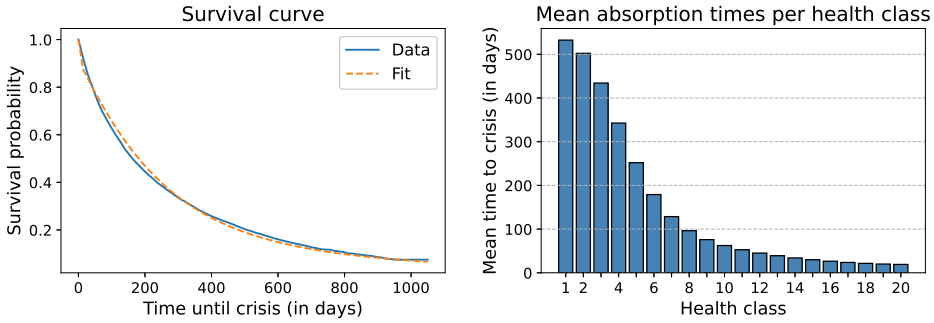


Figure 4.4: Left: Empirical and fitted survival curves based on observed time-to-crisis data. Right: Mean absorption times to crisis per health state for the fitted health model.

4.5.2 Cost parameters

The structure of the cost parameters used in the MDP is as follows. First, it is assumed that the cost of home care increases linearly with health status, as patients with a worse health status often require more intense home care. That is, for $i = 1, \ldots, N$,

$$\mathrm{K_{HC}}(i) = \mathrm{K_{HC}^{-}} + \frac{i}{N}\left(\mathrm{K_{HC}^{+}} - \mathrm{K_{HC}^{-}}\right), \tag{4.7}$$

with $\mathrm{K_{HC}^{-}}$ and $\mathrm{K_{HC}^{+}}$ the smallest and highest home care costs, respectively. Second, the cost of the NH represents the expected cumulative cost throughout the rest of the patient's stay. Patients with a better health status will typically have a longer length of stay in an NH. Hence, the NH cost is assumed to decrease linearly with the health status, that is, for $i = 1, \ldots, N$,

$$\mathrm{K_{L}}(i) = \mathrm{K_{L}^{+}} - \frac{i}{N}\left(\mathrm{K_{L}^{+}} - \mathrm{K_{L}^{-}}\right), \tag{4.8}$$

with $\mathrm{K_{L}^{-}}$ and $\mathrm{K_{L}^{+}}$ the smallest and highest NH costs, respectively. The NH cost for admission in crisis state is assumed to be equal to that of health state N.

Estimating the required cost parameters, including the costs of crisis and checks, in this healthcare setting is inherently difficult. Based on [205], we derived the most plausible estimates for use in the model; see Table 4.2.

4.5.3 Optimal policy

The results are based on the fitted health model from Section 4.5.1, the cost parameters from Section 4.5.2 and $T = 26$ corresponding with at most 26 weeks between checks. The MDP was solved using value iteration, with $\varepsilon = 1 \cdot 10^{-6}$; the optimal policy is depicted in Figure 4.5. This policy reveals that patients are admitted to the NH only when they reach the final health state 20; monitoring

becomes more intensive as they approach this critical threshold. The total expected cost under the optimal policy is 209,376.06, which is mainly driven by the high cost of NH care. Interestingly, 54.48% of patients arrive at the NH due to an unavoidable crisis. This can be attributed to the fact that experiencing a crisis is relatively inexpensive compared to the NH cost when proactively admitting a patient in a healthier state. By strategically admitting patients in the health state 20, the model takes advantage of the lowest possible cost of admission to the NH and aims to avoid a crisis by more intensive monitoring as health deteriorates.

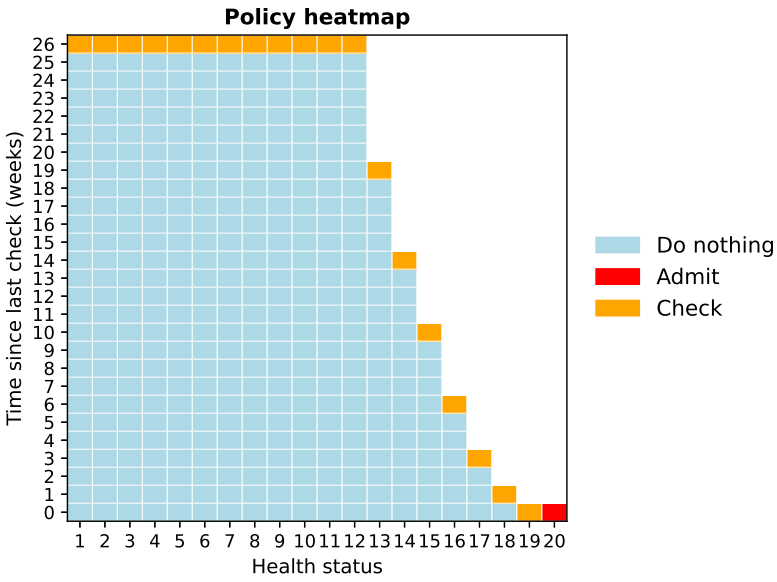


Figure 4.5: Optimal policy under the base case scenario.

To compare this approach with current practice, several benchmark policies were constructed. In these benchmarks, monitoring is time-based, with checks performed every τ_{check} time step, while admission is condition-based in which a patient with a health status equal or worse than θ_{admit} is admitted to the NH. Figure 4.6 visualizes two examples of such a policy. In Figure 4.7 the total expected costs and risk of crisis of the optimal policy are compared with the benchmark policies with $\tau_{\text{check}} \in \{0, 2, 4, 26\}$ and $\theta_{\text{admit}} \in \{1, \ldots, 20\}$. The optimal policy achieves the lowest total expected costs. Interestingly, several benchmark strategies yield a lower crisis fraction, but this comes at the expense of higher overall costs. Furthermore, the results suggest that the monitoring frequency parameter τ_{check} has a minor influence on both the cost and the crisis fraction when patients are admitted in relatively healthy states. However, as the admission threshold θ_{admit} shifts to less healthy states, the impact of τ_{check} becomes more pronounced. This indicates that frequent monitoring is only required when patients are at higher risk and can otherwise be scheduled infrequently without substantially affecting outcomes. In contrast to the optimal policy, benchmark policies cannot dynamically adjust the

monitoring frequency depending on the health state.

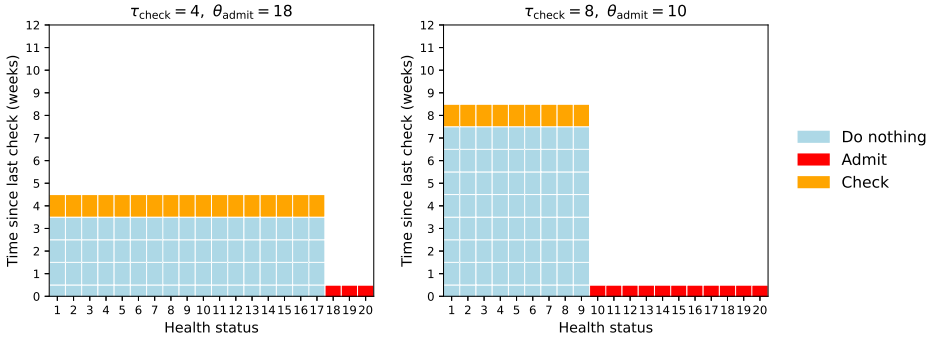


Figure 4.6: Two examples of time-based monitoring and condition-based admission policies.

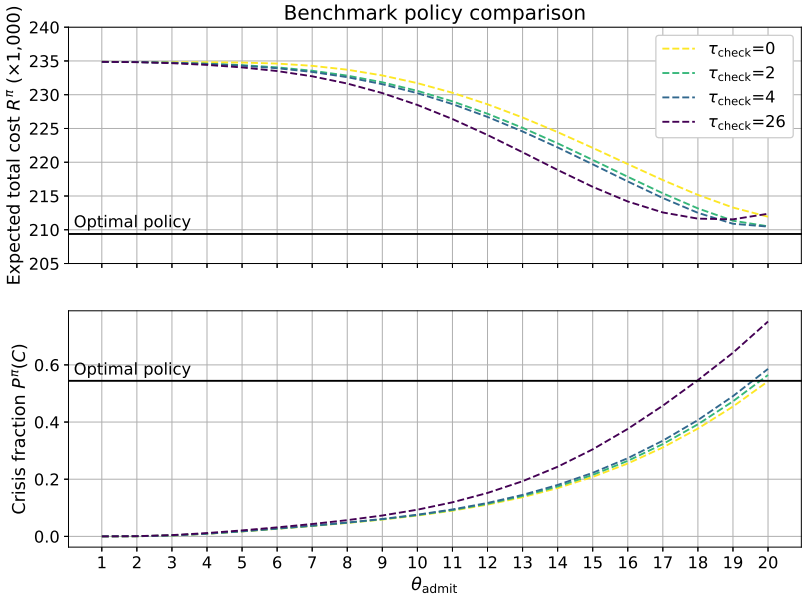


Figure 4.7: Comparison of the optimal policy to policies with time-based monitoring and condition-based admissions.

The optimal policy was also compared with a benchmark policy that admits older adults a fixed number of time units after entry, under various check frequencies and admission times. A time-based policy with an admission time of 54 weeks,

approximately equal to the average admission time under the optimal policy, resulted in an expected total cost that was 13,822 higher than that of the optimal policy, and a higher crisis fraction (70% versus 54%). Extending the admission time to three years (156 weeks) reduced the expected total cost to 218,949 but further increased the crisis fraction to 93%. In the extreme case where all admissions occurred through crises, the expected cost decreased slightly to 216,279.60. These results highlight that relying solely on elapsed time for admission decisions is inefficient: incorporating information about individuals' health states into monitoring and admission strategies leads to substantially better outcomes in both cost and crisis prevention.

The results show that policies that reduce total costs also lead to a considerable crisis fraction. This is due to the relatively high costs of early admissions to NHs. Despite uncertainty in cost estimates, the analysis reveals a structural issue: preventive NH admissions hardly provides a monetary benefit. From a policy measure perspective, these results suggest that preventive NH admissions should be justified by objectives beyond cost reduction, such as improving quality of care or preventing crises, or that financial incentives should be implemented to promote preventive care. In the following sections, we use sensitivity analysis to examine the effect of different strategic choices. Specifically, we consider NH costs, crisis costs, level of uncertainty in health status, and the effect of admission delays.

4.5.4 Impact of cost parameters

In this section, we explore how the optimal policy is affected by changes in cost parameters. We focus mainly on NH cost, while we address the cost of crisis in Section 4.5.5 below. Varying NH cost may reflect changes in policy measures or changes in patient preferences regarding NH residence; higher costs mimic reluctance to enter a NH, while lower costs represent scenarios in which NH care is preferred. For the sensitivity analysis, we modify the NH cost as $\tilde{\mathrm{K}}_{\mathrm{L}}(i) = \alpha \cdot \mathrm{K}_{\mathrm{L}}(i)$, for $i \in \{1, \ldots, N\}$, with $\alpha > 0$ a cost multiplier. Here, $\alpha < 1$ represents scenarios in which patients prefer NH care, while $\alpha > 1$ indicates a stronger aversion to it. For each value of α, the optimal policy, associated costs, and the crisis fraction are determined.

Figure 4.8 shows the total expected costs, based on the original cost parameters, (vertical left axis) and the crisis fraction (vertical right axis) for various NH cost multipliers α. As α decreases, implying a greater preference for NH care, patients are admitted earlier, leading to a significant reduction in crisis admissions. In contrast, when the NH becomes increasingly expensive, the policy tends to converge back to the original solution. Specifically, for $\alpha > 0.6$, the results are comparable, suggesting that the policy is robust to increases in NH cost. In other words, a considerable reduction in NH cost is required before less crisis admissions become financially viable. It should be noted that the results are based on the model assumption of a linear structure for NH costs across health states.

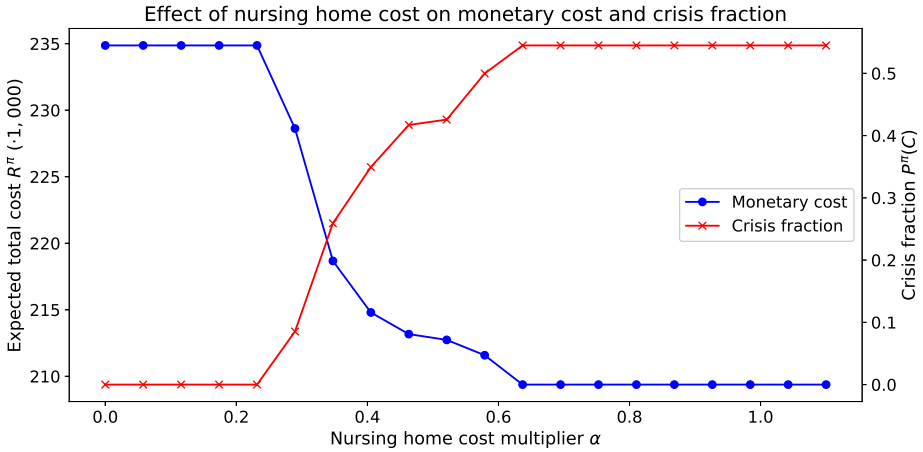


Figure 4.8: Impact of NH cost multiplier α on total expected cost (blue) and fraction of patients experiencing a crisis (red).

Changes in home care costs ($K_{HC}(i)$) have effects that are largely opposite to those of changes in NH costs ($K_L(i)$); due to similarity, we omit the corresponding results. When home care becomes more expensive, early NH admission becomes relatively more attractive. The optimal policy will shift admissions to earlier health states in order to avoid prolonged periods of costly home care. In contrast, when home care is relatively inexpensive, the optimal policy favors postponing admission to the NH, thereby increasing the risk of a crisis but reducing the total cost. Similarly to the impact of NH costs, this trade-off reflects how the model balances the cost of waiting at home against the risk and cost of a potential crisis.

The cost of monitoring (i.e., checking) has a surprisingly limited influence on the optimal policy and the corresponding performance. For a wide range of checking costs (K_{check}), the optimal policy remains largely unchanged, with regular checks occurring as patients approach critical health states. However, when checking becomes very costly, the model begins to avoid it altogether. In such cases, even the final check at time T is then often circumvented by preemptive NH admission. These results indicate that monitoring influences decisions only when its cost exceeds that of a crisis or an early admission, highlighting the robustness of the optimal policy to moderate variations in checking cost.

4.5.5 Controlling the risk of a crisis

A crisis can severely affect the quality of life of an older adult. As such, the impact can extend beyond the monetary costs, which typically result from ED visits and hospitalizations. In this section, we investigate the impact on the total expected costs in the event of a maximum acceptable risk of a crisis. In view of Equation (4.5), the corresponding constrained optimization is equivalent to our MDP by modifying the cost of a crisis (K_C). To assess the sensitivity of the model to

the risk of crisis situations, we modify the crisis cost as $\tilde{K}_C = \alpha \cdot K_C$, with $\alpha > 0$ a crisis cost multiplier. Again, we determine the optimal policy for each value of α.

The fraction of admissions due to a crisis (vertical right axis) and the total expected costs, based on the original cost parameters (vertical left axis), are shown in Figure 4.9. Initially, the multiplier α should be chosen well above 2 for any impact on the risk of a crisis. With an increase in total cost from 209,376 to 221,461, i.e. an additional 12,084 per patient, the crisis fraction can be reduced from 54% to 18%. Further increasing the cost to 234,868 yields a crisis fraction of 0%. Although this increase in cost appears to be relatively small, it should be noted that every patient has at least a cost of $K_L(N) = 192{,}383.10$ due to an NH admission. Nevertheless, this shows that a relatively modest increase in cost can lead to a substantial decrease in preventable crises.

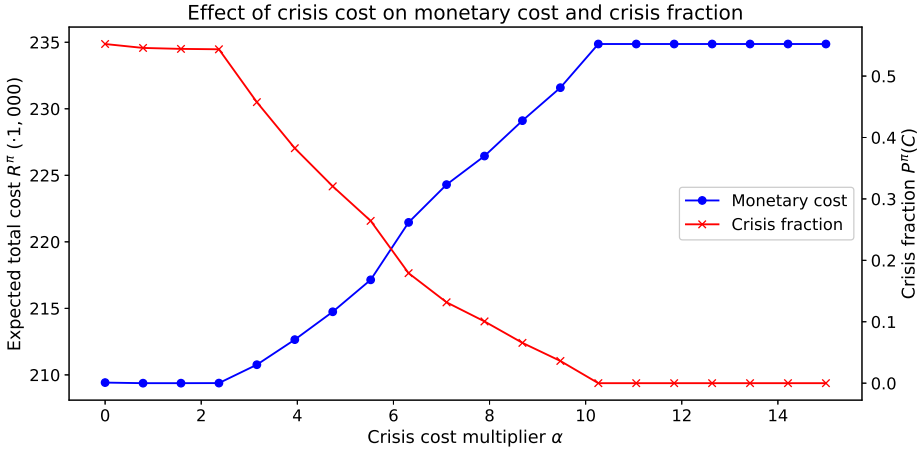


Figure 4.9: Effect of decreasing the fraction of patients experiencing a crisis (red) on the total expected cost (blue) using the crisis cost multiplier α.

4.5.6 Impact of health state granularity

The granularity of the representation of the health state is dictated by the number of health classes N. A decrease in N leads to a less precise classification, which occurs particularly in contexts where subtle health states are difficult for professionals to differentiate. This section illustrates how a coarser description of health impacts both the admission and the monitoring strategy, as well as the overall effectiveness. The base case involves $N = 20$, whereas we constructed alternative models with $N = 10$ and $N = 5$ health classes for coarser discretizations. For comparability, each class with $N = 5$ aggregates four consecutive classes from the $N = 20$ case. A similar pattern is assumed for $N = 10$ where two consecutive classes in the $N = 20$ model are aggregated into one class. The optimal policies for the $N = 5$ and $N = 10$ models are obtained from the corresponding MDPs. For a visual comparison, the resulting policies are projected back onto the original

system with 20 classes; the projected policies for $N = 5$ and $N = 10$ are shown in Figure 4.10.

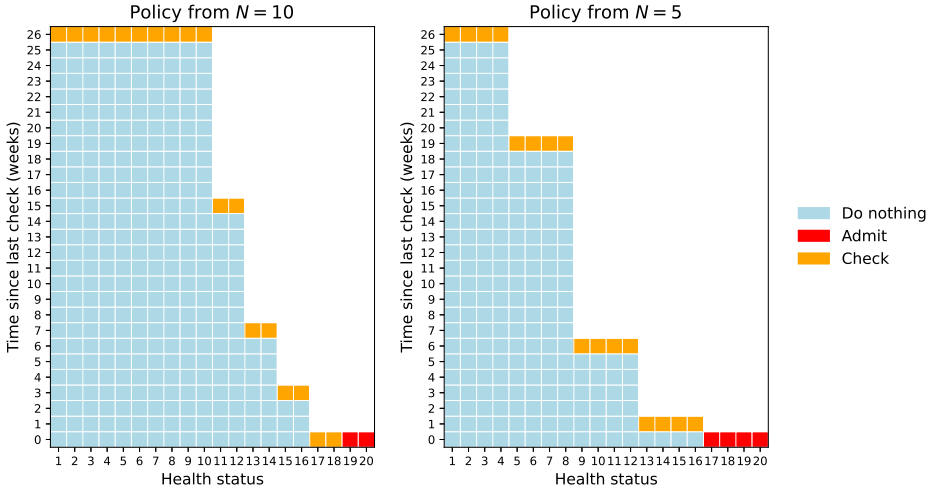


Figure 4.10: Comparison of policies obtained from models with $N = 10$ (left) and $N = 5$ (right) health classes, mapped onto the original $N = 20$ state space.

Interestingly, the overall structure of the optimal policies remains consistent between different values of N. In all cases, patients are admitted to a NH in the last health class, exploiting the lower admission cost in this state while attempting to avoid crisis events. In addition, the frequency of monitoring increases as patients approach this final class, allowing for timely intervention. However, when comparing the effectiveness of the policies in terms of expected cost and crisis fraction, see Figure 4.11, notable differences emerge. The transition from $N = 20$ to $N = 10$ results in only a marginal increase in cost, from 209,376 to 210,935, and a slight reduction in crisis admissions, from 54% to 45%, indicating that most of the relevant information is retained. In contrast, reducing to $N = 5$ health classes leads to a more substantial increase in cost to 215,092, and a crisis fraction of 32%. This is due to the inability of the coarser model to capture subtle health dynamics, which results in less refined policies. These findings suggest that increasing the number of health classes enhances policy flexibility and that moderate granularity (e.g., $N = 10$) already yields meaningful improvements over very coarse representations.

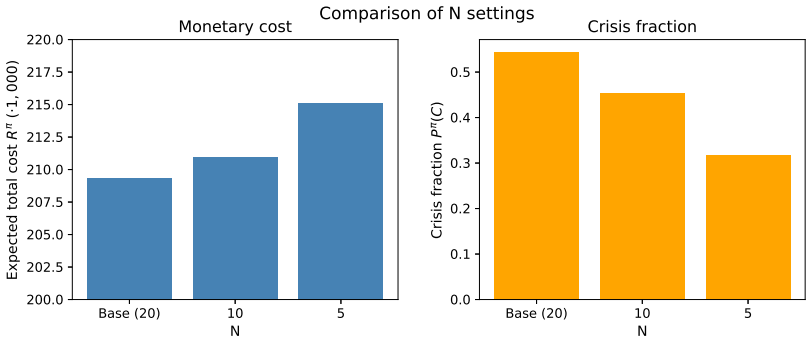


Figure 4.11: Comparison of expected cost and crisis fraction for different values of N.

4.5.7 Impact of admission delay

In practice, there is often a delay between the decision to admit a patient to a NH and the actual admission, due to capacity limitations, administrative procedures, or logistical constraints. During this delay, a patient's health can still deteriorate or a crisis may occur. This section evaluates how such delays influence the policy and its ability to preemptively mitigate risks. We assume a fixed admission delay ranging from 0 to 18 weeks between the decision to admit and the actual transition to the NH. During the delay, patients incur a weekly cost representing home care or other support needs, with values $\{0, 200, 1000\}$. Crucially, patients are still at risk of experiencing a crisis during this delay. If a crisis occurs, the crisis cost is incurred and the patient is immediately admitted, halting any further delay costs.

We observe that as the delay increases, the optimal policy is to admit patients in a healthier state. This is intuitive, as delay introduces an additional opportunity for a crisis, while admitting patients in a healthier state reduces the risk of a crisis during the delay. As the delay increases, it is observed that it is beneficial to move the admission threshold to an even healthier state. Due to the possibility of crises during delay, the crisis fraction still increases compared to the base case of no delay and delay cost. Interestingly, the crisis fraction decreases as the delay increases from 2 to 4 weeks, and can be explained by the fact that the policy changes from admitting patients in health status 20 to admitting patients already in health status 19. If the delay and the corresponding cost is high (here a delay cost of 1,000 and a delay larger than 14), the combined cost of waiting and the risk of a crisis becomes so large that it becomes optimal to admit patients only after a crisis, resulting in lower overall costs. This effect can be observed in Figure 4.12, where the crisis fraction gradually increases at first and eventually reaches 1 as the delay cost is 1,000 and the delay length increases. These findings highlight how even modest operational delays can influence optimal strategies, reinforcing the importance of minimizing both delay duration and associated costs in practice.

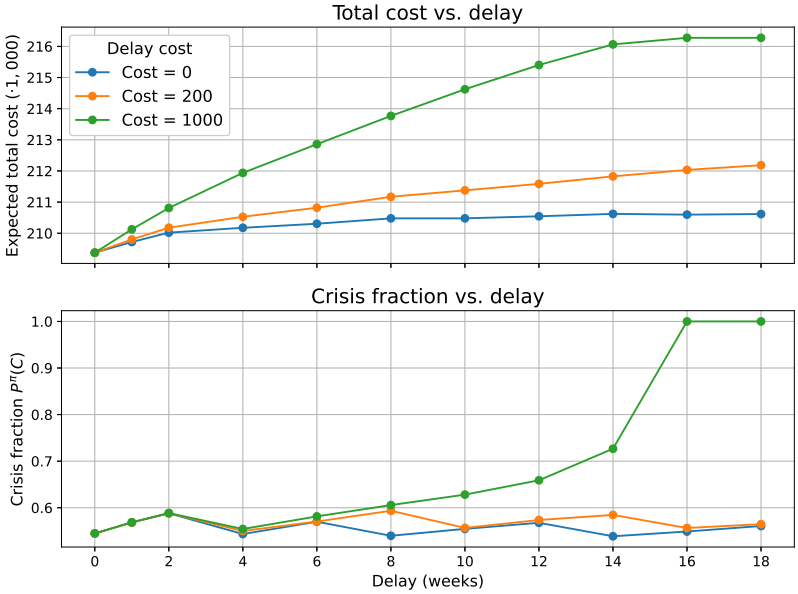


Figure 4.12: The total cost, including delay costs, and crisis fraction for varying admission delays and weekly delay costs.

4.6 Conclusion and Discussion

This study introduced an MDP framework to optimize preventive admission and monitoring strategies for NH care. The model combines realistic cost estimates with a DTMC for health deterioration to determine monitoring and admission decisions with the objective of minimizing total costs. The resulting policy indicates that the intensity of the monitoring increases as health deteriorates, whereas admission occurs only in the (few) worst health states. Calibrating the health model with real-world data, we find that the fraction of unavoidable crisis admissions remains almost 55%. This results from the relatively high estimated NH costs, causing the trade-off to favor shorter NH stays rather than preventive admissions. Sensitivity analysis shows that the fraction of crisis admissions can be reduced by introducing stronger incentives for preventive admissions, although this requires substantial changes in the cost (or health model) parameters. Overall, the MDP provides a natural and interpretable structure for admission and monitoring decisions. When combined with our health model, this constitutes an important first step toward modeling preventive care, which is a crucial element to the future of our long-term care system.

4.6.1 Limitations

Although the findings provide valuable insights, several limitations should be considered when interpreting the results, particularly those related to the model parameters. For example, obtaining reliable cost estimates is inherently challenging; the current estimates are based on the Dutch context. In addition, the health status model was constructed from generic and administrative billing data, which lacks detailed information on health status. A more detailed modeling of health deterioration would require access to medical records or functional status assessments, which are currently unavailable or only contain static information; for example, frailty scores are being recorded, but those scores are typically not updated over time and therefore not useful for this chapter. Furthermore, the data set covers only three years (2017-2019). Due to the relatively long period until a crisis occurs, many observations were censored, limiting the information about the underlying distribution.

Our model abstracts from several real-world complexities. For instance, we assume that a check reveals a patient's true health status, although clinical assessments inevitably involve judgment and noise. This uncertainty could be captured in modified transition rates or by introducing additional states to represent potential misclassifications. Mortality was excluded from the model dynamics, as incorporating it would require estimating the associated costs or utilities (that is, we essentially focus on the patient population that is admitted to an NH). In this chapter, the patient population is treated as homogeneous. In future practice, more accurate results may be achieved with models that incorporate patient-specific characteristics. The above limitations mainly relate to model parameter choices, whose accuracy can be enhanced as more data become available.

4.6.2 Future research and practical implications

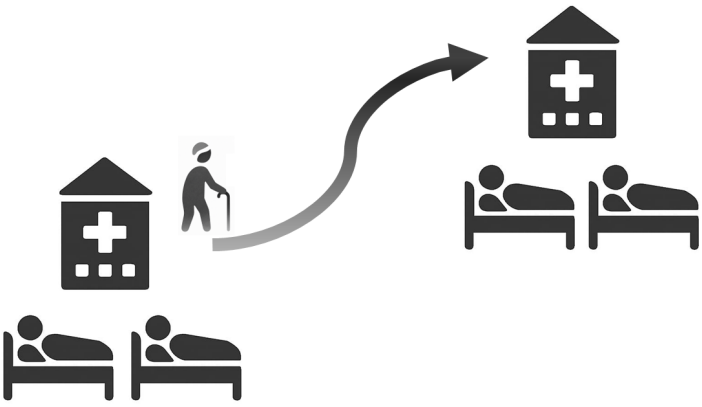
Despite uncertainty in the model parameters, the results indicate that the costs associated with a preventive NH admission are too high to outweigh the risk of a crisis admission. For clinical policy making, an interesting direction for future research is to investigate the extent to which hidden costs may be associated with crisis events. Similarly, the high costs associated with NH care reflect not only medical services but also housing and other forms of support. Reducing crisis admissions likely requires a stronger incentive than the financial mechanisms currently in place; however, the appropriate allocation of financial costs remains a matter of discussion. Another direction for future research is to use patient-specific features in the health model. By improving the predictability of health trajectories, the fraction of crises events can be reduced. Other interesting model extensions are incorporating mortality, explicitly including the capacity constraints of NHs, and the impact on other care facilities.

More broadly, the importance of preventive care is increasingly recognized. To quantify the impact of preventive actions, a model for the health trajectory of a patient is indispensable. To the best of our knowledge, no detailed models have

been proposed to address this issue. Our health model is specifically intended as a first step in this direction. By dynamically capturing the evolution of patient health, the model can subsequently inform preventive strategies and adaptive, health-based admission and monitoring policies. The proposed MDP explicitly incorporates patient health trajectories, a feature rarely addressed in NH admission models. This framework enables decision-makers to assess the cost-effectiveness of interventions, such as preventive care, and to weigh trade-offs between early admission, crisis management, and patients' well-being.

Part II

Capacity-Based Perspective



5

After-Service Blocking in Tandem Queues

Based on:
de Boer, T. R., Bekker, R., & van der Mei, R. D. (2024). After-service blocking in tandem queues. *In Proceedings of Valuetools (Milan, Italy, December 2024).*

Abstract

Bed-blocking and wrong-bed-days are a well-known problem in healthcare settings. Motivated by this phenomenon, this chapter examines tandem queues with zero buffers. These types of systems find many applications in systems where jobs are processed via a predefined number of sequential processing steps. From an analytic point of view, the analysis of such systems is highly complicated, caused by the phenomenon of After-Service Blocking (ASB). ASB occurs when a completed job occupies a server while waiting for the next phase. This challenge is addressed by introducing and comparing heuristics to quantify the impact of ASB on the performance of the tandem queue. Our proposed heuristics offer significant advantages over existing methods, delivering more insightful and accurate performance estimations. This allows for faster evaluations and their use in optimization, making it ideal for real-time decision making and capacity allocation. This chapter presents a valuable tool for queueing system designers and managers to make informed decisions regarding capacity allocation, resource management, and service level optimization, particularly in scenarios with limited budgets.

5.1 Introduction

In healthcare systems, patients often move through multiple stages of care, for example, from Emergency Departments (EDs) to hospital wards and then to Nursing Homes (NHs), where capacity limitations at one stage can delay or block flow at another [26]. Such processes can be modeled using tandem queues with zero or limited buffers, typically represented by M/M/c/c queues. Similar dynamics are also found in manufacturing [194] and telecommunications [177]. In these systems, limited or absent buffer space means that jobs or patients cannot wait, or only a few can do so. Inefficiencies in queueing systems with finite buffers often arise from the limited buffer capacity, particularly due to the phenomenon known as ASB [1], in which a server becomes blocked while waiting for available downstream capacity.

An example of ASB can be observed in healthcare systems, where a patient remains in the ED or a hospital due to a lack of beds available in a downstream care facility, such as a NH [15]. This situation not only incurs higher costs due to more expensive acute care, but also contributes to system-wide congestion, reducing overall patient throughput, and potentially exacerbating patient conditions due to delayed care [92]. More generally, ASB occurs when jobs, after completing service, are directed to another queue but are forced to remain at their current server due to capacity constraints. This results in the server being occupied by completed jobs, preventing it from serving new ones [49]. This scenario not only impedes the efficiency of the server, but can also lead to cascading delays, as subsequent jobs are held up by those that are blocked. As a result, costs can increase, especially when the next stage of service is more appropriate and cost-effective.

Motivated by this, the aim of this chapter is to approximate the impact of ASB on the throughput in tandem queues. Previous research has indicated that it is possible to approximate the effects of ASB in specific scenarios [26], or through machine learning techniques [62]. However, given the anticipated increase in the demand for healthcare and the limited availability of healthcare professionals, there is a pressing need for an approximation method that can handle a wide range of scenarios. This is especially critical in scenarios where the demand is close to the maximum capacity that the servers can manage and where budget constraints necessitate optimal allocation of capacity.

The contribution of this chapter is two-fold. First, we introduce a heuristic that provides more precise estimates compared to existing methods in approximating the effects of ASB in tandem queues with zero buffers across a wide range of scenarios. The approximations provide insight in how ASB is affected by the system parameters. As a byproduct of developing the heuristic, we also present continuous extensions of M/M/c/$c+k$ performance formulas for non-integer values of c. Second, our heuristic allows for fast evaluations of system performance, making it an ideal tool for optimization contexts. Unlike simulation-based performance evaluations, which are often more computationally intensive, the heuristic significantly accelerates the optimization process. We demonstrate this using a capacity

allocation problem.

The organization of the chapter is as follows. Section 5.2 reviews the related literature. Section 5.3 provides a detailed description of the model. Section 5.4 discusses the approximations employed in this research and indicates how these approximations can be applied to optimization problems. Section 5.5 shows the results of the approximation, as well as the optimization. Finally, Section 5.6 presents the conclusion and discussion.

5.2 Literature

In this section, we review previous studies that address ASB in tandem queues. An exact analysis of ASB in tandem systems is hardly available in the literature. An exception is Akyildiz and Von Brand [1]. They show that exact solutions are possible, but only for a small number of scenarios, such as two-node single-server systems and closed networks. However, the computation time of these exact solutions increases dramatically when more servers are considered or when the system size increases, as this directly affects the state space. Other papers [1, 11, 12, 81, 103, 131] focusing on exact analyses encountered similar challenges, involving the extensive computation time required and the limited set of scenarios in which they are applicable. This underscores the need for more efficient approximation methods.

Next, we discuss various approximation methods from the existing literature. Koizumi et al. [123] introduced a simple approximation, which adjusts the service time of queue i by incorporating the estimated waiting time to access queue $i+1$. Bretthauer et al. [26], motivated by bed-blocking problems in healthcare settings, improved the heuristic of [123] by developing an approximation for throughput rates in tandem queues. Their approach extends [123] by adjusting not only the service rate, but also the number of servers and the flow rate. However, their study does not address issues such as how the calculations of the heuristic are adapted to accommodate non-integer modified numbers of server, and is limited to tandem queues where the load of every node is strictly smaller than one.

Dallery and Frein [48] introduced a decomposition method customized for single-server tandem queues. Their method entails dissecting the tandem network into individual subsystems, with buffers positioned between a pair of servers. Their approach is not applicable to networks with multiple servers. Similarly, Van Vuuren et al. [227] introduced a decomposition method for tandem server pools and buffers. Their iterative method approximates throughput and sojourn time for various buffer sizes under the assumption that the system is never starved. This assumption implies that there is always a new arrival at the first queue, ensuring that any available service capacity is immediately utilized. The assumption that the system is never starved renders this approach impractical for realistic scenarios.

Osorio and Bierlaire [167] presented a heuristic that distinguishes between arrival, service, blocking, and unblocking rates. This method approximates the blocking probabilities for different networks, but its application is limited to specific networks and assumes immediate resolution of *deadlocks*. A deadlock occurs when jobs are blocked directly or indirectly by themselves. Alternatively, Dieleman et al. [62] used Neural Networks (NNs) to predict throughput for stable tandem queues. Their research shows that NNs informed by queueing-theoretic knowledge outperform those based solely on input parameters, although the complexity of NNs complicates the interpretation of their methods.

Apart from approximating the effect of ASB, there are also other papers considering ASB in queueing networks. The impact of ASB on *open* queueing networks and the modeling of deadlocks within these systems has been explored by Palmer et al. [171]. This study focuses on the time until a deadlock occurs, noting that once a system is in a deadlock state, it remains so without external intervention. El-Taha and Wolff [49] developed a simulation model to analyze the effects of ASB in healthcare, to understand the capacity requirements in such settings, highlighting the complexity of ASB.

From the literature, we observe that most ASB-related studies focus on approximating the ASB effect, but they are typically limited in the number of scenarios in which the approximations provide an accurate estimation. This chapter addresses this research gap by introducing a new heuristic for rapidly quantifying the impact of ASB on the performance of tandem queues. This heuristic is required to accommodate a wider range of tandem queues. Moreover, the heuristic should be both explainable and transparent to illustrate how different parameters interact.

5.3 Model Description

We consider a tandem of n zero-buffer multi-server nodes, as illustrated in Figure 5.1. Arrivals to the first node occur according to a Poisson process with rate λ. The service times are exponentially distributed, whereas the service rates are denoted by $(\mu_1, \mu_2, \ldots, \mu_n)$. The number of servers at each node is given by the vector $(c_1, c_2, \ldots, c_n)$. We denote the buffer size at node i by k_i, but the buffer size is set to zero in all experiments. Jobs that arrive at the first node when all servers are busy (and the buffer is full) are rejected. Upon completing the service at node i, the job moves to the next node $i+1$. If node $i+1$ lacks available server space, the job must wait in the current node i, thus blocking other jobs from entering service. We denote by P_i the probability that all servers are busy (and the buffer is full) at node i, for $i = 1, \ldots, n$. Note that P_1 is also the fraction of jobs denied access to the first node. The ASB mechanism is visualized in Figure 5.2. In this figure, we can see that a new job, say job A, arrives at the first queue, enters service, is blocked after completing service, and is finally unblocked and can start service at the second node. This job can only move to the second node if server capacity is available at the second node. The jobs are served and unblocked on a

First-Come-First-Served (FCFS) basis.

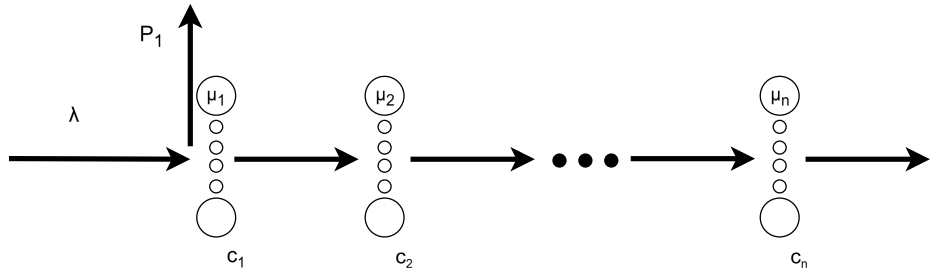


Figure 5.1: Visual representation of a tandem of n M/M/c/c queues.

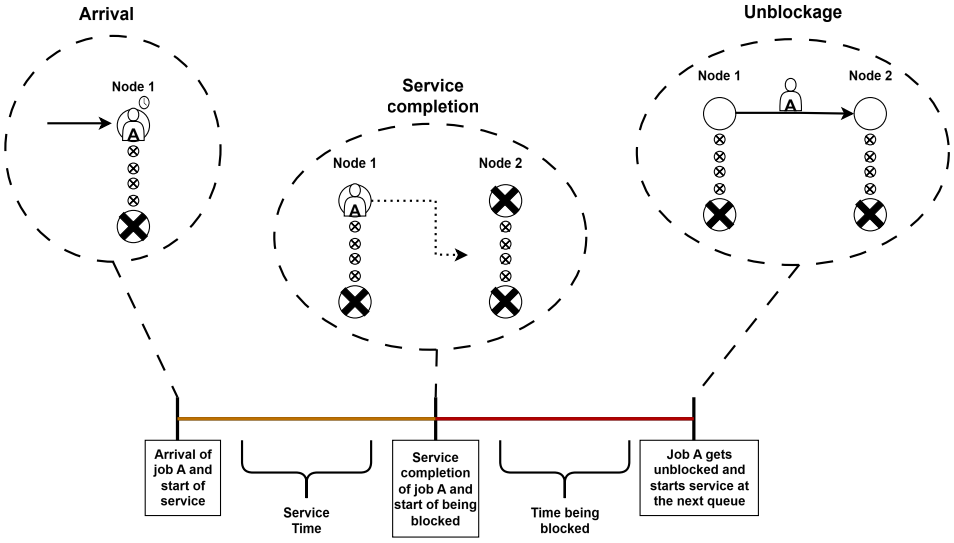


Figure 5.2: Visual representation of ASB for job A in a two-queue tandem model.

5.4 Methodology

In this section, we discuss the methodology employed to address ASB in tandem queues. Section 5.4.1 presents four approximation techniques: two established methods from the literature and two newly developed approaches. In Section 5.4.2, we apply these methods within an optimization framework, focusing on optimizing capacity allocation.

5.4.1 Approximations

This section presents a detailed analysis of four approximation techniques: one is a benchmark approximation, one is taken from the literature, and two are new approximations. The key idea of each method is to decompose the network into a series of individual queues, where the impact of ASB is incorporated by a modified service capacity in terms of service rate and number of servers. The strengths of such approximations is that they are fast, whereas they also provide insight in how subsequent queues affect each other's behavior. It should be noted that although the buffer size is zero ($k_i = 0$) for all nodes in our scenarios, the approximations are formulated without imposing this restriction on k_i to enhance their generalizability.

M/M/c/$c+k$ based approximation (Loss)

The first approximation is the most straightforward and is based on the M/M/c/$c+k$ queue, i.e., the classical Poisson-driven finite-buffer c-server queue with exponential service times. We refer to the M/M/c/$c+k$ approximation as 'Loss' throughout the remainder of this chapter. For such a queue, arriving jobs finding $c+k$ jobs present are lost. For this approximation, the fraction of lost jobs is approximated by only considering the first queue. This implies that additional blocking that occurs when jobs are delayed at a server due to congestion in downstream queues is not taken into account. Consequently, the blocking probability derived from this approximation represents a lower bound since ASB will only increase the blocking probability. Let $a = \frac{\lambda}{\mu}$ denote the *offered load* and let $\rho = \frac{\lambda}{c\mu}$ be the *load per server*. The stationary blocking probability, denoted as π_{c+k}, follows directly from the steady-state distribution of the number of customers in the system (cf. [127])

$$\pi_{c+k}(\lambda, \mu, c, k) = \frac{\rho^{c+k} c^c}{\Gamma(c+1)} \pi_0, \tag{5.1}$$

with $\Gamma(c+1)$ and $\Gamma(c,a)$ the Gamma function and the upper incomplete Gamma function, respectively, see also Appendix 5.A. The above is a reformulation of the well-known steady-state probabilities, as presented in the Appendix, where c is no longer required to be integer. This expression is essential to improve the speed of calculations as well as for the other heuristics where the number of servers may be non-integer. The probability of an empty system, π_0, is given by:

$$\pi_0 = \begin{cases} \left[e^a \frac{\Gamma(c,a)}{\Gamma(c)} + \frac{a^c}{\Gamma(c+1)}(k+1) \right]^{-1}, & \text{if } \rho = 1, \\ \left[e^a \frac{\Gamma(c,a)}{\Gamma(c)} + \frac{a^c}{\Gamma(c+1)} \frac{1-\rho^{k+1}}{1-\rho} \right]^{-1}, & \text{otherwise.} \end{cases} \tag{5.2}$$

The first approximation of the blocking probability is then given by $\pi_{c_1}(\lambda, \mu_1, c_1, k_1)$.

The heuristic by Bretthauer (BR)

We use the heuristic proposed in [26], from now on referred to as 'BR', as a benchmark because it is closely aligned with our method and is tested on scenarios

that are particularly relevant. BR introduces a flow rate $F_{1,2}$, which represents the flow from queue 1 to queue 2. Since no jobs are lost after queue 1, the flow rate is equal for each pair of subsequent queues i and $i+1$, and is denoted by F. The flow rate is calculated using:

$$F = F_{i,i+1} = F_{1,2} = \lambda \cdot (1 - P_1). \tag{5.3}$$

In addition to the flow rate, the heuristic modifies the number of servers, denoted by c_i^*, and the service rate, denoted by μ_i^*, to account for the blocking effects of subsequent queues. The *modified number of servers* captures how many of the servers are not blocked due to ASB and the modified service rate can be used to capture an extra time a server is blocked due to ASB. Specifically, the expected number of customers waiting for service at a stage with c servers, service rate μ and flow rate F, is approximated using the M/M/c queue, cf. [110] for the case of any positive c, including fractional values,

$$L(F, \mu, c) = \frac{\left(\frac{F}{\mu}\right)^c F\mu}{\Gamma(c)(c\mu - F)^2} \left[\frac{e^{\frac{F}{\mu}} \Gamma(c, \frac{F}{\mu})}{\Gamma(c)} + \frac{\left(\frac{F}{\mu}\right)^c c\mu}{\Gamma(c+1)(c\mu - F)} \right]^{-1}. \tag{5.4}$$

Observe that when $F \geq c\mu$ the M/M/c queue becomes unstable, such that the expected number of jobs waiting for service explodes and $L(F, \mu, c)$ grows without bound.

The modified number of servers at station i is then calculated by the number of servers that are not blocked:

$$c_i^* = [c_i - L_{i+1}]^+, \tag{5.5}$$

and the modified service rate is determined by

$$\mu_i^* = \left[\frac{c_i^*}{c_i} \frac{1}{\mu_i} + \frac{c_i - c_i^*}{c_i} \frac{1}{c_{i+1}\mu_{i+1}} \right]^{-1}. \tag{5.6}$$

The rationale behind this adaptation is that blocked jobs must wait for the service completion at the next node before they can be unblocked (see [26] for a more elaborate discussion). The pseudo-code for heuristic BR can be found in Algorithm 5.1.

Algorithm 5.1 Heuristic Algorithm by Bretthauer et al.

```
Initialize m = 0, P_1^0 = 0, μ_i^0 = μ_i, c_i^0 = c_i for i = 1, ..., n
while |P_1^m − P_1^{m−1}| ≥ δ do
    m = m + 1
    Calculate the flow rate F^m = λ(1 − P_1^{m−1})
    for i = 1, ..., n do
        Calculate the modified number of servers c_i^m using Equation (5.5)
        Calculate the modified service rate μ_i^m using Equation (5.6).
    end for
    Update the blocking probability P_1^m = π_i(F^m, μ_1^m, c_1^m)
end while
```

The modified service rate heuristic (MS)

One primary limitation of BR lies in its dependence on calculations derived from M/M/c formulas. For example, these formulas are only valid when the system workload remains below 1, that is, when $\rho < 1$. As a result, BR cannot provide accurate approximations in scenarios where the arrival rate leads to a workload larger than or equal to 1. This constraint significantly narrows the range of scenarios in which the heuristic can be applied effectively, limiting its usefulness in higher workload situations. Furthermore, we also observed a significant decrease in the performance of the heuristic in [26] when the number of servers becomes prohibitively small.

We address the limitations by focusing on a different adaption of the service rate, while we do not change the number of servers. To this end, we first notice that while nodes $2, 3, ..., n$ concern queues without buffer, jobs cannot be directly denied access to the system as these jobs wait at another node. We incorporate the ASB impact by including the waiting time in a suitable finite-buffer multi-server system. This differs from [26], which uses the expected number of waiting jobs. The adaptations are further explained below.

The service rate is modified to reflect the effect of ASB, as jobs remain longer at the nodes. The modified service rate is determined using the sojourn time, where we define the expected sojourn time at node i as:

$$\mathbb{E}[S_i] = \frac{1}{\mu_i} + \mathbb{E}[B_{i+1}], \tag{5.7}$$

where $\mathbb{E}[B_{i+1}]$ is the expected time being blocked by node $i+1$. This is straightforward for the last queue, yielding $\mathbb{E}[S_n] = 1/\mu_n$. For the other queues, the time blocked by the next queue is determined by modeling queue $i+1$ as an M/M/c/$c+k$ queue where a fraction of the servers of queue i are used as a buffer for queue $i+1$. This is done since the servers at node i function as a buffer for node $i+1$, due to the ASB effect. We approximate the time blocked by the expected waiting time of an M/M/c/$c+k$ model with the altered parameters. The arrival rate is the original arrival rate λ and the service rate is the modified service rate μ^*_{i+1}. The number of servers remains c_{i+1}, and the number of buffer spaces is the number of buffer spaces at queue $i+1$ plus the average number of servers at queue i that do not serve a job, yielding $k_{i+1} + \left[c_i - \frac{\lambda(1-P_1)}{\mu_i}\right]^+$. Combining the above, the modified service rate for queue i is determined by

$$\mu_i^* = \begin{cases} \Big[\frac{1}{\mu_i} + W_q\Big(\lambda, \mu^*_{i+1}, c_{i+1}, \\ \qquad k_{i+1} + \big[c_i - \frac{\lambda(1-P_1)}{\mu_i}\big]^+\Big)\Big]^{-1}, & i = 1, \ldots, n-1, \\ \mu_n, & i = n. \end{cases} \tag{5.8}$$

with $W_q(\lambda, \mu, c, k)$ the expected waiting time in the M/M/c/$c+k$ queue with arrival rate λ, service rate μ, c servers and buffer size k, see also Appendix 5.A.

Next, we determine the probability that node 1 is fully occupied, thus blocking newly arriving jobs and denying them access to the system. The probability is based on the M/M/c/c+k queue, with the service rate corresponding to a modified service rate. The probability of queue 1 being fully occupied is then computed as follows:

$$P_1 = \pi_{c_1+k_1}(\lambda, \mu_1^*, c_1, k_1). \tag{5.9}$$

This yields the 'modified service rate' heuristic, shortened to 'MS', as presented in Algorithm 5.2.

Algorithm 5.2 The modified service rate heuristic

1: Initialize $m = 0$, $\mu_i^m = \mu_i$, $P_1^m = 1.0$ for $i = 1, \ldots, n$
2: **while** $|P_1^m - P_1^{m-1}| \geq \delta$ **do**
3: $m = m + 1$
4: $P_1^m = \pi_{c_1+k_1}(\lambda, \mu_1^m, c_1, k_1)$
5: **for** $i = 1$ to n **do**
6: Update the modified service rate using Equation (5.8)
7: **end for**
8: **end while**

The modified service rate and number of servers heuristic (MS&C)

Recall that a key limitation of the heuristic in [26] is its inability to handle systems with a load exceeding one, primarily due to its reliance on the M/M/c model. We propose a modified heuristic that also relies on adapting the service rate and the number of servers, named the modified service rate and number of servers heuristic ('MS&C'). However, instead of relying on the M/M/c model, we use the M/M/c/$c + k$ variant, which is stable for any finite load. Moreover, our adjustments enhance the accuracy of the heuristic. The specific modifications are detailed below.

We modify the service rate, μ_i^*, by conditioning on whether a job experiences blockage. If a job is not blocked, which occurs with probability $1 - P_{i+1}$, the job proceeds with its expected service time of $\frac{1}{\mu_i}$. However, if a job is blocked, with probability P_{i+1}, it must wait until all previously blocked jobs, which is equal to $c_i - c_i^*$, are unblocked. The expected unblocking time for these jobs is given by $1/(c_{i+1}^* \mu_{i+1}^*)$. Since there is no blocking at node n, we have $\mu_n^* = \mu_n$. This leads to the following recursive relation:

$$\mu_i^* = \begin{cases} \left[(1 - P_{i+1})\frac{1}{\mu_i} + P_{i+1}(c_i - c_i^*)\frac{1}{c_{i+1}^* \mu_{i+1}^*}\right]^{-1}, & \text{for } i = 1, \ldots, n-1, \\ \mu_n, & \text{for } i = n. \end{cases} \tag{5.10}$$

We adopt a straightforward modification for the number of servers. The underlying intuition is that only a portion of the servers are actively serving jobs, while the

remaining servers are occupied by blocked jobs. The fraction of blocked servers is represented by P_{i+1}. Consequently, the modified number of servers is denoted by:

$$c_i^* = \begin{cases} c_i(1 - P_{i+1}), & \text{for } i = 1, \ldots, n-1, \\ c_n, & \text{for } i = n. \end{cases} \tag{5.11}$$

The probability that queue i is fully occupied is again estimated using the M/M/c/c+k system, as shown in Equation (5.1). It is important to note that while the first queue experiences an arrival rate of λ, subsequent queues receive only a fraction of these arrivals, $\lambda(1-P_1)$. The blocking probability for queue i is then computed as follows:

$$P_i = \begin{cases} \pi_{c_1^*+k_1}(\lambda, \mu_1^*, c_1^*, k_1), & \text{for } i = 1, \\ \pi_{c_i^*+k_i}(F, \mu_i^*, c_i^*, k_i), & \text{for } i = 2, \ldots, n. \end{cases} \tag{5.12}$$

Here, F is the number of jobs that complete service at queue 1 and is calculated, similar to [26], by $F = \lambda(1 - P_1)$.

For computational purposes, we have chosen to update the blocking probability using averaging with decreasing weights, such that the blocking probabilities are updated as

$$P_i^m = \begin{cases} P_1^{m-1}\frac{m-1}{m} + \pi_{c_1^*+k_1}(\lambda, \mu_1^*, c_1^*, k_1)\frac{1}{m}, & \text{if } i = 1, \\ P_i^{m-1}\frac{m-1}{m} + \pi_{c_i^*+k_i}(F, \mu_i^*, c_i^*, k_i)\frac{1}{m}, & \text{if } i > 1. \end{cases} \tag{5.13}$$

Combining the above yields Algorithm 5.3 as a second new heuristic.

Algorithm 5.3 The modified service rate and number of servers heuristic

Initialize $m = 1$, $\mu_i^m = \mu_i$, $c_i^m = c_i$ for $i = 1, \ldots, n$
$P_1^m = \pi_{c_1+k_1}(\lambda, \mu_1^m, c_1^m, k_1)$, $F^m = \lambda(1 - P_1^m)$
$P_i^m = \pi_{c_i+k_i}(F^m, \mu_i^m, c_i^m, k_i)$ for $i = 2, \ldots, n$
while $|P_1^m - P_1^{m-1}| \geq \delta$ **do**
    $m = m + 1$
    for $i = n, \ldots, 1$ **do**
        Update the modified service rate using Equation (5.10)
        Update the modified number of servers using Equation (5.11)
        Update the blocking probability of queue i using Equation (5.13)
    end for
    Update $F^m = \lambda(1 - P_1^m)$
end while

5.4.2 Optimization

There is a vast increase in the number of scenarios for capacity allocation as the number of queues grows. Therefore, a fast performance evaluation is crucial when optimizing the capacity over the different nodes. Previously, these evaluations

had to be done by simulation or by a flawed approximation. The heuristics in Section 5.4.1 enable us to speed up these evaluations, reducing the time of an evaluation from minutes to seconds. For optimization, various methods can be employed, such as (i) grid search [68], (ii) evolutionary algorithm [13], and (iii) marginal allocation [183]. This section illustrates the optimization method that combines our heuristic with grid search in the case of *budget-constrained server allocation*.

In the optimization problem, each server at node i has a cost of b_i. The problem is to determine a server allocation that minimizes the rejection probability at the first node, $P_1(c_1, \ldots, c_n)$, where the total cost should not exceed a budget B. The mathematical formulation of the optimization problem is then given by:

$$\begin{aligned} \min_{(c_1,\ldots,c_n)} \quad & P_1(c_1, \ldots, c_n) \\ \text{s.t.} \quad & \sum_{i=1}^{n} b_i c_i \leq B \\ & c_i \in \mathbb{Z}^+ \quad \text{for } i = 1, \ldots, n. \end{aligned}$$

The algorithm to approximate the optimal allocation of servers can be found in Algorithm 5.4 and involves three main steps. First, in line 3 of the pseudocode, all possible combinations of server allocations are generated, where a server allocation is represented by $(c_1, c_2, \ldots, c_n)$, with c_i the number of servers allocated to node i. Each combination must satisfy the budget constraint, ensuring that the total cost does not exceed B, that is, $\sum_{i=1}^{n} b_i c_i \leq B$. In line 4 we exclude server allocations where the remaining budget allows for adding at least one more server. To further limit the number of evaluations needed, we exclude server allocations that are unlikely to be optimal as a result of large speed differences, that is, $\frac{\mu_i c_i}{\mu_1 c_1} < 0.5$ or $\frac{\mu_i c_i}{\mu_1 c_1} > 1.5$. This step helps to eliminate suboptimal allocations.

For each of the remaining server allocations, we evaluate the blocking probability P_1 using a heuristic. Finally, we compare the blocking probabilities for all evaluated allocations and select the one with the lowest P_1. This allocation is optimal according to the heuristic, as it minimizes the likelihood of job blocking at the first node, thereby enhancing the overall performance of the queueing system.

Other optimization problems, such as the *cost-minimizing server allocation*, can also be addressed using our new heuristic evaluations. This is a similar optimization problem, but instead of minimizing the blocking probability under a budget constraint, we aim to minimize the total cost while adhering to a maximum allow-

Algorithm 5.4 Optimal budget-constrained server allocation based on heuristic

Input: Budget B, cost of a server at node i as b_i for $i = 1, \ldots, n$
Output: Optimal server allocation $(c_1^{\#}, \ldots, c_n^{\#})$ with the lowest P_1
Generate all possible server allocations $(c_1, \ldots, c_n)$ such that $\sum_{i=1}^{n} b_i c_i \leq B$
Remove all server allocations where $\sum_{i=1}^{n} b_i c_i < B - \min_i(b_i)$
Remove all server allocations where $\frac{\mu_i c_i}{\mu_1 c_1} < 0.5$ or $\frac{\mu_i c_i}{\mu_1 c_1} > 1.5$
Initialize $(c_1^{\#}, \ldots, c_n^{\#}) = \emptyset$ and $P_1^{\#} = \infty$
for each server allocation $(c_1, \ldots, c_n)$ **do**
    Evaluate P_1 using a heuristic for allocation $(c_1, \ldots, c_n)$
    if $P_1 < P_1^{\#}$ **then**
        $P_1^{\#} = P_1$
        $(c_1^{\#}, \ldots, c_n^{\#}) = (c_1, \ldots, c_n)$
    end if
end for
return $(c_1^{\#}, \ldots, c_n^{\#})$

able blocking probability. This yields the following formulation:

$$\begin{aligned} \min_{(c_1,\ldots,c_n)} \quad & \sum_{i=1}^{n} b_i c_i \\ \text{S.t.} \quad & P_1(c_1, \ldots, c_n) \leq P_1^{max} \\ & c_i \in \mathbb{Z}^+ \quad \text{for } i = 1, \ldots, n. \end{aligned}$$

The optimal server allocation for this problem can be determined using a method similar to Algorithm 5.4.

5.5 Results

In this section, we present an evaluation of the four approximation methods for ASB within tandem queues. Our analysis focuses on identifying which approximation provides the most accurate representation of blocking probabilities when jobs are held up by subsequent nodes. In Section 5.5.1, we compare these approximations by evaluating their performance in various scenarios. In Section 5.5.2 we provide preliminary results that demonstrate the application of our proposed heuristic to the budget-constrained server allocation problem.

5.5.1 Accuracy approximations

The heuristics are tested on various scenarios of n nodes in tandem with $n \in \{2, \ldots, 5\}$. The arrival process is Poisson with rate λ to the first node. Each node i has a number of servers c_i and a service rate μ_i. The aggregate service rate of the queue with the smallest service capacity is $\min_i\{c_i\mu_i\}$. Our primary interest is in

scenarios where roughly $\lambda \approx \min_i\{c_i\mu_i\}$, as these correspond to realistic settings that are also challenging to approximate. The generated scenarios are presented in Table 5.1. In this context, P_1 is determined by simulation, denoted P_1^{sim}, and the mean value of P_1 is obtained by averaging across all scenarios of the respective row in Table 5.1. The number of possible scenarios quickly increases with more nodes; therefore, we have chosen to limit the number of parameter settings when increasing the number of queues. This explains why the mean P_1 decreases when moving from the $n = 3$ scenarios to the $n = 4$ scenarios, despite an expected increase caused by the additional strain of after-service blocking from an extra queue.

n	c_i	μ_1	$\frac{c_i\mu_i}{c_1\mu_1}$	$\frac{\lambda}{min(c_i\mu_i)}$	**# scenarios**	**Mean P_1^{sim},** (min; max)
2	$\{1, 5, 10, 20\}$	1.0	$\{0.7, 0.8, \ldots, 1.3\}$	$\{0.7, 0.8, \ldots, 1.3\}$	784	$0.29, (0.00; 0.63)$
3	$\{1, 5, 10, 20\}$	1.0	$\{0.8, 0.9, \ldots, 1.2\}$	$\{0.8, 0.9, \ldots, 1.2\}$	8000	$0.29, (0.02; 0.63)$
4	$\{5, 10, 20\}$	1.0	$\{0.9, 1.0, 1.1\}$	$\{0.9, 1.0, 1.1\}$	6561	$0.22, (0.07; 0.38)$
5	$\{5, 20\}$	1.0	$\{0.9, 1.0, 1.1\}$	$\{0.9, 1.0, 1.1\}$	7776	$0.23, (0.07; 0.39)$

Table 5.1: Overview of simulation scenarios.

Each of the 23,121 scenarios has a warm-up period of 50,000 service completions. The simulation ends when P_1 has reached a stable average, which we have chosen such that the confidence interval (CI) of the blocking probability is less than 5% of the mean value of P_1. If this condition is not met, then the simulation is continued for another 10,000 service completions, and repeated until the stability condition is met. For the convergence of the heuristic algorithms, we take $\delta = 10^{-6}$. In certain exceptional instances, convergence may not occur at all, this phenomenon was notably observed with the BR heuristic. The heuristic occasionally oscillates between a high blocking probability at the first queue, due to a reduced modified number of servers or service rate, and a low blocking probability for similar reasons. The error of heuristic H is then given by

$$\Delta^{(H)} = |P_1^{(H)} - P_1^{\text{sim}}|.$$

The errors of the heuristics can be found in Table 5.2 and Figure 5.3. Figure 5.3 shows only the three best-performing heuristics to allow a clearer and more accurate comparison of these approximations. As can be seen, MS&C performs the best over all scenarios. This heuristic has both the smallest mean error and the smallest maximum error. From Figure 5.3, we observe that MS&C is able to maintain small errors, also when the number of queues, and thus the complexity of the system, increases. The Loss approximation turns out to perform quite well. This approximation naturally breaks down when the service rate imbalance between queues increases, for instance, when the speed of other queues is significantly lower than that of the first queue. Consequently, this approximation is not suitable for optimization as it is based on only the first queue and ignores the effect

of ASB. The effectiveness of BR notably lags behind that of the other heuristics. As the benchmark of [26] is constructed for cases with $\rho < 1$, we split the scenarios between $\rho^{\#} < 1$ and $\rho^{\#} \geq 1$, with $\rho^{\#} = \min_{i=1,\ldots,n}\{\frac{\lambda}{c_i \mu_i}\}$, see Figures 5.4 and 5.5. The boxplots for $\rho^{\#} < 1$, in Figure 5.4, reveal that the new heuristics clearly outperform the Loss approximation and BR. Moreover, for $\rho^{\#} \geq 1$ it also becomes clear that MS&C overall has the best performance.

n	**Loss**	**BR**	**MS**	**MS&C**
	(min; max)	(min; max)	(min; max)	(min; max)
2	0.03, (0.00; 0.16)	0.10, (0.00; 0.22)	0.02, (0.00; 0.15)	0.02, (0.00; 0.12)
3	0.04, (0.00; 0.30)	0.39, (0.00; 0.83)	0.03, (0.00; 0.31)	0.02, (0.00; 0.13)
4	0.04, (0.00; 0.15)	0.42, (0.02; 0.82)	0.03, (0.00; 0.15)	0.01, (0.00; 0.08)
5	0.04, (0.00; 0.18)	0.43, (0.02; 0.82)	0.03, (0.00; 0.18)	0.02, (0.00; 0.12)

Table 5.2: Mean error for different scenarios.

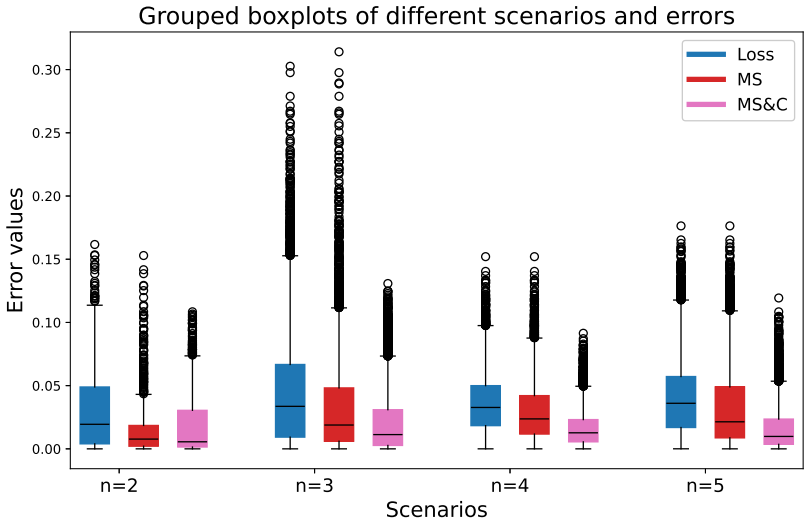


Figure 5.3: Box plots of the errors of the three best-performing approximations.

The impact of the load on the blocking probability, both for MS&C and the simulation, is visualized in Figures 5.6 and 5.7, for a two- and three-node tandem system, respectively. The parameters for the scenario of Figure 5.6 are $c_1 = 10$, $c_2 = 10$, $\mu_1 = 1.0$, with μ_2 equal to 1.0 and 0.8 for the left- and right-hand sides

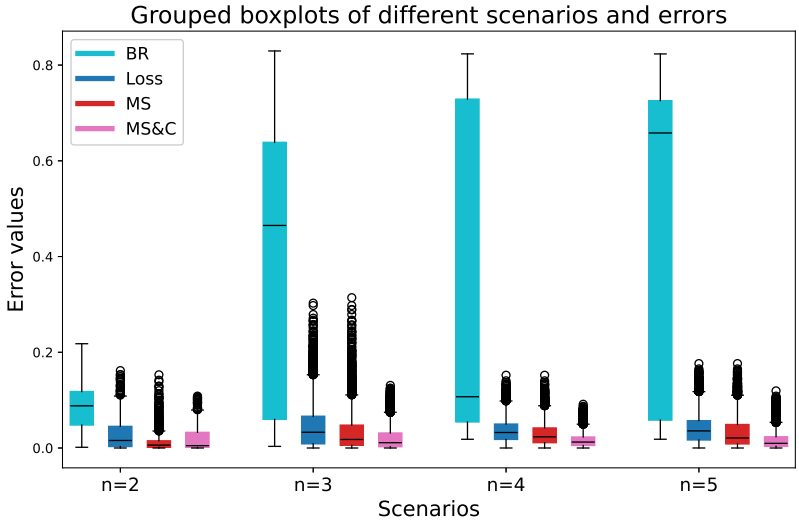


Figure 5.4: Box plots of the errors of approximations for $\rho^{\#} < 1.0$.

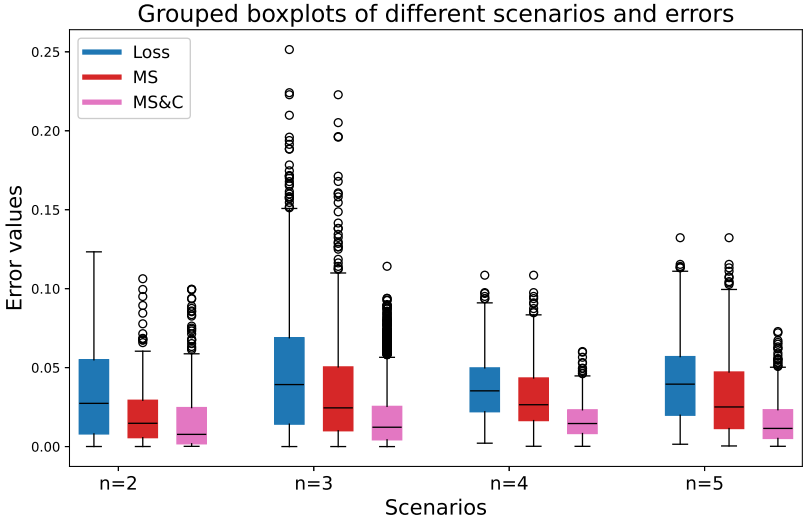


Figure 5.5: Box plots of the errors of the three best-performing approximations for $\rho^{\#} \geq 1.0$.

of the figure, respectively. The parameters of Figure 5.7 are $c_1 = 10$, $c_2 = 10$, $c_3 = 10$, $\mu_1 = 1.0$, with μ_2 and μ_3 equal to 1.0 or 0.8 for the different scenarios. The load varies due to λ ranging from 0 to 40. The results show that the impact

of the load is similar for both the heuristic and simulation. When queue 2, queue 3, or both become a bottleneck (that is, for $\mu_2 = 0.8$ and/or $\mu_3 = 0.8$), we see that the error increases somewhat with the arrival rate, although the functional impact of the load on the heuristic is similar to the simulation.

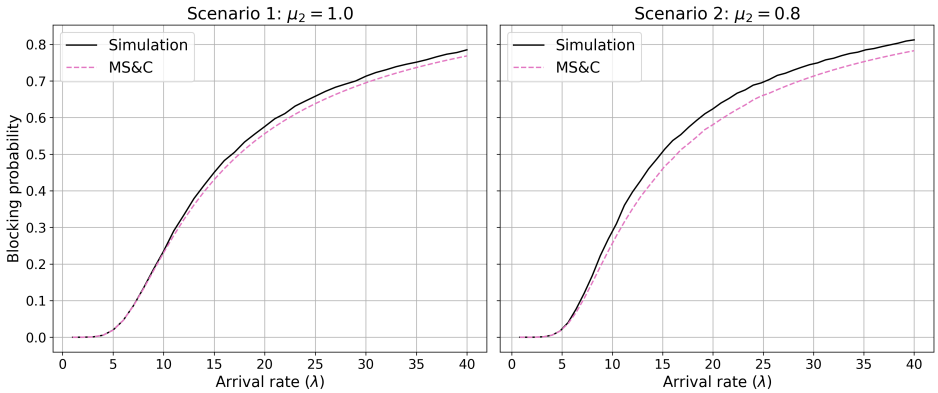


Figure 5.6: Comparison of the simulation and the heuristic for a system with two nodes with a varying arrival rate.

Runtime

To illustrate the runtime efficiency of the heuristics, we tested four simple scenarios, with $n = 2, 3, 4$ or 5. Each queue has the same number of servers and service rate with $c_i = 10$ and $\mu_i = 1.0$, while fixing the arrival rate at $\lambda = 10.0$. The resulting runtimes and iteration counts required for convergence are presented in Table 5.3. As observed, all approximations significantly reduce runtime compared to the simulation approach for the same scenario. Among the heuristics, MS achieves the shortest runtime, followed by MS&C, with both providing a clear runtime advantage over the heuristic BR. The iteration counts needed for convergence, also shown in Table 5.3, further support this result. While MS&C is not the fastest overall, it still offers a substantial runtime reduction, making it a viable option for optimization, especially with respect to simulation times.

n	**Simulation**	**M/M/c/$c+k$**	**BR**	**MS**	**MS&C**
	sec (iter)	sec (iter)	sec (iter)	sec (iter)	sec (iter)
2	54.470 (-)	0.000 (-)	0.099 (767)	0.002 (9)	0.097 (369)
3	56.587 (-)	0.000 (-)	1.679 (8590)	0.003 (10)	0.136 (250)
4	78.707 (-)	0.000 (-)	1.780 (8600)	0.004 (12)	0.212 (231)
5	97.804 (-)	0.000 (-)	1.776 (8551)	0.005 (13)	0.245 (228)

Table 5.3: Runtime in seconds and number of iterations for different methods.

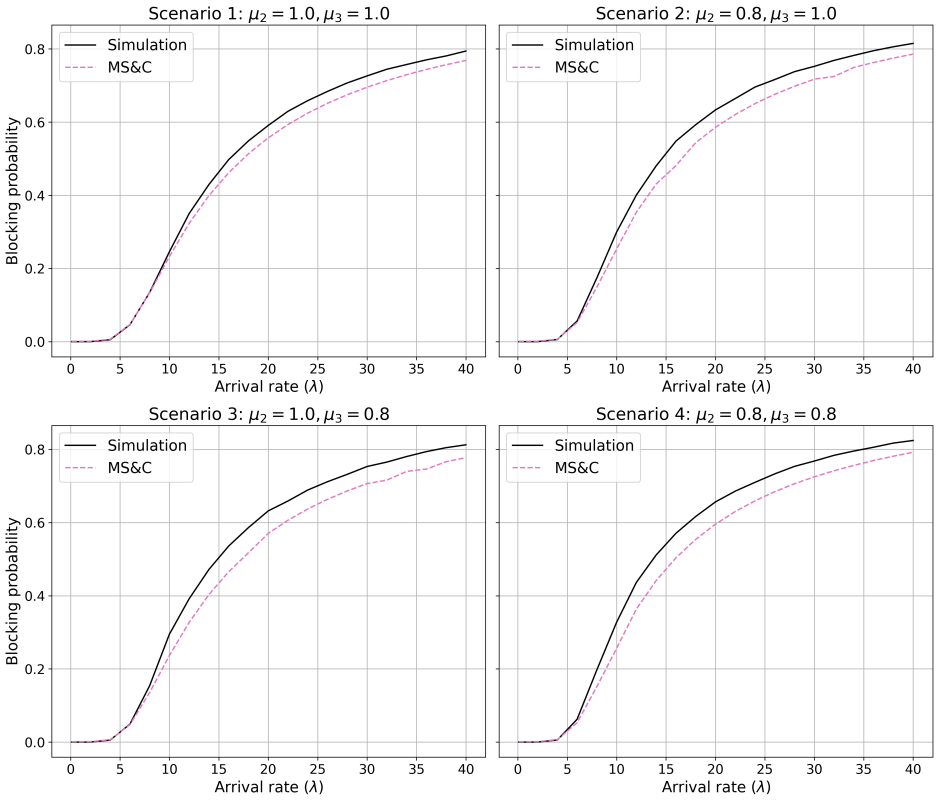


Figure 5.7: Comparison of the simulation and the heuristic for a system with three nodes with a varying arrival rate.

5.5.2 Optimization

We illustrate the optimization based on a relatively small tandem system of three queues. For such a system, we are able to evaluate the different server configurations using simulation. For larger systems, this approach becomes impractical; however, the heuristic remains feasible, as it significantly accelerates the evaluation process. Specifically, we chose the following parameters: $\lambda = 10, \mu_1 = 2.0, \mu_2 = 1.0, \mu_3 = 0.5$, each server having a cost of 1 ($b_1 = b_2 = b_3 = 1.0$), and a budget B ranging from 10 to 50. For example, $B = 10$ represents that there is a budget for 10 servers in total. Table 5.4 shows the optimization results when the evaluations are based on both the simulation and MS&C. The optimal allocation according to the heuristic is close to that of the optimal allocation according to the simulation; the blocking probabilities typically differ only by a few percentage points. Overall, the optimization results indicate that the simulation leads to somewhat more server capacity further downstream of the tandem. Moreover, $c_1\mu_1 \geq c_2\mu_2 \geq c_3\mu_3$, which implies that it is desirable to have the largest service capacity upstream.

B	$(c_1^{*sim}, c_2^{*sim}, c_3^{*sim})$	$(c_1^{*heur}, c_2^{*heur}, c_3^{*heur})$	P_1^{*sim}	P_1^{*heur}	$P_1^{*heur} - P_1^{*sim}$
10	(2,3,5)	(2,4,4)	0.79	0.81	0.02
20	(4,6,10)	(4,7,9)	0.55	0.57	0.02
30	(6,9,15)	(7,9,14)	0.32	0.34	0.02
40	(8,12,20)	(10,10,20)	0.11	0.14	0.03
50	(12,14,24)	(12,14,24)	0.02	0.02	0.00

Table 5.4: Server allocations based on optimization using simulation and heuristic.

5.6 Conclusion and Discussion

This chapter investigated the impact of ASB in tandem queues and introduced two new heuristics to approximate its effects. The best performing heuristic is MS&C and offers several advantages over the existing methods. First, it delivers more accurate estimations compared to previous approaches. Second, it enables rapid evaluations and adjustments, making it suitable for real-time decision-making and capacity allocation. Finally, the heuristic is transparent, providing insight in the impact of different parameters on system performance.

With the improved heuristic, there is clear potential for practical applications. For instance, the heuristic can be easily applied in a decision support system for real-world queueing systems such as manufacturing lines, telecommunications networks, and service industries. It enhances the ability to optimize server allocations, thereby improving operational efficiency and without being reliant on time-consuming simulations, or black-box prediction models. This makes it a valuable tool for professionals who want to effectively balance cost and service quality.

5.6.1 Future work

While our heuristic shows promising performance, there are some limitations. For example, the model assumes a specific structure of tandem queues and may not be directly applicable to networks with more complex topologies. Extending the heuristic to handle more complex queueing networks is an interesting direction for future research. We envisage that the first extension should focus on feedforward networks. In networks with feedback loops, the phenomenon of *deadlocks* will be crucial, which occur when blocked jobs are directly or indirectly blocked by themselves. These deadlocks need to be resolved, adding another layer of complexity to the system. Furthermore, the heuristic relies on assumptions about the arrival (Poisson process) and service processes (exponentially distributed), which may not hold in all real-world scenarios. Although the impact of the service time distribution is usually small for systems with a considerable number of servers, it is of interest to determine how the heuristic needs to be adapted to handle general arrival and service processes.

Apart from the approximation, future research could also focus on the optimization part of this research. During experiments it seemed that the following rule seems to hold for optimal allocations: $c_1\mu_1 \geq c_2\mu_2 \geq \ldots \geq c_n\mu_n$. This observation is as expected; upstream queues require more capacity due to the cumulative effect of strain and stochasticity introduced by ASB in downstream queues. As jobs progress through the system, any delays or blockages in later queues propagate backward, increasing the workload and variability experienced by earlier queues. This requires additional capacity in upstream queues to manage the compounded effects of downstream inefficiencies. In contrast, downstream queues require comparatively less capacity, since each subsequent queue only needs to handle the additional workload generated by the ASB of the preceding upstream queues. Therefore, the capacity of the i^{th} queue must be sufficient to process the increased workload resulting from blockages in the $n - i$ queues after it. This could be further researched to determine whether this optimization process can be sped up without the need for evaluations.

5.A Appendix: Queues with Fractional Number of Servers

In this appendix, we present the performance analysis for the M/M/c and M/M/c/c+k systems in case c is not necessarily integer. The results rely on the Gamma function, given by,

$$\Gamma(c) = \int_0^\infty t^{c-1} e^{-t} \mathrm{d}t, \tag{5.14}$$

and the upper incomplete Gamma function,

$$\Gamma(c, x) = \int_x^\infty t^{c-1} e^{-t} \mathrm{d}t. \tag{5.15}$$

M/M/c Queue

The M/M/c queue is a queue with Poisson arrivals, with rate λ, and exponential service times, with rate μ, and c servers. While not resembling our ASB situation due to the fact that all jobs are accepted in this case, we can still use the knowledge from these formulas in approximations. The steady-state probabilities of the number of customers in the system are given by [127], with $\rho = \frac{\lambda}{\mu c}$,

$$\theta_i = \begin{cases} \theta_0 \frac{(c\rho)^i}{i!}, & \text{for } 0 < i < c, \\ \theta_0 \frac{(c\rho)^i c^{c-i}}{c!}, & \text{for } i \geq c, \\ \left[\sum_{j=0}^{c-1} \frac{(c\rho)^j}{j!} + \frac{(c\rho)^c}{c!} \cdot \frac{1}{1-\rho}\right]^{-1}, & \text{for } i = 0. \end{cases} \tag{5.16}$$

The probability of waiting $\theta_{i \geq c}$ is then given by:

$$\theta_{i \geq c} = \sum_{i=c}^{\infty} \theta_i. \tag{5.17}$$

To improve the speed of the approximations, these calculations are rewritten, using [110], such that the value c does not have to be integer. It is rewritten to:

$$\theta_0 = \left[\frac{e^{c\rho}}{\Gamma(c)} \Gamma(c, c\rho) + \frac{(c\rho)^c}{\Gamma(c+1)} \frac{1}{(1-\rho)}\right]^{-1} \tag{5.18}$$

and

$$\theta_{i \geq c} = 1 - \theta_{i<c} = 1 - \theta_0 \frac{e^{c\rho}}{\Gamma(c)} \Gamma\left(c, c\rho\right). \tag{5.19}$$

M/M/c/$c+k$ Queue

The steady-state probabilities of the number of customers in the system in the M/M/c/$c+k$ queue are given by, with $a = \frac{\lambda}{\mu}$:

$$\pi_i = \begin{cases} \pi_0 \frac{a^i}{i!}, & \text{for } 0 < i < c, \\ \pi_0 \frac{a^i}{c^{i-c}c!}, & \text{for } c \leq i \leq c+k, \\ \left[\sum_{j=0}^{c} \frac{a^j}{j!} + \frac{a^c}{c!} \sum_{j=c+1}^{c+k} \left(\frac{a}{c}\right)^{j-c}\right]^{-1}, & \text{for } i = 0. \end{cases} \tag{5.20}$$

5

For the heuristics we also require the expected waiting time $\mathbb{E}[W_q]$ in the M/M/c/$c+k$ model. This is computed using

$$\mathbb{E}[W_q] = \frac{\mathbb{E}[L_q]}{\lambda \cdot (1 - \pi_{c+k})}, \tag{5.21}$$

where the expected number of customers in the queue $\mathbb{E}[L_q]$ is given by

$$\mathbb{E}[L_q] = \sum_{i=c}^{c+k} (i-c)\pi_i. \tag{5.22}$$

To handle cases where c is not required to be an integer, Equation (5.22) can be rewritten as:

$$\mathbb{E}[L_q] = \begin{cases} \frac{c^c \pi_0 (k+1)}{\Gamma(c+1)}, & \text{for } \rho = 1, \\ \frac{\pi_0}{\Gamma(c+1)} (c\rho)^c \left[\frac{\rho - \rho^{k+1}}{(1-\rho)^2} - \frac{k\rho^{k+1}}{(1-\rho)}\right], & \text{otherwise.} \end{cases} \tag{5.23}$$

Moreover, π_0 can be rewritten as (5.2) using a similar argument as for the M/M/c queue.

6

Deadlock Detection and Resolution in Queueing Networks with Blocking

Based on:
de Boer, T. R., Bekker, R., & van der Mei, R. D. (2025). Deadlock detection and resolution in queueing networks with blocking. *Submitted for publication.*

Abstract

Queueing networks with after-service blocking and limited or no buffer space are prone to deadlocks, i.e., situations in which jobs become directly or indirectly blocked by themselves. Such networks are frequently used to model real-world scenarios, particularly in healthcare. This chapter addresses deadlock *detection* and *resolution* in queueing networks with blocking, as a resolution mechanism is essential to prevent permanent deadlocks. We derive closed-form expressions for the number of customers in a single-node network to illustrate the effect of resolution frequency. For more complex settings, we propose a dynamic deadlock resolution strategy based on Integer Linear Programming, which adaptively determines job transitions. The effectiveness of this method is demonstrated using a system dynamics model, and its impact on blocking probabilities is evaluated through simulation. This approach substantially improves system performance: increasing the resolution frequency reduces job rejections and blocked servers, especially in systems with few servers or coarse time discretization. A case study on elderly care in Amsterdam shows that coordinated deadlock resolution can reduce patient rejections by up to 10% and wrong-bed days by up to 20% over a ten-year horizon, improving patient flow without expanding capacity.

6.1 Introduction

Healthcare is fundamental to societal well-being and quality of life across all age groups. As populations age, healthcare systems face increasingly complex challenges, particularly in elderly care, where demand continues to increase [72, 185]. Figure 6.1, similar to Figure 1.6, provides a stylized representation of the elderly healthcare system, which includes four interconnected types of care: Home Care (HC), Hospital Care (Hos), Short-Term (Residential) Care (STC), and Long-Term Care (LTC). Each care setting operates under finite capacity, such as available beds or care professionals' time, making the system vulnerable to *bed blocking*. Bed blocking occurs when a patient cannot transition to the required level of care due to capacity constraints and therefore remains in their current setting [15]. A typical example is a hospital patient waiting for a bed in LTC, occupying a costly hospital bed in the meantime [75, 257]. This not only increases costs, as hospital care is generally more expensive than aftercare, but also adversely affects patient outcomes due to prolonged hospitalization and subsequent complex care needs [43, 92]. The days during which a patient receives care in a setting that does not align with their medical requirements are referred to as *wrong-bed days* [60].

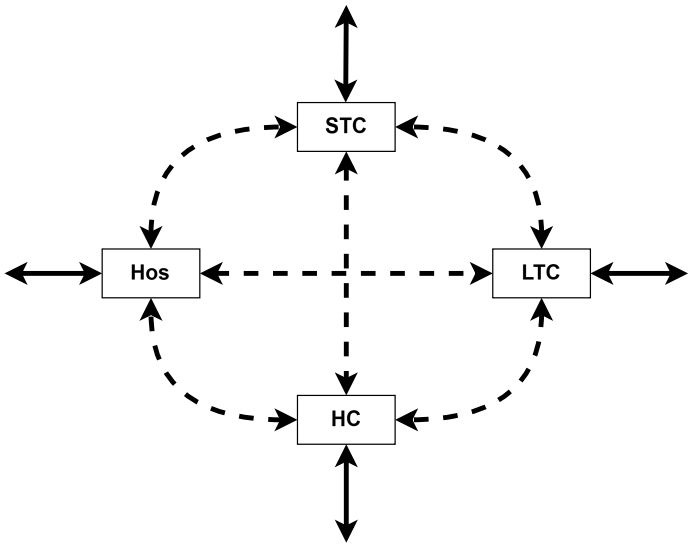


Figure 6.1: Visual representation of patient flows for older adults between various types of healthcare.

Bed blocking can result in patients (in)directly blocking their own progression through the healthcare system [171]; for instance, a hospital patient may be waiting for a bed in the LTC, while an LTC patient simultaneously waits for a hospital bed. These situations are known as deadlocks and can trigger a snowball effect of congestion if not addressed. Currently, these deadlocks are often resolved manually, placing additional pressure on an already stretched workforce. There-

fore, an effective and coordinated deadlock resolution method is vital to support operational decision making, ease the burden on healthcare staff, and build accurate models of real-world systems. The concept of bed blocking extends beyond healthcare and appears in domains such as manufacturing and communication networks [1, 48, 123, 194, 227], where it is generally referred to as *After-Service Blocking* (ASB). In these contexts, as well, effectively resolving deadlocks can lead to better capacity utilization.

To illustrate the impact of deadlocks, this chapter considers a network of multi-server queues without buffer space and with ASB, where external arrivals are lost if all servers are occupied. We show that when routing induces feedback loops in the queueing network, deadlocks inevitably arise, and ASB traps the system in an absorbing state if blocking is not resolved. Therefore, resolving deadlocks is essential when modeling realistic ASB systems with feedback loops. To illustrate the impact of the deadlock resolution frequency, we derive closed-form expressions for the blocking probability in a single-node continuous-time model. This analysis reveals that a relatively coarse resolution strategy is typically sufficient, in particular for queues with many servers.

For more complex queueing networks, we introduce a novel and efficient method for deadlock detection and resolution. Our approach combines wait-for graphs for deadlock detection with a flexible Integer Linear Programming (ILP) model for deadlock resolution. The proposed method is intuitive, flexible, and computationally efficient. We demonstrate the ILP-based approach using a System Dynamics (SD) model. The SD model illustrates that even in deterministic settings deadlocks may arise when capacity is insufficient, unless resolution mechanisms are applied. Resolution mechanisms can either eliminate deadlocks or reduce the capacity required to avoid them. Finally, to evaluate the impact on blocking probabilities, the resolution method is incorporated in a discrete-time simulation, reflecting the nature of many real-world systems. The simulation results indicate that coarse time discretization increases blocking probabilities, highlighting the importance of timely deadlock resolution.

The main contributions of this chapter are threefold:

(i) We show that deadlock resolution is essential in queueing networks. We provide insight into the impact of the deadlock resolve frequency and show that networks with few servers and coarse time discretization are especially vulnerable to the occurrence of deadlocks.

(ii) We propose a scalable ILP-based approach to resolve deadlocks in queueing networks with limited or no buffer space and ASB, easily adaptable to various objectives and constraints.

(iii) We demonstrate the effectiveness of the method in a realistic healthcare-inspired case study, showing a significant increase in throughput and a decrease in the number of wrong-bed days.

The remainder of this chapter is organized as follows. Section 6.2 reviews related work on ASB and deadlock phenomena in queueing networks. Section 6.3 provides a formal definition of the problem and the concept of deadlocks. Deadlocks and the impact of their resolution are illustrated using the analysis of Markov chains and SD models in Section 6.4. Section 6.5 describes the proposed deadlock detection and resolution method in a discrete-time simulation setting, including the detection and resolution processes. The simulation results are given in Section 6.6, and the application of the method to the elderly healthcare setting in Amsterdam, the Netherlands, is presented in Section 6.7. Section 6.8 provides a comprehensive conclusion to this study, summarizing the key findings and insights discussed throughout the chapter.

6.2 Literature

Queueing networks with blocking mechanisms have been extensively studied, with ASB receiving particular attention (see Chapter 5). Onvural and Perros [165] distinguish three forms of blocking: ASB, *Before-Service Blocking* (BSB), and *Repetitive Service Blocking* (RSB). ASB occurs when a task completes its service at node i and requires a transition to node j. If no server or buffer space is available at node j, the task must remain at the server of node i until space at node j becomes available. BSB resembles ASB, but the job only begins processing at queue i once a transition to node j is possible, otherwise it continues to occupy a server at node i. Finally, under RSB, jobs repeat their service in the current node i until a space at node j is available. Although all three types of blocking have the potential to cause deadlocks, ASB is the most widely studied and is frequently used to model realistic scenarios. Consequently, this chapter focuses specifically on that type of blocking.

Research on exact solutions for ASB in queueing networks [1, 11, 12, 81, 103, 131] demonstrates that the impact of ASB on system performance, such as throughput or number of blocked jobs, can be determined analytically in some cases. These articles also emphasize that such analyses are often computationally intensive and are generally restricted to particular simplified networks. These models mainly involve tandem queues or, more broadly, feed-forward networks [93] (i.e., networks with acyclic routing), which are unable to enter deadlocks.

The complexity and limited applicability of exact methods have motivated the development of approximations to assess the impact of ASB. Various studies have explored approximation techniques, ranging from decomposing the system into subsystems with adjusted parameters [1, 26], using neural networks as metamodels [62], or simulating the system [49]. These approaches substantially reduce computational effort compared to exact methods. However, they are typically limited to feed-forward networks [48, 62, 123, 227], disregard deadlocks [26], or assume that they are resolved instantly without constraints on the resolution process [167]. Thus, both exact and approximate methods to ASB typically assume

that deadlocks can be readily resolved or ignored, thereby underestimating their effect on system performance.

Deadlocks have been studied primarily in the context of concurrent computing, a well-established area in computer science [44, 67, 90, 108, 115, 179, 180]. In this context, deadlocks arise when processes hold locks on specific resources and cannot proceed because each is waiting for another, possibly itself, to take an action, such as sending a message or releasing a lock. Research in this area has focused mainly on ignoring deadlocks [206], detecting them [209], or preventing and avoiding them [130]. However, these approaches have limited applicability to deadlocks in queueing networks, particularly in realistic applications like healthcare systems.

The study of deadlocks within queueing networks remains relatively limited. According to Palmer et al. [171], permanent deadlocks in feedback loop queueing networks lead to absorbing states. Their study determines the expected time until absorption, highlighting the necessity of incorporating deadlocks into simulation models. Incorporating an absorbing state hinders the computation of long-term averages, which are frequently utilized in queueing theory. The studies [128, 133] show how topology and capacity allocation can prevent deadlocks in a closed queueing network. However, they do not address the resolution of deadlocks in open networks, an important limitation given the practical prevalence of open queueing networks.

Deadlock detection and resolution methods identify and resolve deadlocks as they arise, often using wait-for graphs [2, 44] to represent jobs waiting for specific resources. Using these graphs, the presence of a deadlock in the current configuration of the system can be determined. Resolution methods are then applied to guide the system from a deadlock configuration to a resolved one. Research on this topic remains limited and often assumes that swaps are trivial [61]. In practice, swaps can be subject to constraints and priorities; for example, bed blocking in a hospital can incur higher costs than in a nursing home, and the number of swaps may be restricted by personnel availability.

We conclude that the amount of research on deadlock phenomena in queueing networks is limited. Moreover, a key challenge in complex systems, such as healthcare networks, lies not only in resolving deadlocks but also in accurately modeling their occurrence. Understanding *when* and *why* deadlocks arise is crucial for designing effective interventions and improving system efficiency. Many existing approaches lack adaptability to dynamic real-world scenarios, limiting their effectiveness. To address these gaps, we introduce a flexible deadlock resolution method that accounts for system constraints and varying operational conditions. By integrating structured modeling with a centralized resolution approach, our method offers a more scalable and systematic solution to handling deadlocks in complex environments than ad hoc strategies.

6.3 Model Description and Deadlocks

In this section, we present the model and some preliminary results related to deadlocks. We consider a queueing network with n nodes, each having zero buffer space and multiple servers; examples with $n = 1$ and $n = 2$ nodes can be found in Figure 6.3. External arrivals at node i follow a Poisson process with rate λ_i. The service times are exponentially distributed with rate μ_i for node i. The number of servers at node i is given by c_i. The examples in this chapter assume zero buffer; the methods are also applicable to networks with finite buffers. Upon service completion at node i, jobs request service from queue j with probability r_{ij}, with R the corresponding routing matrix. The probability of a job leaving the system after receiving service at queue i is $r_{i0} = 1 - \sum_{j=1}^{n} r_{ij}$. External arriving jobs finding all servers occupied at the corresponding node are rejected, whereas ASB applies to jobs within the system. That is, a finished job at queue i with destination queue j remains at node i and continues to occupy a server at node i, until a server at j becomes available.

Let $L_i^s(t)$ denote the number of jobs in service at node i at time t, and let $L_{ij}^b(t)$ represent the number of jobs at node i that are blocked from moving to node j at time t. Define $\mathbf{L}^s(t)$ as the corresponding vector of jobs in service and $\mathbf{L}^b(t)$ as the corresponding matrix of blocked jobs. If $L_{ij}^b(t) > 0$ for multiple i, and a server at node j becomes available at time t, an unblocking mechanism must be applied. If this unblocking mechanism depends solely on $\mathbf{L}^b(t)$, then the process $\{\bar{\mathbf{L}}(t), t > 0\}$, with $\bar{\mathbf{L}}(t) = (\mathbf{L}^s(t), \mathbf{L}^b(t))$, forms a Markov process with state space

$$\left\{ (\mathbf{L}^s(t), \mathbf{L}^b(t)) \in \mathbb{N}_0^n \times \mathbb{N}_0^{n \cdot n} : L_i^s(t) + \sum_{j=1}^{n} L_{ij}^b(t) \leq c_i, i = 1, \ldots, n \right\}.$$

Let L_i^s and L_{ij}^b denote the corresponding stationary versions of these random variables. The primary performance measures we consider for any node i are the throughput $\bar{\gamma}_i$, the fraction of rejected external arrivals B_i, and the expected number of jobs blocked $\mathbb{E}[L_i^b]$, with $L_i^b = \sum_j L_{ij}^b$. Since external arrivals follow a Poisson process, PASTA yields that B_i corresponds to the probability that all servers are busy at node i, i.e., $B_i = \mathbb{P}(L_i^s + L_i^b = c_i)$. The throughput $\bar{\gamma}_i$ and fraction of rejected arrivals B_i are related via the routing equations

$$\bar{\gamma}_i = \lambda_i(1 - B_i) + \sum_{j=1}^{n} \bar{\gamma}_j r_{ji}. \tag{6.1}$$

The aim of this chapter is to illustrate the necessity of incorporating a control policy and to propose a simple and efficient strategy for determining job transitions. In particular, the ASB mechanism imposes the risk of deadlocks, which occurs when one job is hindered by another, which in turn is also obstructed by the initial job. We define deadlock as follows, cf. [171]:

Definition 6.1. *When there is a subset of jobs that are blocked directly or indirectly by themselves, then the system is said to be in deadlock.*

Here, a deadlock refers to a situation in which at least one server is not serving a job but unable to serve a new job, because that job is blocked by a circular dependency within the system. Figure 6.2 illustrates such a scenario: both nodes are fully occupied, with each server actively serving a job. Job A completes service first and requests service at node 2, but remains at node 1 due to capacity constraints. Subsequently, job B finishes service and requests service at node 1, but is similarly blocked at node 2. This results in a deadlock, where job A is obstructed by B, and B is similarly hindered by A.

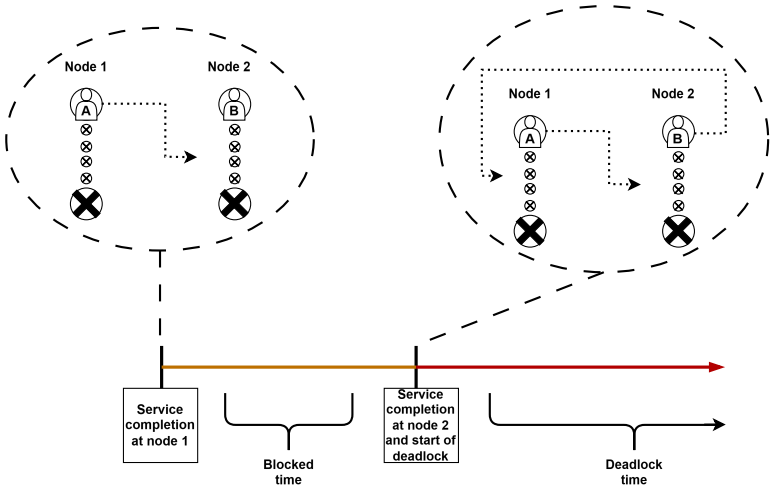


Figure 6.2: Visual representation of a simple deadlock situation.

We distinguish between a deadlock and a *permanent* deadlock, where the latter is defined by:

Definition 6.2. *When all servers are blocked and there are no transitions possible, without deadlock resolve, the system is said to be in a permanent deadlock.*

Deadlock situations occur if the ASB queueing network is not feed-forward. More formally, consider the directed graph $G = (V, E)$ induced by the routing matrix R, where there is a directed edge from i to j if and only if $r_{ij} > 0$. Let $J \subseteq V$ be the subset of nodes forming a *strongly connected component* of G [161]. This means it is a subset of nodes where each node is reachable from every other node via a directed path. Define $A \subseteq V \setminus J$ as the set of *ancestors* of J, i.e., nodes with directed paths leading into J.

Theorem 6.1. *As time progresses, without deadlock resolve options, all servers in $A \cup J$ become permanently blocked. If $A \cup J = V$, the whole system inevitably converges to a permanent deadlock.*

Proof. Consider the continuous-time Markov process $\{\bar{\mathbf{L}}(t), t > 0\}$ and let $\bar{\boldsymbol{\ell}} = (\boldsymbol{\ell}^s, \boldsymbol{\ell}^b)$ and $\bar{\boldsymbol{\ell}}' = (\boldsymbol{\ell}^{s'}, \boldsymbol{\ell}^{b'})$ denote the current state and pre-action state, respectively, where $\boldsymbol{\ell}^s$ is a vector containing the number of jobs in service at each node,

and $\boldsymbol{\ell}^b$ is a matrix specifying the number of jobs blocked at each node while waiting for service at another node. A *pre-action state* is a state after a transition to which an unblocking mechanism may still be applied resulting in a *post-action state*. Such an unblocking mechanism is often trivial, but is more involved when jobs at different nodes can be unblocked due to a service completion at their destination node; as we do not need the unblocking mechanism here, we leave it unspecified. Let e_i denote the i-th unit vector and e_{ij} be a matrix with zeroes and a 1-entry in the i-th row and the j-th column. We then have the following transition rates

$$q_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\ell}}'} = \begin{cases} \lambda_i & \text{if } \boldsymbol{\ell}^{s'} = \boldsymbol{\ell}^s + e_i, \boldsymbol{\ell}^{b'} = \boldsymbol{\ell}^b, \ell_i^s + \sum_j \ell_{ij}^b < c_i, \\ \ell_i^s \mu_i r_{i0} & \text{if } \boldsymbol{\ell}^{s'} = \boldsymbol{\ell}^s - e_i, \boldsymbol{\ell}^{b'} = \boldsymbol{\ell}^b, \\ \ell_i^s \mu_i r_{ij} & \text{if } \boldsymbol{\ell}^{s'} = \boldsymbol{\ell}^s - e_i + e_j, \boldsymbol{\ell}^{b'} = \boldsymbol{\ell}^b, \ell_j^s + \sum_k \ell_{jk}^b < c_j, \\ \ell_i^s \mu_i r_{ij} & \text{if } \boldsymbol{\ell}^{s'} = \boldsymbol{\ell}^s - e_i, \boldsymbol{\ell}^{b'} = \boldsymbol{\ell}^b + e_{ij}, \ell_j^s + \sum_k \ell_{jk}^b = c_j, \\ 0 & \text{otherwise.} \end{cases}$$

The first line corresponds to an arrival at node i, provided there is sufficient capacity. The subsequent three lines relate to a service completion at node i, followed by a system departure, a transition to node j, or a job being blocked due to a lack of available capacity at node j, respectively.

Denote the set C of states in which all servers in $A \cup J$ are blocked by jobs at nodes $A \cup J$, that is

$$C = \left\{ (\boldsymbol{\ell}^s, \boldsymbol{\ell}^b) : \sum_{j \in A \cup J} \ell_{ij}^b = c_i, i \in A \cup J \right\}. \tag{6.2}$$

Clearly, $\ell_i^s = 0$ for $i \in A \cup J$, and therefore no service completions are possible. From the specified transition rates, it follows directly that C is an absorbing class, since $\sum_j \ell_{ij}^b = c_i$ yields that no external arrivals at node i are accepted. Moreover, it may be readily verified that there is a directed path from states in the complement of C to states in C. As the state space is finite, it follows from standard results on Markov processes that the limiting distribution has only positive probabilities for states in C. Finally, if $A \cup J = V$, then all servers in the system are blocked and the system is in a permanent deadlock. □

For modeling practical scenarios, we need to define some deadlock resolution mechanism, as situations in which all servers are permanently blocked do not occur.

Remark 6.1. The model in this section is formulated in continuous time. However, in many real-world settings, especially healthcare, events such as patient transfers and discharges typically occur at fixed intervals (for example, daily), making discrete-time models realistic and practical. Consequently, the SD model (Section 6.4.2) and the simulation model (Section 6.5) are based on this discrete-time framework. Note that Theorem 6.1 can be directly extended to the discrete-time equivalent.

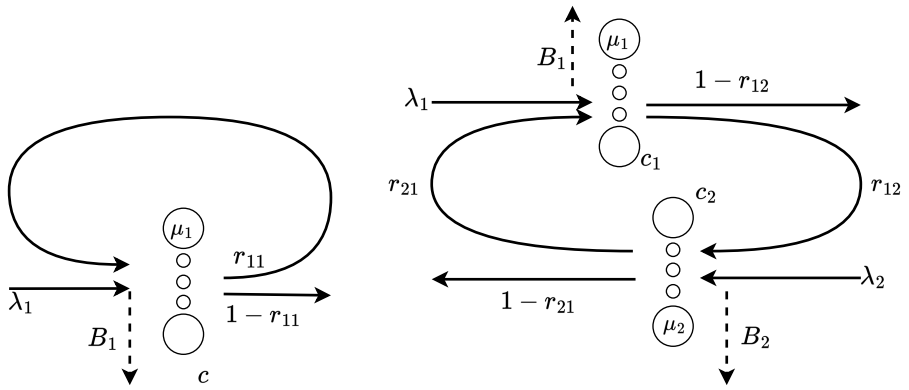


Figure 6.3: One- and two-node queueing network with ASB and potential deadlocks.

6.4 Deadlocks: Motivating Examples

This section presents examples that illustrate the impact of deadlocks and the effectiveness of resolution strategies. We analyze one- and two-node Markovian systems with a deadlock resolution mechanism in Section 6.4.1. Section 6.4.2 introduces an SD model, demonstrating that deadlocks can arise even in deterministic systems with limited capacity. We also propose a strategy for coordinating job transfers to mitigate this effect. Together, these examples provide crucial insights into when deadlocks occur and how they can be effectively resolved.

6.4.1 Performance analysis of Markovian systems

In this section, we analyze the Markovian systems with one node and two nodes, as illustrated in Figure 6.3. Although our main focus is on the one-node system, since it allows for closed-form expressions, we also include a two-node system.

Single-node system Consider a one-node queueing network ($n = 1$) with c servers, see the left panel in Figure 6.3. Upon service completion, a job rejoins service with probability r_{11}, and leaves the system otherwise. The system exhibits ASB, meaning that a job completing service is blocked from re-entering service if no (other) server is immediately available. In a single-node system, ASB and deadlock are identical. In order to prevent permanent deadlocks, an unblocking mechanism is employed following a Poisson process with a rate of ν, unblocking all servers in deadlock simultaneously. Although the unblocking mechanism is straightforward in this system, this example provides insight in the impact of the frequency of unblocking events.

Similar to [171], we define the state as the total number of occupied ($L^s(t) + L^b(t)$) *plus* the blocked ($L^b(t)$) servers, i.e., $\{X(t), t > 0\}$ with $X(t) = L^s(t) + L^b(t) + L^b(t) = L^s(t) + 2L^b(t)$, constitutes a Markov process on the state space

$\{0, 1, \ldots, 2c\}$. The state $i < c$ indicates that i servers are active and delivering service, while $c + j$ (with $j \geq 0$) implies that all servers are occupied, with j servers blocked due to jobs waiting to rejoin service, and $c - j$ servers still providing service. The transition rates of this Markov process are as follows, cf. Figure 6.4,

$$q_{i_1,i_2} = \begin{cases} \lambda_1 & \text{if } i_2 = i_1 + 1, i_1 < c, \\ i_1\mu_1(1 - r_{11}) & \text{if } i_2 = i_1 - 1, i_1 \leq c, \\ (2c - i_1)\mu_1 r_{11} & \text{if } i_2 = i_1 + 1, i_1 \geq c, \\ (2c - i_1)\mu_1(1 - r_{11}) & \text{if } i_2 = c - 1, i_1 > c, \\ \nu & \text{if } i_2 = c, i_1 > c, \\ 0 & \text{otherwise.} \end{cases}$$

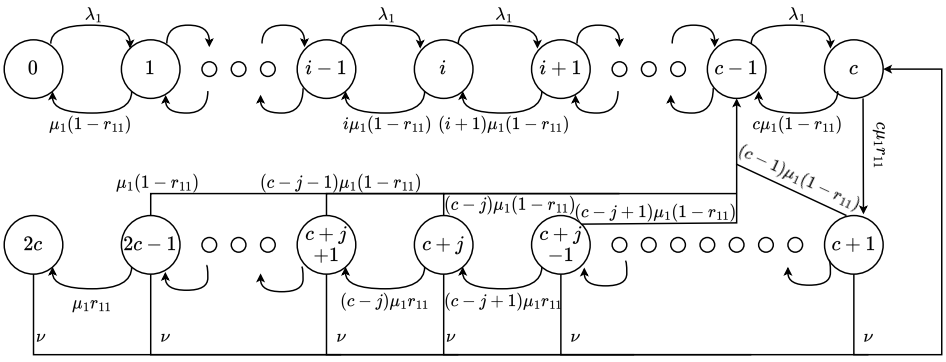


Figure 6.4: Continuous-time Markov chain of one-node network with ASB and deadlock resolution.

To obtain the limiting distribution of the total number of occupied and blocked servers in closed form, define

$$a := \frac{\lambda_1}{\mu_1(1 - r_{11})}, \qquad \tilde{a}_k := \frac{c!}{(c-k)!} \frac{(\mu_1 r_{11})^k}{\prod_{j=1}^{k}((c-j)\mu_1 + \nu)}, \quad k = 0, 1, \ldots, c,$$

where an empty product is understood to be 1 (i.e., $\tilde{a}_0 = 1$).

Theorem 6.2. *For a single-node system with $r_{11} < 1$, the limiting distribution of $L^s(t) + 2L^b(t)$ is given by $\pi_i = a^i/i!\,\pi_0$, for $i = 0, \ldots, c-1$, and, for $k = 0, \ldots, c$,*

$$\pi_{c+k} = \tilde{a}_k \cdot \frac{\lambda_1}{c\mu_1 - \nu \sum_{j=1}^{c} \tilde{a}_j} \frac{a^{c-1}}{(c-1)!} \pi_0,$$

with

$$\pi_0 = \left[\sum_{i=0}^{c-1} \frac{a^i}{i!} + \frac{a^{c-1}}{(c-1)!} \frac{\lambda_1}{c\mu_1 - \nu \sum_{j=1}^{c} \tilde{a}_j} \sum_{k=0}^{c} \tilde{a}_k \right]^{-1}.$$

The loss fraction is

$$B_1 = \sum_{k=0}^{c} \tilde{a}_k \cdot \frac{\lambda_1}{c\mu_1 - \nu \sum_{j=1}^{c} \tilde{a}_j} \frac{a^{c-1}}{(c-1)!} \pi_0.$$

Proof. First, consider states $i = 0, 1, \ldots, c-1$. Using the balance equations $i\mu_1(1-r_{11})\pi_i = \lambda_1 \pi_{i-1}$, we directly obtain $\pi_i = a^i/i!\ \pi_0$. For states $c+k$, $k = 1, \ldots, c$, the balance equations $\pi_{c+k}((c-k)\mu_1 + \nu) = (c-k+1)\mu_1 r_{11} \pi_{c+k-1}$ yield, for $k = 1, \ldots, c$,

$$\pi_{c+k} \ = \ \frac{c!}{(c-k)!} \frac{(\mu_1 r_{11})^k}{\prod_{j=1}^{k}((c-j)\mu_1 + \nu)} \pi_c \ = \ \tilde{a}_k \pi_c.$$

For state c, the balance equation reads $c\mu_1\pi_c = \lambda_1 \pi_{c-1} + \nu \sum_{k=1}^{c} \pi_{c+k}$. Using the expression above, we can find π_c in terms of π_0:

$$\pi_c = \frac{\lambda_1}{c\mu_1 - \nu \sum_{k=1}^{c} \tilde{a}_k} \frac{a^{c-1}}{(c-1)!} \pi_0.$$

Using normalization, we obtain the limiting distribution. The loss fraction follows directly from PASTA and the fact that $B_1 = \sum_{k=0}^{c} \pi_{c+k}$. □

If there is no unblocking mechanism (i.e., $\nu = 0$), then state $2c$ is absorbing and corresponds with a system in permanent deadlock, in line with Theorem 6.1. In case $\nu = \infty$, any blocking is instantly resolved, representing a scenario without ASB. In that case, the system corresponds to an Erlang loss system with service rate $\mu_1(1 - r_{11})$. This follows from Theorem 6.2, using $\tilde{a}_k \to 0$, for $k = 1, \ldots, c$, $\nu\tilde{a}_1 \to c\mu_1 r_{11}$, and $\nu\tilde{a}_k \to 0$, for $k = 2, \ldots, c$, as $\nu \to \infty$.

Another appealing special case is $\nu = \mu_1$. In that case, $\tilde{a}_k = r_{11}^k$, indicating that the probability of having k blocked servers decreases geometrically. Moreover, using $\tilde{a}_k = r_{11}^k$, the loss fraction may be written as

$$B_1 = \frac{c(1-r_{11})(1-r_{11}^{c+1})}{c(1-r_{11}) - r_{11}(1-r_{11}^c)} \times \frac{a}{c} \times \frac{a^{c-1}}{(c-1)!} \pi_0,$$

with

$$\pi_0 = \left[\sum_{i=0}^{c-1} \frac{a^i}{i!} + \frac{c(1-r_{11})(1-r_{11}^{c+1})}{c(1-r_{11}) - r_{11}(1-r_{11}^c)} \times \frac{a^c}{c!} \right]^{-1}.$$

As the number of servers grows, the first factor in B_1 will quickly approach 1, implying that the system approximately behaves as an Erlang loss system with offered load a. The same applies in case $\nu > 0$, with c becoming large, as formalized in the following corollary (see Section 6.A.1 for the proof).

Corollary 6.3. *Let $B(a, c)$ be the loss fraction in the Erlang loss model with offered load a and c servers. For fixed μ_1, $r_{11} < 1$, and $\nu > 0$, it holds that $B_1/B(a, c) \to 1$ as $c \to \infty$.*

These examples show that the impact of ASB will be small if the number of servers becomes large. This indicates that an unblocking mechanism is crucial to avoid permanent deadlocks, but there is no need to invoke it too often for larger systems.

In Figure 6.5, we visualize the behavior of the blocking probability B_1 and the expected number of blocked servers $\mathbb{E}[L_1^b]$. The base parameters are $\lambda_1 = 0.5c$, $\mu_1 = 1$, $r_{11} = 0.5$, $\nu = 0.25$, and $c = \{1, 4, 7\}$, unless stated otherwise; in each panel one parameter is varied. Panel (a) illustrates how the blocking probability decreases with a larger deadlock resolution rate ν. In line with Theorem 6.1, $B_1 = 1$ in case $\nu = 0$. As ν increases, the blocking probability converges to that of an Erlang loss system, indicated by the dotted lines. The plot also indicates that even a relatively small value of ν is sufficient to mitigate the effects of deadlock, particularly in systems with a larger number of servers c.

The impact of the service rate μ_1 on the expected number of blocked jobs is visualized in panel (b). For $c = 1$, we observe that $\mathbb{E}[L_1^b]$ in increasing with μ_1. This is intuitive: higher service rates accelerate transitions into deadlock states, which then require the resolution mechanism to recover. For $c = 4$ and $c = 7$, we observe a non-monotonic behavior with $\mathbb{E}[L_1^b]$ initially increasing with μ_1, and then decreasing. This pattern reflects the interplay between deadlock resolution and service departures. For small μ_1, resolution dominates and clears deadlocks relatively quickly. As μ_1 increases, more service completions trigger more deadlocks. At high μ_1, departures occur frequently enough to resolve deadlocks naturally, reducing the expected number of blocked jobs. The impact of the arrival rate λ_1 on B_1 and $\mathbb{E}[L_1^b]$ can be found in panels (c) and (d), respectively. In all scenarios, the blocking probability tends to 1 as λ_1 increases, reflecting a growing load. Also, $\mathbb{E}[L_1^b]$ increases with λ_1, albeit $\mathbb{E}[L_1^b]$ stabilizes in highly overloaded scenarios, as additional arrivals are simply rejected from the system.
Finally, we elaborate on the findings of [171] for the single-node situation. In particular, [171] considers $\nu = 0$ and consequently numerically determines the time till absorption. Define the expected time till absorption from state i as $m_i = \mathbb{E}[T_{2c} \mid X(0) = i]$, with $T_j = \inf\{t \geq 0 : X(t) = j\}$. A closed-form expression for m_0 can be found in the lemma below; the proof is given in Section 6.A.1. Although the expression appears somewhat involved, the term $1/r_{11}^c$ indicates that the absorption time grows at least geometrically with the number of servers (when $r_{11} < 1$).

Lemma 6.4. *Assume $\nu = 0$. Then the expected time till absorption from an empty system is*

$$m_0 = \frac{1}{\lambda_1} \sum_{i=0}^{c-2} \sum_{k=0}^{i} \frac{i!}{(i-k)!} a^{-k} + \frac{1}{r_{11}^c} \left(\frac{1}{\mu_1} \sum_{k=0}^{c-1} \frac{r_{11}^k}{c-k} + \frac{1}{\lambda} \sum_{k=0}^{c-1} \frac{(c-1)!}{(c-1-k)!} a^{-k} \right),$$

with the convention that a sum over an empty set is defined to be 0.

Two interacting nodes Next, we consider a two-node queueing network ($n = 2$), as visualized in the right panel of Figure 6.3. In this example, $r_{11} = r_{22} = 0$,

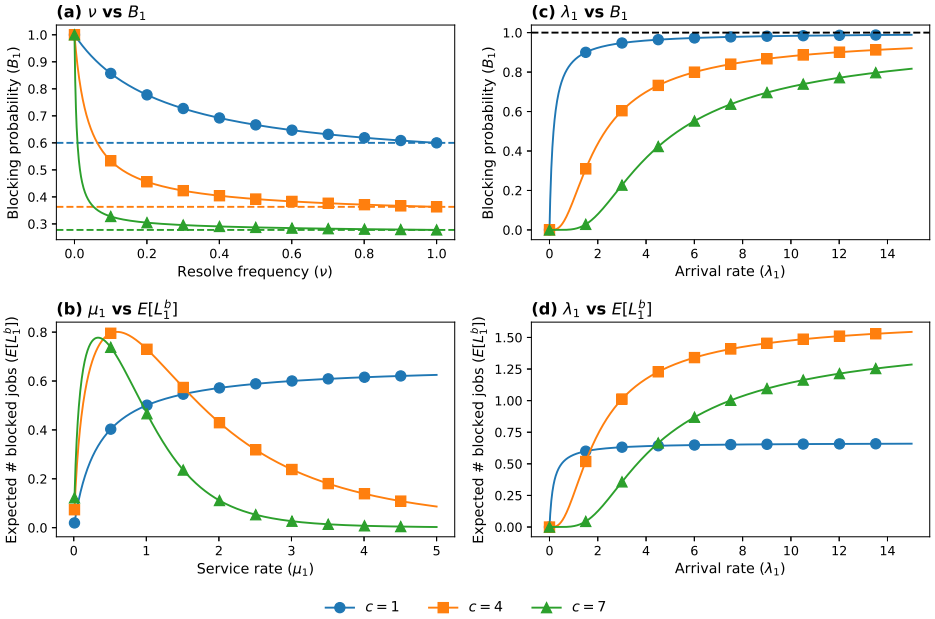


Figure 6.5: The impact of ν, μ_1 and λ_1 on B_1 and $\mathbb{E}[L_1^b]$ in various values of c. The dashed lines correspond to the values of B_1 as ν or λ_1 tend to ∞.

meaning that jobs will either seek service at the alternate node or exit the system once service is finished. Now, ASB occurs as a job can be blocked at node 1 from entering node 2 and vice versa. The deadlock resolution mechanism occurs again according to a Poisson process with rate ν. We define the state as a combination of the number of occupied servers and the number of jobs requesting service at node 1, and at node 2, resulting in the vector $Y(t) = \big(L_1^s(t) + L_{12}^b + L_{21}^b(t),\ L_2^s(t) + L_{21}^b(t) + L_{12}^b(t)\big)$. A more detailed description of the Markov process $\{Y(t), t > 0\}$ can be found in Section 6.A.2.

Section 6.A.2 also provides insight in the impact of various parameters (similar to Figure 6.5 in case $n = 1$) on the blocking probability and expected number of blocked servers. These results are in line with our findings from the previous single-node system, showing that again a relatively small ν is sufficient to avoid the effect of deadlocks on the system. For $n = 2$, we also study the impact of an asymmetric system. Finally, we note that a similar approach may also be applied in case $r_{ii} > 0$, $i = 1, 2$. However, we then need to extend the state description to keep track of which job is blocked where and in what order, which is not currently captured in $Y(t)$.

6.4.2 System dynamics

Theorem 6.1 shows that permanent deadlocks occur in the absence of deadlock resolution options due to randomness in demand and service times. In this section, we illustrate that deadlocks may also occur in the absence of randomness. More specifically, we introduce an SD model, where the random arrival and service processes are replaced by deterministic flows having the same means. SD models are frequently employed as a modeling tool in healthcare [24], with applications ranging from simulating hospital department operations [140] to analyzing the evolving care needs of the population for older adults [25]. In the SD literature, $L_i^s(t)$ and $L_{ij}^b(t)$ are typically referred to as *stocks* (just before time t) and can now take fractional values.

Many SD models assume time to be discrete, with the time step denoted by dt. This aligns with practical situations, as transfers typically do not occur continuously but rather at discrete intervals (e.g., on a daily basis). Note that the number of external arrivals at node i is given by the average number of arrivals in a time interval of size dt, $\lambda_i dt$, whereas the number of service completions at node i that are ready to join node j is $\mu_i dt \cdot L_i^s(t-dt) \cdot r_{ij}$, for dt sufficiently small. The actual flow may depend on the control policy if capacity is not sufficient. Let $F_{ij}(t)$, for $i = 0, \ldots, n$, $j = 1, \ldots, n$, denote the flow at time t, which will be the result of the control policy. For convenience, also define $F_{i0}(t) = \mu_i dt \cdot L_i^s(t-dt) \cdot r_{i0}$ as the flow from node i leaving the network. Now, the changes in stocks in terms of $F_{ij}(t)$ are described by

$$L_i^s(t) = L_i^s(t-dt) + \sum_{j=0}^{n} F_{ji}(t) - \mu_i L_i^s(t-dt) \cdot dt, \tag{6.3}$$

$$L_{ij}^b(t) = L_{ij}^b(t-dt) + \mu_i L_i^s(t-dt) r_{ij} \cdot dt - F_{ij}(t). \tag{6.4}$$

The actual flows $F_{ij}(t)$ should be such that $L_i^s(t) + L_i^b(t) \in [0, c_i]$, $L_{ij}^b(t) \geq 0$, and $F_{0i}(t) \leq \lambda_i dt$. Observe that the flow control policy at time t is based on the pre-action state, i.e., at the stock levels at time t^-. Hence, the available capacity at that instant is

$$c_i^{\text{left}}(t^-) = c_i - L_i^s(t-dt) \cdot (1 - \mu_i r_{i0} \cdot dt) - \sum_{j=1}^{n} L_{ij}^b(t-dt).$$

For the system in equilibrium, we have $L_{ij}^b(t) = L_{ij}^b(t-dt)$ and $L_i^s(t) = L_i^s(t-dt)$. Let $L_i^s = \lim_{t\to\infty} L_i^s(t)$ denote the equilibrium stock at node i, and F_{ij} the equilibrium flow from node i to j. Due to (6.4) and (6.3), we obtain, respectively, for $i, j = 1, \ldots, n$,

$$F_{ij} = \mu_i L_i^s r_{ij} \cdot dt, \qquad \text{and} \qquad \sum_{j=0}^{n} F_{ij} = \mu_i L_i^s \cdot dt. \tag{6.5}$$

Assuming that all external arrivals are accepted, combining the above and dividing by dt gives for $i = 1, \ldots, n$

$$\lambda_i + \sum_{j=1}^{n} \mu_j L_j^s r_{ji} = \mu_i L_i^s.$$

This yields $L_i^s = \gamma_i/\mu_i$, with γ_i the solution of the standard routing equations: $\gamma_i = \lambda_i + \sum_j \gamma_j r_{ji}$. Observe that $\gamma_i = \bar{\gamma}_i$ in the absence of rejected arrivals, i.e., for $B_j = 0$ for $j = 1, \ldots, n$.

Priority to external flows Consider the two-node system from Section 6.4.1, see also the right panel from Figure 6.3, with the arrival and service processes replaced by their deterministic counterparts. Assume that external flows receive priority. In practice, this may be due to urgency or financial incentives. In this priority case, we have

$$F_{0i}(t) = \min\left\{\lambda_i dt, c_i^{\text{left}}(t^-)\right\},$$

$$F_{ij}(t) = \min\left\{c_j^{\text{left}}(t^-) - F_{0j}(t), \mu_i L_i^s(t - dt) r_{ij} dt + L_{ij}^b(t - dt)\right\}.$$

Now, if $\gamma_i > c_i\mu_i$ for some $i \in \{1, 2\}$ (or both), then the system will get into a permanent deadlock, i.e., $L_{12}^b = c_1$ and $L_{21}^b = c_2$. This can be intuitively illustrated by considering the evolution of the stocks. Let all stocks start at level 0 and assume that $\gamma_1 > c_1\mu_1$, whereas $\gamma_2 < c_2\mu_2$. Then, $L_i^s(t)$ will initially increase until node 1 is no longer able to handle all flows, which happens at say t_1. Then, during some time interval $[t_1, t_2]$, there is ASB at node 2, i.e., $L_{21}^b(t)$ increases, whereas $L_1^s(t)$ remains roughly constant. Finally, as $L_{21}^b(t)$ increases, the capacity of node 2 will be reached, such that $L_{12}^b(t)$ starts to increase as well, and $L_1^s(t)$ and $L_2^s(t)$ start to drop and ultimately approach 0. The evolution of the four stock levels is visualized in the left panel of Figure 6.6 for the parameter setting $c_1 = c_2 = 10$, $\mu_1 = \mu_2 = 0.5$, $r_{12} = r_{21} = 0.4$, $\lambda_1 = 4$, and $\lambda_2 = 1.5$, such that $\gamma_1 = 5\frac{10}{21}$ and $\gamma_2 = 3\frac{29}{42}$.

Capacity requirements In this example, we determine the capacity required to satisfy all demand in the SD model, that is, the minimum capacity needed to prevent deadlocks. From the equilibrium analysis, we have $L_i^s = \gamma_i/\mu_i$. In an uncoordinated system, arrivals and departures occur sequentially, implying that the capacity must be sufficient to accommodate all incoming external and internal flows $\sum_{j=0}^{n} F_{ij}(t)$ at time t, in addition to the existing stock in the pre-action state at time t^-. In equilibrium, this stock is given by $L_i^s - \mu_i L_i^s r_{i0} \cdot dt$. Based on (6.3), the required capacity in equilibrium for the uncoordinated case is

$$\begin{aligned} c_i^{\text{unco}} &= L_i^s - \mu_i L_i^s r_{i0} \cdot dt + \lambda_i dt + \sum_{j=1}^{n} (L_j^s \mu_j r_{ji} dt) \\ &= L_i^s \left(1 - \mu_i r_{i0} \cdot dt\right) + \left(\lambda_i + \sum_{j=1}^{n} \gamma_j r_{ji}\right) dt = L_i^s \cdot (1 + \mu_i(1 - r_{i0})dt)\,. \end{aligned}$$

Coordinating arrivals with departures, meaning they occur at the same time so that capacity freed by departing jobs can immediately be used by new arrivals, leads to more efficient capacity requirements: $c_i^{\text{co}} = L_i^s$, which is independent of the time step size. Hence, the relative difference between scenarios with and without coordination amounts to $(c_i^{\text{unco}} - c_i^{\text{co}})/c_i^{\text{co}} = \mu_i(1 - r_{i0})dt$. Clearly, this factor grows linearly both in the time step size dt, as well as the service completion rate, revealing the impact of the time discretization dt.

Flow determination with coordination To achieve an efficient use of capacity as sketched above, we introduce a Linear Programming (LP) model named *Flow determination with coordination* that simultaneously considers all possible flows and enforces capacity constraints after all transfers have occurred. In particular, we have:

$$\max \sum_{i=0}^{n}\sum_{j=1}^{n} F_{ij}(t) \tag{6.6}$$

$$\text{s.t. } F_{0i}(t) \le \lambda_i dt \qquad i = 1,\dots,n, \tag{6.7}$$

$$F_{ij}(t) \le \mu_i L_i^s(t-dt) r_{ij} dt + L_{ij}^b(t-dt) \qquad i = 1,\dots,n, j = 1,\dots,n, \tag{6.8}$$

$$L_i^s(t-dt) + \sum_{j=0}^{n}\big(F_{ji}(t) - F_{ij}(t)\big) + \sum_{j=1}^{n} L_{ij}^b(t-dt) \le c_i \qquad i = 1,\dots,n, \tag{6.9}$$

$$F_{ij}(t) \in \mathbb{R}_{\ge 0} \qquad i = 0,\dots,n,\ j = 1,\dots,n. \tag{6.10}$$

The objective of this LP is to maximize the total flow (6.6), where the decision variables $F_{ij}(t)$ represent the actual flows between nodes. This is subject to the constraint that actual flows do not exceed potential flows (6.7)–(6.8), and that post-transition stock levels remain within capacity limits (6.9). The above LP is also intuitively appealing in case it is applied to the equilibrium situation. Using (6.5), the objective function becomes $\sum_{i=0}^{n}\sum_{j=1}^{n} F_{ij} = \sum_{j=1}^{n} \mu_j L_j^s \cdot dt = \sum_{j=1}^{n} \bar{\gamma}_j \cdot dt$, with $\bar{\gamma}_j = L_j^s \mu_j$. Hence, the objective is to maximize total throughput. Also, we may write $F_{0i} = \lambda_i(1 - B_i)dt$. Due to the second part of (6.5), combining the above and dividing by dt yields the routing equation (6.1) as a constraint in equilibrium. Using that the constraint (6.9) can now be written as $\gamma_i \le c_i\mu_i$ (in equilibrium, we may enforce $L_i^b = 0$), we have the following LP in equilibrium, with decision variables B_i and dummy variables $\bar{\gamma}_i$,

$$
\begin{aligned}
\max \quad & \sum_{i=1}^{n} \bar{\gamma}_i \\
\text{s.t.} \quad & \bar{\gamma}_i = \lambda_i(1 - B_i) + \sum_{j=1}^{n} \bar{\gamma}_j r_{ji} && i = 1, \ldots, n, \\
& \bar{\gamma}_i \leq c_i \mu_i && i = 1, \ldots, n, \\
& B_i \in [0, 1] && i = 1, \ldots, n.
\end{aligned}
$$

The LP not only reduces the capacity needed to prevent ASB, but also efficiently resolves deadlocks and prevents them even when $\gamma_i > c_i\mu_i$ for at least one node. In the example above with priority for external flows (cf. the left panel of Figure 6.6), the LP will accept external arrivals until the capacity of node 1 is reached. At that point, swapping flows between nodes 1 and 2 leads to a better aggregate flow, leading to the trajectory depicted in the right panel of Figure 6.6. We note that the LP admits multiple solutions with identical objective values once the values of $L_i^s(t)$ have stabilized, all leading to the same long-run average of accepted flows. The right panel of Figure 6.6 illustrates one such realization, in which external flows are fully accepted until the capacity of node 2 is reached, after which $L_{21}^b(t)$ is entirely depleted. In this sense, the system does not converge to a (static) equilibrium. If desired, this behavior can be mitigated by introducing a modified objective function that assigns (slightly) different weights to internal flows.

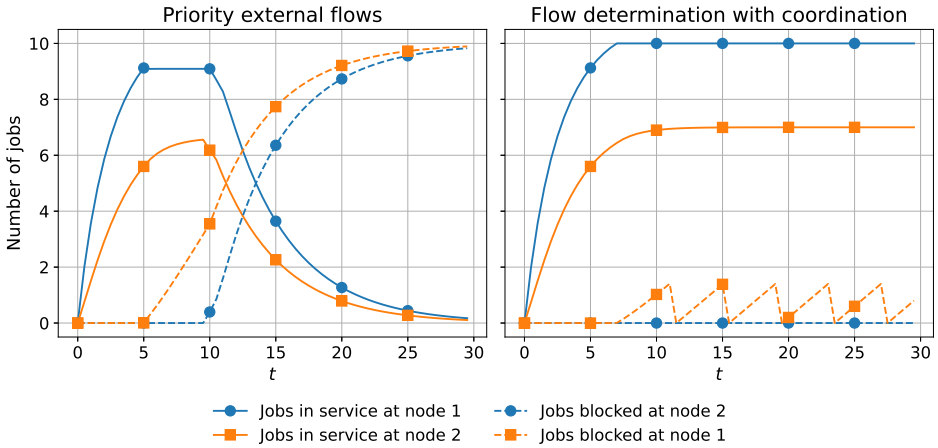


Figure 6.6: Evolution of SD stocks given priority to external flows (left) vs LP deadlock resolution (right).

6.5 Deadlock Resolution in Simulation Environment

This section introduces the framework for detecting and resolving deadlocks in open queueing networks with ASB. The proposed approach consists of two core components: (i) a deadlock *detection* mechanism based on wait-for graphs to identify deadlock states, and (ii) a deadlock *resolution* strategy using an ILP to efficiently resolve such situations. The design of these components is strongly inspired by the insights from Section 6.4. To evaluate the framework's effectiveness in practical settings, we integrate it into a simulation environment.

Consistent with many practical settings, we consider a discrete-time simulation model that serves as the stochastic counterpart of the SD model introduced in Section 6.4.2. To reduce overhead, both in practical coordination and computational effort, we implement a deadlock detection schedule that checks for deadlocks every Δ_t time step. When $\Delta_t = 1$, deadlock detection is performed after every time step, whereas $\Delta_t = \infty$ implies that no detection is performed. Figure 6.7 shows a high-level flow diagram of the simulation process. The key steps are discussed in more detail below.

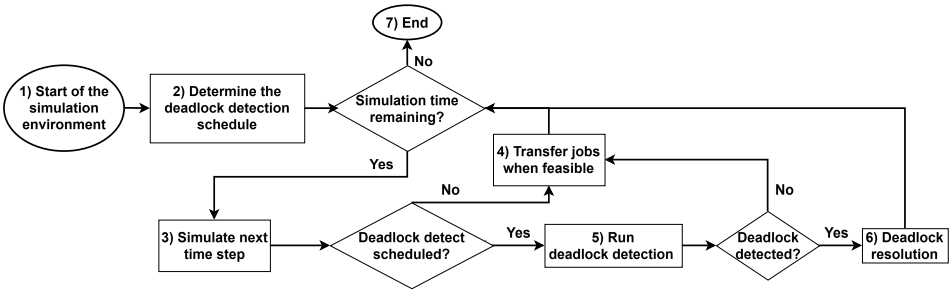


Figure 6.7: Flow diagram illustrating the simulation process, including deadlock detection and resolution.

Step 1: Start of the simulation The simulation proceeds through a sequence of structured steps. In step 1, the simulation environment is initialized, and the number of jobs present is set to zero unless specified otherwise.

Step 2: Determine the deadlock detection schedule In step 2, deadlock detection is scheduled at intervals of Δ_t time steps.

Step 3: Simulate next time step At each time step, the simulation checks whether the current time exceeds the predefined end time or other termination condition. If not, it advances to the next time step in step 3. Suppose that the current time epoch is $t - dt$, at which the state is $(\mathbf{L}^s(t - dt), \mathbf{L}^b(t - dt))$. During

time interval $(t-dt, t)$, external arrivals and service completions occur, yielding the pre-action state at time t^-. In the pre-action state, jobs that complete service at node i and require service of node j are included in $L^b_{ij}(t^-)$ as potentially blocked jobs at time t. Hence, for $i, j = 1, \ldots, n$,

$$\begin{aligned} L^s_i(t^-) &= L^s_i(t-dt) - D_i(t-dt, t), \\ L^b_{ij}(t^-) &= L^b_{ij}(t-dt) + D_{ij}(t-dt, t), \\ L^b_{0i}(t^-) &= A_i(t-dt, t), \end{aligned}$$

where $A_i(t-dt, t)$ denotes the number of arrivals at node i, $D_i(t-dt, t)$ denotes the number of service completions at node i, and $D_{ij}(t-dt, t)$ denotes the number of service completions at node i requiring subsequent service at node j.

Here, $A_i(t-dt, t)$ follows a Poisson distribution with rate $\lambda_i dt$, whereas $D_i(t-dt, t)$ follows a Binomial$\big(L^s_i(t-dt), 1-e^{-\mu_i dt}\big)$ distribution. Finally, $(D_{i0}(t-dt, t), D_{i1}(t-dt, t), \ldots, D_{in}(t-dt, t))$ follows a Multinomial$(D_i(t-dt, t), r_{i0}, r_{i1}, \ldots, r_{in})$ distribution. The actual transfers are based on the pre-action state in the subsequent steps 4 or 6.

Step 4: Transfers jobs when feasible If there is no deadlock detection mechanism at time t (or if there is no deadlock detected), then jobs are accepted and transferred based on the available capacity at time t^-. Let $N_i(t^-)$ denote the total unmet demand for node i at time t^-. The remaining capacity and demand can then be expressed as

$$c^{\text{left}}_i(t^-) = c_i - L^s_i(t^-) - \sum_{j=1}^{n} L^b_{ij}(t^-), \qquad \text{and} \qquad N_i(t^-) = \sum_{j=0}^{n} L^b_{ji}(t^-).$$

In case $N_i(t^-) \le c^{\text{left}}_i(t^-)$, there will be no ASB to node i, such that the actual number of jobs that are transferred to node i is $F_{ji}(t) = L^b_{ji}(t^-)$, for all $j = 0, \ldots, n$. Otherwise, the $N_i(t^-)$ jobs randomly compete for the remaining $c^{\text{left}}_i(t^-)$ positions. More specifically, $(F_{0i}(t), F_{1i}(t), \ldots, F_{ni}(t))$ then follows a Multinomial $\big(N_i(t^-) - c^{\text{left}}_i(t^-), \tilde{p}_0, \tilde{p}_1, \ldots, \tilde{p}_n\big)$ distribution, with $\tilde{p}_j = L^b_{ji}(t^-)/N_i(t^-)$.

Step 5: Run deadlock detection If a deadlock detection event occurs at time t, the deadlock detection algorithm is applied to efficiently identify deadlocks (step 5). Detection relies on the system's pre-action state (at time t^-), from which a *wait-for graph* [44] is constructed. This directed graph represents which jobs are blocked by which nodes. Our approach generates a *slim* wait-for graph, incorporating only essential information for deadlock detection. In this graph, a directed edge from vertex i to vertex j indicates that at least one job at node i is blocked by node j, i.e., $L^b_{ij}(t^-) \ge 1$ and $N_j(t^-) > c^{\text{left}}_j(t^-)$. Figure 6.8 illustrates the construction of the wait-for graph. Initially, with no blockages, there are no edges. If a job completes service at node 1 and is blocked by node 2, an edge is added from vertex 1 to vertex 2. When another job completes service but is

blocked at node 1 by node 2, another edge is added from vertex 1 to vertex 2. The presence of a cycle in the graph signals a deadlock, which is detected through depth-first search.

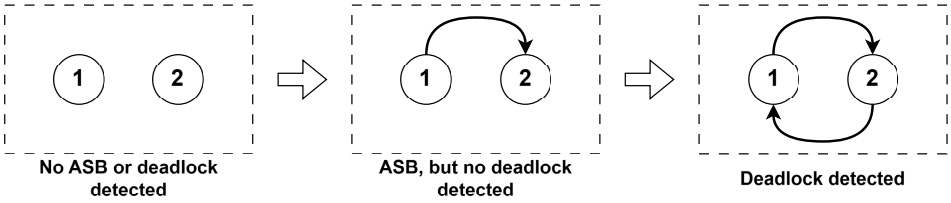


Figure 6.8: The construction of wait-for graph.

Step 6: Deadlock resolution If a deadlock is detected at time t, we apply a *deadlock resolution strategy* to mitigate its impact. We opt for a relatively simple, greedy approach due to the complexity of open queueing networks with ASB and the need for real-time decision-making. Greedy policies are known to perform well in complex stochastic systems, particularly when decisions occur on a fast timescale or when network dynamics are highly uncertain [6, 89, 121].

Our strategy maximizes the one-step throughput using an ILP, closely related to the linear program defined in Section 6.4.2, but adapted to integer decisions. Specifically, we determine the number of jobs $F_{ij}(t) \in \mathbb{Z}_{\geq 0}$ transferred from node i to node j at time t (with $i = 0$ representing external arrivals). The ILP is defined as follows:

$$\max \quad \sum_{i=0}^{n} \sum_{j=1}^{n} F_{ij}(t) \tag{6.11}$$

$$\text{s.t.} \quad F_{ij}(t) \leq L_{ij}^{b}(t^{-}) \qquad i = 0, \ldots, n,\ j = 1, \ldots, n, \tag{6.12}$$

$$L_i^{s}(t^{-}) + \sum_{j=0}^{n} F_{ji}(t) + \sum_{j=1}^{n} \left(L_{ij}^{b}(t^{-}) - F_{ij}(t) \right) \leq c_i \quad i = 1, \ldots, n, \tag{6.13}$$

$$F_{ij}(t) \in \mathbb{Z}_{\geq 0} \qquad i = 0, \ldots, n,\ j = 1, \ldots, n. \tag{6.14}$$

The objective (6.11) maximizes total job throughput, where the decision variables $F_{ij}(t)$ represent the actual flows between nodes. Constraint (6.12) ensures that transfers do not exceed the available (potentially blocked or external) number of jobs, while (6.13) guarantees that node capacities are not violated after the transition. The objective can also be rewritten as:

$$\sum_{i=0}^{n}\sum_{j=1}^{n} F_{ij}(t) = \underbrace{\sum_{j=1}^{n} F_{0j}(t)}_{\text{(I) external arrivals}} \underbrace{- \sum_{i=1}^{n}\sum_{j=1}^{n} \left(L_{ij}^{b}(t^{-}) - F_{ij}(t)\right)}_{\text{(II) remaining blocked jobs}} + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij}^{b}(t^{-})}_{\text{(III) 'constant'}}.$$

Since term (III) is independent of the decision at time t, the optimization balances maximizing external job admissions (I) and minimizing the number of remaining blocked jobs (II), with equal weights assigned to each. Hence, the objective aligns well with the key performance criteria of the system.

This ILP provides a foundational framework that can be adapted to specific scenarios. Below, we present several such adaptations. Given the method's versatility, a wide range of alternative constraints or objectives can be incorporated as needed.

- A natural extension is to consider a weighted objective function, $\sum_{i=0}^{n}\sum_{j=1}^{n} w_{ij}F_{ij}(t)$, where the weights w_{ij} reflect the relative importance of different nodes or express preferences between admitting external jobs (I) and unblocking internal jobs (II).
- The maximum number of jobs in a 'swap' can be limited to Δ_s by adding the constraint: $\sum_{i=1}^{n}\sum_{j=1}^{n} F_{i,j}(t) \leq \Delta_s$.
- Upper bounds (say Δ_{max}) can be imposed on transitions, such as: a maximum number of new arrivals at node i, $\sum_{j=1}^{n} F_{ji}(t) \leq \Delta_{max}$; a maximum number of departures from node i, $\sum_{j=1}^{n} F_{ij}(t) \leq \Delta_{max}$; or a bound on individual flows between nodes, $F_{ij}(t) \leq \Delta_{max}$.

Finally, the *post-action* states follow directly from the number of job transfers $F_{ij}(t)$, whether determined by the transfer rules without deadlock detection or through the ILP-based deadlock resolution. This yields the following post-action state updates:

$$L_i^s(t) = L_i^s(t^{-}) + \sum_{j=0}^{n} F_{ji}(t), \qquad \text{and} \qquad L_{ij}^b(t) = L_{ij}^b(t^{-}) - F_{ij}(t).$$

Step 7: End When the simulation ends, either due to the end time or another termination condition, results are collected to compute performance metrics. This is done by removing a specified warmup period and collecting the performance metrics based on batch-means.

6.6 Simulation Experiments

In this section, we consider the simulation results of the deadlock detection and resolution mechanisms, as described in Section 6.5. The two-node and four-node systems are addressed in Sections 6.6.1 and 6.6.2, respectively. The simulated period is taken so that the confidence intervals are smaller than 5% of the corresponding mean.

6.6.1 Systems consisting of two nodes

For the two-node system, we use the same system setup as the SD model described in Section 6.4.2; that is, we choose parameters $c_1 = c_2 = 10, \mu_1 = \mu_2 = 0.5, r_{12} = r_{21} = 0.4, \lambda_1 = 4, \lambda_2 = 1.5$, and a time step of $dt = 1$. To evaluate the effect of the deadlock resolution frequency, we tested various Δ_t values: $\infty, 80, 40, 20, 10, 5, 2$, and 1. Figure 6.9 shows the blocking probability and expected number of blocked servers at node 1. Without resolution of deadlocks (i.e., $\Delta_t = \infty$), the system remains in a permanent deadlock with all servers blocked. The detection and resolution of deadlocks scheduled each 80 time steps significantly reduces both the probability of blocking and the number of blocked servers compared to no deadlock resolution. Increasing the frequency of deadlock resolution further enhances performance, significantly reducing both the blocking probability and the number of blocked servers. Specifically, reducing the resolution frequency from $\Delta_t = 5$ to $\Delta_t = 1$ yields significant improvements. In continuous-time systems, raising the frequency of deadlock resolution yields diminishing improvements in blocking probability. In contrast, in discrete-time systems, frequent resolution of deadlocks provides more significant benefits compared to continuous-time scenarios. This is intuitively clear; in continuous-time models, service completions can occur at any moment, potentially resolving deadlocks immediately. In contrast, Section 6.4.2 showed that discrete-time models require additional capacity to maintain a smooth flow in the absence of deadlock resolution. During periods of high congestion, frequent resolution opportunities are needed to prevent blocking and deadlocks.

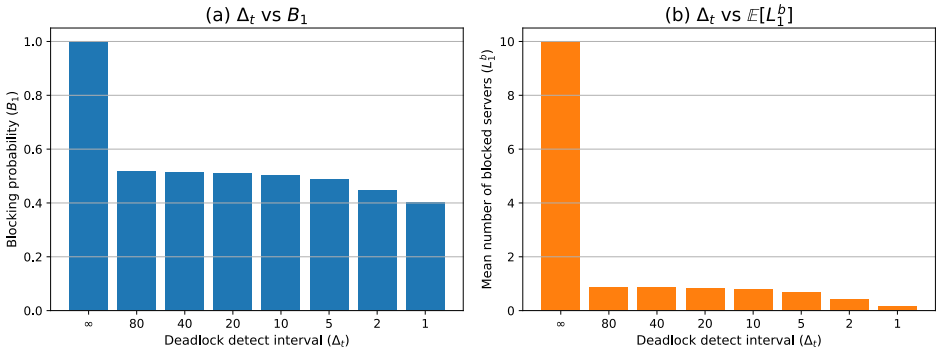


Figure 6.9: The impact of Δ_t on B_1 and $\mathbb{E}[L_1^b]$ in the simulation of a two-node system.

6.6.2 System consisting of four nodes

To examine the impact of deadlocks in a more complex setting, we now analyze a four-node network. We begin with a symmetric configuration, using the following parameters: $\lambda_i = \frac{c}{16}$, $\mu_i = 0.25$, and $c_i = c \in \{4, 10, 20\}$ for all $i = 1, 2, 3, 4$, with routing probabilities $r_{ij} = 0.25$ for all $i \neq j$. We explore various values for Δ_t and compare two time discretizations, $dt = 1$ and $dt = 4$, to highlight the effect of coarser time steps. As larger time steps lead to more simultaneous

arrivals and transfers competing for capacity, we expect deadlock effects to be more pronounced. The results are presented in Figure 6.10.

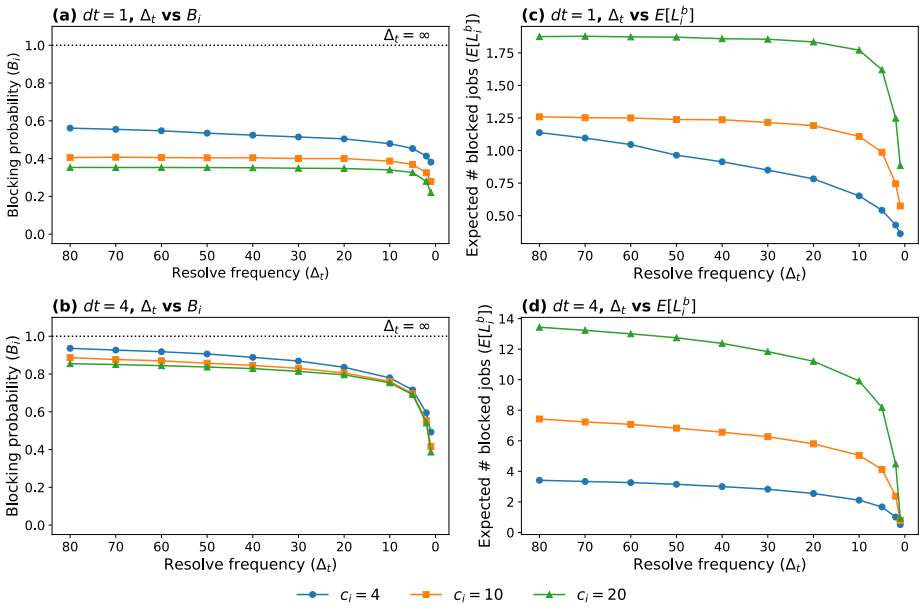


Figure 6.10: Impact of Δ_t on B_i and L_i^b in a symmetric four-node system for time discretizations $dt = 1$ (top) and $dt = 4$ (bottom).

Subfigures (a) and (b) in Figure 6.10 show how the frequency of deadlock resolution affects the blocking probability for two time-discretizations: $dt = 1$ and $dt = 4$. For both cases, we see that frequent deadlock resolution options greatly improve system performance. As in the two-node system, reducing Δ_t from 5 to 2 or 1 leads to a substantial drop in blocking probability. This drop is even more pronounced for coarser time discretization ($dt = 4$). In case $\Delta_t = 1$, the blocking probabilities for $dt = 1$ and $dt = 4$ are somewhat comparable. However, as Δ_t increases (for example, $\Delta_t \geq 10$), the blocking probability reaches values greater than 0.8 when $dt = 4$. This illustrates that coarser discretizations are more susceptible to deadlock build-up, necessitating more frequent resolution to sustain performance.

Subfigures (c) and (d) in Figure 6.10 present similar trends for the expected number of blocked servers. However, note the different scales on the vertical axis of the two panels: with larger time steps ($dt = 4$), the system is more prone to simultaneous blocking events, resulting in a higher average number of blocked servers. This further highlights that deadlock resolution becomes increasingly critical under coarser time discretization. Moreover, across all subplots, a higher number of servers consistently mitigates deadlock effects, reducing both blocking probability and the number of blocked servers, regardless of resolution frequency or time step

size.

We now proceed by considering an asymmetric four-node system. Similarly to Section 6.A.2, we consider three distinct scenarios. The first one is for reference and corresponds to the symmetric case described above with $c_i = 4$. The second is the unbalanced load scenario in which $\lambda_i = 0.25$ for $i \in \{1, 2, 3, 4\}$ and $c_i = 6$ for $i \in \{1, 2\}$ while $c_i = 3$ for $i \in \{3, 4\}$, leading to underutilization of the first two nodes and the latter two being heavily loaded. The third scenario is the unbalanced capacity scenario, where $\lambda_i = 0, c_i = 4$ for $i \in \{1, 2\}$ and $\lambda_i = 0.625, c_i = 6$ for $i \in \{3, 4\}$, creating equal load per server across all nodes but granting higher capacity to nodes 3 and 4. The simulation results are shown in Figure 6.11.

The results confirm the previous finding concerning the impact of the deadlock resolution frequency and the time discretization. Interestingly, the blocking probabilities for the asymmetric system with unbalanced capacity are very close to those of the symmetric case. In case of unbalanced loads, the blocking probabilities of the heavily loaded nodes are evidently significantly higher. In fact, the higher B_i cannot be fully compensated for by the blocking probabilities of the lightly loaded nodes, leading to a worse aggregate performance compared to a balanced scenario. This underscores the value of evenly distributed load among nodes to mitigate deadlocks. However, the main conclusion remains that frequent deadlock resolution is essential, especially in systems with coarse time discretization, which contrasts with the behavior observed in continuous-time models.

6.7 Case Study: Elderly Healthcare System

To evaluate the impact of our proposed method for resolving deadlocks, we apply it to a case study of Amsterdam's elderly healthcare system in the Netherlands, consisting of 17 distinct healthcare forms depicted in Figure 6.12. Acronyms are provided in Table 6.1. All nodes in Figure 6.12, except HC, represent bed-based care. Short-Term Residential Care (STC) and Short-Term Residential Palliative Care (STC Pall) both provide temporary residential care, typically post-hospitalization. Palliative care is for patients with limited life expectancy. Geriatric Rehabilitation (GR) offers inpatient rehabilitation focused on functional recovery following hospital discharge. Home Care (HC) entails professional nursing and support at the patient's residence, varying in form and intensity. Long-Term Care (LTC) is intended for patients who are expected to require permanent care, divided into four categories: Somatic (SOM), Psychogeriatric (PG), LTC at Home, and Rehabilitation. Additionally, we include the eight primary hospital departments for older adults: cardiology (Card), surgery (Surg), neurology (Neur), urology (Uro), orthopedics (Ortho), internal medicine (Int Med), lung diseases (Lung), and gastroenterology (Gastro). All other departments are aggregated into one node.

As long as patients are in the system, they need care. This implies that ASB

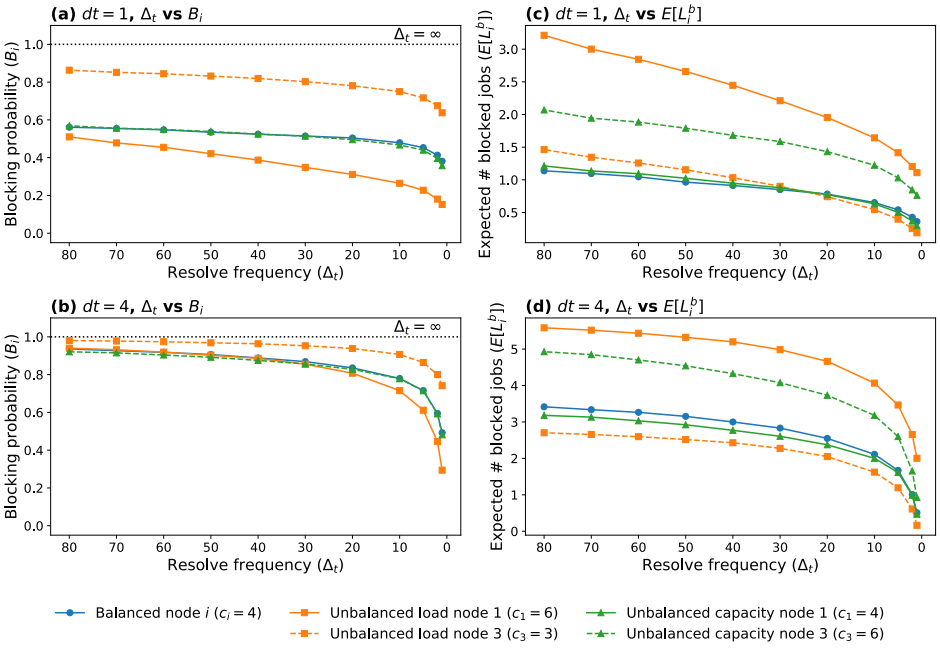


Figure 6.11: Impact of Δ_t on B_i and L_i^b in a four-node asymmetric system for time discretizations $dt = 1$ (top) and $dt = 4$ (bottom).

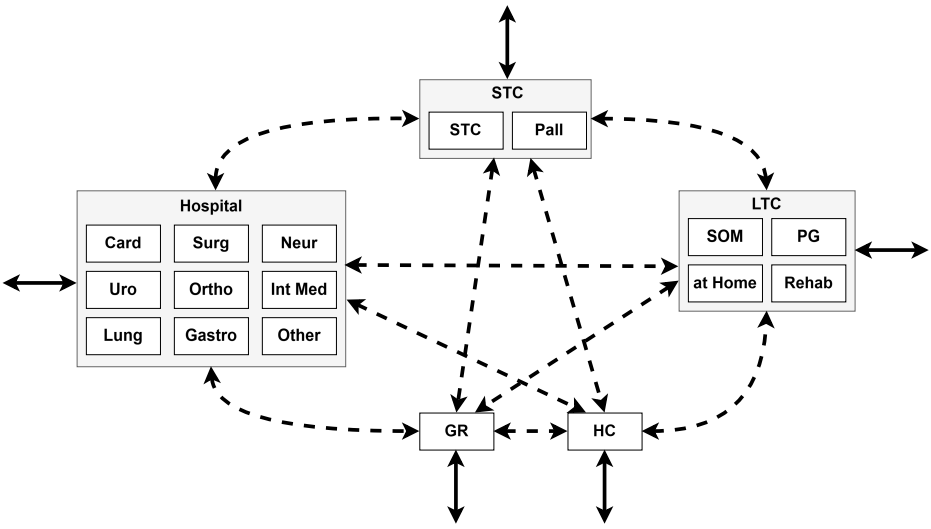


Figure 6.12: The system representing the different healthcare forms of the elderly care system in Amsterdam, the Netherlands.

occurs when a patient must transfer to another form of care but no capacity is available at the destination. In turn, this provides the possibility of deadlocks, as the system contains feedback loops. Providing an accurate model of the complete elderly care system is challenging, as multiple organizations can provide similar type of care, and patient flows do not strictly adhere to geographical boundaries. However, older adults typically prefer to receive care close to home. In the model, each type of care is represented by a single node that provides that service within the corresponding neighborhood or city district. We do not account for patients who relocate to areas outside the simulated system, and any resulting ASB for those patients is not considered. Consequently, the actual impact of deadlocks and ASB may even be more significant than those suggested by the simulation.

System parameters Most of the system parameters are derived from non-public data from Statistics Netherlands (CBS) [139]. The data used involve healthcare declarations from 2017 to 2019, from older adults 65 and older, in Amsterdam; the data set is based on Chapter 2. A single patient may have multiple journeys through the system. In our data analysis, a patient is considered new if no care services have been used for a consecutive period of four weeks. Departures from the system occur when patients do not need care for at least four weeks, or have died. Note that older adults may simultaneously utilize multiple types of care; in these instances, only the most intensive form is considered.

Abbreviation	Healthcare Form	λ_i	γ_i	μ_i	$c_i^{\#}$
GR	Geriatric Rehabilitation	0.0707	0.7543	0.0274	28
HC	Home Care	0.1055	0.1988	0.0062	33
Short-Term (Residential) Care					
STC	-	0.2204	0.4571	0.0269	17
STC Pall	Palliative	0.024	0.0671	0.0323	3
Long-Term Care					
LTC SOM	Somatic	0.0060	0.0398	0.0028	15
LTC PG	Psychogeriatric	0.0174	0.071	0.0014	51
LTC at Home	at Home	0.0447	0.0924	0.0028	33
LTC Rehab	Rehabilitation	0.0003	0.0057	0.0137	1
Hospital Departments					
Card	Cardiology	4.3317	4.7967	0.1608	30
Surg	Surgery	6.9852	7.73	0.1595	49
Neur	Neurology	2.8884	3.1133	0.1718	19
Uro	Urology	2.1451	2.3967	0.25	10
Ortho	Orthopedics	2.2051	2.32	0.2004	12
Int Med	Internal Medicine	5.5406	7.0633	0.1344	53
Lung	Lung Diseases	2.5991	3.17	0.1464	22
Gastro	Gastroenterology	1.5360	1.8267	0.1733	11
Other	Other Departments	10.9911	12.3333	0.1866	67

Table 6.1: Healthcare forms with abbreviations, external arrival rates (λ_i), *effective* arrival rates γ_i, service rates (μ_i), and required capacities ($c_i^{\#}$).

The arrival and service rates are given in Table 6.1, while the transition probabilities can be found in Section 6.B.3; these values are derived from the analysis of the CBS data sets described above. The parameters indicate that external arrivals occur most frequently in hospital departments, as expected, since patients typically require only a single hospital visit before returning home without further care. These departments also show a significantly shorter average length of stay compared to other locations. Most transition probabilities are small. Nonetheless, deadlocks can still significantly impact system performance; for example, a patient awaiting transfer from the hospital to an LTC bed may experience prolonged blocking due to the long length of stay in nursing homes. The aggregate arrival rate γ_i, calculated from the external arrival rates and transition probabilities, is also presented in Table 6.1. Information about capacities is lacking. Therefore, we consider three scenarios representing varying levels of pressure on the healthcare system. Define $c_i^{\#} = \lceil \gamma_i/\mu_i \rceil$ as the minimum number of servers required to handle the offered load at node i, rounded up to the nearest integer. The actual capacities used in the simulation are scaled versions: $c_i = c_i^{\#} \cdot l$, for $l \in \{1.0, 0.9, 0.8\}$.

Simulation results Due to the considerable variation in lengths of stay and the large scale of the healthcare system, we simulate over a ten-year time horizon and repeat this for 1000 runs. To reflect the sustained capacity pressure seen

in practice, the simulation is initialized with all resources fully occupied. Again, we focus on two main performance metrics: the number of rejected patients over a ten-year horizon ($\approx 10 \cdot 365 \cdot \sum_{i=1}^{n} \lambda_i B_i$) and the number of wrong-bed days ($\approx 10 \cdot 365 \cdot \sum_{i=1}^{n} \mathbb{E}[L_i^b]$), reflecting situations where a server remains occupied by a patient who no longer needs its services.

A comprehensive overview of the simulation results is presented in Figure 6.13; the results for individual nodes in the system are provided in Section 6.B.4. In all three examined capacity configurations, implementing deadlock resolution improves system performance, reducing both the number of rejected patients and wrong-bed days even without increasing total capacity. Observe that the simulation timespan of ten years is typically too short for the system to reach a permanent deadlock state. However, we see improvements when the deadlock resolution option is provided regularly. Even fairly infrequent resolution options, such as every five time steps ($\Delta_t = 5$), contribute to performance gains. These benefits become more pronounced as deadlock resolution is applied more frequently, particularly at intervals of $\Delta_t = 2$ or $\Delta_t = 1$.

Compared to a system without deadlock resolution ($\Delta_t = \infty$), applying resolution at every time step ($\Delta_t = 1$) yields substantial performance improvements. Over a ten-year simulation period, the number of rejected patients reduces from 23,607 to 21,111 at $l = 1$ (a reduction of 10.57%), and from 43,629 to 40,885 at $l = 0.8$ (a reduction of 6.29%). The reduction of wrong-bed days is even more pronounced, reducing from 12,785 to 11,785 at $l = 1$ (a reduction of 7.83%) and from 43,629 to 40,885 at $l = 0.8$ (a reduction of 20.33%). At higher capacity levels ($l = 1.0$), the main benefit of deadlock resolution lies in reducing patient rejections, indicating improved utilization of available capacity. In more constrained settings ($l = 0.8$), the primary gains are observed in the reduction of wrong-bed days, suggesting that deadlock resolution helps maintain access by not only admitting more patients but also more effectively transferring them to appropriate levels of care. This shows that solving deadlocks can help maintain access and flow even during periods of high system pressure.

Surprisingly, when $l = 1.0$, $\Delta_t = 2$ results in fewer wrong-bed days than $\Delta_t = 1$, although the latter achieves a greater reduction in patient rejections. This suggests that the coordination mechanism can sometimes prolong the occupancy of resources by patients who no longer need that service, increasing wrong-bed days but reducing rejections. Although the overall blocking probability decreases with more frequent deadlock resolution, blocking at individual nodes may vary; reduced blocking at one node can lead to higher external inflow, causing congestion elsewhere. These findings highlight the complex effects of coordination and the need for a system-wide perspective when designing interventions. In any case, deadlock resolution is especially beneficial in highly loaded systems, where it serves as a crucial buffer against systemic collapse.

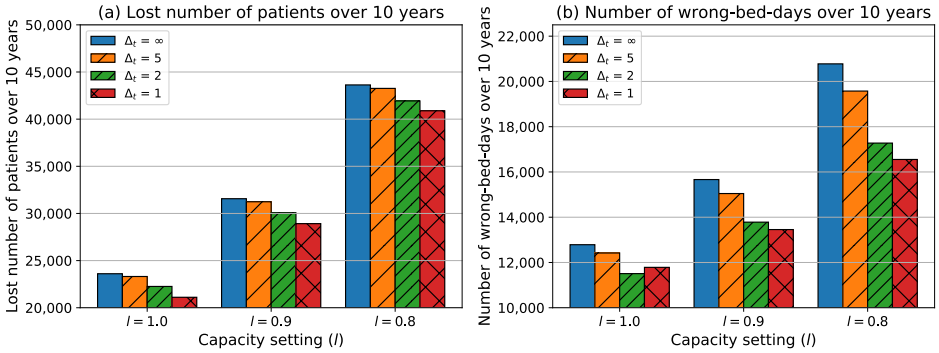


Figure 6.13: The impact of deadlock resolution on the number of rejected patients and wrong-bed days for the case study of older adults in Amsterdam.

6

6.8 Conclusion and Discussion

This study investigates deadlock resolution in queueing networks with ASB and feedback loops, where resolving deadlocks is essential to escape from absorbing states. We illustrate that even in deterministic SD models, permanent deadlocks may occur. As permanent deadlocks are typically not observed in practice, there is an evident need to take deadlock resolve into account when modeling ASB systems with feedback. For continuous-time models, we demonstrate that systems with few servers are significantly more vulnerable to deadlocks than larger systems. In fact, in larger systems, even a low resolution rate, implying infrequent opportunities to resolve deadlocks, is sufficient to mitigate their impact. For a single node with feedback, we derived closed-form expressions for the blocking probability and showed that, with any positive resolution rate, the system approaches the Erlang loss model as the number of servers grows.

In contrast, discrete-time models are more sensitive to the frequency of deadlock resolution opportunities. We propose an approach that detects deadlocks using a wait-for graph and resolves them through a novel coordination mechanism based on ILP. This approach is intuitive, flexible, and computationally fast. Simulations of discrete-time systems revealed that performance is highly sensitive to the resolution frequency. Unlike continuous-time models, resolving deadlocks at every time step significantly reduced the blocking probability, especially in systems with few servers and coarse time steps. This underscores how system granularity influences the effectiveness of deadlock mitigation.

We applied our method to a case study of Amsterdam's elderly care system, modeled as a 17-node network representing different types of care. Despite the relatively small transition probabilities between care forms, the proposed coordination mechanism, when implemented daily, led to a notable decrease in the number of rejections by 6% to 10% and a reduction in wrong-bed days by 7% to 20%.

These simulation results emphasize the practical relevance of regional coordination mechanisms, particularly in aging societies, where increasing demand should be delivered with limited care resources. In addition, we envision that automated regional cooperation allows care workers and organizations to save valuable time on administrative placement tasks.

Several directions remain open for future work. Although the case study on elderly care shows great potential, implementing regional coordination in practice remains a significant challenge, including data sharing and aligning incentives. Depending on practical limitations, our model can be adapted to incorporate elements such as coordination delays, cooling periods after blocking, and decentralized decision making. In addition, the applicability of the model to large-scale systems must be tested. Furthermore, while our ILP-based mechanism is flexible and effective, it is essentially based on a greedy approach. Future work could explore deadlock resolution policies that account for predicted future behavior and develop dynamic strategies that balance performance, fairness, and coordination cost.

6.A Appendix: Supplementary Details for Markov Chain Models

Section 6.A.1 covers the technical proofs for the single-node model. Section 6.A.2 describes the Markov chain formulation and computational insights for the two-node system.

6.A.1 Proofs for single-node system

Proof of Corollary 6.3. For convenience, define $\alpha = \nu/\mu_1$. Using the Gamma function $\Gamma(\cdot)$, we may write

$$\tilde{a}_k = r_{11}^k \cdot \frac{\Gamma(c+1)}{\Gamma(c+\alpha)} \cdot \frac{\Gamma(c-k+\alpha)}{\Gamma(c-k+1)}.$$

Using the asymptotic approximation for the Gamma function ratio [215]

$$\frac{\Gamma(x+\alpha)}{\Gamma(x+1)} \sim x^{\alpha-1},$$

with $\sim$ meaning that their ratio tends to 1, we obtain

$$\sum_{k=0}^{\epsilon c} \tilde{a}_k \;\sim\; \sum_{k=0}^{\epsilon c} r_{11}^k \left(1 - \frac{k}{c}\right)^{\alpha-1},$$

for some $\epsilon \in (0,1)$. The tail $\sum_{k=\epsilon c}^{c} \tilde{a}_k \to 0$ as $r_{11}^k \leq r_{11}^{\epsilon c}$ decays exponentially fast as $c \to \infty$, whereas $\Gamma(c-k+\alpha)/\Gamma(c-k+1) = \mathcal{O}(c^{\alpha-1})$. The expressions in the sum above are dominated by the geometric sequence r_{11}^k, with the term in brackets converging to 1. Hence, from dominated convergence, it follows that $\sum_{k=0}^{c} \tilde{a}_k \to 1/(1-r_{11})$ as $c \to \infty$. Define $A(c) = (1-r_{11}) \sum_{k=0}^{c} \tilde{a}_k$ and $B(c) = c\mu_1/(c\mu_1 - \nu \sum_{i=1}^{c} \tilde{a}_j)$. Observe that $A(c) \to 1$ and $B(c) \to 1$ as $c \to \infty$. Moreover, we can write $B_1 = A(c) \cdot B(c) \cdot a^c/c! \cdot \left[\sum_{i=0}^{c}(a^i/i!) + (a^c/c!) \cdot (A(c)B(c)-1)\right]^{-1}$. Recalling that

$$B(a,c) = \frac{a^c/c!}{\sum_{i=0}^{c} a^i/i!},$$

we obtain by rearranging terms

$$\frac{B_1}{B(a,c)} = \frac{A(c) \cdot B(c) \cdot \sum_{i=0}^{c} a^i/i!}{\sum_{i=0}^{c} a^i/i! + \frac{a^c}{c!}(A(c)B(c)-1)}.$$

The result now follows by letting $c \to \infty$ and using that $A(c) \to 1$ and $B(c) \to 1$ as $c \to \infty$. □

Proof of Lemma 6.4. Let $\tilde{m}_i = \mathbb{E}[T_{i+1} \mid X(0) = i]$. Using standard results (see e.g. [184, Section 6.3]), we have, for $i = 0, 1, \ldots, c-1$,

$$\tilde{m}_i = \frac{1}{\lambda} \sum_{k=0}^{i} \frac{i!}{(i-k)!} \left(\frac{\mu_1(1-r_{11})}{\lambda_1}\right)^k = \frac{1}{\lambda} \sum_{k=0}^{i} \frac{i!}{(i-k)!} a^{-k}.$$

Now, for $i = 1, \ldots, c$, we have

$$m_{2c-i} = \frac{1}{i\mu_1} + (1 - r_{11})m_{c-1} + r_{11}m_{2c-i+1}.$$

Iterating the equation above, we obtain

$$m_{2c-i} = \sum_{k=0}^{i-1} \frac{r_{11}^k}{(i-k)\mu_1} + (1 - r_{11}^i)m_{c-1}.$$

Applying the above for $i = c$ and using that $m_{c-1} = \tilde{m}_{c-1} + m_c$, we find

$$m_c = \frac{1}{r_{11}^c} \left(\frac{1}{\mu_1} \sum_{k=0}^{c-1} \frac{r_{11}^k}{c-k} + (1 - r_{11}^c)\tilde{m}_{c-1} \right).$$

As $m_0 = \sum_{i=0}^{c-1} \tilde{m}_i + m_c$, we obtain the result after some elementary rewriting. □

6.A.2 Two interacting nodes

Here, we elaborate on the Markov chain $\{Y(t), t > 0\}$ of a two-node system, see the right panel of Figure 6.3. The state space is given by the set of all pairs $(i, j) \in \mathbb{N}_0^2$ such that $0 \leq i, j \leq c_1 + c_2$, $i - c_1 \leq \min(j, c_2)$, and $j - c_2 \leq \min(i, c_1)$. In case $i < c_1$, there are i servers at node 1 currently serving; similarly if $j < c_2$ there are j active servers at node 2. If $i > c_1$, all servers at node 1 are occupied, and there $i - c_1$ jobs at node 2 being blocked from entering service at node 1, implying that $i - c_1 \leq \min(j, c_2)$. Let $\ell_k^s(i, j)$ denote the number of jobs in service at node k, $k = 1, 2$, as a function of state (i, j):

$$\ell_1^s(i,j) = \min(i, c_1) - \max(j - c_2, 0), \qquad \ell_2^s(i,j) = \min(j, c_2) - \max(i - c_1, 0).$$

Define $\phi_1(i) = \min(i - c_1, 0)$ and $\phi_2(j) = \min(j - c_2, 0)$. The transitions rates are as follows:

$$q_{(i_1,j_1),(i_2,j_2)} = \begin{cases} \lambda_1 & \text{if } i_2 = i_1 + 1,\ j_2 = j_1,\ i_1 < c_1, \\ \lambda_2 & \text{if } i_2 = i_1,\ j_2 = j_1 + 1,\ j_1 < c_2, \\ \ell_1^s(i_1,j_1)\mu_1(1-r_{12}) & \text{if } i_2 = i_1 - 1,\ j_2 = j_1,\ i \le c_1, \\ \ell_1^s(i_1,j_1)\mu_1(1-r_{12}) & \text{if } i_2 = i_1 - 1,\ j_2 = j_1 - 1, i > c_1, \\ & \quad j_1 \le c_2, \\ \ell_1^s(i_1,j_1)\mu_1(1-r_{12}) & \text{if } i_2 = i_1 - \min(\phi_1(i_1)+1, \phi_2(j_1)+1), \\ & \quad j_2 = j_1 - \min(\phi_1(i_1), \phi_2(j_1)+1), \\ & \quad i_1 > c_1,\ j_1 > c_2, \\ \ell_2^s(i_1,j_1)\mu_2(1-r_{21}) & \text{if } i_2 = i_1,\ j_2 = j_1 - 1,\ j_1 \le c_2, \\ \ell_2^s(i_1,j_1)\mu_2(1-r_{21}) & \text{if } i_2 = i_1 - 1,\ j_2 = j_1 - 1, i \le c_1, \\ & \quad j_1 > c_2, \\ \ell_2^s(i_1,j_1)\mu_2(1-r_{21}) & \text{if } i_2 = i_1 - \min(\phi_1(i_1)+1, \phi_2(j_1)), \\ & \quad j_2 = j_1 - \min(\phi_1(i_1)+1, \phi_2(j_1)+1), \\ & \quad i_1 > c_1,\ j_1 > c_2, \\ \ell_1^s(i_1,j_1)\mu_1 r_{12} & \text{if } i_2 = i_1 - 1,\ j_2 = j_1 + 1,\ i_1 \le c_1, \\ & \quad j_1 < c_2, \\ \ell_1^s(i_1,j_1)\mu_1 r_{12} & \text{if } i_2 = i_1 - 1,\ j_2 = j_1,\ i_1 > c_1,\ j_1 < c_2, \\ \ell_1^s(i_1,j_1)\mu_1 r_{12} & \text{if } i_2 = i_1,\ j_2 = j_1 + 1,\ j_1 \ge c_2, \\ \ell_2^s(i_1,j_1)\mu_2 r_{21} & \text{if } i_2 = i_1 + 1,\ j_2 = j_1 - 1,\ i_1 < c_1, \\ & \quad j_1 \le c_2, \\ \ell_2^s(i_1,j_1)\mu_2 r_{21} & \text{if } i_2 = i_1,\ j_2 = j_1 - 1,\ i_1 < c_1,\ j_1 > c_2, \\ \ell_2^s(i_1,j_1)\mu_2 r_{21} & \text{if } i_2 = i_1 + 1,\ j_2 = j_1,\ i_1 \ge c_1, \\ \nu & \text{if } i_2 = i_1 - \min(\phi_1(i_1), \phi_2(j_1)), \\ & \quad j_2 = j_1 - \min(\phi_1(i_1), \phi_2(j_1)), \\ & \quad i_1 > c_1,\ j_1 > c_2, \\ 0 & \text{otherwise.} \end{cases} \tag{6.15}$$

Numerical results In Figure 6.A.1, we demonstrate the impact of various parameters on performance metrics B_i and $\mathbb{E}[L_i^b]$, similar to the single-node system in Section 6.4.1. We use three scenarios: (i) a balanced symmetric scenario, (ii) an unbalanced load scenario, in which the load $(\gamma_i/(c_i\mu_i))$ of node 1 is smaller than the load of node 2, and (iii) an unbalanced capacity scenario in which nodes 1 and 2 have identical loads, but node 2 has twice as many servers. The parameters can be found in Table 6.A.1. Observe that in scenario (iii), when node i does not receive any external arrivals, consequently, there can be no external arrivals denied entry into the system at that node. Hence, B_i is omitted for such nodes.

Scenario	λ	λ_1	λ_2	μ_1	μ_2	c_1	c_2	r_{12}	r_{21}	ν
Balanced	4	0.5λ	0.5λ	1	1	4	4	0.5	0.5	0.025
Unbalanced load	4	0.5λ	0.5λ	1	1	6	3	0.5	0.5	0.025
Unbalanced capacity	6	0	λ	1	1	4	8	0.5	0.5	0.025

Table 6.A.1: Parameter settings for each case in a two-node system.

Panel (a) shows that relatively small values of the deadlock resolution rate ν suffice to avoid the impact of deadlocks on B_1 (and B_2, not shown here due to symmetry), akin to the situation in a single-node network. In the unbalanced load scenario, $B_2 > B_1$, which follows directly from the fact that node 1 has a smaller load. The scenario of unbalanced capacity yields a somewhat smaller blocking probability than the balanced scenario, due to increased scale of node 2. In panel (b), the expected number of blocked jobs $\mathbb{E}[L_i^b]$ initially increases with μ_i, and subsequently decreases, corresponding to the 1-node situation. The node with the lowest load typically exhibits a higher expected number of blocked jobs, since these jobs are blocked from entering the other node. Nodes with higher capacity c_i can accommodate more blocked jobs overall. In panels (c) and (d), the probability of blocking and the expected number of blocked jobs increase again with the overall arrival rate λ. In the scenario of unbalanced capacity, the expected number of blocked jobs at node 1, $\mathbb{E}[L_1^b]$, remains relatively small. Since node 1 receives no external arrivals and only gets jobs from node 2, blocking at node 2 prevents node 1 from overloading - even at high arrival rates.

6.B Appendix: Supplementary Details for the Elderly Care Case Study

Section 6.B.3 describes the transition probabilities utilized in the case study, whereas Section 6.B.4 presents more detailed results from the case study.

6.B.3 Transition probabilities of elderly care case study

Here, we specify the transition probabilities of the elderly care case study described in Section 6.7.

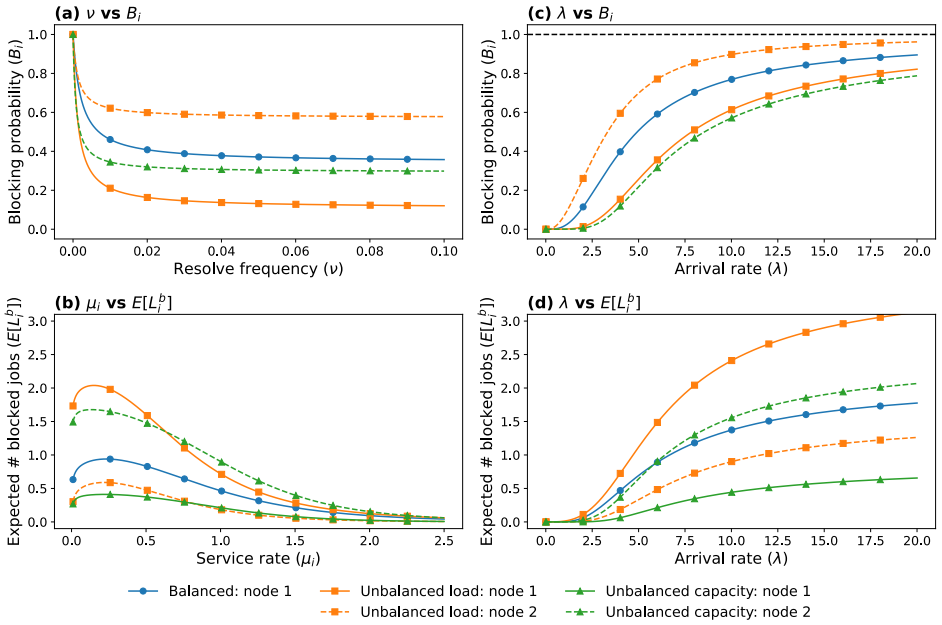


Figure 6.A.1: The effect of ν, μ, and λ on B_i and $\mathbb{E}[L_i^b]$ in a two-node system. In panel (c), the dashed line indicates the value of B_i as $\lambda \to \infty$.

$r_{ij}(\times 10^{-4})$	GR	STC	STC Pall	HC	LTC SOM	LTC PG	LTC at Home	LTC Rehab
GR	69	235	58	123.5	76.7	33.3	44.1	0
STC	206	116	165	108.1	125.1	162	88.6	0
STC Pall	0	0	0	12.1	36	0	0	0
HC	49	250	33	8	7	44	384	0
LTC SOM	0	0	0	17.8	0	411	297	71
LTC PG	0	0	0	13.5	54	91	482	75
LTC at Home	30	0	0	44.8	595	2339	284	30
LTC Rehab	0	0	0	0	1793	1873	1833	0
Card	96.4	35.6	6	25.9	3.4	2.7	10.4	0.4
Surg	248.6	55.7	3.9	15.7	2.6	5.9	4.6	2.4
Neur	289.3	51.4	6	13.3	4.6	4.7	6.3	2.1
Uro	24.4	16.3	5.6	19.9	3.7	2.8	4.1	0
Ortho	609	86.1	0	18.8	2.3	6.6	3.4	2.5
Int Med	141	91.7	15.9	28.8	7.4	6.9	9	0.7
Lung	149.6	44.6	17.6	30.1	5.4	2.6	8.5	0.7
Gastro	58.7	38.6	14.1	22.7	4.2	3	6.5	0
Other	27.9	18.4	1.7	5.1	0.7	1.7	1.6	0.2

Table 6.B.1: R matrix (Part 1). All values are scaled by 10^{-4}.

$r_{ij}(\times 10^{-4})$	**Card**	**Surg**	**Neur**	**Uro**	**Ortho**	**Int Med**	**Lung**	**Gastro**	**Other**
GR	185	360	142	43	137	486	178	103	146
STC	116	310	75	52	148	516	162	87	128
STC Pall	0	0	0	0	0	194	0	194	0
HC	479	533	225	185	137	914	390	155	244
LTC SOM	375	517	312	276	127	1537	482	227	184
LTC PG	273	792	187	102	241	744	150	96	262
LTC at Home	622	670	386	120	141	897	419	132	260
LTC Rehab	0	398	398	0	0	398	398	0	0
Card	497	49	36	15	7	104	57	29	143
Surg	31	532	17	18	15	111	30	39	68
Neur	30	72	206	8	8	91	27	10	176
Uro	36	62	13	754	13	125	13	14	22
Ortho	19	44	13	0	244	45	17	0	13
Int Med	70	126	36	32	11	1493	58	93	100
Lung	80	61	29	10	13	108	1259	18	67
Gastro	45	267	23	17	0	326	38	773	59
Other	48	32	49	5	2	57	18	7	834

Table 6.B.2: R matrix (Part 2). All values are scaled by 10^{-4}.

6.B.4 Results of elderly care case study

More detailed results concerning the elderly care case study are presented in this section.

Node i	B_i				$\mathbb{E}[L_i^b]$			
	$\Delta_t = \infty$	$\Delta_t = 5$	$\Delta_t = 2$	$\Delta_t = 1$	$\Delta_t = \infty$	$\Delta_t = 5$	$\Delta_t = 2$	$\Delta_t = 1$
GR	0.174	0.179	0.182	0.19	0.184	0.186	0.196	0.228
STC	0.217	0.217	0.219	0.219	0.199	0.197	0.213	0.246
STC Pall	0.263	0.26	0.263	0.264	0.005	0.004	0.004	0.005
HC	0.117	0.117	0.117	0.119	0.05	0.057	0.066	0.091
LTC SOM	0.156	0.152	0.16	0.164	0.024	0.024	0.026	0.028
LTC PG	0.075	0.08	0.08	0.081	0.046	0.042	0.046	0.057
LTC at Home	0.131	0.134	0.135	0.136	0.211	0.223	0.247	0.309
LTC Rehab	0.168	0.16	0.166	0.167	0.024	0.021	0.019	0.02
Card	0.167	0.164	0.159	0.153	0.239	0.219	0.203	0.21
Surg	0.144	0.141	0.136	0.129	0.538	0.516	0.493	0.491
Neur	0.174	0.172	0.168	0.162	0.218	0.216	0.217	0.245
Uro	0.242	0.236	0.229	0.221	0.126	0.099	0.072	0.058
Ortho	0.222	0.218	0.214	0.209	0.214	0.208	0.207	0.219
Int Med	0.146	0.141	0.133	0.121	0.668	0.637	0.588	0.536
Lung	0.187	0.181	0.174	0.165	0.281	0.247	0.21	0.195
Gastro	0.227	0.222	0.215	0.207	0.134	0.111	0.098	0.105
Other	0.136	0.133	0.127	0.12	0.341	0.305	0.247	0.185

Table 6.B.3: Expected values of B_i and L_i^b for different values of Δ_t, with $l = 1$.

Node i	B_i				$\mathbb{E}[L_i^b]$			
	$\Delta_t = \infty$	$\Delta_t = 5$	$\Delta_t = 2$	$\Delta_t = 1$	$\Delta_t = \infty$	$\Delta_t = 5$	$\Delta_t = 2$	$\Delta_t = 1$
GR	0.205	0.211	0.218	0.228	0.225	0.211	0.236	0.269
STC	0.242	0.243	0.246	0.248	0.234	0.225	0.265	0.297
STC Pall	0.242	0.238	0.249	0.243	0.005	0.005	0.006	0.006
HC	0.167	0.167	0.172	0.173	0.069	0.073	0.088	0.118
LTC SOM	0.179	0.168	0.187	0.181	0.029	0.03	0.031	0.037
LTC PG	0.117	0.12	0.123	0.124	0.051	0.044	0.048	0.064
LTC at Home	0.184	0.183	0.188	0.192	0.275	0.287	0.347	0.397
LTC Rehab	0.156	0.149	0.155	0.151	0.028	0.024	0.025	0.025
Card	0.211	0.208	0.203	0.197	0.294	0.268	0.246	0.243
Surg	0.202	0.199	0.194	0.186	0.637	0.602	0.572	0.531
Neur	0.207	0.204	0.2	0.195	0.266	0.26	0.268	0.3
Uro	0.241	0.235	0.229	0.222	0.142	0.113	0.091	0.079
Ortho	0.274	0.271	0.267	0.262	0.252	0.244	0.247	0.263
Int Med	0.216	0.21	0.201	0.188	0.832	0.754	0.665	0.55
Lung	0.248	0.24	0.232	0.224	0.339	0.286	0.238	0.211
Gastro	0.283	0.277	0.27	0.262	0.162	0.133	0.116	0.12
Other	0.202	0.198	0.192	0.184	0.451	0.381	0.284	0.175

Table 6.B.4: Expected values of B_i and L_i^b for different values of Δ_t, with $l = 0.9$.

Node i	B_i				$\mathbb{E}[L_i^b]$			
	$\Delta_t = \infty$	$\Delta_t = 5$	$\Delta_t = 2$	$\Delta_t = 1$	$\Delta_t = \infty$	$\Delta_t = 5$	$\Delta_t = 2$	$\Delta_t = 1$
GR	0.276	0.283	0.295	0.307	0.307	0.294	0.293	0.344
STC	0.325	0.328	0.329	0.331	0.302	0.311	0.333	0.37
STC Pall	0.215	0.217	0.22	0.221	0.009	0.008	0.01	0.012
HC	0.22	0.217	0.224	0.228	0.089	0.096	0.113	0.156
LTC SOM	0.249	0.252	0.26	0.264	0.037	0.031	0.033	0.038
LTC PG	0.156	0.162	0.16	0.169	0.059	0.05	0.058	0.074
LTC at Home	0.237	0.241	0.246	0.251	0.363	0.415	0.467	0.57
LTC Rehab	0.145	0.155	0.143	0.142	0.041	0.04	0.037	0.041
Card	0.312	0.309	0.303	0.297	0.392	0.338	0.285	0.252
Surg	0.285	0.281	0.275	0.268	0.845	0.782	0.716	0.642
Neur	0.28	0.278	0.274	0.269	0.349	0.345	0.35	0.377
Uro	0.37	0.362	0.354	0.348	0.193	0.147	0.105	0.078
Ortho	0.334	0.33	0.327	0.323	0.344	0.337	0.344	0.367
Int Med	0.294	0.287	0.276	0.264	1.107	0.991	0.813	0.633
Lung	0.314	0.307	0.298	0.29	0.436	0.371	0.302	0.26
Gastro	0.344	0.337	0.33	0.322	0.21	0.171	0.144	0.143
Other	0.288	0.283	0.276	0.269	0.609	0.49	0.328	0.175

Table 6.B.5: Expected values of B_i and L_i^b for different values of Δ_t, with $l = 0.8$.

7

Queueing and Forecasting Models for Online Mental Health Support Systems

Based on:
de Boer, T. R., Mérelle, S., Bhulai, S., Gilissen, R., & van der Mei, R. D. (2023). Forecasting call and chat volumes at online helplines for mental health. *BMC Public Health*, 23(1), 984.

de Boer, T. R., Mérelle, S., Bhulai, S., & van der Mei, R. D. (2022). Performance modeling for call centers providing online mental health support. *International Journal on Advances in Life Sciences*, volume 14, number 3&4, pages 120-129.

Abstract

Each year, many help seekers in need contact mental health helplines for support. It is crucial that they receive support and waiting times are minimal. In order to minimize delays, helplines must have adequate staffing levels, especially during peak hours. This has raised the need for means to predict required staffing levels, by forecasting call and chat volumes and combining them with a model to determine staffing levels. In this chapter, we present a study aimed at improving both the operational efficiency and predictive capabilities of such helplines, using real-world data from 113 Suicide Prevention‘[170] in the Netherlands. First, we develop and validate a detailed call center model that captures the unique features of mental health support systems, including triage, abandonment, and multi-phase service durations. This model is validated through trace-driven simulation and shown to reproduce real-world waiting times with high accuracy. Second, we present a forecasting framework to predict daily and hourly call and chat volumes using historical arrival data. We evaluate several forecasting models, including (S)ARIMA, linear regression, and LSTM, and find that (S)ARIMA consistently provides the best short-term forecasts, while linear regression performs better for longer time horizons. Finally, we incorporate counselor feedback collected through shift-based questionnaires to explore the relationship between experienced workload and objective indicators such as chat and call volumes. This approach provides a robust basis for dynamic staffing decisions, aiming to reduce waiting times and improve counselor well-being.

7.1 Introduction

Mental health helplines are helplines that are concerned with helping or assisting help seekers requiring mental health help, instead of physical care such as emergency helplines. There are many forms of mental health, with many countries having one or multiple helplines. Examples of mental health helplines in the Netherlands are 113 Suicide Prevention [170], the helpline for help-seekers with suicidal thoughts, the listen helpline (Dutch: luisterlijn) [59], and the Kindertelefoon [100] for children. In recent years, mental-support services have received a lot of attention due to the increase in call volumes related to (partial) lockdown to combat the spread of the corona [27, 190]. This chapter focuses on call center modeling for suicide prevention helplines. However, we emphasize that the results can also be used for modeling the waiting-time performance of other (mental) health helplines.

Suicide is a worldwide health problem. In 2020, on average, five persons died each day by suicide in the Netherlands alone. Suicide is a leading cause of death among adolescents [253]; worldwide, more than 700,000 people die from suicide every year [207]. In many countries, people struggling with suicidal thoughts can contact a helpline to get support to prevent and reduce suicidal thoughts [82]. In the Netherlands, persons with suicidal thoughts can contact 113, either by telephone or by chat [148], and are helped by volunteers and professionals. It is crucial that help seekers are answered swiftly. Therefore, it is important that adequate staffing is present to answer telephone calls and chat requests, minimizing the waiting times and abandonments (i.e., sudden termination of ongoing calls or chats while waiting). In order to calculate proper staffing levels (e.g., [109]), a good understanding of the processes of the call center system for mental health is required. It is also crucial to gain insight into the arrival process of help seekers, because these insights can contribute to good predictions of call volumes, and hence adequate staffing levels.

Mental health helplines differ in various aspects from classical call centers. First, the subjects of conversations are mental health concerns, e.g., loneliness and substance use [178, 188, 247]. Second, agents often have to handle complex conversations with the patient-in-need, and may themselves require support during or after a difficult conversation [208]. Therefore, an important aspect is that agents often need some time to cool down after emotionally difficult conversations. Third, when a chat or telephone call enters the system, it has to be determined which agent is best capable of handling the call, which results in a warm-up time before the call is taken into service. Lastly, chat conversations that enter the system first go through a triage phase. The inclusion of triage plays an important role, and functions as a filter [41] to chat requests, checking whether these help seekers are at the right helpline or might require emergency care. The triage agent can also estimate the conversation's difficulty to best match an agent with enough experience to handle the chat. So, the model for such a mental health helpline needs to meet the following requirements: (1) the possibility of abandonment, (2) a service time consisting of multiple phases, (3) a warm-up and cool-down time,

(4) inclusion of chats and telephone calls, and (5) triage.

Classic call center models were considered for modeling this helpline, see [73] for an excellent overview of queueing models for call centers. The Erlang C model is proven to be inaccurate when modeling call centers with abandonments [181]. The Erlang A model includes abandonments [74], but does not include multiple skills, triage, and warm-up times. Multi-skill call center models, such as an N-configuration [73], were also considered. The N-configuration model includes two types of arrivals, which in this case would be telephone and chat, and assumes that new chat arrivals can be picked up before triage. However, in the mental health helpline of [170], it is crucial that chats first go through triage. The importance of modeling triage is already shown in the emergency domain [223, 224]. Therefore, in this chapter, we modify the N-configuration model in such a way that it includes the specifics of mental health helplines.

This chapter introduces a new queueing model to include the possibility of triage. The model is built based on anonymized real-life call and chat data, made available by [170]. The records used cannot be traced back to individual callers, as only timestamps and durations are used. The anonymous data is also used to validate the model. This data is also utilized to determine different patterns that can be helpful for forecasting demand. various factors were considered vital for the number of help seekers per day, such as the historic trend, daily, weekly, monthly, and yearly patterns, and the effect of large news items or events discussed in various media forms. Whitley et al. [245] and Niederkrotenthaler et al. [159] studied the influence of media events on suicides and found that the number of suicides increased after the suicide of a well-known celebrity.

Data for this research are provided by 113 Suicide Prevention [170], referred to as '113'. Its mission is to prevent suicides and break the taboo surrounding suicide. Help seekers struggling with suicidal thoughts can contact 113 24/7 by either chat or telephone. These help seekers are then helped by counselors consisting of volunteers and paid employees. Apart from that, 113 also provides online therapy, self-tests, and self-help courses [231].

Research has been done on forecasting call volumes at helplines (e.g., [211]). In particular, Gijo et al. demonstrated the added benefit of applying (S)ARIMA models to forecast call volumes in emergency services [78]. (S)ARIMA models can identify possible trends and cycles, Gijo et al. show that forecasting can be done effectively using historical data only. In contrast, research on helplines for mental health is often focused on the conversation topics or the types of callers, rather than on call volumes and waiting times. For example, Salmi et al. showed the change in conversation topics during COVID-19 [188]. Grigorash et al. have studied the caller type of mental health helplines [85]. In this context, this chapter aims to fill the gap between these studies by combining a forecasting approach mostly seen in general call centers on the one hand with the specifics of the mental health helpline context where media events might affect the demand on the other

hand.

This chapter aims to test and compare different forecasting models on the anonymous call-volume data provided by 113. Assumptions about the helpline are validated using data analysis, especially a possible trend, cycle effects (over different time scales), and exogenous factors, such as the effect of media events. A cycle is defined as a seasonal effect, which repeats over time on a yearly, monthly, weekly, or even daily basis. The trend shows the general tendency of the data to increase or decrease during a longer period. More specifically, we address the following hypotheses:

- Weekly, daily, and yearly cycles are important for predicting the number of arrivals.
- The number of incoming phone calls and chats increases during and after a large media event (such as the suicide of a celebrity).
- Accurately forecasting the call volumes is possible using models that use historical data with possible exogenous factors. The historical data is used to incorporate possible cycles and trends, while exogenous factors are added to account for the possible effect of media events.
- The workload that counselors experience increases when the average waiting time for phone or chat is higher than usual. This applies to situations where there are more help seekers than counselors can help or more help seekers than expected.

Lastly, a questionnaire was used to determine when and why counselors experienced a higher workload. This is then combined with the model of the helpline and the forecasting process to determine adequate staffing levels is explained for this specific model.

This chapter is organized as follows. Section 7.2 discussed related work done in call center models or suicide prevention helplines. Section 7.3 describes the different patterns found in the data. In Section 7.4, the proposed model is described. In Section 7.5, the model is validated using a combination of data and trace-driven simulation. Next, Section 7.6 explains how demand call volumes can be predicted based on historical data. Section 7.7 describes how the model and the forecast can be used to construct staffing advice. A questionnaire to determine how counselors experience workloads is discussed in Section 7.8. Finally, in Section 7.9, the conclusion and discussion are given.

7.2 Literature

Call centers and/or helplines have been researched in different fields of science. For this chapter, we will mention some of the research done on the modeling of call centers and research done on specifically suicide prevention helplines. Various

research has been done on modeling call centers, Garnett et al. [74] have shown how and why Erlang B and Erlang C lack the feature of impatient customers. They introduce an Erlang C model, where customers abandon the system after not being answered after their impatience has run out. Here, impatience is drawn from an exponential distribution. This research, however, lacks the inclusion of multi-skill. Gans et al. [73] provide an overview of different types of skill routings, where the skill may be based on training, compensation, or time restrictions. Forecasting of arrivals has also been researched, for example, Gijo et al. [78] have shown how (S)ARIMA models can be used in forecasting the call volume. Research done specifically at mental health helplines is often limited to conversation topics and different types of callers. Salmi et al. [188] show the change in conversation topics during COVID-19. While Grigorash et al. [85] and O'Neill et al. [163] have both studied different caller types and how these caller types can be predicted. This chapter aims to fill the gap between research done on standard call centers and that done on mental health helplines, specifically by applying modeling and forecasting, that are normally seen in standard call centers, to mental health helplines.

7.3 Data Analysis

Call data is provided over the period 2016-2021 and consists of around 250,000 chats and 175,000 telephone calls. The distribution of the arrivals over the years can be seen in Figures 7.1 and 7.2 for phone and chat, respectively. It should be noted that data for 2016 and 2021 were incomplete at the time, explaining the low number of arrivals in 2016 and 2021. However, omitting 2016 and 2021, it can still be seen that there is an increasing trend present in the number of arrivals of phone calls as well as chats. The number of telephone arrivals increased from almost 20,000 in 2017 to more than 50,000 in 2020, and chat arrivals increased from close to 40,000 in 2017 to almost 70,000 in 2020. Each call or chat has the following elements: (1) the arrival time, (2) the time entering the queue, (3) the time of acceptance by an agent, (4) the disconnection time, and (5) the completion time, all elements are denoted using a combination of date and time, using the following formats dd-mm-yyyy and hh:mm:ss for date and time, respectively. Each call also has a contact id and an *initial* contact id, which may be different if a call is a forwarded telephone call or chat.

Missing data
Not all data were useful in its original form. Therefore, we identified missing values in the data and determined whether any imputations were required. Missing values are handled differently based on context: a missing value in waiting time often meant that the help seeker abandoned the queue. In some cases, it could also have been due to the help seeker being accepted before being queued. These values were filled in based on these conditions. Finally, the data were aggregated to obtain call and chat volumes per day and hour. We had two days for which the call and chat volumes were both zero, due to a technical issue. These two volumes were estimated using linear interpolation.

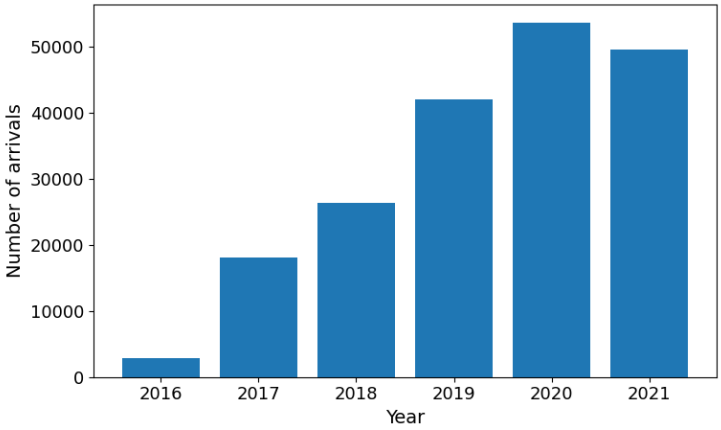


Figure 7.1: The number of phone arrivals per year.

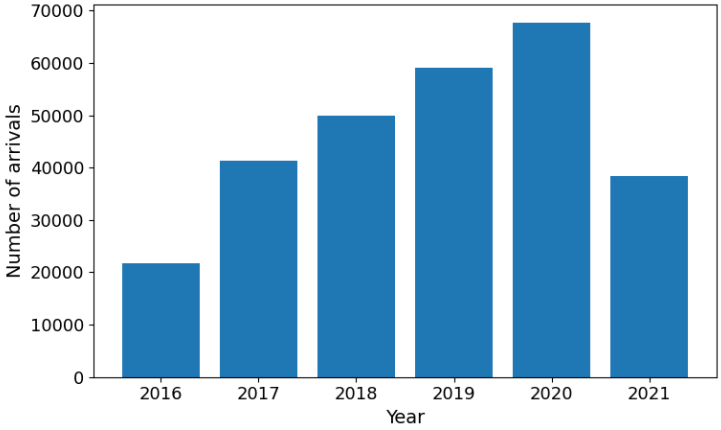


Figure 7.2: The number of chat arrivals per year.

Different features were added to the data to obtain a better understanding of the different durations found in the data. The following durations were calculated: *pre-queue duration*, *waiting time*, *call duration* and *wrap-up time*. The *pre-queue duration* is the time between the arrival of a call and the time entering the queue, and is the time spent by the caller in a selection menu. This phase does not require resources from the helpline and, therefore, falls outside the scope of this chapter. The *waiting time* is defined as the time between entering the queue and the time that an agent accepts the call. The *call duration* where the agent and the help seeker are actually connected is defined as the time between acceptance

of a call and the disconnection time. Finally, after each call, the agent has to fill in a wrap-up form, which is the time between disconnection and completion. A visual representation of this timeline is given in Figure 7.3.

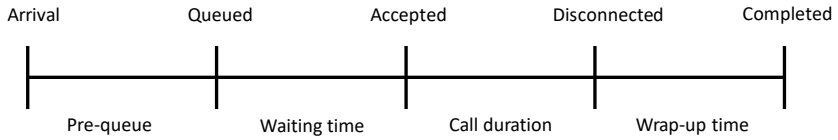


Figure 7.3: Timeline of a call from the perspective of the caller.

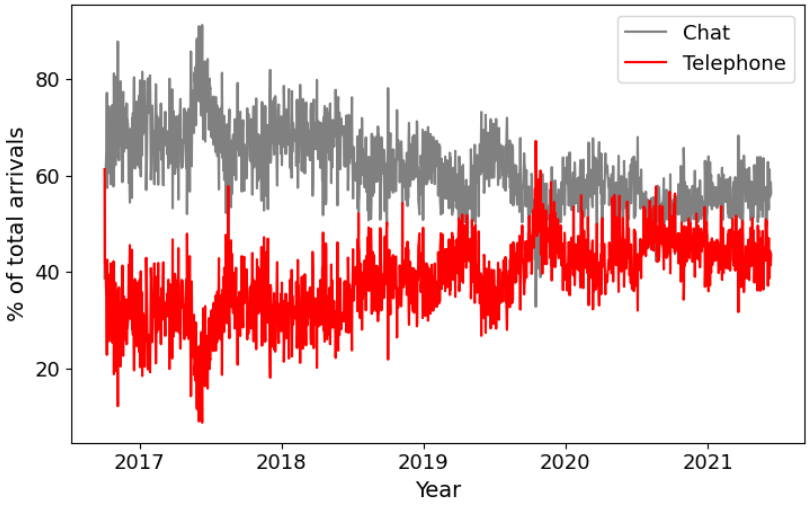


Figure 7.4: Decomposition of incoming calls and chats.

Recall that there are two options for help seekers to contact the helpline: via *telephone* or *chat*. These two contact types are mostly handled by the same type of agents, but differ in some important aspects. Traditionally, there used to be more chats than telephone calls. However, the difference has diminished in recent years. This can be seen in Figure 7.4, which shows that in 2017 and 2018, most of the arrivals were chats. In 2019, the difference between chat and phone diminished, while since 2020, there are still more chats, but the difference is not as large as before. However, phone calls and chat follow different patterns, the weekly patterns are shown in Figures 7.5 and 7.6 and the daily patterns in Figures 7.7 and 7.8. The weekly patterns in both cases show that the weekends see a lower number of calls. However, for chats, we see a clear dip on Saturday and a small increase on

Sunday. We also observe that most telephone calls arrive between 12:00 and 20:00, while chats have a clear peak at 20:00. Both telephone and chat call arrivals show a decrease during the night and early morning until around 5:00 in the morning. Yearly and monthly cycles were also analyzed. However, both were found to not significantly vary over time, possibly due to limited available data in the case of yearly cycles.

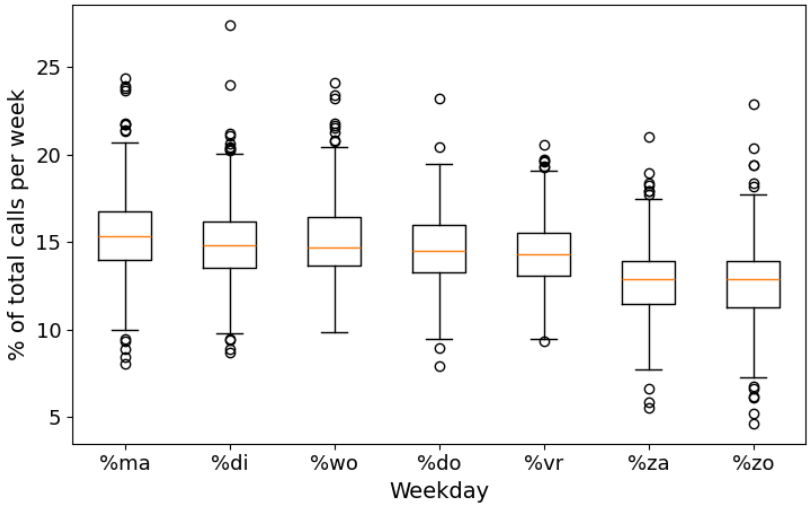


Figure 7.5: Weekly pattern of incoming telephone calls.

Media events

To assess whether media events might influence the number of call arrivals at the helpline, the periods before and after large media events were analyzed, and briefly outlined below. In the period studied, one celebrity in the Netherlands died by suicide. Figure 7.9 (lower graph) shows the effect of the suicide of a well-known Dutch author on the number of chat arrivals. In this period, only the data of chats are available. It is clear that the news significantly affected the number of chats, especially in the week after the news. Figure 7.9 (upper graph) also shows the absence of the effect the suicide of an internationally well-known artist had on the arrivals at the helpline. Lastly, Figure 7.10 shows the effect national political news had on the arrivals, as an example of the effect of media events other than the suicide of well-known persons. We observe that in most cases the effects of these events were limited, or only short-term (one or two days). Figure 7.9 (lower graph) shows that there are events that have a larger or long-term effect, but most events do not have this large or long-term effect. Together with the fact that this type of event cannot be predicted ahead of time, we chose not to include these events in our modeling or forecasting models.

Incoming chats are first handled by triage, this is needed to filter chats that are at

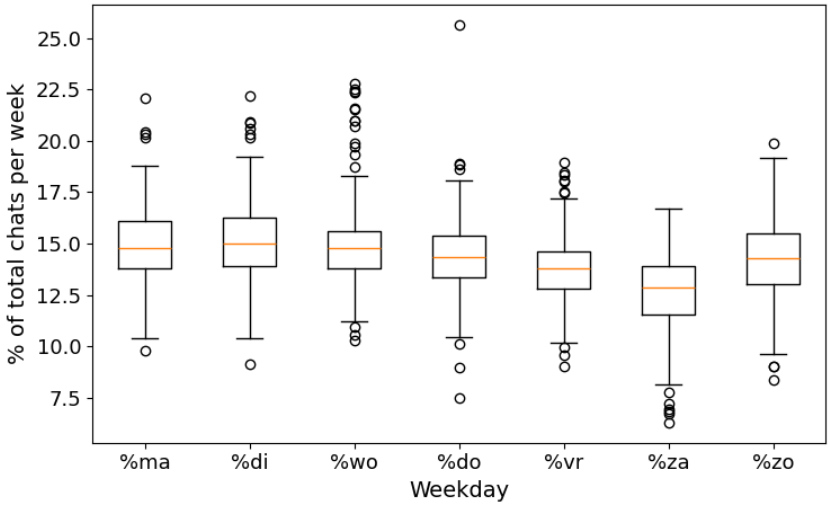


Figure 7.6: Weekly pattern of incoming chats.

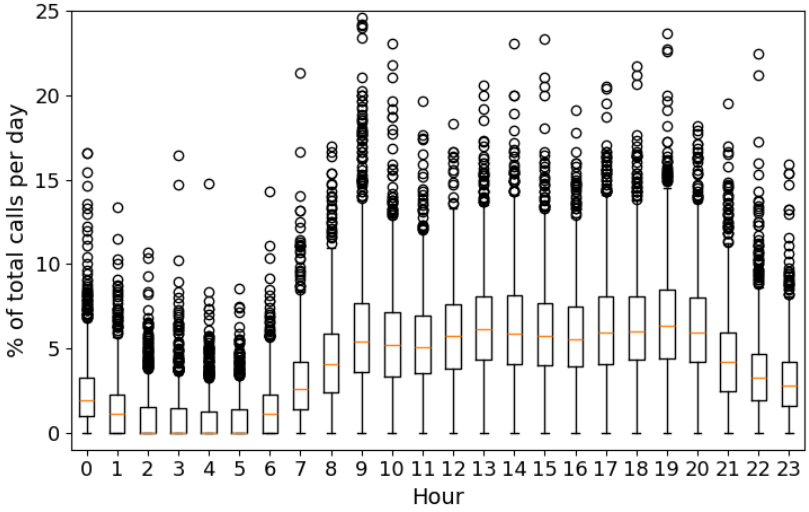


Figure 7.7: Daily pattern of incoming telephone calls.

the wrong helpline. Only part of the incoming chats after triage gets sent to another agent. The percentage of chats that an agent handles after triage is around 50% during day shifts, but this differs over the day. During the night shifts, fewer chats are forwarded to an agent. The different nature of night conversations may

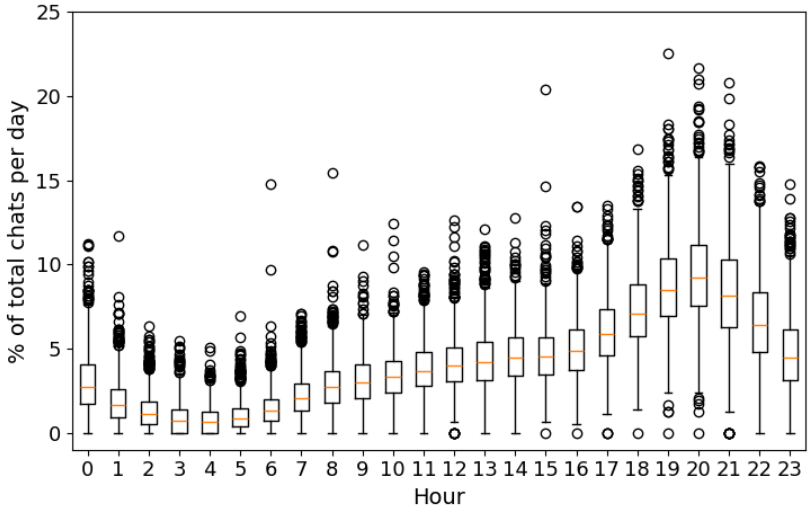


Figure 7.8: Daily pattern of incoming chats.

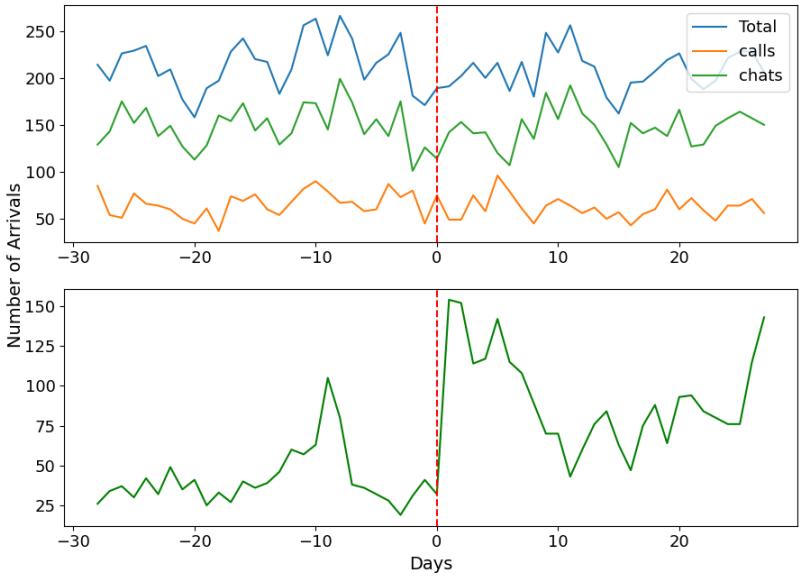


Figure 7.9: Suicide of an internationally well-known artist on day 0 (upper graph) and suicide of a nationally well-known author on day 0 (lower graph).

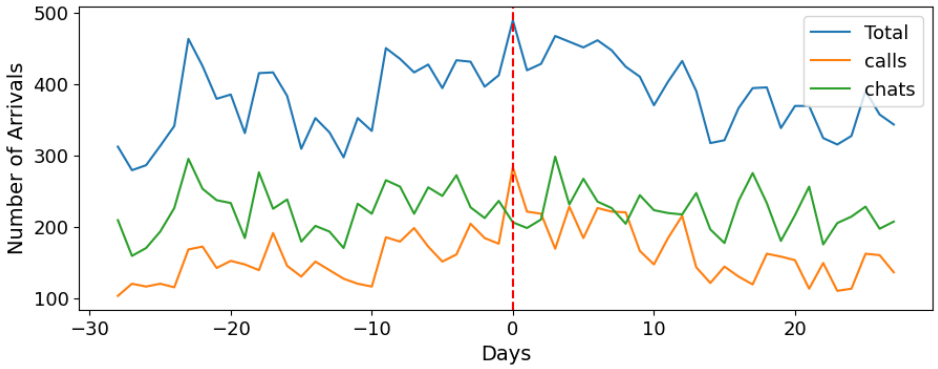


Figure 7.10: 113 receiving media attention due to political news, peak on day 0.

cause this. The triage plays an important role in filtering chats, as seen by the percentage of chats that get through triage. Chats are filtered out due to various reasons. For example, the chatter may be at the wrong helpline and or is identified as a prank chatter [240].

In the data, we distinguish three service time distributions: for the duration of (1) telephone calls, (2) chats *during* triage, and (3) chats *after* triage. The empirical distribution of phone call durations can be found in Figure 7.11. On average, a telephone call takes around 18 minutes, with the longest phone call in the data taking multiple hours. The duration of chats in triage can be seen in Figure 7.12, and the duration of chat conversations after triage can be seen in Figure 7.13. As can be seen, the chats that have gone through triage tend to take much longer than the chats during triage: on average, chats in triage take 19 minutes versus 34 minutes after triage. This is also clearly visible in the histograms with durations of chats in triage having a clear peak at around 400 seconds, while chats after triage have a peak in durations at 4000 seconds.

Remark 7.1. Mental health helplines often use a mix of paid professionals and volunteers, where an agent's responsibilities depend on experience and other factors. The assumption is that experienced agents can handle more difficult conversations alone, while inexperienced agents might require assistance when handling a more complex conversation (this could be either a phone or chat conversation), resulting in a longer service time. Analysis of our data pointed out that there seems to be no significant difference in service time distribution between different volunteers and professionals. There might be a difference, but this could be obscured due to many other factors, such as experienced agents handling the more complex conversations and assisting or coaching the less experienced agents. Based on this observation, the call duration distributions are assumed to be the same for all agent experience levels.

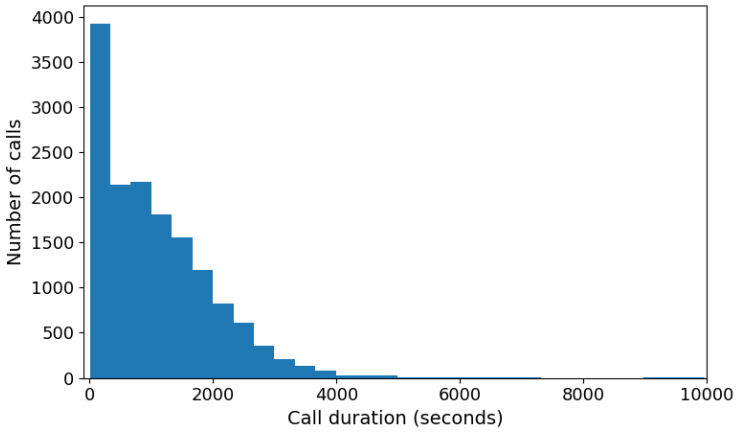


Figure 7.11: Histogram of telephone call durations.

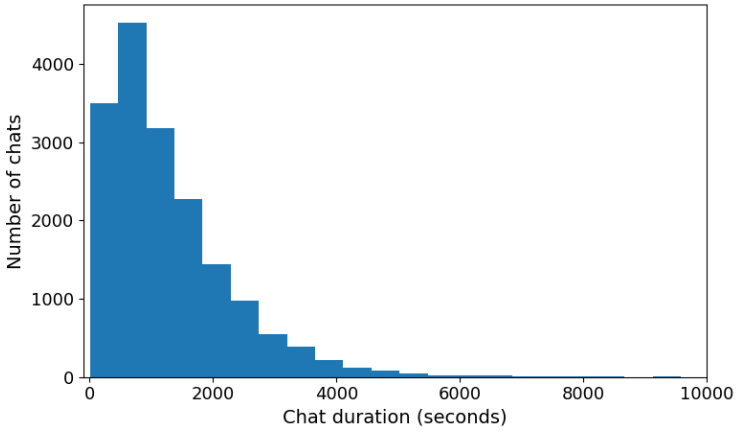


Figure 7.12: Histogram of chat triage durations.

7.4 Model Description

For the modeling step, we distinguish between two different types of calls, namely: (1) *chat calls*, and (2) *telephone calls*, which are handled differently. The incoming chat calls are first handled by a triage system. The triage system consists of c_{triage} triage agents. Each triage agent can handle n_{triage} chats simultaneously without perceivable slow-down. Together, there are $c_{\text{triage}}n_{\text{triage}}$ 'triage slots' for incoming chats. Arriving chats enter service at triage immediately if a triage slot is available. If no triage slot is available, chats enter an infinitely sized queue and are

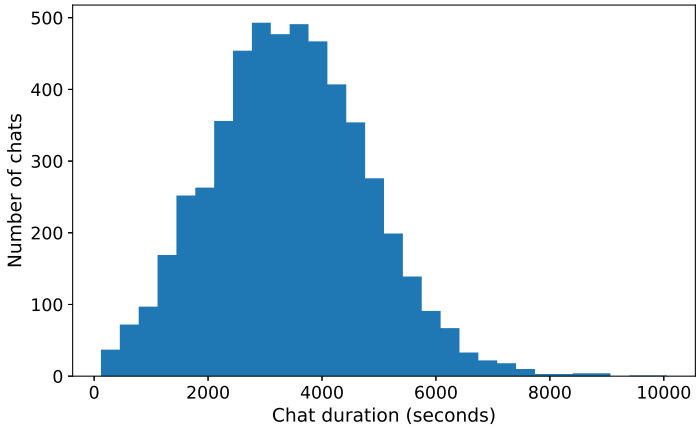


Figure 7.13: Histogram of durations of chat conversations after triage.

helped First-Come-First-Served (FCFS) or abandon the queue if the waiting time is longer than their impatience. The service time of a chat call in triage, denoted B_{chat}, is the convolution of four random variables:

$$B_{\text{chat}} = B_{\text{warm-up}} + B_{\text{conversation}} + B_{\text{wrap\_up}} + B_{\text{cool-down}}.$$

All variables are drawn from exponential probability distributions that can be estimated from the data. A visual representation of the service time can be seen in Figure 7.14 and differs from the timeline in Figure 7.3, since Figure 7.14 shows the time an agent is occupied by an arrival, while Figure 7.3 shows the perspective of the help-seeker.

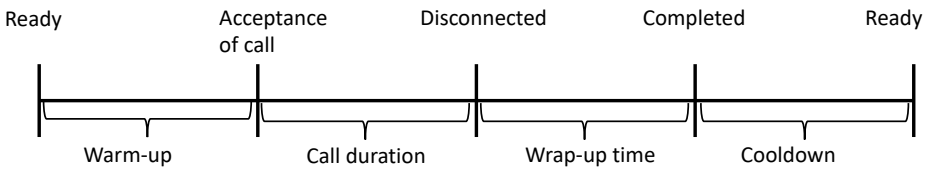


Figure 7.14: Timeline of call from the perspective of the agent.

A key aspect of the model is the inclusion of *impatience*, i.e., the maximum amount of time that a help seeker is willing to wait before they abandon the system. The

impatience of a help seeker who enters the system via a chat is modeled as an independent sample drawn from an exponential distribution with mean $\mu_{\text{chat-impatience}}$. After service completion at the triage center, a chat is either sent through to the helpline (HL) for assistance (with probability $p_{\text{sent-through}}$), or the chat leaves the system. Note that during the warm-up period $B_{\text{warm-up}}$, the agent is busy, but the help seeker is not yet answered. Therefore, help seekers may abandon the queue during that period.

Different from chat calls, incoming telephone calls do *not* go through a triage phase and arrive directly at the HL. This is the core part of the system where most of the service processing occurs. The HL is equipped with c_{HL} agents, each of which can handle both chat calls and telephone calls, not more than one call at a time. When a telephone call finds an HL-agent available, he enters service immediately. If the telephone call arrives and no agent is available, the call enters an infinite-sized queue that is handled on an FCFS basis. Here, phone calls are able to abandon the queue if their waiting time is longer than their impatience.

The HL processes both telephone calls and forwarded chat calls (i.e., those that have passed through the triage phase). Here, chat calls have *non-preemptive priority* over telephone calls. Thus, when an HL-agent becomes idle, he first checks whether there is a chat call pending (while keeping a triage slot occupied), and if so, starts to service the longest-waiting chat call. If no chat call is pending, the agent checks if telephone calls are pending.

Similar to the modeling of chat sessions in triage, the duration of phone calls and chats after triage both consists of four subsequent independent phases: (1) warm-up, (2) conversation, (3) wrap-up, and (4) cool-down, where each phase has its own probability distribution differently for phone and chat after triage. The impatience of help seekers via telephone is modeled as a sample from some probability distribution that can be obtained from the data. Calls abandon the queue if their waiting time exceeds the impatience. When service is completed, the calls exit the system. Note that agents are busy during the warm-up period, similar to abandonments at the triage, but the help seeker does not perceive this and can, therefore, still abandon the queue during the warm-up phase. See Figure 7.15 for an illustration of the model.

7.5 Model Validation

The model described in the previous section has to be validated to confirm our modeling choices. The validation was done using trace-driven simulation. For the simulation of this model, the following values are needed for each arrival: (1) an arrival time, (2) service duration (consisting of warm-up, conversation, wrap-up and cooldown), (3) impatience, and (4) call type (chat or telephone) and if the call type is chat, then also information about triage (duration and if the chat is at

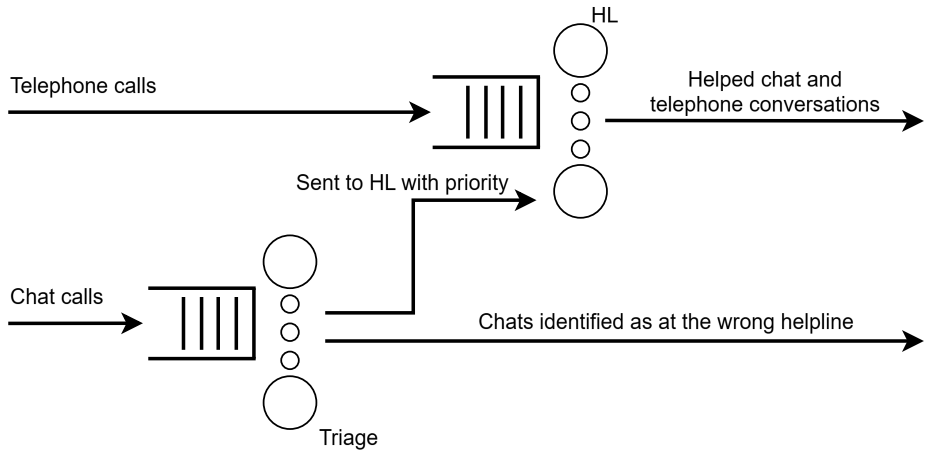


Figure 7.15: Illustration of the model.

the right helpline) are also needed. Trace-driven simulation uses values obtained exactly from the data, thereby giving a precise comparison between reality and simulation. The data contained some missing values, and these need to be filled in before simulating, namely: (a) the conversation and wrap-up duration of calls that were unanswered, (b) impatience of help seekers, and (c) warm-up and cool-down durations.

The missing values of conversation and wrap-up durations were filled in using hot-deck-imputation [233]. This method samples from the known values to fill in the missing values. The distinction was made between the different types of conversations: telephone, chats during triage, and chats after triage. The impatience of help seekers is mainly unknown due to the limited availability. Only a small percentage of telephone calls were unanswered, and for chats, even fewer impatience data was available. Warm-up and cool-down durations were not present in the data. Therefore, the missing values of impatience, warm-up, and cool-down periods were all drawn from exponential distributions.

The parameters were estimated using expert opinions from paid professionals and volunteers. Impatience of chat conversation is determined by the sum of a constant 300 seconds and a duration drawn from an exponential distribution with a mean of 300 seconds. The impatience of a telephone caller is drawn from an exponential distribution with a mean of 240 seconds. The warm-up for both chat and telephone is drawn from exponential distributions with means of 60 and 45 seconds for telephone and chat, respectively. Lastly, the cool-down durations of chat and telephone are both drawn from an exponential distribution with a mean of 120 seconds.

Moreover, based on current practice at 113, for the experiments, we assume that

$n_{\text{triage}} = 5$. Thus, each triage agent can handle a maximum of five triage chats simultaneously. Further, the simulations are trace-driven, and follow the realizations of the time variations of (1) the arrival processes for chat and telephone, (2) the number of triage- and HL-agents, and (3) the fraction of triage call that is forwarded to the HL system.

The simulation ran for almost 12,000 arrivals consisting of around 6,500 chat arrivals and 5,500 phone calls during a period of 1 month. The average waiting time of the simulation and the data were then compared. Figure 7.16 shows both the simulated and realized average waiting times for telephone calls as a function of the cumulative number of call arrivals. The results show that for a small number of calls, the average waiting time is rather sensitive to outliers but quickly stabilizes over time. The average waiting time of the data and simulation both converge to the same value, around 80 seconds. This validation experiment is repeated for chat call arrivals, which show similar convergence, see Figure 7.17. This process was repeated over different months and showed similar convergences; an example can be seen in Figures 7.18 and 7.19, which shows the experiment repeated for September.

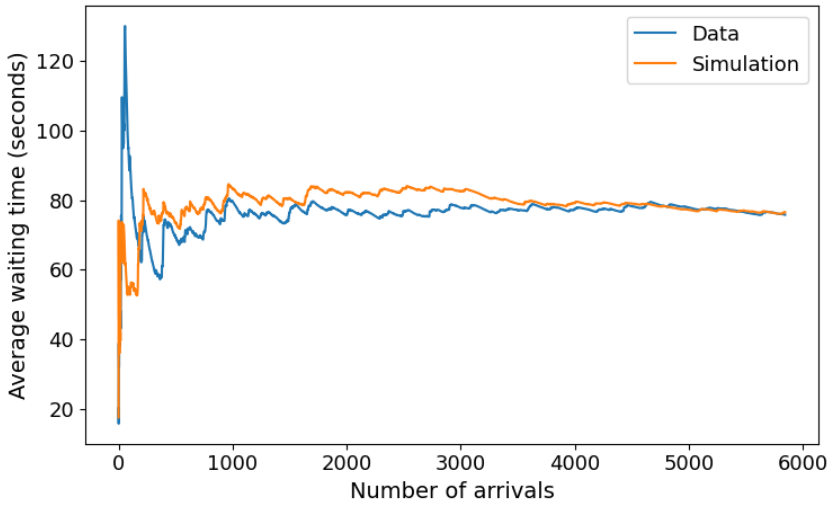


Figure 7.16: The average waiting time of telephone calls in the simulation and the data of August 2021.

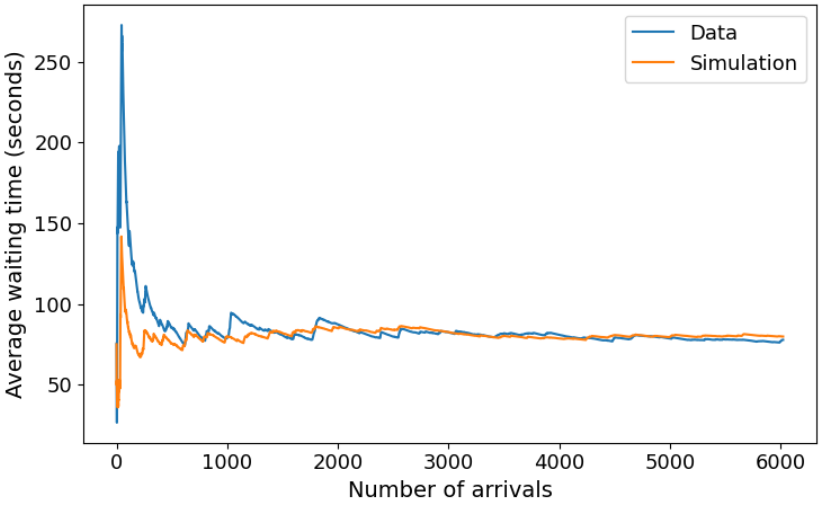


Figure 7.17: The average waiting time of chat calls in the simulation and the data of August 2021.

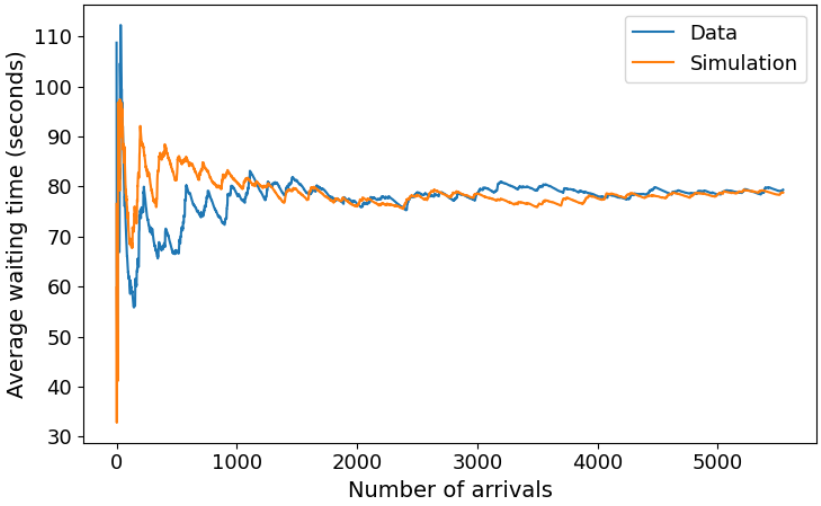


Figure 7.18: The average waiting time of telephone calls in the simulation and the data of September 2021.

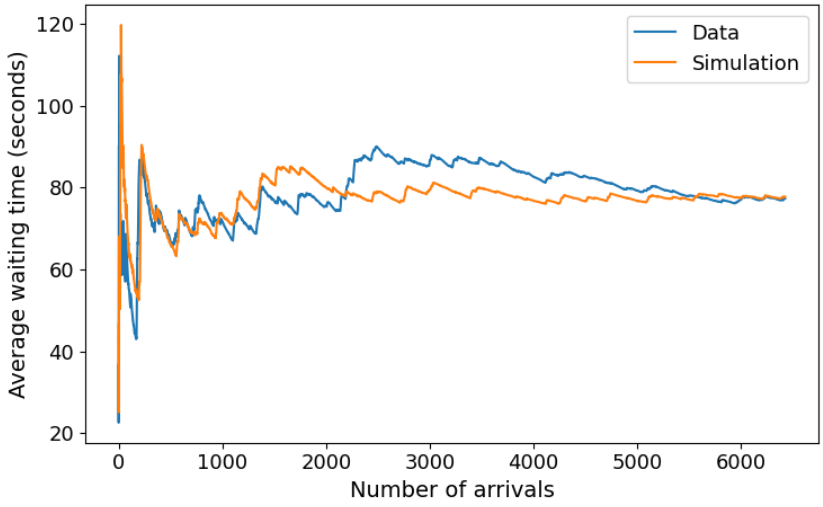


Figure 7.19: The average waiting time of chat calls in the simulation and the data of September 2021.

In summary, these validation results show that the model works well in simulating waiting times and confirm the modeling choices made in Section 7.4.

7.6 Demand Forecasting

Demand forecasting is concerned with predicting the call volumes for both telephone and chat calls. The performance of the forecasts is dependent on several choices: the time window and the aggregation level. The time window concerns itself with how far ahead the forecast is. Forecasting daily volumes for tomorrow is often easier than the daily volume over 8 weeks. For this chapter, we chose to predict 1 day ahead until 8 weeks ahead, since the schedule of the agents is made 8 weeks ahead but can be adapted over time. The aggregation level concerns itself with the kind of volumes to be forecasted. Hourly volumes are harder to predict than monthly volumes, but are less useful for scheduling. In this chapter, we chose to predict daily volumes. However, later in this section, it is explained how these daily volumes can be adapted to hourly volume predictions. The choice was made to forecast chat and telephone arrivals independently due to the difference in arrival processes (also explained in Section 7.3) and the difference in handling.

The following forecasting models were considered: Long Short-Term Memory (LSTM), (S)ARIMA [195], linear regression [174], and different baseline models. (Seasonal) Autoregressive Integrated Moving Average (shortly, (S)ARIMA) models are well-known time series models [78] used for forecasting. ARIMA uses previously measured values for forecasting future values. SARIMA is similar to

ARIMA, but here a seasonal component is added, for an overview of (S)ARIMA models see [47]. The parameters of (S)ARIMA models are chosen using auto-ARIMA [104], which are $(5, 1, 1)$ for ARIMA and $(1, 1, 1)(0, 1, [1, 2], 7)$ for SARIMA. AutoArima is an R-method that determines the best parameters based on the Akaike Information Criterion (AIC), which is an estimator of the prediction error. Linear regression is seen as a simple Machine Learning (ML) approach, and was used to fit a linear trend with a weekly effect on the data. These models fit a linear relation between various factors and the outcome, in this case, the number of arrivals. In formula form, this model looks as follows:

$$\mathbf{F} = \mathbf{X}\beta + \epsilon,$$

where $\mathbf{F}$ is a vector that contains forecasts, $\mathbf{X}$ is a vector that contains input variables, β is the vector that contains parameters, and ϵ is noise.

The LSTM model is a more sophisticated ML-model used for forecasting in time series, and is a special kind of Recurrent Neural Network (RNN). Lastly, these models are compared to two baseline models, referred to as *Baseline 1* and *Baseline 2*, where the forecast of day i is taken as the measurement of day $i - 7$ or $i - 56$, respectively, given by:

$$F_i = A_{i-7} \quad \& \quad F_i = A_{i-56},$$

where F_i[1] is the forecast of day i and A_i is the actual value of day i. The data provided show that there is an increasing trend present and that weekly cycles seem to be predominant. Therefore, it is chosen to focus on forecasting using historical data on call volumes. The models will be evaluated using the Mean Absolute Percentage Error (MAPE), the Mean Squared Error (MSE), and the Mean Absolute Error (MAE). The MAPE is calculated using the following formula:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{F_i - A_i}{A_i}\right| * 100\%.$$

Here n is the number of forecasts, F_i is the forecast of day i, and A_i is the actually recorded number of arrivals on day i. The found MAPE results can be seen in Figures 7.20 and 7.21. Note that *Baseline 1* can only forecast up to 7 days ahead, as it relies on observations from 7 days earlier. The LSTM model already shows inferior performance at a 7-day forecasting horizon and is therefore not evaluated for longer horizons. We find that (S)ARIMA models perform best (in terms of the MAPE), especially when forecasting for five weeks (or six in the case of chats) ahead or less. For longer forecasting windows, it turns out that a simple linear regression model might provide more accurate forecasts in the case of telephone arrivals. However, both (S)ARIMA and linear regression models have an MAPE that is lower than the MAPE of the baseline models. The LSTM model has the highest error term in this situation.

[1] F_i is also used in Chapters 5 and 6 to denote a flow level; despite the identical notation, it represents a different variable in this context.

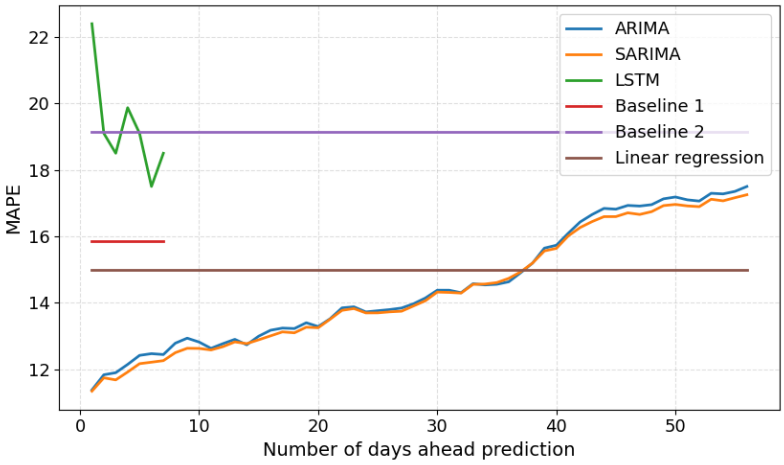


Figure 7.20: The MAPE of telephone call volume forecasts.

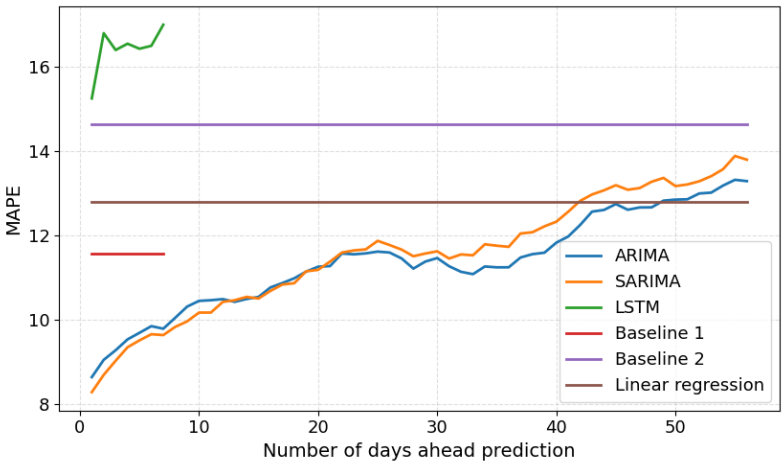


Figure 7.21: The MAPE of chat volume forecasts.

Next, the models are evaluated using MSE, which gives more weight to large prediction errors. The MSE is calculated as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(A_i - F_i)^2.$$

The found MSE values can be found in Figures 7.22 and 7.23. The results show that in terms of MSE, (S)ARIMA models have the lowest error. However, for forecasting more than three weeks ahead, the linear regression seems to have a lower error term. This differs from the finding when comparing the MAPE error terms, meaning that the (S)ARIMA models likely have more large prediction errors than the linear regression.

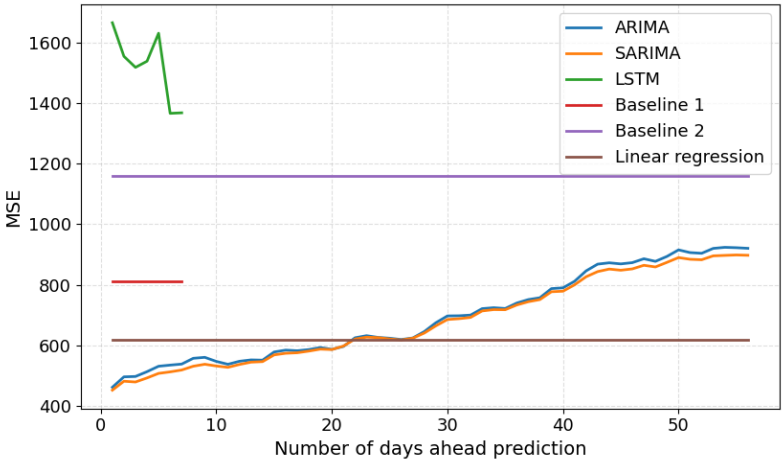


Figure 7.22: The MSE of telephone call volume forecasts.

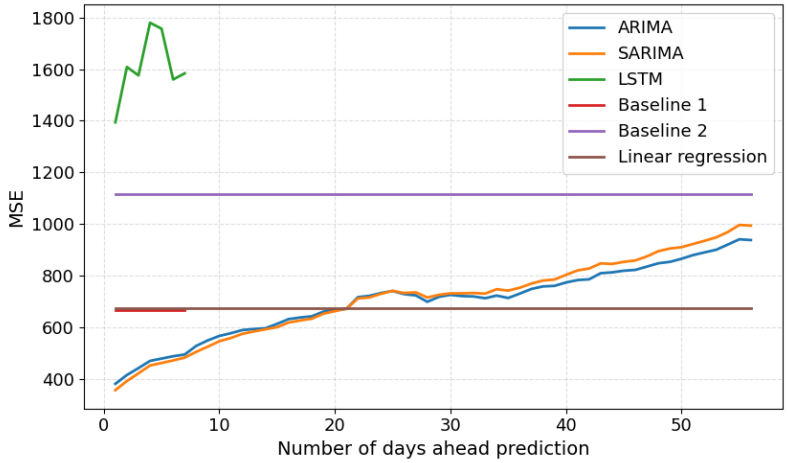


Figure 7.23: The MSE of chat volume forecasts.

Lastly, for a complete comparison, the models are also evaluated using MAE, which is calculated by the following formula:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|A_i - F_i|.$$

The found MAE values can be found in Figures 7.24 and 7.25. The graphs show similar findings when evaluating based on MAPE, in the telephone calls as well as chat, (S)ARIMA models perform best when forecasting for short-term windows, and linear regression for longer-term forecasts. The MAE shows again that forecasting chats is more accurate, and the window when (S)ARIMA performs best is longer (around 40 days for chats versus around 30 days for telephone).

Combining the evaluation of the three different error terms, we can come to the following findings:

1. All error terms agree that for short-term forecasting (S)ARIMA models perform best, long-term forecasting can best be done using linear regression, and LSTM seems to perform the worst.

2. The different error terms differ in the time window of when (S)ARIMA model performs the best. The MSE graph of phone forecasting shows a shorter time window in which (S)ARIMA performs best when compared to that of the MAPE graph. This tells us that it is likely that phone forecasting has more large errors influencing the MSE. A similar picture can be seen for chat forecasting.

3. All evaluations agree that the LSTM model performs the worst of all evaluated models at the moment.

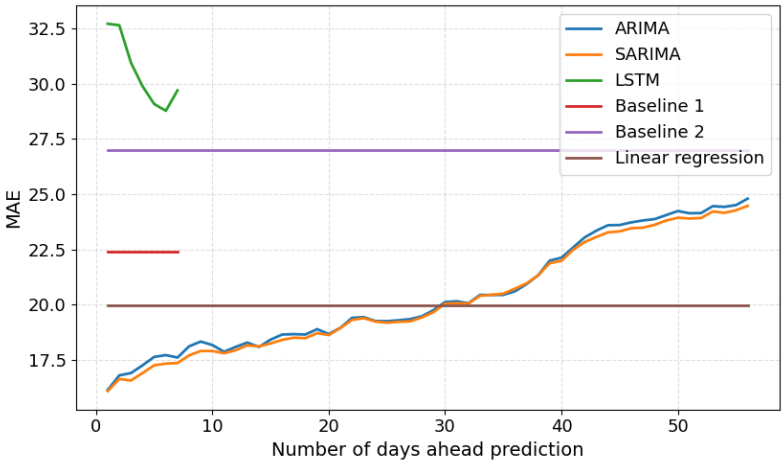


Figure 7.24: The MAE of telephone call volume forecasts.

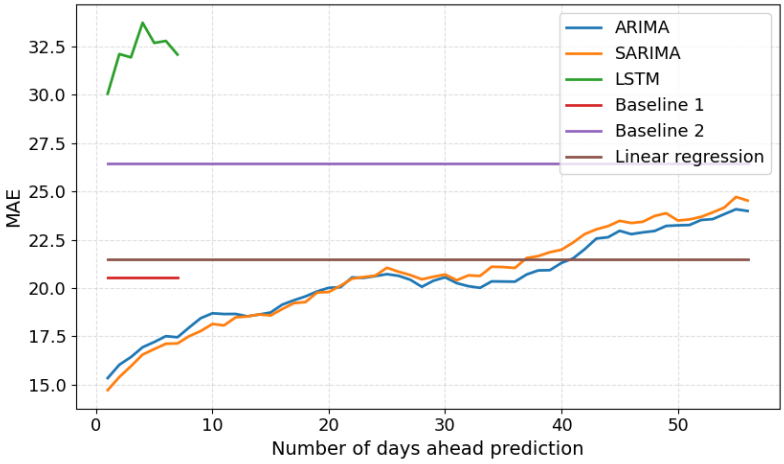


Figure 7.25: The MAE of chat volume forecasts.

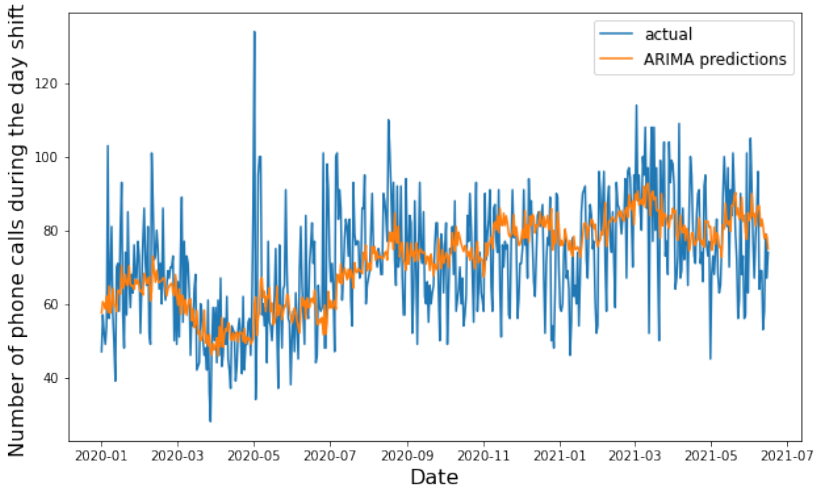


Figure 7.26: Comparison between the actual and predicted number of phone calls during the day shift (from 8.00 until 16.00).

The performance of (S)ARIMA models can be attributed to the flexibility of the models. Nonetheless, it is possible that, given additional time and further optimization, the LSTM could achieve better performance. However, it is questionable if the model would perform better than the (S)ARIMA models.

Hourly predictions
The results above concern the prediction of daily volumes. Call centers, however, often use hourly volumes for staffing purposes. Predicting hourly volumes can be done in several ways, namely forecasting hourly volumes directly or indirectly by using the predicted daily volumes. Forecasting hourly volumes directly introduces more uncertainty in the predictions. Therefore, it was chosen to forecast using the daily volumes. The call volume of hour h on day i is calculated as follows:

$$F_{i,h} = F_i \times P_h.$$

Here P_h is the mean percentage of arrivals on a day that arrives in hour h. This method can be expanded to also include smaller or larger intervals, such as shifts. An example of this can be seen in Figure 7.26, again the prediction is able to follow the actual number of arrivals.

7.7 Staffing

This section will briefly describe how the model of the helpline can be used together with demand forecasting to construct staffing advice. Staffing aims to determine

the required number of agents per time interval and is, therefore, different from a work schedule or planning of agents, for a schedule often staffing advice is first needed. One of the most well-known and often used rules is the square-root staffing rule [84]. The square-root staffing principle can be used to determine the staffing level c, which balances efficiency and service quality by adding a safety buffer proportional to the square root of the offered load. The formula is given by:

$$c = a + \beta\sqrt{a},$$

where a is the offered load calculated by multiplying the arrival rate with the mean service time, β is a parameter reflecting the service level, a higher β corresponds to a higher quality of service.

The given formula is based on the Erlang C queue and can be adapted to other queues. The model described in Section 7.4 needs two different staffing levels at triage and at HL. The staffing at triage receives only chat arrivals, a_{triage} is, therefore, calculated by multiplying the chat arrivals with the average service time of chats in triage. Since staffing is done ahead of time, the chat arrivals are predicted chat arrivals. Together with the fact that agents at triage can handle five conversations simultaneously, the square-root staffing rule therefore becomes:

$$c_{triage} = \frac{a_{triage} + \beta \times \sqrt{a_{triage}}}{5}.$$

The staffing at HL can be determined using the square-root staffing rule together with the facts that each agent can handle one conversation and agents handle chats after triage and new phone calls. The formula then becomes:

$$c = a_{HL} + \beta\sqrt{a_{HL}}.$$

Here, a_{HL} is calculated by summing the chats after triage multiplied by the average chat service duration, and the phone arrivals with average phone service duration, where again the arrivals are done by forecasting.

7.8 Questionnaire

Next, a questionnaire was given to senior counselors to determine when and why they experienced a high workload. This process consists of a questionnaire based on a quick scan for workload and adapted to the situation of the helpline. Senior counselors were asked to fill in a questionnaire after each shift [202]. Here, we examine whether the workload of the senior counselors is related to the number of calls or chats, and are so able, together with the forecasts, to understand adequate staffing levels. Data collection took place from 16 February to 30 April 2022. Table 7.1 gives an overview of the 10-item questionnaire that uses a 5-point Likert scale [134]. We transformed all questions such that a score of "5" indicates a high workload and a score of "1" indicates a low workload. Questionnaires that reported technical problems were excluded from analyses since high experienced

workload can then be related to the technical issues. We can identify whether there were any technical issues based on positive answers to question 7.

Number	Question	Answers
1	Did you have many tasks to do during the shift?	Very busy, busy, neither, unbusy, very unbusy
2	Did you have to work hard to do everything?	Always, often, usually, sometimes, never
3	Did you have to rush during your shift?	Always, often, usually, sometimes, never
4	Was there a large backlog, many missed conversations, or a high waiting time?	Always, often, usually, sometimes, never
5	Did you have issues with the pace?	Always, often, usually, sometimes, never
6	Could you show interest for colleagues?	Always, often, usually, sometimes, never
7	Were there any technical issues during your shift?	Always, often, usually, sometimes, never
8	Could you forward the chats in triage?	Very well, well, neither, bad, very bad
9	Did you have energy left at the end of your shift?	Very much, much, neither, little, very little
10	Were there enough counselors besides interns?	Yes, neutral, no

Table 7.1: Questions of the questionnaire to measure workload.

The sum scale that represents the counselor's experienced workload is calculated by summing all questions, except question 7, since this question is used to filter out questionnaires with technical issues. The questions used in the sum scale do not include questions concerning only chats or phone calls, except for question 8, which only concerns incoming chats. However, this question is included, since the question is about the process of forwarding chats after triage, chats after triage are handled by the same counselors who handle phone calls. The process of forwarding chats can therefore be obstructed by a large workload of phone calls. The results using the sum score excluding and including question 8 and the resulting conclusions were similar. Therefore, we chose to use the collected data of this item. This sum scale was then compared to features of the objective workload captured by the data. These variables are: the number of chats and phone calls during the shift, mean waiting time of chats and phone calls, missed percentage of chats, and phone calls. The Cronbach's alpha of the questionnaire was 0.72 [210], which

tells us that using the questionnaire in its current form is acceptable. Pearson's correlation coefficient was then used to measure the strength of the relationship, where 0 means no relation and 1 or -1 means a perfect correlation [191]. P values smaller than 0.05 were considered statistically significant.

The senior counselors filled in 88 questionnaires ($n = 32$ day, $n = 41$ evening, $n = 13$ night), two were excluded from analyses since respondents reported technical problems during the shift. Descriptive statistics are given in Table 7.2. Most questions are filled in with a mean of around 3, except for Questions 5,6, and 10, which all have a mean below 2. These answers indicate that senior counselors, in general, experience fewer problems with the pace, can still show interest in their colleagues, and have enough counselors besides interns in the shift. In contrast, they experience more problems due to the multitude of tasks, which can be seen by the mean of question 1, the highest mean score of all questions.

Question	Mean	Standard Deviation
1	3.03	0.82
2	2.99	0.86
3	2.28	1.01
4	2.52	0.94
5	1.85	0.77
6	1.98	0.53
8	2.52	1.06
9	2.99	0.79
10	1.91	0.79

Table 7.2: Descriptive statistics of the usable questionnaires.

Next, it was checked whether the questionnaire data contained some busy shifts; this was done by checking the number of arrivals of each shift. It is found that the questionnaires filled in by the evening shift contain the most variability in workload. The correlations found in the evening shift are presented in Figure 7.27. We found that most correlations were significant, except for the correlation between the sum score and the number of phone calls (see Table 7.3), albeit moderately correlated (i.e., around or below 0.5). The two strongest correlations are between the number of chats and the total sum score. The correlations of the percentage of unanswered chats are omitted since all chats were answered. However, it should be noted that the results of the questionnaire showed that, on average, senior counselors do not experience a high workload or seem able to work with a high workload. A higher variability in experienced workload during shifts is needed to determine the relationship between call volumes and waiting times more precisely, such that the staffing advice could be altered, and determining which β is most appropriate.

Outcome variable	Correlations with the sum score	P value
Mean waiting time phone	0.49	0.001
Mean waiting time chat	0.50	0.001
% Unanswered phone calls	0.47	0.002
Number of phone calls	0.29	0.069
Number of chats	0.61	Smaller than 0.001

Table 7.3: Correlations between the outcome variables and the questionnaires of the evening shift.

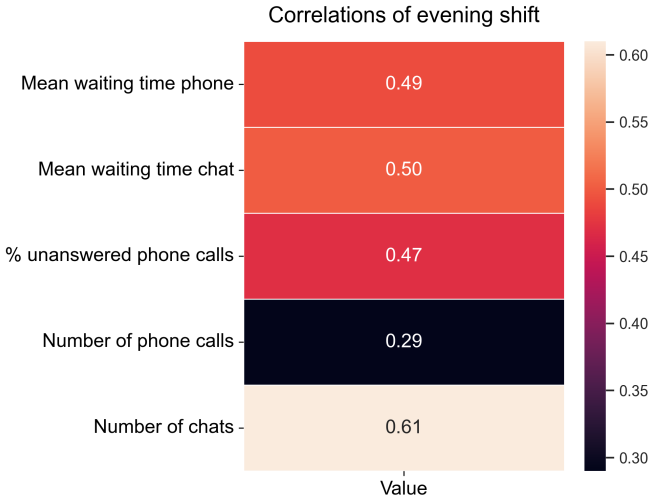


Figure 7.27: Correlations between the outcome variables and the questionnaires of the evening shift.

7.9 Conclusion and Discussion

This chapter presented a study on forecasting and modeling call and chat volumes at mental health helplines, based on real-world data from 113 in the Netherlands. The work offers several key contributions to both the academic understanding and practical improvement of helpline operations.

First, we developed a validated simulation model of the helpline that captures its essential operational features, such as triage, warm-up and cool-down periods, and abandonment behavior. The model accurately reproduces waiting-time performance for both telephone and chat arrivals using trace-driven simulation, and can be applied to other helplines with similar structures, possibly with minor adjustments. The parameters for warm-up and cool-down durations were estimated

through counselor experience; while difficult to validate, these values could in future work be linked to activity log data for greater precision.

Second, we provided a forecasting framework to predict daily and hourly chat and phone call volumes. Among the tested models, (S)ARIMA consistently achieved the best short-term forecasts, while linear regression performed better for long-term forecasting. Interestingly, LSTM models did not outperform traditional statistical models, suggesting that the arrival processes are largely driven by historical patterns and seasonal trends rather than complex nonlinear structures. These findings align with earlier studies on emergency service forecasting and support the use of time-series approaches in similar settings.

Third, we demonstrated how to integrate these two contributions for operational decision-making. Forecasts can feed directly into the simulation model to adjust staffing levels using the square-root staffing rule, enabling dynamic workforce planning. In addition, forecasts and staffing advice can be embedded into a user-friendly dashboard for use by the helpline's planning department. Several additional strategies used by 113, such as using agents in other time zones, optimizing IVR menus, and targeting shorter conversations, can further reduce waiting times and improve service availability.

Our analysis of shift-level questionnaire data revealed that counselor workload is more closely associated with chat volume than phone call volume, suggesting that chats may require a different type of attention or emotional labor. While senior counselors overall did not report high average workload, a wider range of responses would be helpful to better model strain across varying traffic conditions.

7.9.1 Limitations

Several limitations should be acknowledged. The availability and scope of the data constrain the depth of some analyses. For example, yearly cycles could not be accurately analyzed due to the relatively short data window and rapid growth in demand. The warm-up and cool-down times used in the model were based on qualitative estimates, which could be refined through more granular log data. In the questionnaire analysis, the sample size of 41 limits statistical power, especially in detecting subtle relationships between workload and staffing.

Regarding media events, we found only short-term or limited effects on arrivals. While some studies suggest that the suicide of a public figure may trigger increases in help-seeking, the impact of such events may vary based on national relevance and recency.

7.9.2 Implications

The forecasting and modeling tools developed in this study have direct operational implications. Forecasting allows for proactive staffing, helping to reduce waiting

times and counselor overload. Improved planning not only enhances service availability, but also offers counselors better pacing, including sufficient recovery time after emotionally intense conversations. Together, these tools support the long-term sustainability and effectiveness of helplines.

7.9.3 Future research

Several directions are proposed for future work. First, incorporating different help-seeker types and modeling their specific arrival patterns could yield more refined forecasts and deeper behavioral insight. Second, improving the estimation of warm-up and cool-down durations using system log data could increase model accuracy. Third, further investigation is warranted into the effects of counselor experience and professional background on service time and perceived workload, especially with larger and more diverse samples.

Lastly, the integration of modeling, forecasting, and real-time data visualization into an operational dashboard remains a promising avenue for implementation research, offering the potential for real-time decision support in helpline staffing and performance management.

8 Implications and Outlook

The thesis concerns itself with combining two different fields, quantitative methods, such as data science and queueing theory with a healthcare application: the elderly care domain. Using various methods such as process mining, simulation, queueing theory, and optimization, it examines patient flows, admission policies, and capacity allocation across multiple levels and forms of care. This thesis addresses challenges such as bed-blocking, home care utilization, and capacity allocation, showing that quantitative methods can be of tremendous value to the care domain for older adults. The need for such solutions is already crucial, but will become more important as the population will age. This section presents the implications of this thesis for the elderly care domain, as well as for other care domains, and discusses possible future directions for research or implications.

8.1 Implications

The thesis focuses on two different perspectives, namely *flow-based* and *capacity-based*. The flow-based perspective chapters demonstrate the importance of distinguishing between different types of patients and understanding the origins of care demand. The capacity-based perspective chapter demonstrates how capacity introduces issues concerning bed-blocking.

8.1.1 From risk identification to crisis prevention

Chapters 2, 3 and 4 are the chapters with a flow-based perspective. The most important implication of these chapters is the advice to *classify patients* according to different characteristics. Chapter 2 shows that different sub-populations have different healthcare usage and can be targeted in different ways, for example, the general population is easily targeted with hospital or ED care, however, the 85+ sub-population, while smaller in population size, has longer and more complex journey, reducing transitions of this sub-population could have a larger effect.

Chapter 3 further reinforces the value of differentiation by showing that a person's home care status is strongly linked to the risk of adverse outcomes. By classifying patients according to home care utilization it becomes possible to identify sub-populations with the highest risk and to focus preventive or supportive interventions where they are most needed. An example of preventive intervention based on status is demonstrated in Chapter 4. This chapter uses the health status of an individual to determine the optimal timing of nursing home admission to avoid a crisis. This example highlights how the use of patient classes can guide the setting of priorities and help reduce the overall demand for care by preventing crises and avoiding unnecessary transitions.

8.1.2 Collaboration as the key to capacity management

Chapters 5, 6 and 7 are the chapters with a capacity-based perspective. The main takeaway from these chapters for the elderly healthcare is the importance of *collaboration*. This does not mean that all the capacity should be in *one* location, but it suffices to collaborate between different locations. This advice arises mostly from Chapters 5 and 6: both chapters show that bed-blocking is a problem for the healthcare sector, but that the effect of this issue diminishes with a large number of servers (beds, for example). This large number of servers can be obtained by strong collaboration between different locations, where the shortage of capacity at one can be balanced by a surplus of capacity at the other. Chapter 6 also shows how collaboration between different healthcare forms is able to diminish the possibility of deadlocks. Lastly, Chapter 7 shows how adapting the capacity based on the demand is able to diminish costs and costly rejections.

8.2 Other Care Domains

This thesis focuses on the elderly care domain; however, many of the challenges found here are not unique to this domain. Issues such as patient classifications, which patient to give priority and bed-blocking can be found in different types of care. Acute care hospitals, mental health services, and rehabilitation programs often face similar bottlenecks, such as overcrowded emergency departments, limited bed availability, and the need to coordinate care across different facilities and specialties. The modeling approaches and analytical frameworks presented here, from process mining to queueing theory and Markov decision processes, therefore provide insights that extend well beyond elderly care.

8.3 Outlook

Future research and policy development should start with a better understanding of the older adults population. Accurately and quickly determining health status is crucial for anticipating demand and designing preventive interventions. Studies that integrate clinical data with social and demographic factors can help identify

high-risk sub-populations and reveal how care needs evolve over time. Such insights will guide the development of targeted strategies to reduce crises, prevent unnecessary transitions, and improve long-term outcomes.

Exploring the potential of pooling resources among various organizations and care settings presents a promising opportunity. By examining models for shared hospital beds, nursing home spaces, and specialized personnel, it could be possible to identify ways to achieve economies of scale that reduce bed-blocking and waiting times. Simulation studies and pilot projects might measure the effects of regional or national pooling on system resilience and patient flow.

These future directions rely on strong collaboration between organizations and types of healthcare. Future work should examine mechanisms for regional or national coordination, inspired by the lessons of the COVID-19 pandemic. Research might explore the feasibility of a "healthcare control tower," analogous to air-traffic management, to monitor real-time capacity, forecast surges, and guide patient transfers.

Bibliography

[1] I. F. Akyildiz and H. von Brand. 'Exact solutions for networks of queues with blocking-after-service'. In: *Theoretical Computer Science* 125.1 (1994), pages 111–130.

[2] G. E. Allen, P. E. Zucknick and B. L. Evans. 'A distributed deadlock detection and resolution algorithm for process networks'. In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. Volume 2. IEEE. (2007), pages II–33.

[3] F. Aminzadeh and W. B. Dalziel. 'Older adults in the emergency department: a systematic review of patterns of use, adverse outcomes, and effectiveness of interventions'. In: *Annals of Emergency Medicine* 39.3 (2002), pages 238–247.

[4] J. F. Andersen and B. F. Nielsen. 'A comparative study of time-based maintenance and condition-based maintenance for multi-component systems'. In: *Reliability Engineering & System Safety* 256 (2025), page 110759.

[5] S. Andradóttir. 'A review of random search methods'. In: *Handbook of Simulation Optimization* (2014), pages 277–292.

[6] R. J. Arntzen, R. Bekker and R. D. van der Mei. 'Preference-based allocation of patients to nursing homes'. In: *Operations Research for Health Care* 42 (2024), page 100442.

[7] R. J. Arntzen, R. Bekker, O. S. Smeekes, B. M. Buurman, H. C. Willems, S. Bhulai and R. D. van der Mei. 'Reduced waiting times by preference-based allocation of patients to nursing homes'. In: *Journal of the American Medical Directors Association* 23.12 (2022), pages 2010–2014.

[8] K. Arts, R. van Gaalen, J. van der Laan, F. Linder, J. Mol, J. van Rooijen and C. Siermann. *Berekenwijze sociaal economische status scores*. Report. Centraal Bureau voor de Statistiek (CBS), 2021.

[9] R. Assis and P. C. Marques. 'A dynamic methodology for setting up inspection time intervals in conditional preventive maintenance'. In: *Applied Sciences* 11.18 (2021), page 8715.

[10] V. Atella, A. Piano Mortari, J. Kopinska, F. Belotti, F. Lapi, C. Cricelli and L. Fontana. 'Trends in age-related disease burden and healthcare utilization'. In: *Aging Cell* 18.1 (2019), e12861.

[11] B. Avi-Itzhak and M. Yadin. 'A sequence of two servers with no intermediate queue'. In: *Management Science* 11.5 (1965), pages 553–564.

[12] J. M. A. Baber. *Queues in series with blocking*. Cardiff University (United Kingdom), 2008.

[13] T. Bäck and H.-P. Schwefel. 'An overview of evolutionary algorithms for parameter optimization'. In: *Evolutionary Computation* 1.1 (1993), pages 1–23.

[14] P. Bakx, B. Wouterse, E. van Doorslaer and A. Wong. 'Better off at home? Effects of nursing home eligibility on costs, hospitalizations and survival'. In: *Journal of Health Economics* 73 (2020), page 102354.

[15] A. Bamezai, G. Melnick and A. Nawathe. 'The cost of an emergency department visit and its relationship to emergency department volume'. In: *Annals of Emergency Medicine* 45.5 (2005), pages 483–490.

[16] *Beeld van de wijkverpleging in 2022 — igj.nl*. `https://www.igj.nl/publicaties/publicaties/2023/04/13/beeld-van-de-wijkverpleging-2022`. [Accessed 13-08-2025].

[17] D. I. Ben-Tovim, M. L. Dougherty, T. J. O'Connell and K. M. McGrath. 'Patient journeys: the process of clinical redesign'. In: *Medical Journal of Australia* 188.S6 (2008), S14–S17.

[18] P. Bergeron, J. Courteau and A. Vanasse. 'Proximity and emergency department use: multilevel analysis using administrative data from patients with cardiovascular risk factors'. In: *Canadian Family Physician* 61.8 (2015), e391–e397.

[19] G. Bertazzoni, M. Cristofani, A. Ponzanetti, A. Trabalzini, H. Attalla, C. De Vito and P. Villari. 'Scant justification for interhospital transfers: a cause of reduced efficiency in the emergency department'. In: *Emergency Medicine Journal* 25.9 (2008), pages 558–561.

[20] *Bijlage bronnen en methoden marktscan Acute zorg 2017 - Nederlandse Zorgautoriteit — puc.overheid.nl*. `https://puc.overheid.nl/nza/doc/PUC_13389_22/`. [Accessed 13-08-2025].

[21] *Blijvende zorg aanvragen vanuit de Wet langdurige zorg — ciz.nl*. `https://www.ciz.nl/`. [Accessed 08-09-2025].

[22] K. Bloomfield, Z. Wu, A. Tatton, C. Calvert, N. Peel, R. Hubbard, H. Jamieson, J. Hikaka, M. Boyd, D. Bramley et al. 'An interRAI derived frailty index predicts acute hospitalizations in older adults residing in retirement villages: a prospective cohort study'. In: *PLoS One* 17.3 (2022), e0264715.

[23] S. Bottery. *Home care in England: Views from commissioners and providers*. King's Fund, 2018.

[24] S. C. Brailsford and N. A. Hilton. 'A comparison of discrete event simulation and system dynamics for modelling health care systems'. In: *Planning for the Future: Health Service Quality and Emergency Accessibility*. Edited by J. Riley. Operational Research Applied to Health Services. Glasgow Caledonian University. Glasgow, UK, 2001.

[25] S. C. Brailsford, M. Evandrou, R. Luff, R. Shaw, J. Viana, A. Vlachantoni and R. M. Willis. 'Using system dynamics to model the social care system: simulation modeling as the catalyst in linking demography to care delivery'. In: *Proceedings of the 2012 Winter Simulation Conference*. IEEE. 2012.

[26] K. M. Bretthauer, H. S. Heese, H. Pun and E. Coe. 'Blocking in healthcare operations: a new heuristic and an application'. In: *Production and Operations Management* 20.3 (2011), pages 375–391.

[27] M. Brülhart, V. Klotzbücher, R. Lalive and S. K. Reich. 'Mental health concerns during the COVID-19 pandemic as revealed by helpline calls'. In: *Nature* 600.7887 (2021), pages 121–126.

[28] *bupaR Docs — bupaverse.github.io*. `https://bupaverse.github.io/docs/index.html`. [Accessed 16-06-2025].

[29] J. Burkell, A. Wright, B. Hoffmaster and K. Webb. 'A decision-making aid for long-term care waiting list policies: modelling first-come, first-served vs. needs-based criteria'. In: *Healthcare Management Forum*. Volume 9. 1. Elsevier. 1996, pages 35–39.

[30] S. Cardin, M. Afilalo, E. Lang, J.-P. Collet, A. Colacone, C. Tselios, J. Dankoff and A. Guttman. 'Intervention to decrease emergency department crowding: does it have an effect on return visits and hospital readmissions?' In: *Annals of Emergency Medicine* 41.2 (2003), pages 173–185.

[31] L. A. Carrasco-Ribelles, A. Roso-Llorach, M. Cabrera-Bean, A. Costa-Garrido, E. Zabaleta-del-Olmo, P. Toran-Monserrat, F. O. Pernas and C. Violan. 'Dynamics of multimorbidity and frailty, and their contribution to mortality, nursing home and home care need: a primary care cohort of 1456052 ageing people'. In: *EClinicalMedicine* 52 (2022).

[32] M. G. Carta, G. Cossu, E. Pintus, R. Zoccheddu, O. Callia, G. Conti, M. Pintus, C. I. A. Gonzalez, M. V. Massidda, G. Mura et al. 'Active elderly and health—can moderate exercise improve health and wellbeing in older adults? Protocol for a randomized controlled trial'. In: *Trials* 22.1 (2021), page 331.

[33] G. Castaneda, M. J. Lee and J. Kang. 'Risk factors for institutionalization among the elderly: a systematic literature review'. In: *Archives of Physical Medicine and Rehabilitation* 103.3 (2022), e4.

[34] *CBS Statline — azwstatline.cbs.nl*. `https://azwstatline.cbs.nl/#/AZW/nl/`. [Accessed 24-09-2025].

[35] R. D. Cebul, J. B. Rebitzer, L. J. Taylor and M. E. Votruba. 'Organizational fragmentation and care quality in the US healthcare system'. In: *Journal of Economic Perspectives* 22.4 (2008), pages 93–113.

[36] F. Cegri, F. Orfila, R. M. Abellana and M. Pastor-Valero. 'The impact of frailty on admission to home care services and nursing homes: eight-year follow-up of a community-dwelling, older adult, Spanish cohort'. In: *BMC Geriatrics* 20.1 (2020), page 281.

[37] M. Cesari, G. Gambassi, G. Abellan van Kan and B. Vellas. 'The frailty phenotype and the frailty index: different instruments for different purposes'. In: *Age and Ageing* 43.1 (2014), pages 10–12.

[38] N. L. Chappell, B. H. Dlitt, M. J. Hollander, J. A. Miller and C. McWilliam. 'Comparative costs of home care and residential care'. In: *The Gerontologist* 44.3 (2004), pages 389–400.

[39] D. Chen and K. S. Trivedi. 'Optimization for condition-based maintenance with semi-Markov decision process'. In: *Reliability Engineering & System Safety* 90.1 (2005), pages 25–29.

[40] J. Chen, M. Zhao, R. Zhou, W. Ou and P. Yao. 'How heavy is the medical expense burden among the older adults and what are the contributing factors? A literature review and problem-based analysis'. In: *Frontiers in Public Health* 11 (2023), page 1165381.

[41] M. D. Christian. 'Triage'. In: *Critical Care Clinics* 35.4 (2019), page 575.

[42] J. C. Close, S. R. Lord, E. J. Antonova, M. Martin, B. Lensberg, M. Taylor, J. Hallen and A. Kelly. 'Older people presenting to the emergency department after a fall: a population with substantial recurrent healthcare use'. In: *Emergency Medicine Journal* 29.9 (2012), pages 742–747.

[43] J. K. Cochran and A. Bharti. 'Stochastic bed balancing of an obstetrics hospital'. In: *Health Care Management Science* 9 (2006), pages 31–45.

[44] E. G. Coffman, M. Elphick and A. Shoshani. 'System deadlocks'. In: *ACM Computing Surveys* 3.2 (1971), pages 67–78.

[45] D. Contandriopoulos, M. Breton, J. Knopp-Sihota, K. McGilton, W. B. Dalziel, J. Ploeg, C. Herbert and K. McGrail. 'A realist review of the home care literature and its blind spots'. In: *Journal of Evaluation in Clinical Practice* 28.4 (2022), pages 680–689.

[46] M. Crotty, C. H. Whitehead, R. Wundke, L. C. Giles, D. Ben-Tovim and P. A. Phillips. 'Transitional care facility for elderly people in hospital awaiting a long term care bed: randomised controlled trial'. In: *BMJ* 331.7525 (2005), page 1110.

[47] J. D. Cryer and K.-S. Chan. *Time series analysis: with applications in R*. Springer, 2008.

[48] Y. Dallery and Y. Frein. 'On decomposition methods for tandem queueing networks with blocking'. In: *Operations Research* 41.2 (1993), pages 386–399.

[49] E. El-Darzi, C. Vasilakis, T. Chaussalet and P. Millard. 'A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department'. In: *Health Care Management Science* 1 (1998), pages 143–149.

[50] J. de Almeida Mello, A. Declercq, S. Cès, T. van Durme, C. van Audenhove and J. Macq. 'Exploring home care interventions for frail older people in Belgium: a comparative effectiveness study'. In: *Journal of the American Geriatrics Society* 64.11 (2016), pages 2251–2256.

[51] T. R. de Boer, S. Mérelle, S. Bhulai and R. D. van der Mei. 'A call center model for online mental health support'. In: *Proceedings of the International Conference on Prediction Solutions for Technical and Societal Systems*. Nice, France, July 2022.

[52] T. R. de Boer, R. J. Arntzen, R. Bekker, B. M. Buurman, H. C. Willems and R. D. van der Mei. 'Process mining on national health care data for the discovery of patient journeys of older adults'. In: *Journal of the American Medical Directors Association* 26.1 (2025), page 105333.

[53] T. R. de Boer, R. Bekker and R. D. van der Mei. 'After-service blocking in tandem queues'. In: *EAI International Conference on Performance Evaluation Methodologies and Tools*. 2024.

[54] T. R. de Boer, R. Bekker and R. D. van der Mei. 'Deadlock detection and resolution in queueing networks with blocking'. Submitted for publication. 2025.

[55] T. R. de Boer, R. Bekker, R. D. van der Mei and J. M. van den Nieuwenhof. 'Too soon or too late? A Markov decision process approach for preventive nursing home admission'. Submitted for publication. 2025.

[56] T. R. de Boer, S. Mérelle, S. Bhulai, R. Gilissen and R. D. van der Mei. 'Forecasting call and chat volumes at online helplines for mental health'. In: *BMC Public Health* 23.1 (2023), page 984.

[57] T. R. de Boer, S. Mérelle, S. Bhulai and R. D. van der Mei. 'Performance modeling for call centers providing online mental health support'. In: *International Journal on Advances in Life Sciences* 14.3-4 (2023), pages 152–162.

[58] J. de Gelder, J. A. Lucke, B. de Groot, A. Fogteloo, S. Anten, K. Mesri, E. Steyerberg, C. Heringhaus, G. J. Blauw and S. Mooijaart. 'Predicting adverse health outcomes in older emergency department patients: the APOP study'. In: *The Netherlands Journal of Medicine* 74.8 (2016), pages 342–52.

[59] *De Luisterlijn | — deluisterlijn.nl*. `https://www.deluisterlijn.nl/`. [Accessed 24-07-2025].

[60] Y. Deniz. 'Inpatient discharge process in a hospital; a case-study at Medisch Spectrum Twente'. Master's thesis. University of Twente, 2019.

[61] B. L. Deuermeyer, G. L. Curry, A. T. Duchowski and S. Venkatesh. 'An automatic approach to deadlock detection and resolution in discrete simulation systems'. In: *INFORMS Journal on Computing* 9.2 (1997), pages 195–205.

[62] N. Dieleman, J. Berkhout and B. Heidergott. 'A neural network approach to performance analysis of tandem lines: the value of analytical knowledge'. In: *Computers & Operations Research* 152 (2023), page 106124.

[63] Directorate-General for Employment, Social Affairs and Inclusion (European Commission). *Long-term care report: trends, challenges and opportunities in an ageing society. Volume II, Country profiles*. Technical report. Joint report prepared by the Social Protection Committee and the European Commission. Luxembourg: Publications Office of the European Union, 2021.

[64] P. Dolan. 'Modelling valuations for health states: the effect of duration'. In: *Health Policy* 38.3 (1996), pages 189–203.

[65] W. Duan-Porter, K. Ullman, C. Rosebush, L. McKenzie, K. E. Ensrud, E. Ratner, N. Greer, T. Shippee, J. E. Gaugler and T. J. Wilt. 'Interventions to prevent or delay long-term nursing home placement for adults with impairments—A systematic review of reviews'. In: *Journal of General Internal Medicine* 35 (2020), pages 2118–2129.

[66] M. Elkjær, D. L. Wolff, J. Primdahl, C. B. Mogensen, M. Brabrand and B. Gram. 'Older adults who receive homecare are at increased risk of readmission and mortality following a short ED admission: a nationally register-based cohort study'. In: *BMC Geriatrics* 21.1 (2021), page 696.

[67] D. Engler and K. Ashcraft. 'RacerX: Effective, static detection of race conditions and deadlocks'. In: *ACM SIGOPS Operating Systems Review* 37.5 (2003), pages 237–252.

[68] K. B. Ensor and P. W. Glynn. 'Stochastic optimization via grid search'. In: *Lectures in Applied Mathematics-American Mathematical Society* 33 (1997), pages 89–100.

[69] E. Flaherty and S. J. Bartels. 'Addressing the community-based geriatric healthcare workforce shortage by leveraging the potential of interprofessional teams'. In: *Journal of the American Geriatrics Society* 67.S2 (2019), S400–S408.

[70] L. P. Fried, R. A. Kronmal, A. B. Newman, D. E. Bild, M. B. Mittelmark, J. F. Polak, J. A. Robbins and J. M. Gardin. 'Risk factors for 5-year mortality in older adults: the Cardiovascular Health Study'. In: *JAMA* 279.8 (1998), pages 585–592.

[71] L. P. Fried, C. M. Tangen, J. Walston, A. B. Newman, C. Hirsch, J. Gottdiener, T. Seeman, R. Tracy, W. J. Kop, G. Burke et al. 'Frailty in older adults: evidence for a phenotype'. In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 56.3 (2001), pages M146–M157.

[72] V. Fuster. 'Changing demographics: a new approach to global health care due to the aging population'. In: *Journal of the American College of Cardiology* 69.24 (2017), pages 3002–3005.

[73] N. Gans, G. Koole and A. Mandelbaum. 'Telephone call centers: tutorial, review, and research prospects'. In: *Manufacturing & Service Operations Management* 5.2 (2003), pages 79–141.

[74] O. Garnett, A. Mandelbaum and M. Reiman. 'Designing a call center with impatient customers'. In: *Manufacturing & Service Operations Management* 4.3 (2002), pages 208–227.

[75] J. Gaughan, H. Gravelle and L. Siciliani. 'Testing the bed-blocking hypothesis: Does nursing and care home supply reduce delayed hospital discharges?' In: *Health Economics* 24 (2015), pages 32–44.

[76] N. Genet, W. Boerma, M. Kroneman, A. Hutchinson and R. Saltman. *Home Care Across Europe. Current Structure and Future Challenges*. Jan. 2013.

[77] *Geriatrische revalidatiezorg (Zvw) — zorginstituutnederland.nl*. `https://www.zorginstituutnederland.nl/Verzekerde+zorg/geriatrische-revalidatiezorg-zvw`. [Accessed 16-06-2025].

[78] E. Gijo and N. Balakrishna. 'SARIMA models for forecasting call volume in emergency services'. In: *International Journal of Business Excellence* 10.4 (2016), pages 545–561.

[79] T. Gilbert et al. 'Development and validation of a Hospital Frailty Risk Score focusing on older people in acute care settings using electronic hospital records: an observational study'. In: *The Lancet* 391.10132 (2018), pages 1775–1782.

[80] J. Gonçalves, L. Filipe and C. H. van Houtven. 'Trajectories of disability and long-term care utilization after acute health events'. In: *Journal of Aging & Social Policy* (2023), pages 1–24.

[81] W. J. Gordon and G. F. Newell. 'Cyclic queuing systems with restricted length queues'. In: *Operations Research* 15.2 (1967), pages 266–277.

[82] M. S. Gould, J. Kalafat, J. L. HarrisMunfakh and M. Kleinman. 'An evaluation of crisis hotline outcomes. Part 2: suicidal callers'. In: *Suicide and Life-Threatening Behavior* 37.3 (2007), pages 338–352.

[83] L. C. Gray, N. M. Peel, A. P. Costa, E. Burkett, A. B. Dey, P. V. Jonsson, P. Lakhan, G. Ljunggren, F. Sjostrand, W. Swoboda et al. 'Profiles of older patients in the emergency department: findings from the interRAI Multinational Emergency Department Study'. In: *Annals of Emergency Medicine* 62.5 (2013), pages 467–474.

[84] L. V. Green, P. J. Kolesar and W. Whitt. 'Coping with time-varying demand when setting staffing requirements for a service system'. In: *Production and Operations Management* 16.1 (2007), pages 13–39.

[85] A. Grigorash, S. O'Neill, R. Bond, C. Ramsey, C. Armour and M. D. Mulvenna. 'Predicting caller type from a mental health and well-being helpline: analysis of call log data'. In: *Journal of Medical Internet Research Mental Health* 5.2 (2018), e9946.

[86] *Grijze druk — cbs.nl*. `https://www.cbs.nl/nl-nl/onze-diensten/methoden/begrippen/grijze-druk`. [Accessed 03-09-2025].

[87] R. Gualandi, C. Masella, D. Viglione and D. Tartaglini. 'Exploring the hospital patient journey: what does the patient experience?' In: *PLoS One* 14.12 (2019), e0224899.

[88] C. Guo and Z. Liang. 'A predictive Markov decision process for optimizing inspection and maintenance strategies of partially observable multi-state systems'. In: *Reliability Engineering & System Safety* 226 (2022), page 108683.

[89] V. Gupta. 'Greedy algorithm for multiway matching with bounded regret'. In: *Operations Research* 72.3 (2024), pages 1139–1155.

[90] A. N. Habermann. 'Prevention of system deadlocks'. In: *Communications of the ACM* 12.7 (1969), pages 373–377.

[91] L. Hallal. 'Queuing models for long term care wait time reduction & capacity optimization: a Nova Scotia study'. Master's thesis. Halifax, NS, Canada: Saint Mary's University, 2015.

[92] C. Haraden and R. Resar. 'Patient flow in hospitals: Understanding and controlling it better'. In: *Frontiers of Health Services Management* 20.4 (2004), page 3.

[93] J. M. Harrison and R. J. Williams. 'Brownian models of feedforward queueing networks: quasireversibility and product form solutions'. In: *The Annals of Applied Probability* 2.2 (1992), pages 263–293.

[94] *Health workforce — who.int*. `https://www.who.int/health-topics/health-workforce#tab=tab_1`. [Accessed 16-06-2025].

[95] G. A. Heckman, C. Tannenbaum, A. P. Costa, K. Harkness and R. S. McKelvie. 'The journey of the frail older adult with heart failure: implications for management and health care systems'. In: *Reviews in Clinical Gerontology* 24.4 (2014), pages 269–289.

[96] *Hoe interpreteer je de SES-WOA-scores en hoe zijn deze bepaald? — cbs.nl*. `https://www.cbs.nl/nl-nl/faq/infoservice/hoe-interpreteer-je-de-ses-woa-scores-en-hoe-zijn-deze-bepaald-`. [Accessed 13-08-2025].

[97] S. Hoeck, G. François, J. Geerts, J. van der Heyden, M. Vandewoude and G. van Hal. 'Health-care and home-care utilization among frail elderly persons in Belgium'. In: *The European Journal of Public Health* 22.5 (2012), pages 671–677.

[98] G. Hoitinga, J. Macneil-Vroomen, D. Kolk, S. Rijkenberg, K. Melkert and B. M. Buurman. 'Perceptions of Frail Older Adults on Contributing Factors Causing the Onset of Crises Leading to Hospital Admissions: A Qualitative Study'. In: *Journal of Advanced Nursing* 81.8 (2025), pages 4883–4895.

[99] S. Holterman, M. Lahr, K. Wynia, M. Hettinga and E. Buskens. 'Integrated care for older adults: a struggle for sustained implementation in Northern Netherlands'. In: *International Journal of Integrated Care* 20.3 (2020), page 1.

[100] *Homepage — kindertelefoon.nl*. `https://www.kindertelefoon.nl/`. [Accessed 24-07-2025].

[101] E. O. Hoogendijk, J. Afilalo, K. E. Ensrud, P. Kowal, G. Onder and L. P. Fried. 'Frailty: implications for clinical practice and public health'. In: *The Lancet* 394.10206 (2019), pages 1365–1375.

[102] J. S. Hopstaken, L. Verweij, C. J. van Laarhoven, N. M. Blijlevens, M. W. Stommel and R. P. Hermens. 'Effect of digital care platforms on quality of care for oncological patients and barriers and facilitators for their implementation: systematic review'. In: *Journal of Medical Internet Research* 23.9 (2021), e28869.

[103] G. C. Hunt. 'Sequential arrays of waiting lines'. In: *Operations Research* 4.6 (1956), pages 674–683.

[104] R. J. Hyndman and Y. Khandakar. 'Automatic time series forecasting: the forecast package for R'. In: *Journal of Statistical Software* 27 (2008), pages 1–22.

[105] S. Ilinca and S. Calciolari. 'The patterns of health care utilization by elderly Europeans: frailty and its implications for health systems'. In: *Health Services Research* 50.1 (2015), pages 305–320.

[106] *Infographic De weg naar het verpleeghuis | ActiZ — actiz.nl*. `https://www.actiz.nl/publicaties/infographic-de-weg-naar-het-verpleeghuis`. [Accessed 15-09-2025].

[107] S. K. Inouye, Y. Zhang, R. N. Jones, P. Shi, L. A. Cupples, H. N. Calderon and E. R. Marcantonio. 'Risk factors for hospitalization among community-dwelling primary care older patients: development and validation of a predictive model'. In: *Medical Care* 46.7 (2008), pages 726–731.

[108] S. S. Isloor and T. A. Marsland. 'The deadlock problem: an overview.' In: *Computer* 13.9 (1980), pages 58–78.

[109] N. Izady and D. Worthington. 'Setting staffing requirements for time dependent queueing networks: the case of accident and emergency departments'. In: *European Journal of Operational Research* 219.3 (2012), pages 531–540.

[110] A. Jagers and E. A. van Doorn. 'On the continued Erlang loss function'. In: *Operations Research Letters* 5.1 (1986), pages 43–46.

[111] Y.-H. Jeon, J. M. Simpson, T. Comans, M. Shin, J. Fethney, H. McKenzie, T. Crawford, C. Lang and M. Inacio. 'Investigating community-based care service factors delaying residential care home admission of community dwelling older adults and cost consequence'. In: *Age and Ageing* 52.10 (2023), afad195.

[112] S. Johnson, J. Bacsu, H. Abeykoon, T. McIntosh, B. Jeffery and N. Novik. 'No place like home: a systematic review of home care for older adults in Canada'. In: *Canadian Journal on Aging/La Revue Canadienne Du Vieillissement* 37.4 (2018), pages 400–419.

[113] A. Jones, S. E. Bronskill, H. Seow, M. Junek, D. Feeny and A. P. Costa. 'Associations between continuity of primary and specialty physician care and use of hospital-based care among community-dwelling older adults with complex care needs'. In: *PLoS One* 15.6 (2020), e0234205.

[114] A. Jones, A. P. Costa, A. Pesevski and P. D. McNicholas. 'Predicting hospital and emergency department utilization among community-dwelling older adults: statistical and machine learning approaches'. In: *PLoS One* 13.11 (2018), e0206662.

[115] P. Joshi, M. Naik, K. Sen and D. Gay. 'An effective dynamic analysis for detecting generalized deadlocks'. In: *Proceedings of the eighteenth ACM SIGSOFT international symposium on Foundations of Software Engineering*. 2010, pages 327–336.

[116] R. Jouini, C. Houaidia and L. A. Saidane. 'Hidden Markov model for early prediction of the elderly's dependency evolution in ambient assisted living'. In: *Annals of Telecommunications* 78.9 (2023), pages 599–615.

[117] M. Kallestrup-Lamb, A. O. Marin, S. Menon and J. Søgaard. 'Aging populations and expenditures on health'. In: *The Journal of the Economics of Ageing* 29 (2024), page 100518.

[118] E. L. Kaplan and P. Meier. 'Nonparametric estimation from incomplete observations'. In: *Journal of the American Statistical Association* 53.282 (1958), pages 457–481.

[119] A. R. Kaushik, B. Celler and E. Ambikairajah. 'A methodology to monitor the changing trends in health status of an elderly person by developing a Markov model'. In: *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE. 2006, pages 2171–2174.

[120] A. Kelly, J. Conell-Price, K. Covinsky, I. S. Cenzer, A. Chang, W. J. Boscardin and A. K. Smith. 'Length of stay for older adults residing in nursing homes at the end of life'. In: *Journal of the American Geriatrics Society* 58.9 (2010), pages 1701–1706.

[121] S. Kerimov, I. Ashlagi and I. Gurvich. 'On the optimality of greedy policies in dynamic matching'. In: *Operations Research* 73.1 (2025), pages 560–582.

[122] A. Kihlgren, H. Sunvisson, K. Ziegert and A.-G. Mamhidir. 'Referrals to emergency departments: the process and factors that influence decision-making among community nurses'. In: *Open Journal of Nursing* 4.5 (2014), pages 366–374.

[123] N. Koizumi, E. Kuno and T. E. Smith. 'Modeling patient flows using a queuing network with blocking'. In: *Health Care Management Science* 8 (2005), pages 49–60.

[124] D. Kolk, A. F. Kruiswijk, J. L. MacNeil-Vroomen, M. L. Ridderikhof and B. M. Buurman. 'Older patients' perspectives on factors contributing to frequent visits to the emergency department: a qualitative interview study'. In: *BMC Public Health* 21.1 (2021), page 1709.

[125] *Kosten ouderenzorg, ggz en gehandicaptenzorg stijgen in 2025 | Zorginstituut Nederland — zorginstituutnederland.nl*. `https://www.zorginstituutnederland.nl/actueel/nieuws/2025/07/04/kosten-ouderenzorg-ggz-en-gehandicaptenzorg-stijgen-in-2025`. [Accessed 03-09-2025].

[126] M. Krogseth, S. Rostoft, J. Š. Benth, G. Selbæk and T. B. Wyller. 'Frailty among older patients receiving home care services'. In: *Tidsskrift for Den Norske Legeforening* (2021).

[127] V. G. Kulkarni. *Introduction to modeling and analysis of stochastic systems*. Volume 1. 3.3. Springer, 2011.

[128] S. Kundu and I. F. Akyildiz. 'Deadlock free buffer allocation in closed queueing networks'. In: *Queueing Systems* 4 (1989), pages 47–56.

[129] M. Kyrarini, F. Lygerakis, A. Rajavenkatanarayanan, C. Sevastopoulos, H. R. Nambiappan, K. K. Chaitanya, A. R. Babu, J. Mathew and F. Makedon. 'A survey of robots in healthcare'. In: *Technologies* 9.1 (2021), page 8.

[130] S.-D. Lang. 'An extended banker's algorithm for deadlock avoidance'. In: *IEEE Transactions on Software Engineering* 25.3 (1999), pages 428–432.

[131] G. Latouche and M. F. Neuts. 'Efficient algorithmic solutions to exponential tandem queues with blocking'. In: *SIAM Journal on Algebraic Discrete Methods* 1.1 (1980), pages 93–106.

[132] W. Liao, E. Pan and L. Xi. 'Preventive maintenance scheduling for repairable system with deterioration'. In: *Journal of Intelligent Manufacturing* 21 (2010), pages 875–884.

[133] J. Liebeherr and I. F. Akyildiz. 'Deadlock properties of queueing networks with finite capacities and multiple routing chains'. In: *Queueing Systems* 20.3 (1995), pages 409–431.

[134] R. Likert. 'A technique for the measurement of attitudes'. In: *Archives of Psychology* 22.140 (1932), pages 1–55.

[135] Z. Lin, J. Kang, Y. Wei and B. Zou. 'Maintenance management strategies for medical equipment in healthcare institutions: a review'. In: *BME Horizon* 2.3 (2024), pages 135–135.

[136] B. Liu, S. Wu, M. Xie and W. Kuo. 'A condition-based maintenance policy for degrading systems with age-and state-dependent operating cost'. In: *European Journal of Operational Research* 263.3 (2017), pages 879–887.

[137] T. E. Locker, S. Baston, S. M. Mason and J. Nicholl. 'Defining frequent use of an urban emergency department'. In: *Emergency Medicine Journal* 24.6 (2007), pages 398–401.

[138] F. Lundin Gurne, S. Jakobsson, E. Lidén and I. Björkman. 'District nurses' perspectives on health-promotive and disease-preventive work at primary health care centres: a qualitative study'. In: *Scandinavian Journal of Caring Sciences* 37.1 (2023), pages 153–162.

[139] *Maatwerk en microdata — cbs.nl*. `https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata`. [Accessed 16-06-2025].

[140] M. Maliapen and B. C. Dangerfield. 'A system dynamics-based simulation study for managing clinical governance and pathways in a hospital'. In: *Journal of the Operational Research Society* 61.2 (2010), pages 255–264.

[141] R. S. Mans, M. Schonenberg, M. Song, W. M. van der Aalst and P. J. Bakker. 'Application of process mining in healthcare–a case study in a Dutch hospital'. In: *International Joint Conference on Biomedical Engineering Systems and Technologies*. Springer. 2008, pages 425–438.

[142] *Marktscan acute zorg 2017 - Nederlandse Zorgautoriteit — puc.overheid.nl*. `https://puc.overheid.nl/nza/doc/PUC_3650_22/1/`. [Accessed 13-08-2025].

[143] J. Martinsson and S. Gustafsson. 'Modeling the effects of telephone nursing on healthcare utilization'. In: *International Journal of Medical Informatics* 113 (2018), pages 98–105.

[144] J. Maybin, A. Charles and M. Honeyman. 'Understanding quality in district nursing services'. In: *London: Kings Fund* (2016).

[145] K. McKenzie. 'Aging with intellectual and developmental disabilities: the effect of frailty and health instability in home care users on admission to long-term care'. In: *Ontario, Canada: Department of Public Health Sciences. Queen's University Kingston* (2015).

[146] F. Meiland, J. Danse, J. Wendte, L. Gunning-Schepers and N. Klazinga. 'Urgency coding as a dynamic tool in management of waiting lists for psychogeriatric nursing home care in the Netherlands'. In: *Health Policy* 60.2 (2002), pages 171–184.

[147] D. T. Michaeli, J. C. Michaeli, S. Albers and T. Michaeli. 'The healthcare workforce shortage of nurses and physicians: practice, theory, evidence, and ways forward'. In: *Policy, Politics, & Nursing Practice* 25.4 (2024), pages 216–227.

[148] J. K. Mokkenstorm, M. Eikelenboom, A. Huisman, J. Wiebenga, R. Gilissen, A. J. Kerkhof and J. H. Smit. 'Evaluation of the 113Online suicide prevention crisis chat service: outcomes, helper behaviors and comparison to telephone hotlines'. In: *Suicide and Life-Threatening Behavior* 47.3 (2017), pages 282–296.

[149] *Monitor Langdurige Zorg*. `https://www.monitorlangdurigezorg.nl/kerncijfers/indicatie`. [Accessed 29-01-2025].

[150] *Monitor Ouderenzorg Trendanalyses deel 2 - Nederlandse Zorgautoriteit — puc.overheid.nl*. `https://puc.overheid.nl/nza/doc/PUC_709708_22/1/`. [Accessed 16-06-2025].

[151] F. I. Mowbray, K. Aryal, E. Mercier, G. Heckman and A. P. Costa. 'Older emergency department patients: does baseline care status matter?' In: *Canadian Geriatrics Journal* 23.4 (2020), page 289.

[152] R. Moynihan, S. Sanders, Z. A. Michaleff, A. M. Scott, J. Clark, E. J. To, M. Jones, E. Kitchener, M. Fox, M. Johansson et al. 'Impact of COVID-19 pandemic on utilisation of healthcare services: a systematic review'. In: *BMJ Open* 11.3 (2021), e045343.

[153] A. M. Mudge, P. McRae, R. E. Hubbard, N. M. Peel, W. K. Lim, A. G. Barnett and S. K. Inouye. 'Hospital-associated complications of older people: a proposed multicomponent outcome for acute care'. In: *Journal of the American Geriatrics Society* 67.2 (2019), pages 352–356.

[154] S. Musich, S. Wang, K. Hawkins and A. Klemes. 'The impact of personalized preventive care on health care quality, utilization, and expenditures'. In: *Population Health Management* 19.6 (2016), pages 389–397.

[155] E. D. Mynatt, I. Essa and W. Rogers. 'Increasing the opportunities for aging in place'. In: *Proceedings on the 2000 conference on Universal Usability*. 2000, pages 65–71.

[156] J. M. Nagurney, W. Fleischman, L. Han, L. Leo-Summers, H. G. Allore and T. M. Gill. 'Emergency department visits without hospitalization are associated with functional decline in older persons'. In: *Annals of Emergency Medicine* 69.4 (2017), pages 426–433.

[157] M. Naseer, K. J. McKee, A. Ehrenberg, P. Schön and L. Dahlberg. 'Individual and contextual predictors of emergency department visits among community-living older adults: a register-based prospective cohort study'. In: *BMJ Open* 12.2 (2022), e055484.

[158] T. P. Ng, L. Feng, M. S. Z. Nyunt, A. Larbi and K. B. Yap. 'Frailty in older persons: multisystem risk factors and the Frailty Risk Index (FRI)'. In: *Journal of the American Medical Directors Association* 15.9 (2014), pages 635–642.

[159] T. Niederkrotenthaler, K.-w. Fu, P. S. Yip, D. Y. Fong, S. Stack, Q. Cheng and J. Pirkis. 'Changes in suicide rates following media reports on celebrity suicide: a meta-analysis'. In: *Journal of Epidemiology & Community Health* 66.11 (2012), pages 1037–1042.

[160] R. Noori, A. Sadegheih and M. Lotfi. 'Integrated inspection planning and preventive maintenance for a Markov deteriorating system under scenario-based demand uncertainty'. In: *International Journal of Engineering* 33.4 (2020), pages 607–620.

[161] E. Nuutila and E. Soisalon-Soininen. 'On finding the strongly connected components in a directed graph'. In: *Information Processing Letters* 49.1 (1994), pages 9–14.

[162] A. O'Cathain, E. Knowles, J. Turner, R. Maheswaran, S. Goodacre, E. Hirst et al. 'Explaining variation in emergency admissions: a mixed-methods study of emergency and urgent care systems. Southampton (UK): NIHR Journals Library; 2014'. In: *Health Services and Delivery Research* 2.48 (2014).

[163] S. O'Neill, R. R. Bond, A. Grigorash, C. Ramsey, C. Armour and M. D. Mulvenna. 'Data analytics of call log data to identify caller behaviour patterns from a mental health and well-being helpline'. In: *Health Informatics Journal* 25.4 (2019), pages 1722–1738.

[164] W. M. Ondersteuning. 'Social Support Act'. In: *Ministry of Health, Welfare & Sport (VWS). http://wetten. overheid. nl/BWBR0035362/2016-08-01* (2015).

[165] R. O. Onvural and H. Perros. 'On equivalencies of blocking mechanisms in queueing networks with blocking'. In: *Operations Research Letters* 5.6 (1986), pages 293–297.

[166] W. H. Organization. *World report on ageing and health*. World Health Organization, 2015.

[167] C. Osorio and M. Bierlaire. 'An analytic finite capacity queueing network model capturing the propagation of congestion and blocking'. In: *European Journal of Operational Research* 196.3 (2009), pages 996–1007.

[168] *Ouderen — cbs.nl*. `https://www.cbs.nl/nl-nl/visualisaties/dashboard-bevolking/leeftijd/ouderen`. [Accessed 16-06-2025].

[169] *Our World in Data — ourworldindata.org*. `https://ourworldindata.org/`. [Accessed 29-08-2025].

[170] *Over ons | 113 Zelfmoordpreventie — 113.nl*. `https://www.113.nl/over-113/over-ons`. [Accessed 24-07-2025].

[171] G. I. Palmer, P. R. Harper and V. A. Knight. 'Modelling deadlock in open restricted queueing networks'. In: *European Journal of Operational Research* 266.2 (2018), pages 609–621.

[172] K. G. Papakonstantinou and M. Shinozuka. 'Planning structural inspection and maintenance policies via dynamic programming and Markov processes. Part II: POMDP implementation'. In: *Reliability Engineering & System Safety* 130 (2014), pages 214–224.

[173] J. Partridge, M. Sbai and J. Dhesi. 'Proactive care of older people undergoing surgery'. In: *Aging Clinical and Experimental Research* 30.3 (2018), pages 253–257.

[174] B. M. Pavlyshenko. 'Machine-learning models for sales time series forecasting'. In: *Data* 4.1 (2019), page 15.

[175] J. M. Pines, P. M. Mullins, J. K. Cooper, L. B. Feng and K. E. Roth. 'National trends in emergency department use, care patterns, and quality of care of older adults in the United States'. In: *Journal of the American Geriatrics Society* 61.1 (2013), pages 12–17.

[176] P. Pires, L. Mendes, J. Mendes, R. Rodrigues and A. Pereira. 'Integrated e-healthcare system for elderly support'. In: *Cognitive Computation* 8 (2016), pages 368–384.

[177] B. Pourbabai. 'Tandem behavior of a telecommunication system with repeated calls: II, a general case without buffers'. In: *European Journal of Operational Research* 65.2 (1993), pages 247–258.

[178] M. E. Pratt. 'The future of volunteers in crisis hotline work'. PhD thesis. University of Pittsburgh, 2013.

[179] S. Reveliotis. 'Coordinating autonomy: sequential resource allocation systems for automation'. In: *IEEE Robotics & Automation Magazine* 22.2 (2015), pages 77–94.

[180] S. Reveliotis. 'Real-time management of complex resource allocation systems: necessity, achievements and further challenges'. In: *Annual Reviews in Control* 41 (2016), pages 147–158.

[181] T. R. Robbins, D. J. Medeiros and T. P. Harrison. 'Does the Erlang C model fit in real call centers?' In: *Proceedings of the 2010 Winter Simulation Conference*. IEEE. 2010, pages 2853–2864.

[182] K. Rockwood and A. Mitnitski. 'Frailty defined by deficit accumulation and geriatric medicine defined by frailty'. In: *Clinics in Geriatric Medicine* 27.1 (2011), pages 17–26.

[183] A. J. Rolfe. 'A note on marginal allocation in multiple-server service systems'. In: *Management Science* 17.9 (1971), pages 656–658.

[184] S. M. Ross. *Introduction to Probability Models*. Academic press, 2014.

[185] J. W. Rowe, T. Fulmer and L. Fried. 'Preparing for better health and health care for an aging population'. In: *JAMA* 316.16 (2016), pages 1643–1644.

[186] N. Roy, R. Dubé, C. Després, A. Freitas and F. Légaré. 'Choosing between staying at home or moving: a systematic review of factors influencing housing decisions among frail older adults'. In: *PLoS One* 13.1 (2018), e0189266.

[187] M. E. Salive. 'Multimorbidity in older adults'. In: *Epidemiologic Reviews* 35.1 (2013), pages 75–83.

[188] S. Salmi, S. Mérelle, R. Gilissen, R. D. van der Mei and S. Bhulai. 'Detecting changes in help seeker conversations on a suicide prevention helpline during the COVID- 19 pandemic: in-depth analysis using encoder representations from transformers'. In: *BMC Public Health* 22.1 (2022), page 530.

[189] M. Salminen, J. Laine, T. Vahlberg, P. Viikari, M. Wuorela, M. Viitanen and L. Viikari. 'Factors associated with institutionalization among home-dwelling patients of urgent geriatric outpatient clinic: a 3-year follow-up study'. In: *European Geriatric Medicine* 11.5 (2020), pages 745–751.

[190] J. Scerri, A. Sammut, S. Cilia Vincenti, P. Grech, M. Galea, C. Scerri, D. Calleja Bitar and S. Dimech Sant. 'Reaching out for help: calls to a mental health helpline prior to and during the COVID-19 pandemic'. In: *International Journal of Environmental Research and Public Health* 18.9 (2021), page 4505.

[191] P. Schober, C. Boer and L. A. Schwarte. 'Correlation coefficients: appropriate use and interpretation'. In: *Anesthesia & Analgesia* 126.5 (2018), pages 1763–1768.

[192] J. Scott, P. Dawson, E. Heavey, A. De Brún, A. Buttery, J. Waring and D. Flynn. 'Content analysis of patient safety incident reports for older adult patient transfers, handovers, and discharges: do they serve organizations, staff, or patients?' In: *Journal of Patient Safety* 17.8 (2021), e1744–e1758.

[193] S. D. Searle, A. Mitnitski, E. A. Gahbauer, T. M. Gill and K. Rockwood. 'A standard procedure for creating a frailty index'. In: *BMC Geriatrics* 8 (2008), pages 1–10.

[194] D.-W. Seo and H. Lee. 'Stationary waiting times in m-node tandem queues with production blocking'. In: *IEEE Transactions on Automatic Control* 56.4 (2011), pages 958–961.

[195] S. Siami-Namini, N. Tavakoli and A. S. Namin. 'A comparison of ARIMA and LSTM in forecasting time series'. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2018, pages 1394–1401.

[196] D. M. Siclovan, J. T. Bang, O. Yakusheva, M. Hamilton, K. L. Bobay, L. L. Costa, R. G. Hughes, J. Miles, S. J. Bahr and M. E. Weiss. 'Effectiveness of home health care in reducing return to hospital: evidence from a multi-hospital study in the US'. In: *International Journal of Nursing Studies* 119 (2021), page 103946.

[197] Sigra. *Krakende ketens in de zorg voor kwetsbare ouderen (in Dutch)*. Sigra, 2019.

[198] Sigra. *Samen voor betere zorg en welzijn (in Dutch)*. `https://www.sigra.nl/`. Accessed: 2025-09-04. 2025.

[199] O. S. Smeekes, T. R. de Boer, R. D. van der Mei, B. M. Buurman and H. C. Willems. 'Differentiating between home care types to identify older adults at risk of adverse health outcomes in the community'. In: *Journal of the American Medical Directors Association* 25.11 (2024), page 105257.

[200] O. S. Smeekes, T. R. De Boer, R. D. van der Mei, B. M. Buurman and H. C. Willems. 'Receiving home care forms and the risk for emergency department visits in community-dwelling Dutch older adults, a retrospective cohort study using national data'. In: *BMC Public Health* 24.1 (2024), page 1792.

[201] A. A. Smith, S. B. C. Carusone, K. Willison, T. J. Babineau, S. D. Smith, T. Abernathy, T. Marrie and M. Loeb. 'Hospitalization and emergency department visits among seniors receiving homecare: a pilot study'. In: *BMC Geriatrics* 5.1 (2005), page 9.

[202] *Sneltest Werkdruk — sneltestwerkdruk.nl*. `https://sneltestwerkdruk.nl/`. [Accessed 25-07-2025].

[203] *Spoedzorg en spoedopvang bij een crisis — regelhulp.nl*. `https://www.regelhulp.nl/onderwerpen/opvang-en-tijdelijk-verblijf/spoedzorg`. [Accessed 13-08-2025].

[204] M. R. Sterling, L. M. Kern, M. M. Safford, C. D. Jones, P. H. Feldman, G. C. Fonarow, S. Sheng, R. A. Matsouaka, A. D. DeVore, B. Lytle et al. 'Home health care use and post-discharge outcomes after heart failure hospitalizations'. In: *Heart Failure* 8.12 (2020), pages 1038–1049.

[205] T. K. Stijn Peeters, L. Hakkaart-van Roijen and T. Kanters. *Kostenhandleiding voor economische evaluaties in de gezondheidszorg: Methodologie en referentieprijzen*. Technical report. Rotterdam, Nederland: Erasmus School of Health Policy & Management (ESHPM), Institute for Medical Technology Assessment (iMTA), Erasmus Universiteit Rotterdam, 2024.

[206] B. L. Stuart. *Principles of operating systems: design & applications*. Course Technology, 2009.

[207] *Suicide — who.int*. `https://www.who.int/news-room/fact-sheets/detail/suicide`. [Accessed 24-07-2025].

[208] F. Sundram, T. Corattur, C. Dong and K. Zhong. 'Motivations, expectations and experiences in being a mental health helplines volunteer'. In: *International Journal of Environmental Research and Public Health* 15.10 (2018), page 2123.

[209] A. S. Tanenbaum and R. van Renesse. 'Distributed operating systems'. In: *ACM Computing Surveys* 17.4 (1985), pages 419–470.

[210] M. Tavakol and R. Dennick. 'Making sense of Cronbach's alpha'. In: *International Journal of Medical Education* 2 (2011), page 53.

[211] J. W. Taylor. 'Density forecasting of intraday call center arrivals using models based on exponential smoothing'. In: *Management Science* 58.3 (2012), pages 534–549.

[212] N. Tomita, K. Yoshimura and N. Ikegami. 'Impact of home and community-based services on hospitalisation and institutionalisation among individuals eligible for long-term care insurance in Japan'. In: *BMC Health Services Research* 10.1 (2010), page 345.

[213] *Totale zorgkosten Zorgverzekeringswet | Zorgcijfersdatabank.nl*. `https://www.zorgcijfersdatabank.nl/databank?infotype=zvw&label=00-totaal&tabel=B_kost&geg=jjaarNEW&item=home`. [Accessed 29-08-2025].

[214] T. M. Trebble, N. Hansi, T. Hydes, M. A. Smith and M. Baker. 'Process mapping the patient journey: an introduction'. In: *BMJ* 341 (2010).

[215] F. G. Tricomi, A. Erdélyi et al. 'The asymptotic expansion of a ratio of gamma functions'. In: *Pacific J. Math* 1.1 (1951), pages 133–142.

[216] R. Turjamaa, S. Hartikainen, M. Kangasniemi and A.-M. Pietilä. 'Living longer at home: a qualitative study of older clients' and practical nurses' perceptions of home care'. In: *Journal of Clinical Nursing* 23.21-22 (2014), pages 3206–3217.

[217] *Uitgaven en volume — monitorlangdurigezorg.nl*. `https://www.monitorlangdurigezorg.nl/kerncijfers/uitgaven-en-volume`. [Accessed 13-08-2025].

[218] M. Ukkonen, E. Jämsen, R. Zeitlin and S.-L. Pauniaho. 'Emergency department visits in older patients: a population-based survey'. In: *BMC Emergency Medicine* 19.1 (2019), page 20.

[219] A. Ungar, M. Rafanelli, I. Iacomelli, M. A. Brunetti, A. Ceccofiglio, F. Tesi and N. Marchionni. 'Fall prevention in the elderly'. In: *Clinical Cases in Mineral and Bone Metabolism* 10.2 (2013), page 91.

[220] United Nations, Department of Economic and Social Affairs, Population Division. *World population prospects 2022: summary of results*. Technical report. New York: United Nations, 2022.

[221] *United States — commonwealthfund.org*. `https://www.commonwealthfund.org/international-health-policy-center/countries/united-states`. [Accessed 13-08-2025].

[222] N. K. Valtorta, D. C. Moore, L. Barron, D. Stow and B. Hanratty. 'Older adults' social relationships and health care utilization: a systematic review'. In: *American Journal of Public Health* 108.4 (2018), e1–e10.

[223] M. van Buuren, G. J. Kommer, R. D. van der Mei and S. Bhulai. 'A simulation model for emergency medical services call centers'. In: *2015 Winter Simulation Conference (WSC)*. IEEE. 2015, pages 844–855.

[224] M. van Buuren, G. J. Kommer, R. D. van der Mei and S. Bhulai. 'EMS call center models with and without function differentiation: a comparison'. In: *Operations Research for Health Care* 12 (2017), pages 16–28.

[225] R. van Seben, K. E. Covinsky, L. A. Reichardt, J. J. Aarden, M. van der Schaaf, M. van der Esch, R. H. Engelbert, J. W. Twisk, J. A. Bosch and B. M. Buurman. 'Insight into the posthospital syndrome: a 3-month longitudinal follow up on geriatric syndromes and their association with functional decline, readmission, and mortality'. In: *The Journals of Gerontology: Series A* 75.7 (2020), pages 1403–1410.

[226] R. van Seben, L. A. Reichardt, J. J. Aarden, M. van der Schaaf, M. van der Esch, R. H. Engelbert, J. W. Twisk, J. A. Bosch, B. M. Buurman, I. Kuper et al. 'The course of geriatric syndromes in acutely hospitalized older adults: the hospital-ADL study'. In: *Journal of the American Medical Directors Association* 20.2 (2019), pages 152–158.

[227] M. van Vuuren, I. J. Adan and S. A. Resing-Sassen. 'Performance analysis of multi-server tandem queues with finite buffers and blocking'. In: *OR Spectrum* 27 (2005), pages 315–338.

[228] W. M. van der Aalst. *Data science in action*. Springer, 2016.

[229] W. M. van der Aalst. 'Process mining: overview and opportunities'. In: *ACM Transactions on Management Information Systems (TMIS)* 3.2 (2012), pages 1–17.

[230] W. M. van der Aalst. 'Process mining: discovering and improving Spaghetti and Lasagna processes'. In: *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE. 2011, pages 1–7.

[231] M. C. van der Burgt, S. Merelle, A. T. Beekman and R. Gilissen. 'The impact of COVID-19 on the suicide prevention helpline in the Netherlands'. In: *Crisis* (2022).

[232] Vektis. *1 op de 11 65-plussers heeft een ongeplande ziekenhuisopname*. `https://www.vektis.nl/intelligence/publicaties/1-op-de-11-65-plussers-heeft-een-ongeplande-ziekenhuisopname`. Accessed: 2025-09-08. 2022.

[233] P. Verboon and E. S. Nordholt. 'Simulation experiments for hot deck imputation'. In: *Statistical Data Editing, Methods and Techniques* 2 (1997), pages 22–29.

[234] K. J. Verhaegh, J. L. MacNeil-Vroomen, S. Eslami, S. E. Geerlings, S. E. de Rooij and B. M. Buurman. 'Transitional care interventions prevent hospital readmissions for adults with chronic illnesses'. In: *Health Affairs* 33.9 (2014), pages 1531–1539.

[235] D. Verver, H. Merten, P. Robben and C. Wagner. 'Care and support for older adults in the Netherlands living independently'. In: *Health & Social Care in the Community* 26.3 (2018), e404–e414.

[236] D. T. Vinton, R. Capp, S. P. Rooks, J. T. Abbott and A. A. Ginde. 'Frequent users of US emergency departments: characteristics and opportunities for intervention'. In: *Emergency Medicine Journal* 31.7 (2014), pages 526–532.

[237] E. von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche and J. P. Vandenbroucke. 'The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies'. In: *The Lancet* 370.9596 (2007), pages 1453–1457.

[238] J. Wang, M. Ying, H. Temkin-Greener, J. Shang, T. V. Caprio and Y. Li. 'Utilization and functional outcomes among Medicare home health recipients varied across living situations'. In: *Journal of the American Geriatrics Society* 69.3 (2021), pages 704–710.

[239] L. Wang, Y. Tang, F. Roshanmehr, X. Bai, F. Taghizadeh-Hesary and F. Taghizadeh-Hesary. 'The health status transition and medical expenditure evaluation of elderly population in China'. In: *International Journal of Environmental Research and Public Health* 18.13 (2021), page 6907.

[240] A. Weatherall, S. Danby, K. Osvaldsson, J. Cromdal and M. Emmison. 'Pranking in children's helpline calls'. In: *Australian Journal of Linguistics* 36.2 (2016), pages 224–238.

[241] *Welcome to Medicare — medicare.gov*. `https://www.medicare.gov/`. [Accessed 30-10-2025].

[242] *wetten.nl - Regeling - Wet langdurige zorg - BWBR0035917*. `https://wetten.overheid.nl/BWBR0035917/2024-01-01`. [Accessed 13-08-2025].

[243] *wetten.nl - Regeling - Zorgverzekeringswet - BWBR0018450*. `https://wetten.overheid.nl/BWBR0018450/2024-01-01`. [Accessed 13-08-2025].

[244] M. S. White, C. Burns and H. A. Conlon. 'The impact of an aging population in the workplace'. In: *Workplace Health & Safety* 66.10 (2018), pages 493–498.

[245] R. Whitley, D. S. Fink, J. Santaella-Tenorio and K. M. Keyes. 'Suicide mortality in Canada after the death of Robin Williams, in the context of high-fidelity to suicide reporting guidelines in the Canadian media'. In: *The Canadian Journal of Psychiatry* 64.11 (2019), pages 805–812.

[246] *Wijziging Wet langdurige zorg, Wet maatschappelijke ondersteuning 2015, Jeugdwet en Zorgverzekeringswet (34.279) — eerstekamer.nl*. `https://www.eerstekamer.nl/wetsvoorstel/34279_wijziging_wet_langdurige`. [Accessed 16-06-2025].

[247] R. C. W. J. Willems, C. H. C. Drossaert, P. Vuijk and E. T. Bohlmeijer. 'Mental wellbeing in crisis line volunteers: understanding emotional impact of the work, challenges and resources. A qualitative study'. In: *International Journal of Qualitative Studies on Health and Well-Being* 16.1 (2021), page 1986920.

[248] D. Wilson and C. Truman. 'Comparing the health services utilization of long-term-care residents, home-care recipients, and the well elderly'. In: *Canadian Journal of Nursing Research Archive* (2005), pages 138–154.

[249] *Wlz-algemeen: Hoe werkt de Wet langdurige zorg? — zorginstituutnederland.nl*. `https://www.zorginstituutnederland.nl/Verzekerde+zorg/wlz-algemeen-hoe-werkt-de-wet-langdurige-zorg`. [Accessed 29-01-2025].

[250] Y. Yamada, V. Siersma, K. Avlund and M. Vass. 'Formal home help services and institutionalization'. In: *Archives of Gerontology and Geriatrics* 54.2 (2012), e52–e56.

[251] S. Yoo, M. Cho, E. Kim, S. Kim, Y. Sim, D. Yoo, H. Hwang and M. Song. 'Assessment of hospital processes using a process mining technique: outpatient process analysis at a tertiary hospital'. In: *International Journal of Medical Informatics* 88 (2016), pages 34–43.

[252] Y. Young, J. Kalamaras, L. Kelly, D. Hornick and R. Yucel. 'Is aging in place delaying nursing home admission?' In: *Journal of the American Medical Directors Association* 16.10 (2015), 900–e1.

[253] *Zelfdoding in Nederland: een overzicht vanaf 1950 — cbs.nl*. https://www.cbs.nl/nl-nl/longread/statistische-trends/2021/zelfdoding-in-nederland-een-overzicht-vanaf-1950?onepage=true. [Accessed 24-07-2025].

[254] Y. Zhang and W.-J. Jean Yeung. 'Shifting boundaries of care in Asia: an introduction'. In: *International Journal of Sociology and Social Policy* 32.11/12 (2012), pages 612–622.

[255] Y. Zhang, S. H. Baik, A. M. Fendrick and K. Baicker. 'Comparing local and regional variation in health care spending'. In: *New England Journal of Medicine* 367.18 (2012), pages 1724–1731.

[256] *Zorgthermometer: Inzicht in de ouderenzorg — vektis.nl*. https://www.vektis.nl/intelligence/publicaties/zorgthermometer-inzicht-in-de-ouderenzorg. [Accessed 08-09-2025].

[257] N. Zychlinski, A. Mandelbaum, P. Momčilović and I. Cohen. 'Bed blocking in hospitals due to scarce capacity in geriatric institutions—cost minimization via fluid models'. In: *Manufacturing & Service Operations Management* 22.2 (2020), pages 396–411.

Summary

This thesis investigates the healthcare system for older adults using both a flow-oriented and capacity-oriented approach. A brief summary of each chapter is provided below.

Chapter 2 explores how older adults travel through the healthcare network. This chapter combines national healthcare data of older adults from 2017-2019 with a process mining technique. The results show that 44% of the older adults experienced one or more patient journeys during this period of time. These patient journeys were often short and simple, involving hospitalizations or ED visits, with 95% of journeys for those aged 65+, and 90% of those aged 85+, having four or fewer care transitions. Chronic care forms, such as home care and long-term care, account for the majority of time spent in the care system. A small group, often older and with more extensive use of drugs, follows a more complex healthcare journey. Although this sub-population is smaller, reducing the frequency of their transitions can lead to significant improvements within the healthcare system.

Chapter 3 focuses on the risks of older adults with home care on emergency department visits, acute hospitalizations, institutionalization, and mortality. We show that the type of home care influences the risks of these adverse outcomes, even after accounting for age. This is done using data on older adults' healthcare claims, revealing that older adults with personal care are at increased risk for adverse outcomes. Older adults with personal care has more than double the risk of hospitalization and recurrent admissions. The sub-population receiving nursing home care at home faces the greatest risks of institutionalization and death. These results highlight the need for prevention strategies within home care, particularly targeted toward personal care recipients, who make up a significant portion of this population.

Chapter 4 introduces a framework to determine the optimal policy for monitoring and admitting patients to the nursing home to avoid a crisis situation and minimize overall cost. By modeling health deterioration and combining realistic costs for crises, monitoring, and nursing home care, the study finds that under current

cost estimates, the optimal strategy delays nursing home admission until the worst health states. The optimal strategy recommends to increase monitoring frequency as patients deteriorate to worse health states. This framework also enables policy makers to consider the cost of preventive care and avoiding crisis situations. The model provides a basis for data-driven and individualized admission policies that balance crisis prevention and resource use in long-term care.

Chapter 5 focuses on after-service blocking in tandem queues, a mechanism that resembles bed-blocking in healthcare settings. These types of systems are notoriously difficult to analyze; therefore, we introduce a heuristic that enables fast evaluations. Our heuristic outperforms other heuristics and allows to be used in optimization and real-time decision support to support resource allocation to reduce the effect of after-service blocking.

Building on this, Chapter 6 addresses a related but more complex issue: deadlocks in queueing networks. While after-service blocking occurs when downstream capacity is temporarily unavailable, deadlocks are situations in which jobs, or patients, are directly or indirectly blocked by themselves from transitioning to their next care form. Simulation studies show that our deadlock resolution method can reduce the number of blocked servers and job rejections. A case study on elderly care in Amsterdam demonstrates that coordinated deadlock resolution can reduce patient rejections by up to 10% and wrong-bed days by 20% over a ten-year period. These results highlight how this method can directly improve patient flow without the need for additional capacity.

Chapter 7 turns to helplines, with a focus on mental health helplines. In this chapter, data from 113 Suicide Prevention is used to forecast the number of phone calls and chats on their helpline. Seasonality and trend were found to be important factors for forecasting, while the effect of media events was omitted due to the absence of effects. This forecast is then combined with a detailed simulation model of the helpline to generate a staffing advice. Incorporating counselor feedback through surveys further links objective workload measures with perceived strain, supporting better staff scheduling and well-being. The results demonstrate how predictive analytics and queueing models can improve both service quality and staff sustainability in mental health support systems.

Samenvatting

Dit proefschrift richt zich op de gezondheidszorg voor ouderen, benaderd vanuit een flow en capaciteiten perspectief. Hieronder volgt een korte samenvatting van elk hoofdstuk.

Hoofdstuk 2 onderzoekt hoe ouderen door het zorgsysteem reizen. Dit hoofdstuk maakt gebruik van nationale zorgdata van ouderen over de periode 2017–2019 en combineert dit met de techniek process mining. De resultaten laten zien dat 44% van de ouderen minstens één zorgtraject in deze periode had. Deze zorgtraject waren vaak kort en simpel, veelal bestaande uit ziekenhuisopnames of SEH-bezoeken, 95% van de trajecten van 65+'ers en 90% van de trajecten van 85+'ers bevatten vier of minder zorgtransities. Langdurige zorgvormen, zoals wijkverpleging en langdurige zorg, nemen het grootste deel van de tijd in het zorgsysteem in beslag. Een relatief kleine subpopulatie, vaak ouder en met een hoger medicijngebruik, heeft een complexere en langere zorgtrajecten. Hoewel deze groep relatief klein is, kan het verminderen van het aantal transities binnen hun trajecten leiden tot aanzienlijke verbeteringen in het zorgsysteem.

Hoofdstuk 3 richt zich op de risico's op SEH-bezoeken, acute ziekenhuisopnames, verpleeghuisopname en sterfte onder thuiszorg-gebruikers en onderzoekt of het soort thuiszorg invloed heeft op deze risico's. Declaratiedata laat zien dat die bepaalde vormen van thuiszorg ontvangen, een verhoogd risico hebben op één of meerdere van deze nadelige uitkomsten, zelfs na correctie voor leeftijd. Ouderen die ZVW ontvangen hebben tweemaal zo veel kans op acute ziekenhuisopname en heropname. Ouderen die verpleeghuiszorg thuis ontvangen, lopen het grootste risico op verpleeghuisopname en overlijden. Deze resultaten benadrukken de noodzaak van gerichte preventieve zorgstrategieën binnen de thuiszorg, in het bijzonder voor ouderen die gebruik maken van ZVW thuiszorg.

Hoofdstuk 4 introduceert een Markov-beslissingsmodel voor het bepalen van het optimale moment voor een preventieve verpleeghuisopname en een optimale strategie voor gezondheidsmonitoring, met als doel crisissituaties te voorkomen en de totale kosten te minimaliseren. Door gezondheidstoestand te modelleren als

een discrete-tijd Markov-keten en realistische kosten voor crises, monitoring en verpleeghuiszorg te combineren, kan een optimale strategie bepaald worden, deze strategie laat zien dat het optimale beleid is om opname uit te stellen tot de slechtste gezondheidsstatus. Het wordt tegelijkertijd aanbevolen om monitoringsfrequentie te verhogen naarmate de gezondheid verslechtert. Dit model stelt beleidsmakers in staat om de kosten van preventieve zorg en crisispreventie expliciet af te wegen. Het model biedt een basis voor datagedreven en geïndividualiseerde opnamebeslissingen die evenwicht brengen tussen crisispreventie en efficiënt gebruik van zorgcapaciteit.

Hoofdstuk 5 richt zich op blokkeringen na service in tandemwachtrijen, een mechanisme dat vergelijkbaar is met verkeerde bed dagen in de zorg. Deze systemen zijn moeilijk te analyseren, daarom hebben wij een nieuwe heuristiek geïntroduceerd, deze nieuwe heuristiek maakt snelle evaluaties mogelijk. De heuristiek presteert beter dan bestaande methodes en kan worden toegepast in optimalisatie en real-time beslissingsondersteuning om middelen efficiënter toe te wijzen en de effecten van after-service blocking te verminderen.

Voortbouwend hierop introduceert Hoofdstuk 6 een methode om deadlocks in wachtrijnetwerken te detecteren en op te lossen. Deadlocks zijn situaties waarin taken of patiënten, direct of indirect, elkaar blokkeren bij de overgang naar een volgende zorgvorm. Simulatiestudies tonen aan dat deze methode het aantal geblokkeerde servers en geweigerde patiënten aanzienlijk kan verminderen. Een casestudy over ouderenzorg in Amsterdam laat zien dat gecoördineerde deadlock-oplossing het aantal patiëntweigeringen met tot wel 10% en het aantal verkeerde-bed-dagen met 20% kan verlagen over een periode van tien jaar. Deze resultaten illustreren hoe het verbeteren van de doorstroming mogelijk is zonder extra capaciteit toe te voegen.

Hoofdstuk 7 richt zich op hulplijnen, in het bijzonder die voor geestelijke gezondheidszorg. Met behulp van data van 113 Zelfmoordpreventie zijn we in staat om het aantal binnenkomende telefoon- en chatgesprekken te voorspellen. Seizoensinvloeden en trends blijken belangrijke factoren voor de voorspelling, terwijl het effect van media-aandacht niet wordt meegenomen, omdat dit vaak van korte duur of niet te met was. De voorspellingen worden gecombineerd met een gedetailleerd simulatiemodel van de hulplijn om personeelsadviezen te genereren. Door feedback van hulpverleners via enquêtes te integreren, wordt bovendien een verband gelegd tussen objectieve werklast en ervaren werkdruk, wat betere roosterplanning en welzijn ondersteunt. De resultaten laten zien hoe voorspellende analyses en wachtrijmodellen zowel de kwaliteit van de dienstverlening als de duurzaamheid van medewerkers kunnen verbeteren binnen de geestelijke gezondheidszorg.