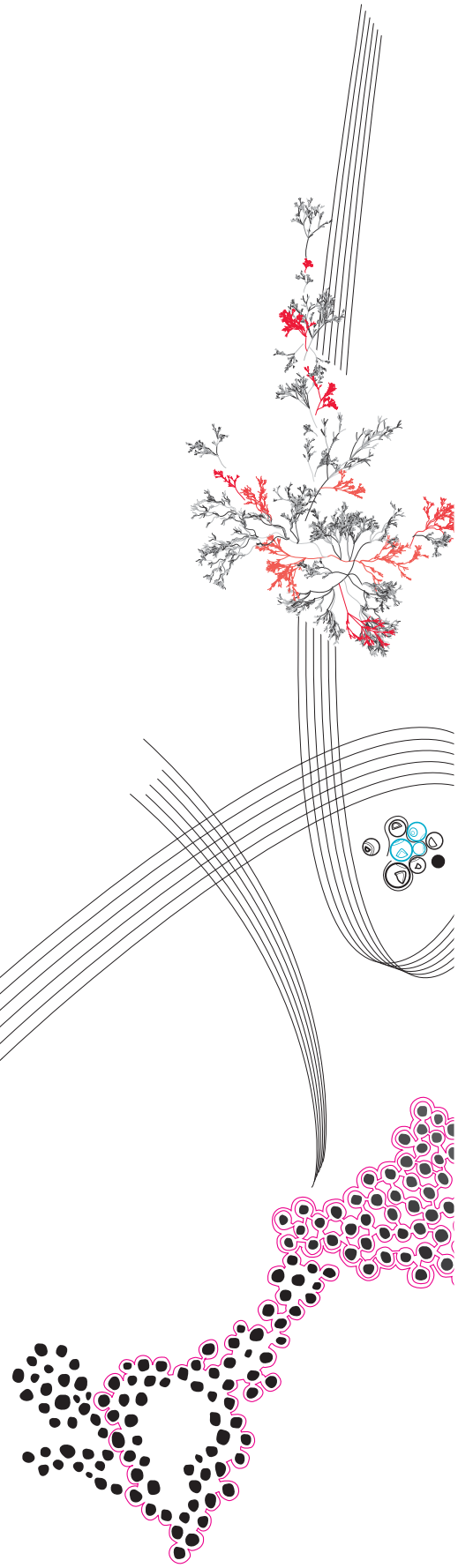


MSc Applied Mathematics
Final Project

Pandemic Preparedness for Elective Care: Managing Spare Resources by Considering Staff Absence Dynamics

Nina Baumgartner

Supervisor:
Prof. Dr. Richard J. Boucherie

Graduation Committee:
Prof. Dr. Richard J. Boucherie
Dr. Eyal Castiel
Dr. Clara Stegehuis

June, 2025

Department of Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente

UNIVERSITY OF TWENTE.

Abstract

The COVID-19 pandemic highlighted the challenge of maintaining elective care during periods of health crisis. While research has focused heavily on forecasting pandemic-related demand, less attention has been given to how hospitals can continue providing elective care when resources are limited and staff availability is uncertain. This thesis addresses that gap by analyzing the interplay between healthcare capacity, waitlist dynamics, and the postponement of elective procedures in a pandemic context.

We focus on spare nursing capacity in general wards and develop a forecasting model that accounts for both pandemic patient load and staff absence. Staff availability is modeled as a Markov chain, assuming independent infection and recovery behavior. Elective care waitlists are represented through a discrete queueing model to derive backlog distributions under varying levels of resource fluctuation. This provides a clearer understanding of how capacity constraints translate into delays in elective care.

We compare multiple elective scheduling strategies using Monte Carlo simulations based on both synthetic data and real data from the first COVID-19 wave. To cope with fluctuating staff availability, overbooking predicted capacity can improve resource utilization. However, this comes at the cost of a higher cancellation risk, which can be mitigated by shorter scheduling horizons.

Our insights offer decision-makers a basis for planning elective care in future pandemic scenarios, aiming to mitigate long-term population health consequences. The forecasting model is designed to be simple and practical. It requires only pandemic demand forecasts, staff infection rates, and observed staff availability. Using these, hospitals can estimate daily capacity for elective care and evaluate scheduling policies tailored to their specific setting.

Keywords: capacity modeling, discrete queueing theory, elective care, fluctuating resources, operations research in healthcare, pandemic preparedness, staff absence

Acknowledgements

I would like to thank my graduation committee for taking the time to read and evaluate my work.

This thesis will not only (hopefully) earn me a Master's degree, but also mark the end of my student life. Over the past years, and especially during the last three in the Netherlands, I've grown into the person I am today. It was a time filled with a lot of joy, adventure, learning (and a fair amount of tears), which I'll look back on with great nostalgia. I owe much of that to the support of many people. Without you, this experience would have been completely different.

First and foremost, I want to thank my supervisor Richard, who gave me the opportunity to work on this highly relevant and interesting topic in the first place. Thank you for all the guidance and feedback you gave me in our weekly discussions. Your critical but pragmatic perspective were invaluable in shaping this research. I appreciate that you made time for me every week - sometimes even after regular working hours - despite your schedule probably being busier than the prime minister's. I'm also thankful for the workplace you provided, which made this otherwise solitary project feel a bit more like working in a team.

On that note, thanks to the people of Zilverling 4006. It was great sharing most of my workdays with you, going on lunch walks together and having nice chats. Thanks for all the coffee your *coffee cards* (aka employee cards) sponsored me! Special thanks to Sander, with whom I not only had valuable discussions about my project, but who also provided live musical entertainment most Fridays.

I'm also grateful to have been part of the CHOIR research group and the pandemic preparedness team within SOR. Your projects are genuinely inspiring, helped me understand where my work fits in the bigger picture, and motivated me to keep going.

I'd also like to take the opportunity to express my gratitude to the UT, not only for the quality education, but for everything they do for their students. Above all, I appreciated the sports facilities, which gave me the chance to blow off steam after long days of thinking. Thanks to all the instructors who made group workouts fun, and to everyone involved in keeping the track, pool, and everything else running. Having all of this right next to my workplace really made a difference during my studies.

Jose, Sophie, Karisma, Renske, Fabio and Steven, you're not only the smartest and nerdiest friend group I've ever had, but also the reason my time in Enschede was so special. It wouldn't have been the same without our international lunches, dinners, or chaotic Secret Santa Christmas editions. Thanks go to *mijo* Jose for catching typos and weird phrasing last-minute and for helping me conquer LaTeX to fix my bibliography. *Gros bisou!*

Thanks also go to my other friends, whether we've known each other for over a decade or just met recently, you certainly made this journey a lot more enjoyable. Thanks also to those, who were there for me at some point, even if we've gone our separate ways.

Ich möchte mich spezifisch bei meiner lieben Freundin Hannah bedanken. Vielen Dank für die stundenlangen FaceTime-Calls über Gott und die Welt, sowie Notfall-Alpaka-Telefonate, wenn der Schuh mal wirklich gedrückt hat. *Donk'sche, guade Frau!*

None of this would have been possible without my parents. Euer Beitrag geht weit über die finanzielle Unterstützung hinaus, die mir dieses Auslandsstudium ermöglicht hat. Vielen Dank, dass ihr immer an mich geglaubt habt, mir all meine Freiheiten gelassen habt und es trotzdem geschafft habt, mir stets das Gefühl zu geben, unterstützt zu werden. Ich kann es nicht erwarten, den Tag meiner Graduation mit euch zu verbringen.

Special thanks go to Olof, who took the biggest hit during this thesis period - dealing with the stressed and emotional version of me. Thank you for standing by me through the whole process. You always knew what I needed, whether it was motivation to keep going or a distraction in the form of a bike ride or a round of UNO Flex (highly recommended!). If there's one person who's almost as happy about this graduation as I am, it's probably you — so congratulations to both of us!

Contents

List of Abbreviations

CI	Confidence interval
CF	Characteristic function
DALY	Disability-adjusted life year
distr.	Distribution
FIFO	First-in-first-out
ICU	Intensive Care Unit
i.i.d.	Independent and identically distributed
LAS	Late arrival system
LoS	Length of Stay
NPR	Nurse-to-patient ratio
NTD	Nursing Time Demand
OR	Operating Room
PALM	Poisson Arrival Location Model
PDF	Probability density function
PGF	Probability generating function
QALY	Quality-adjusted life year
QL	Queue length
QoL	Quality of Life
RIVM	National Institute for Public Health and the Environment *(Rijksinstituut voor Volksgezondheid en Milieu)*
SD	Standard deviation
ST	Sojourn time
WHO	World Health Organization

List of Mathematical Notation

General Mathematical Notation

$\mathbb{N}$	set of natural numbers, includes zero
f'	first derivative of f
f''	second derivative of f
$\mathbb{P}(\cdot)$	probability
$\mathbb{E}[\cdot]$	expectation
$Var[\cdot]$	variance
$G_X(z)$	probability generating function of X
$\phi_X(\theta)$	characteristic function of X
$\forall$	for all
$X \cdot Y$	dot product or matrix multiplication of matrices X and Y
X^i	matrix X raised to the power of i
$X_{i,\cdot}$	i-th row of matrix X
$X_{i,j}$	element in row i, column j of matrix X
$\binom{n}{k}$	Binomial coefficient, "n choose k"
$\lceil x \rceil$	ceiling of x, rounds up to the next integer
i	imaginary unit (also used as an index variable)
$o(\cdot)$	small-o notation
Δx	change in x

Variables

t	time index (e.g., day)
s	scheduling horizon
T	total time horizon
n	total number of nurses
h	average daily working hours per nurse

k	NTD of elective patients
p	staff infection probability
p_t	staff infection probability on day t
q	staff recovery probability

Nursing Capacity

$\hat{X}_t$	health state of a nurse on day t
$\hat{P}$	transition matrix of nurse health states under constant infection probabilities (u.c.i.p.)
$\hat{\pi}$	stationary distribution of nurse health states u.c.i.p.
$\hat{\pi}_h$	stationary probability of a nurse being healthy u.c.i.p.
$\hat{\pi}_s$	stationary probability of a nurse being sick u.c.i.p.
$\tilde{P}^{(t)}$	transition matrix of nurse health states under fluctuating infection probabilities (u.c.f.p.)
$\tilde{\pi}(t)$	distribution of nurse health states on day t u.c.f.p.
$\tilde{\pi}_h(t)$	probability of a nurse being healthy on day t u.c.f.p.
$\tilde{\pi}_s(t)$	probability of a nurse being sick on day t u.c.f.p.
X	state space of Markov chain for available nurse census
X_t	number of available nurse on day t
P	transition matrix for number of available nurses based on p and q
$P^{(t)}$	transition matrix for number of available nurses on day t based on p_t and q
C	number of available nurses/servers
C_t	number of available nurses/servers on day t
S_c	number of nurses getting sick based on current capacity of c available nurses
$S_c^{(t)}$	number of nurses getting sick on day t based on current capacity of c available nurses and infection risk p_t
R_c	number of nurses recovering based on current capacity of c available nurses
π	stationary distribution of the number of available nurses u.c.i.p.
π_i	stationary probability that i nurses are available u.c.i.p.
ρ_t	Binomial success probability in distribution of the number of available nurses on day t

Forecasts and Realized Values

$N_a^{(t,t+s)}$	forecast on day t for number of absent nurses on day $t+s$
$H_a^{(t,t+s)}$	forecast on day t for number of lost nursing hours due to absence on day $t+s$
$N_p^{(t,t+s)}$	forecast on day t for number of pandemic patients on day $t+s$
$H_p^{(t,t+s)}$	forecast on day t for amount of pandemic patient demand in nursing hours on day $t+s$
$\hat{H}_p^{(t,t+s)}$	forecast on day t for lost nursing hours due to pandemic patient demand on day $t+s$
$H_e^{(t,t+s)}$	forecast on day t for number of nursing hours available for elective care on day $t+s$
$h_a^{(t)}$	realized number of lost nursing hours due to absence on day t
$\hat{h}_p^{(t)}$	realized number of lost nursing hours due to pandemic demand on day t
$h_e^{(t)}$	realized number of nursing hours available for elective care on day t
$E^{(t)}$	number of elective patients scheduled for admission on day t
$h(E^{(t)})$	number of nursing hours for elective care scheduled for admission on day t

Queuing Theory Notation

$A/B/c$	Kendall's notation, queueing model with arrival process A, service time distribution B, and c servers
M	Markovian (e.g., Poisson arrivals, exponential service times)
G	General
λ	arrival rate
μ	service rate
ρ	utilization
Q_t	system contents at time t
D_t	departures at time t
A_t	arrivals at time t
Q	stationary system contents
$Q(i)$	stationary probability that i customers are in the system
W	stationary sojourn times

ε	$1 - \lambda$
$\hat{Q}$	εQ
$\hat{W}$	εW
b	mean service time
σ^2	variance of service times
β	exponential rate of system content distribution in heavy traffic
ω	exponential rate of sojourn time distribution in heavy traffic

Chapter 1

Introduction

1.1 Background

Elective surgery patients suffer medical conditions that, while requiring surgery, do not pose an immediate risk to life or limb, and can therefore be scheduled. However, during epidemic outbreaks, limited healthcare capacity can cause disruption of this schedule to protect critical resources or reduce transmission. During the COVID-19 pandemic, a global expert response study estimated that 72.3% of elective surgeries were canceled globally during a 12-week peak disruption of the pandemic [1]. In the Netherlands, retrospective studies revealed a 29.2% decrease of non-cancer procedures during the first COVID-19 wave [2] and a total of around 300,000 operations fewer than anticipated from 2020 to 2021 [3]. Although elective care may be non-urgent, it is not optional or unnecessary. Progression is a key feature of many surgical diseases, meaning that delayed treatment can cause a worse medical outcome. Performing surgery at a later point in time may also come at an increased cost, since more complicated procedures or alternative non-operative treatment may be required during the period of delay. Apart from the purely monetary perspective, patients suffer from a decreased quality of life (QoL) compared to timely treatment, due to worse outcome or simply because they are in the pre-operative state longer, while waiting for surgery [4]. The magnitude of this effect can be highlighted by a Dutch retrospective study that estimated 300,000 lost quality-adjusted life years (QALYs) in the population between 2020-2021 due to delayed surgery [3]. Another study applied to a large academic hospital in the Netherlands determined the extra costs of delayed surgical treatment of the 13 most commonly performed elective surgeries to be over 700,000€ after the first wave of COVID-19 [5].

The emergence of another pandemic is not a question of *if*, but *when*. Given current trends in global travel, urbanization, and increased interaction between humans and animals, outbreaks of infectious diseases are expected to become more frequent [6]. Moreover, the absence of consistent global safety standards increases the risk of accidental releases of dangerous pathogens, as for instance past incidents involving smallpox [7]. In addition to global pandemics, more localized outbreaks, such as recent cases of dengue, Mpox, and Ebola, as well as the annual burden of influenza, regularly push healthcare systems to their limits [8, 9].

Given the failures in preparedness and response during the COVID-19 pandemic [10], it is crucial to take action now to improve our readiness for future outbreaks. One important aspect that should not be neglected is the impact on elective care. Decisions made during health emergencies often have unintended consequences, including delays in surgical treatment for elective patients. Minimizing these delays and the costs associated with them

should be amongst the key goals of better preparedness planning.

1.2 Research Objectives and Contributions

This research aims to improve the care of elective patients during health emergencies by developing models and insights that support more effective planning and resource allocation.

First, we examine the relationship between hospital capacity and the backlog of elective procedures. To this end, we identify the key constraints limiting the accommodation of elective patients and investigate how these resources are affected during pandemics, drawing from existing literature.

Building on these insights, we develop a model of available capacity for elective care by integrating existing approaches to forecasting pandemic-related demand. In particular, we introduce a novel model to account for nurse absenteeism during pandemics, capturing how infection dynamics impact available staffing levels and, consequently, capacity.

Using this extended model, we evaluate policies aimed at maximizing the utilization of available resources. Our focus lies on minimizing the backlog and waiting times for elective surgeries, while also assessing the extent to which patients need to be canceled after being admitted for care.

The main contributions of this work are:

- A clearer understanding of the dynamics between healthcare capacity, backlog, and postponement of elective procedures during pandemics.
- A novel capacity model that introduces nurse absenteeism dynamics during health emergencies to forecast elective care capacity.
- An evaluation of admission policies for elective patients, assessing their impact on spare capacity utilization, waiting times, and cancellation rates after admission.

Together, these contributions offer decision-makers a foundation for planning and prioritizing elective care in future health emergencies, ultimately supporting improved population health.

1.3 Report Structure

This report is structured as follows. Chapter 2 reviews relevant literature, identifies the research gap, and drafts the solution approach. In Chapter 3, we present a model to determine spare capacity for elective care based on a nurse absenteeism model developed in Section 3.4 and pandemic patient demand. Admission policies for elective patients are introduced. Chapter 4 examines the relationship between available nursing resources and backlogs. Using a queueing theory approach, the impact of resource variability on backlog is studied. Chapter 5 evaluates admission policies based on their adaptability to fluctuating resources, measuring performance by realized workload and canceled workload. In Chapter 6, we apply the developed framework to real infection data and capacity forecasts from the first COVID-19 wave. We model a mid-sized teaching hospital in the Netherlands and assess previously developed patient admission policies.

Chapter 2

Related Work and Problem Formulation

This chapter discusses relevant literature. Section 2.1 explores methods used to estimate public health loss due to delayed elective care during the COVID-19 pandemic. Section 2.2 establishes capacity needs of elective surgery patients to identify the resource bottleneck in pandemic circumstances. The behavior of these relevant resources during a pandemic is further researched in Section 2.3.

2.1 Quantifying the Medical Impact of Delayed Surgery

Section 2.1.1 introduces the concept of quality-adjusted life-years (QALYs) as a measure of health outcomes. Section 2.1.2 then examines how the health impact of surgical treatment can be quantified using QALYs. Finally, Section 2.1.3 reviews previous attempts to assess the effects of delayed treatment during the COVID-19 pandemic and the associated population health losses.

2.1.1 The Quality-Adjusted Life-Year (QALY)

To measure cost-effectiveness of treatments in healthcare, the impact on the patient's health is often expressed in QALYs, first introduced in [11]. It uses two factors - a utility measure between 0 (dead) and 1 (full health) and the duration (in years) spent in this utility state. One QALY thus represents one life year spent in perfect health, 0.5 QALYs represent a year lived with a utility of 0.5. The utility score for specific conditions is derived from clinical trials and studies, which examine people's perception of their own health while living with a certain condition [12]. The EQ-5D guidelines constitute a standard questionnaire to determine the utility value [13].

2.1.2 Measuring the QALY Gain of Surgical Treatment

In theory, the calculation of the QALY gain as a result of surgery requires an estimate for the life expectancy *with surgery* versus *without surgery* or *with alternative treatment*. Additionally, utility weights need to be assigned to both scenarios. To calculate expected QALYs for both scenarios, the expected future life years (based on life expectancy estimates) are multiplied with the respective utility value. The difference of these two values then accounts for the QALY gain of surgery opposed to no surgery or alternative treatment [14].

A visualization of this concept can be seen in Figure 2.1. The two step functions represent the utility value over time with and without the intervention. In this graph, an earlier death without the intervention is expected than with intervention (observe the first step function hitting zero at an earlier point in time). The area under the respective curve represents the QALYs associated with the two scenarios. The blue area (A) describes the QALYs without the intervention, and the sum of the blue and yellow one ($A + B$) the QALYs with the intervention. The QALY gain of the treatment is therefore the yellow area (B), which results from the longer duration lived and higher utility during those years.

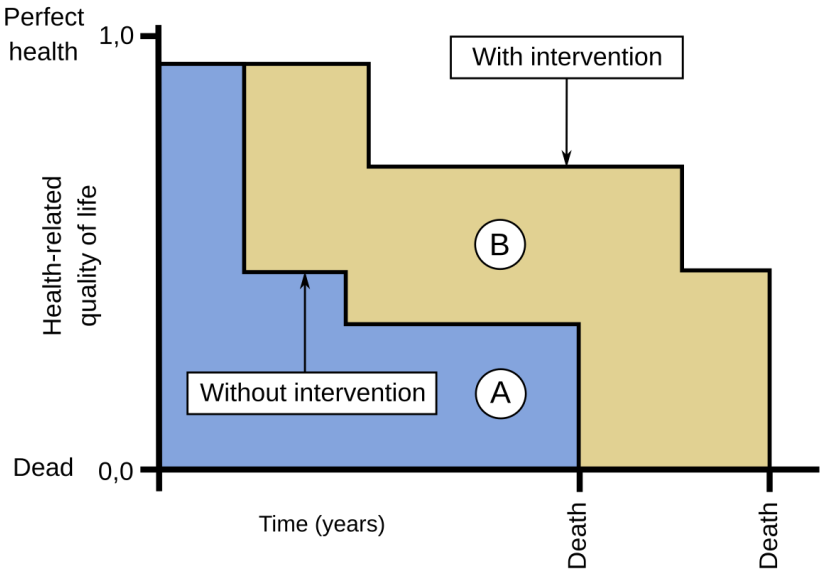


FIGURE 2.1: Schematic illustration of the QALY gain from an intervention. Adapted from [15]; CC BY-SA 3.0.

In practice, however, it is hard to evaluate the exact health gain, since the control group (*no/alternative treatment*) is very small or nonexistent for certain procedures. Thus, the way of measuring the rise in QALYs varies between different procedures. Some studies follow patients for their entire lifetime, as specified above (e.g., in a cost-utility study of of cataract surgery in [16]). Others constrain the time period to a few years (often 5 or less) after treatment, since a relatively immediate effect (after some recovery period) is expected from elective surgeries. Examples of this common approach include the analysis of hip replacement benefit in [17] or of carpal tunnel release in [18].

2.1.3 Health Losses During the COVID-19 Pandemic

The impact of delayed intervention on the QALY gain has hardly been treated in literature. A systematic review [19] found a small number of studies that relate surgical waitlists to lost QALYs and inaccurate calculation thereof in the literature.

In connection to the COVID-19 pandemic, however, there have been some efforts to measure and express the health loss in the population primarily resulting from delayed elective surgery. The authors of [5] determined the gain in QALYs of 13 commonly performed procedures in literature. To calculate the impact of delayed treatment, they defined the loss in QALYs as the difference between the utility after surgery and before surgery during the period of delay, as can be seen in Figure 2.2. That is, if a surgery improving the

utility by 0.1 is postponed for 6 months, the QALY loss due to postponement corresponds to 0.05. The authors also proposed a prioritization scheme of elective patients by weighing the QALY impact and a monetary cost factor by the surgery duration in times of scarce resources.

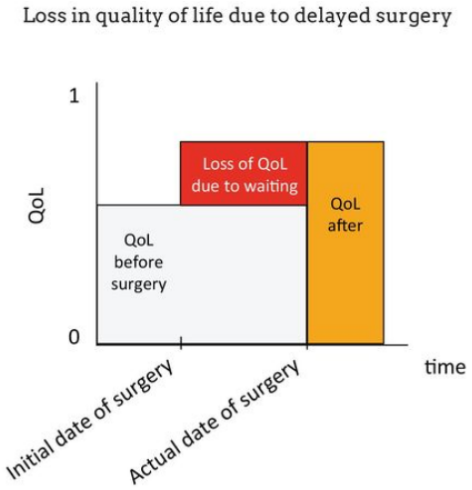


FIGURE 2.2: Definition of lost QALYs due to delayed surgery. Taken from [5].

In a report estimating the lost QALYs from postponed surgery due to COVID-19 in 2020 and 2021 conducted by the RIVM (National Institute for Public Health and the Environment) in the Netherlands [3, 20], QALY gains from all common elective surgeries (defined as surgeries that could be postponed by more than a month) of different specialties were determined from literature. Therefore, they used studies of comparable patient groups and procedures comparing the utility, preferably, to conservative/medical treatment. After determining the difference between the anticipated (based on data from the preceding years) and the performed surgeries of all types, this number was multiplied by the respective QALY gain, and denominated as the lost QALYs. To account for surgeries that were performed later in that period due to excess capacities, some of the lost QALYs could get gained back with a penalty term accounting for the prolonged time of waiting. Therefore, a linear increase in QALYs over time was supposed and the time spent on the waitlist, where patients couldn't benefit from the surgery yet times that increase were deducted from the gain. They unveiled the delay of over 300,000 elective surgeries in the Netherlands between 2020 and 2021 and, corresponding to around 320,000 lost QALYs. An overview over their results over all specialties and the bed census in the research period can be found in Figure 2.3. The contribution of individual treatments to total health losses varied significantly. The largest amount of these health losses occurred in the fields of ophthalmology and orthopedics with the most impactful surgeries in terms of QALY losses being cataract surgery, and knee and hip replacement.

In [21], a model to estimate the impact of postponing semielective surgery (defined as a surgery that ideally should be performed within 3 days to 3 weeks) on health to help prioritization. The authors used a cohort state-transition model, mathematically formulated as a Markov chain, with three states (*preop, postop*, and *dead*), which took the following input for the two states, where the patient is alive: survival rates, QoL and time until no effect of the surgery on survival or QoL. QoL was expressed in terms of disability weights, leading to the output measure of disability-adjusted life-years (DALY). Being closely related to QALY loss, one DALY corresponds to losing a full year of perfect

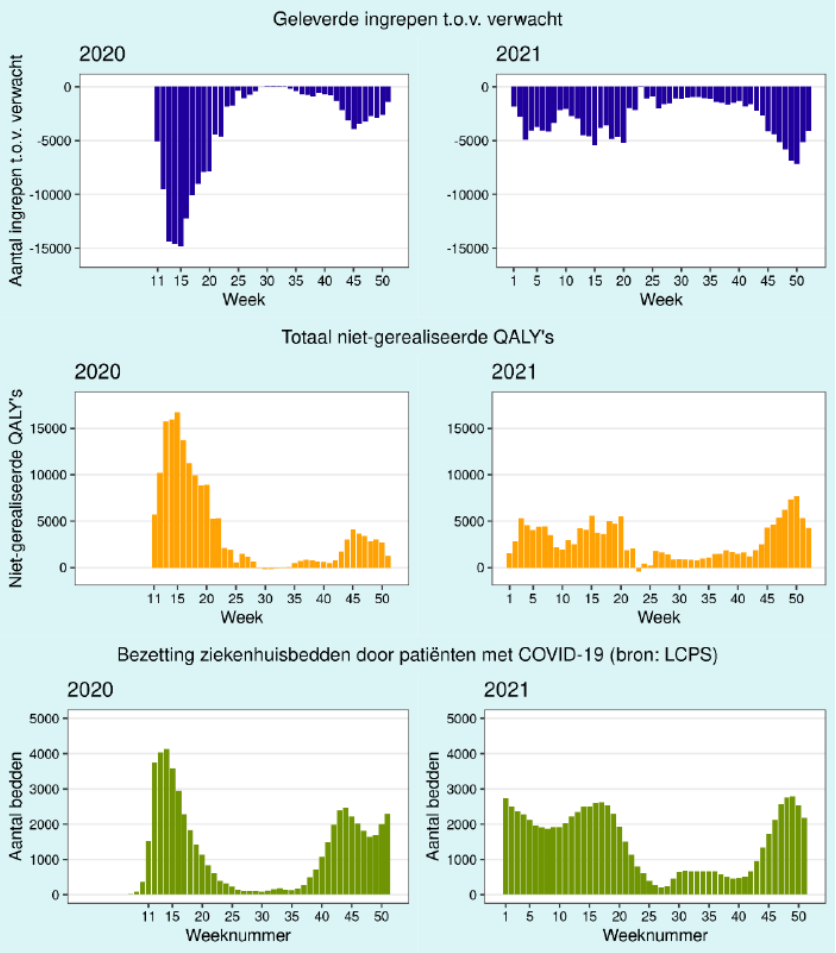


FIGURE 2.3: Difference between anticipated and performed interventions (blue), unrealized QALYs (orange) and occupation of hospital beds by COVID-19 patients (green) by week in 2020 and 2021. Taken from [20].

health. Based on the information of mean age of people undergoing this type of surgery a cohort was simulated and followed for their entire lifespan, with time intervals of one week. In this regard, delaying the surgery by up to one year was evaluated, leading to a DALY value per month delay for each investigated procedure. These insights support prioritization of surgical care in times of scarce capacities, yet have to be combined with the dynamics of capacity according to the authors.

To the best of my knowledge, there are no studies that consider a possible deterioration in the patient's utility while waiting for surgery, nor a different gain in QALYs if the surgery is performed after postponement. It is to be expected that later surgery will result in poorer patient outcomes than timely one. For instance, patients in worse initial condition (lower Oxford Hip Score) benefit more from hip replacement surgery in terms of gain in QALYs, but still don't reach the utility of patients that showed better functionality before surgery [17].

2.2 Capacity Needs of Elective Surgery Patients: Identifying the Bottleneck

In the following, we aim to establish a foundation for modeling available capacity for elective care based on literature. The focus lies on identifying the bottlenecks where pandemic and elective care intersect and compete for the same resources.

The initial decision to suspend elective surgeries during the COVID-19 pandemic primarily intended to free up intensive care unit (ICU) capacities. However, Poeran et al. [22], questioned the usefulness of this measure. They figured that only around 13% of ICU occupancy originates from elective surgery based on data of pre-pandemic years. Moreover, patients from elective care only consumed around 6% of ventilator usage. Therefore, while suspending elective surgeries certainly increased non-ICU bed capacity in hospitals, the analysis indicates that this measure had a relatively limited effect on ICU resource allocation.

In the Netherlands, the total volume of surgical procedures (elective and emergency) dropped by approximately 14% in 2020 compared to the previous year [2]. Studies such as [23] report underutilization of surgical staff, with many surgeons and trainees experiencing redundancy due to the reduced surgical load.

Based on these findings, we focus our capacity analysis on non-ICU inpatient care, more precisely, ward beds. We define capacity in terms of two main resources: beds and nursing staff, excluding other clinical constraints. As seen during the COVID-19 pandemic, increasing the number of physical beds in a healthcare institution is feasible and can be done in a timely manner. Field hospitals were rapidly deployed, such as Leishenshan hospital, which was built within under two weeks in Wuhan, China during early stages of the pandemic and had an initial capacity of 1,600 beds [24]. European hospitals repurposed operating rooms (ORs), expanded private sector capacity, or simply added more beds to rooms [25, 26].

Staffing, however, was the more critical constraint. While temporary increases in staff were achieved through redeployment and recalling retirees [26], there were clear limits to how far this could go, especially to avoid overworking people [27]. Thus, we treat staffed ward beds as the decisive bottleneck for modeling elective care capacity during pandemics.

Our focus on nursing resources in the ward does not fully determine whether an elective patient can be admitted. Other capacities, that elective patients (might) need, must also be considered. Though the probability of needing ICU care after elective surgery is low, some buffer capacity is necessary. The absenteeism model we develop for ward nurses can be extended to ICU nurses to estimate ICU capacity under pandemic conditions, taking into account both staff availability and pandemic-related ICU demand. Similarly, it can be applied to surgical staff (e.g., surgeons, anesthetists, and assistants), where pandemic demand may involve, for instance, anesthetists being reassigned to support ICU nursing teams. While our core model focuses on ward capacity, the approach remains flexible and can be applied to other critical resource types.

2.3 Hospital Capacities During Pandemic

As established in Section 2.2, elective and pandemic patients compete for certain capacities in the ward. Therefore, a reliable estimate of pandemic patient demands poses a crucial requirement to forecast spare capacities for elective patients. Section 2.3.1 explores previous work on predicting pandemic demand. In Section 2.3.2, we examine approaches by hospitals to temporarily increase their capacities to meet extraordinary high demands. Since an

important aspect of my model is the availability of staff, literature on the dynamics of pandemic staff absence is presented in Section 2.3.3.

2.3.1 Predicting Pandemic Demand

An extensive amount of literature exists on mathematical and computational modeling of epidemic outbreaks. The authors of [28, 29] present literature that apply these models to forecast influenza outbreaks, either with statistical or epidemiological techniques, at different geographical levels, and compare the measures used to do so. Reviewed literature provides valuable information for healthcare planners about expected demand, and authors identify the need for good practice and clear indication of the model's applicability.

Based on hospital data, pandemic demand was forecast and used for decision-making during the COVID-19 pandemic. In the Netherlands, the authors of [30] developed a forecast of bed occupancy by pandemic patients in wards and ICUs. They used a Poisson arrival model to estimate arrivals to the two departments, where the rate of arrivals is determined by fitting a Richard's curve to previous hospital arrival data. Lengths of Stay (LoSs) and transfer probabilities between the departments and discharge/death rates were determined data-driven as well. An infinite server queueing model was used to model the demand of beds in the ward and ICU. Using a fairly similar approach, a COVID-19 taskforce at Ghent university hospital predicted bed capacities in different wards [31]. They also assumed a Poisson arrival model, with the rate being determined via additive Poisson modeling. Bed censuses were determined for the ward, and three different levels of ICU (*midcare, standard*, and *ventilated*). In this multistate model, hospital data provided the basis for transition probabilities between the wards and for discharges. Predictions were available for up to 10 days in advance, including best and worst case bounds. The authors reported good predictions, especially for short forecasting intervals. To support the Dutch public health capacity organization, another mathematical model for bed occupancy used linear programming for admissions and queueing approach for bed occupancy [32].

The forecasting approach of [30] was further developed in [33] by an altered arrival rate prediction for the Poisson arrival model. Instead of basing the estimates solely on hospital data, the relationship between positive cases in the population and hospital admissions per day was exploited. By fitting the Richard's curve directly on positive test data and implementing a time-delay and filtration procedure, changing trends could be captured early leading to a more accurate arrival rate prediction. Based on this, they could not only help dedicating resources to pandemic patients within an individual hospital, but also balance demand between hospitals within and across regions, as presented in [33, 34].

2.3.2 Interventions to Increase Capacity

To meet changing demands, hospitals can take certain interventions to temporarily scale up their capacities to meet an unusually high demand.

In [35], authors estimate hospital capacity achieved via hospital interventions implemented during the first wave of the COVID-19 pandemic in England. They focus on capacity in terms of staff and beds for the ward and ICU and ventilators just for the ICU and analyze the effect of multiple interventions on each of these resources. The main motivation of the authors is to evaluate conditions of pandemic demand under which it is feasible to admit elective surgery patients. The interventions are the cancellation of elective surgeries, the use of field hospitals, and deployment of nursing students and retired nurses. It is one of the first studies that does not only take pandemic demand into account, but also staff absence, which was a key factor during the pandemic. While the cancellation

of elective surgeries had the biggest impact in freeing up resources, the study also shows that other measures could have helped to maintain ICU capacity to a level where elective care could continue close to normal for much of the pandemic. For ward-level care, there was even more unused capacity. In fact, even without extra interventions, patients needing only ward care could often still be treated.

This points to an important insight: there was spare capacity, both with and without interventions. This suggests a clear opportunity to improve how we use available resources. As the study highlights, staffing was a key limiting factor during the pandemic, and it's something we should pay particular attention to when planning how to maintain care under pressure.

2.3.3 Absence of Nursing Staff

The global shortage of nursing staff was already a problem in pre-pandemic times [36], and with an aging population, the demand for health care staff will rise in the coming years. With the system running at the upper limit in terms of nursing resources during regular times, we consider staff as the key limiting factor in pandemic circumstances. Multiple studies on staff absenteeism and willingness to work in a pandemic have been conducted. We present an overview of findings on nurse absence and infection dynamics in literature in the following.

In a pandemic setting, the number of absent nurses often exceeds the usual proportion due to a variety of reasons. Firstly, the willingness to work decreases due to, among other reasons, fear of infection (especially if there is a shortage of personal protective gear) [37]. In a survey about willingness to work in a pandemic flu situation, the willingness to work among nurses was determined to be around 90% in [38] given that sufficient protective measures were deployed. During COVID-19, it was also shown that the nursing workforce decreased over the course of the pandemic. A major driver for this was the increased strain put on healthcare workers, leading to decreasing work engagement, job satisfaction and susceptibility to poor mental health and burnout-like conditions [39, 40]. On top of that, a certain proportion of the nursing staff is incapable to work, with the main reasons being that nurses get infected, have to quarantine or must care for a family member [38, 41].

Next to the lower availability, the demand for nursing also changes in a pandemic setting. Because of additional workloads such as preventative safety procedures, more nurses may be needed to care for the same amount of patients compared to a non-pandemic setting. The authors of [37] therefore recommend increasing the nursing staff by 20-25% despite keeping the number of beds constant. Conversely, the ratio of nurses per patient often decreases either out of necessity (due to staff shortage) [37] or because planners specifically aim for that in crisis situations to prevent future shortages [42].

The risk of infection of healthcare workers is a well discussed topic in the literature, however, results broadly vary between studies and opinions about exposure to a pathogen deviate. Many studies report an over-representation of healthcare workers in the COVID-19-infected part of the population. In a study mainly concerning European and American countries, 14% of reported cases were healthcare workers, whereas only 3 to 8% of the population work in healthcare on average [43]. Two Dutch studies researched transmissions among medical staff during the early phase of the pandemic within a region and determined the proportion of infected employees at around 5% at that time, which is far higher than the infection rate in the general population at that time [44, 45]. A study of a single hospital in Poland reported both higher infection and quarantine rates of healthcare workers compared to the general population [46].

Exposure to the virus tends to be categorized into exposure at work and outside work

through regular community transmission [37]. Within work hours, staff are exposed to infectious patients, however, are also in contact among colleagues, where risk awareness is lower and safety measures are often neglected. The authors of [47] argue that with the aid of suitable personal protective equipment and awareness of maintaining safety measures with colleagues, the risk of infection at work can be mitigated. They therefore assume that the highest risk of infection persists during their free time and at home. On the other hand, there have also been efforts to schedule staff in a way to decrease absenteeism due to infection, based on the assumption that transmission predominantly happens during work shifts. For instance, an epidemiological nurse schedule with alternating 80-hour and 0-hour work week is proposed in [48] to reduce sick leaves assuming high infection probabilities at work with patient contact.

In the following, we list some concrete approaches that have been used to model healthcare workers' absenteeism. In [41], the risk of a nurse getting sick on a specific day of a shift is modeled as a geometric distribution with different parameters for three pre-defined risk levels for a hospital. While it is often proposed to consider different probabilities of infection at a work day and a holiday, the authors of [42] assume an infection process similar to the general population and absenteeism to care for somebody at the same rate in their planning application during pandemic flu outbreaks. Depending on the application, either a deterministic quarantine period of 2 to 3 weeks is assumed or some distribution mimicking the recommendation of the World Health Organization (WHO) to quarantine for 10 days plus if applicable 3 days after being asymptomatic. In [48], a period of immunity to infection is assumed for 120 days after COVID-19 infection. Other modeling studies, such as [49], take a parameter indicating the proportion of absent nurse as an input. During the COVID-19 pandemic, the identification of this proportion was the topic of various studies, as described in the systematic literature review [50]. However, these studies mainly determined the mean absenteeism rate of healthcare workers during a specific time period and how certain factors (demographics, vaccination status, etc.) influence it. No generalized measure of absenteeism was reported and numbers of the used measures (e.g., mean absence rate, proportions of being sick/ having suspected or confirmed COVID-19 infection within a year, proportion of staff being absent for multiple weeks) broadly vary across studies. To the best of our knowledge, no study relating the absenteeism of nurses or medical staff in general due to pandemic-related reasons directly to the infection rates in the overall population, exists. A British study, however, shows positive correlations between the weekly COVID-19 community incidence and healthcare-associated infections (infections of patients acquired in the hospital) and between self-reported COVID-19 sickness absence of staff and healthcare-associated infections [51].

This review shows how absenteeism is driven by infection risk, psychological stain and caregiving duties of nurses. To capture this key constraint for maintaining elective care in pandemics, it is required to link the availability of staff to infection trends in the general population.

2.4 Problem Formulation and Solution Approach

The COVID-19 pandemic highlighted the challenges healthcare systems face in maintaining adequate care during periods of crisis. There is broad consensus on the need for better preparedness in the future, and a significant amount of research has focused on forecasting and managing the demand for pandemic-related patients to ensure appropriate treatment. However, while there is evidence that maintaining care for elective patients is crucial to avoid negative effects on population health, less attention has been paid on the continuation

of elective care. Models to support the use of spare resources in pandemics, as well as an understanding of the dynamics of these capacities, remain limited. This study therefore addresses this gap by focusing on understanding how spare resources can be used efficiently for elective care during health crises.

In this thesis, we consider nursing staff in the ward as the key resource bottleneck for elective surgery patients in pandemic circumstances. This capacity is again influenced by two main factors: the demand of care from pandemic patients and staff absence due to unwillingness, illness or other duties. The uncertainty of these factors and the interaction between them impact how much capacity can be used for elective procedures.

To support better use of this limited capacity, we propose a model that forecasts the availability of spare resources for elective care based on the demand for pandemic patients and the availability of ward nursing staff. As a first step, we aim to analyze how uncertainty in staff availability affects elective care delivery, using queueing theory to study backlogs and waiting times. Based on these insights, we explore methods to reduce the impact of this uncertainty and improve planning and scheduling of elective surgeries during times of resource strain.

Chapter 3

Modelling Spare Capacity

We model a hospital system that admits elective patients while managing pandemic-related demand under constrained nursing capacity. A central concept in our model is the nursing time demand (NTD), which captures both the nursing effort and the patient's LoS. It is formally defined in Section 3.1. Section 3.2 presents a conceptual model of spare nursing capacity, showing which factors are taken into account to consume resources. Section 3.3 outlines the key modeling assumptions. Section 3.4 presents a precise description of how nursing staff gets infected and recovers and analyzes the resulting dynamics of the nurse census. Section 3.5 introduces the forecasting model used to predict spare nursing capacity. Finally, Section 3.6 defines the scheduling policies, which we evaluate in this thesis, specifying when and how many patients are admitted under uncertain capacity.

3.1 Definition Nursing Time

We model spare capacity for elective patients based on the availability of staffed beds in the ward. As motivated in Section 2.2, we assume that the availability of physical hospital beds is not a limiting factor and solely the available nursing staff affects how many patients can reside in the ward.

To account for a patient's nurse-level care intensity, we use the nurse-to-patient ratio (NPR), indicating how many nurses this patient needs to care for them. For instance, an NPR of 1:4 for a specific patient group means that one nurse can care for 4 patients of this type simultaneously.

Building on this, we introduce the concept of NTD of a patient, which is a measure that incorporates the NPR and LoS. Expressing these characteristics as a single value simplifies further modeling. The NTD of a patient is given by (3.1).

$$\text{NTD} = \frac{1}{\text{NPR}} \cdot \text{LoS} \tag{3.1}$$

To illustrate, consider an example patient with an NPR of 1:4 and a LoS of six days This corresponds to an NTD of 36 hours. Note that another patient with other characteristics may also require 36 nursing hours, such as a patient that has an NPR of 1:2 staying for three days.

In the following, we express the capacity of staffed beds in terms of nursing time, which represents the total amount of nursing time available to care for patients. Thus, capacity can be expressed by a single numerical value, which we refer to as nursing time capacity.

3.2 Conceptual Model

To determine spare nursing time capacity for elective patients, we consider the nursing time consumed by pandemic patients and capacities *lost* due to absence of staff. A graphical interpretation of this concept, showing how nursing capacity may evolve during a pandemic wave is provided in Figure 3.1.

In the graph, the top dashed line represents the maximum nursing capacity, assuming that all nurses are present and no pandemic patient is currently in the hospital. However, some proportion of the nursing staff is not willing to work in pandemic circumstances, as established in Section 2.3.3). Additionally, emergency patients require treatment that cannot be scheduled or delayed, requiring some nursing resources. These two capacity amounts are represented by the gray area. An additional part of nursing staff is absent despite their willingness to work, due to sickness, quarantine or care duties outside the work environment (e.g., caring for relatives). These lost capacities are shown as the red area, labeled *Absence*. Finally, resources are used by pandemic patients, who require nursing capacity as they are residing in the ward, labeled *Demand of pandemic patients*. This leaves the white area as spare capacity, that may be used for elective care.

Related behavior of nurse absenteeism and demand of pandemic patients can be expected, meaning that if the pandemic demand goes up, also the absent nurse census increases. The reasoning is the following. As infection rates increase in the population, nurses also get infected with a higher probability, resulting in an increase of the amount of absent nurses. Additionally, a larger amount of infections results in more hospital admissions (with some time delay, as established in [33]). As a result, during periods of high pandemic demand, elective care is doubly strained - both by increased pandemic patient load and elevated nurse absenteeism.

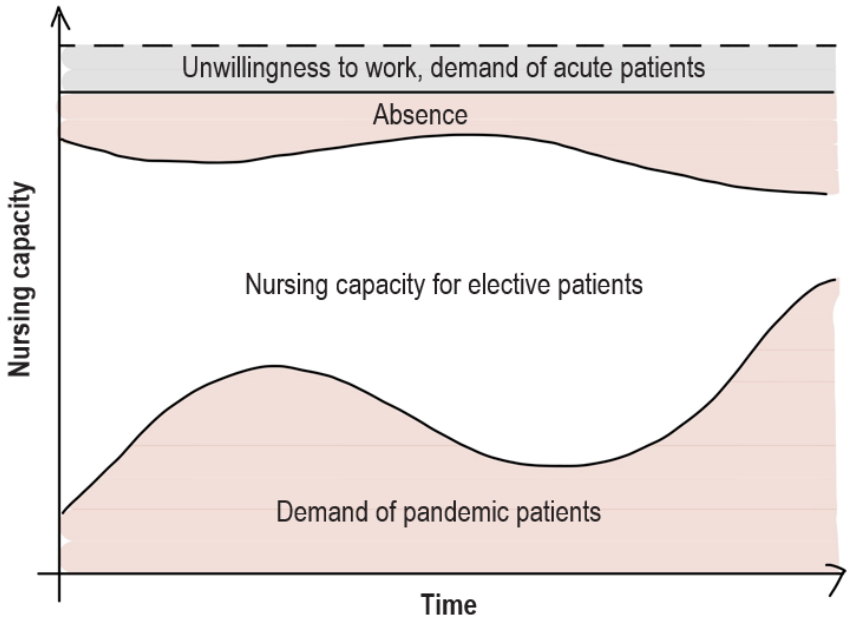


FIGURE 3.1: Conceptual model of spare nursing capacity over time.

3.3 Assumptions

Elective patients arrive according to a stochastic Poisson arrival process. Each patient is characterized by a deterministic NTD of k hours. This results in a batch Poisson arrival

process of workload, where arrivals all require the same amount of nursing time. Upon arrival, patients enter a single first-in-first-out (FIFO) waitlist. We assume a late arrival system (LAS), where patients are only eligible for scheduling the day after they have entered the system.

To determine the available capacity for elective patients, we consider absence dynamics of nurses and pandemic demand. We assume that the patterns of acute patients are only marginally influenced by the pandemic, thus consuming a fixed portion of the nursing resources. Likewise, we assume that the number of nurses, which are not willing to work, stays constant over the course of the pandemic. Thus, we define maximum nursing capacity as the nursing time provided by a nursing workforce, which is responsible for pandemic and elective care only.

We consider a nursing workforce of n full-time nurses, who each work 35 hours per week. We assume a uniform distribution of work over the week, implying that each nurse works an average of $h = \frac{35}{7} = 5$ hours per day. Hence, daily nursing capacity fluctuates in batches of 5 hours per available nurse. The maximum daily nursing capacity is given by $5n$.

Nurses are subject to infection and recovery during a pandemic. We model absenteeism using independent Bernoulli processes. Each nurse has a daily probability p_t of becoming infected and a daily probability q of recovering. We assume that within a day t, the infection probability p_t stays constant, thus being described by a piecewise constant function. We also assume that the recovery dynamics don't change over time, i.e., q remains constant and does not depend on t. In the following, we denote p_t as p, when we assume constant infection risks over time.

Further, we assume that an absence period starts and ends at the end of a given day, i.e., nurses cannot be absent for a fraction of a day. In other words, if a nurse gets infected on day t, they become absent on day $t+1$. Similarly, if they recover on day t, they return to work on day $t+1$. The mathematical model of the infection and recovery patterns is further described in Section 3.4.

For the prediction of future capacities, we assume that future staff infection risks are known, meaning we do not take a forecasting error of infection probabilities into account. This assumption allows us to isolate the dynamic interaction between scheduling policies and fluctuating resource availability.

Nursing demand required by pandemic patients is treated as an input to the model. We assume that daily forecasts of pandemic bed occupancy are available, based on existing models such as the ones described in Section 2.3.1. A fixed NPR of 1:4 is assumed for pandemic care, allowing us to convert predicted bed occupancy into required nursing time. Additionally, a nurse assigned to pandemic care cannot simultaneously work in a non-pandemic ward to prevent infection.

An elective patient can only be admitted if their full NTD fits within the available nursing capacity on a single day. Splitting care across multiple days is not allowed, leading to a service time of one day. If a patient was scheduled for treatment but sufficient capacity is not available on the scheduled day, the patient is canceled on short-notice and returned to the waitlist. Canceled patients become eligible for rescheduling immediately after cancellation, where they have priority over other patients. Since the LoS is captured in the NTD and this represents the *total* nursing effort, this assumption may lead to boundary effects. We mitigate this by an adequate choice of k in our experiments, which is further discussed in Section 5.1.2.

Admissions are planned once per day. If the timing policy is to admit s days in advance of admission day t, then no further planning for day t is possible after day $t-s$. On day t,

the realized nursing capacity becomes known, and any cancellations are carried out at the start of the day. Scheduling decisions for future days are made only after cancellations for the current day, meaning canceled patients can be rescheduled for day $t + s$ onward. An illustration of this time scheme can be seen in Figure 3.2.

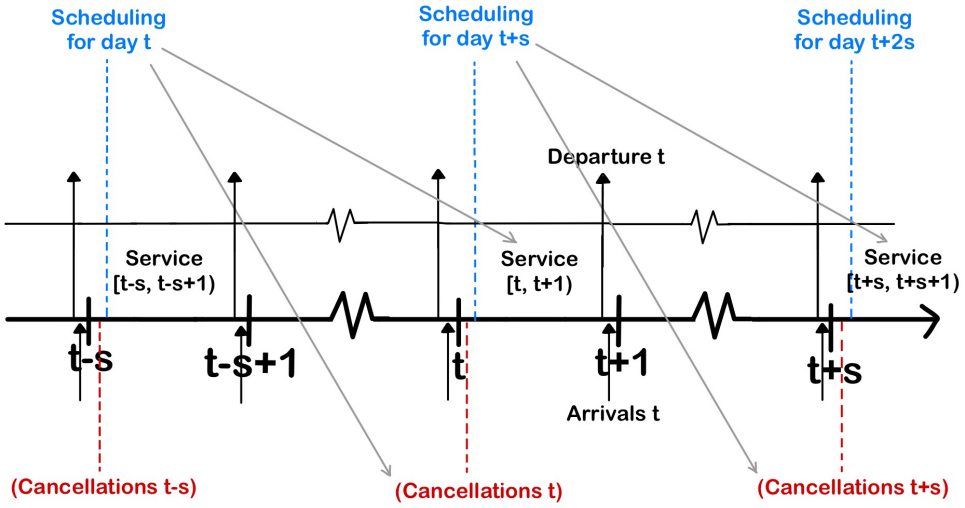


FIGURE 3.2: Time scheme illustrating patient scheduling, arrival, service, discharge, and cancellation events.

3.4 Nurse Absenteeism Model

In the following, we model the health state of a single nurse in Section 3.4.1, which helps us derive the distribution of the available nurse census in Section 3.4.2. We formulate the number of available nurses as a Markov chain, to predict future availabilities in Section 3.4.3. Along the way, examples with specific parameter values illustrate the established theory.

3.4.1 Health State of a Single Nurse

Based on our assumption of independent infection and recovery behavior of nurses, we model the health status of a single nurse as a Markov chain.

Constant Infection Risk

Transitions of a nurse getting sick and recovering are illustrated in Figure 3.3, where p and q are not time-dependent. States are shown as circles, and transitions are represented as arrows labeled with their respective probabilities.

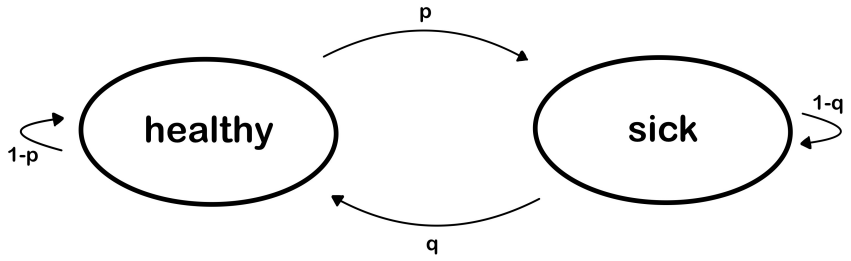


FIGURE 3.3: Markov chain of an individual nurse's health status with transition probabilities.

For the state space $\hat{X} = \{healthy, sick\}$, the corresponding transition probability matrix is

$$\hat{P} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix},$$

where the rows and columns correspond to the states in the order $\{healthy, sick\}$. This defines a homogeneous Markov chain $\{\hat{X}_t\}_{t\in\mathbb{N}}$. We now determine the stationary distribution $\hat{\pi} = [\hat{\pi}_h, \hat{\pi}_s]$, where $\hat{\pi}_h$ and $\hat{\pi}_s$ denote the probability of a nurse being healthy and sick in steady state, respectively.

A distribution $\hat{\pi}$ is stationary for a Markov chain with transition matrix $\hat{P}$ if the distribution does not change over time. Thus, $\hat{\pi}\hat{P} = \hat{\pi}$, from which we get

$$\hat{\pi}_h = \frac{q}{p}\hat{\pi}_s.$$

Using the normalizing condition $\hat{\pi}_h + \hat{\pi}_s = 1$, we obtain

$$\hat{\pi} = (\frac{q}{p+q}, \frac{p}{p+q}), \tag{3.2}$$

meaning that a nurse is healthy with probability $\frac{q}{p+q}$.

Fluctuating Infection Risk

We now consider varying infection probabilities p_t. Similarly to the setting with constant infection probabilities, nurses also get infected and recover independently from each other here. We adapt the transition probability matrix to

$$\tilde{P}^{(t)} = \begin{bmatrix} 1-p_t & p_t \\ q & 1-q \end{bmatrix}$$

This defines a non-homogeneous Markov chain, thus not reaching a stationary distribution.

Let $\tilde{\pi}(t) = [\tilde{\pi}_h(t), \tilde{\pi}_s(t)]$, be the state distribution at time t, where $\tilde{\pi}_h(t)$ and $\tilde{\pi}_s(t)$ denote the probabilities that a nurse is healthy or sick on day t, respectively. These can be calculated recursively by

$$\tilde{\pi}(t+1) = \tilde{\pi}(t) \cdot \tilde{P}^{(t)} \tag{3.3}$$

with

$$\tilde{\pi}(0) = [\tilde{\pi}_h(0), \tilde{\pi}_s(0)]$$

giving the probabilities of starting in each state. Assuming we start at the beginning of the pandemic, where a nurse is certainly healthy, $\tilde{\pi}(0) = [1, 0]$.

3.4.2 Nurse Census Distribution

We define the nursing capacity (i.e., the number of healthy nurses) at time $t+1$ by the recursion given in (3.4).

$$X_{t+1} = X_t - S_{X_t} + R_{X_t} \tag{3.4}$$

Here, X_t describes the nursing capacity at time t, and S_{X_t} and R_{X_t} are the random variables of the number of nurses who get sick and recover based on this capacity, respectively. Since these are Bernoulli processes, the number of infected and recovered nurses at each time step follow a Binomial distribution.

Constant Infection Risk

For constant infection risks, we define

$$S_c \sim Binomial(c, p) \tag{3.5}$$
$$R_c \sim Binomial(n-c, q), \tag{3.6}$$

where S_c describes the number of c healthy nurses getting sick, and R_c describes the number of $n-c$ sick nurses recovering at a certain time.

Since in stable state, every nurse is healthy with probability $\frac{q}{p+q}$ (see (3.2)), the number of healthy nurses C in steady state is Binomial distributed according to (3.7).

$$C \sim Binomial(n, \frac{q}{p+q}) \tag{3.7}$$

We now define $\pi = [\pi_0, \pi_1, ..., \pi_n]$ as the probability mass function of C, where

$$\pi_i = \binom{n}{i} \left(\frac{q}{p+q}\right)^i \left(\frac{p}{p+q}\right)^{n-i}$$

indicates the probability of i present nurses in steady state. The expected nursing capacity in steady state is therefore determined as

$$\mathbb{E}[C] = \sum_{i=0}^{n} \pi_i i = \frac{nq}{p+q}. \tag{3.8}$$

Example 3.1 illustrates the steady state distribution of available nurses for an example setting.

Example 3.1

Let us consider an example scenario, in which we have a total of $n = 100$ nurses and the probabilities of getting sick and recovering are $p = 0.05$ and $q = 0.1$, respectively. Figure 3.4 shows the steady state probability distribution, which is $Binomial(100, \frac{0.1}{0.1+0.05})$.

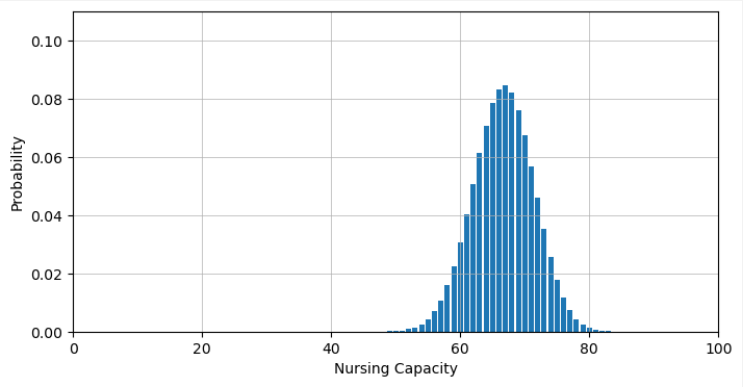


FIGURE 3.4: Steady-state distribution of healthy nurses for example scenario ($n = 100$, $p = 0.05$, $q = 0.1$).

Fluctuating Infection Risk

For time-dependent infection probabilities, the number of nurses getting sick on day t is defined as

$$S_c^{(t)} \sim Binomial(c, p_t), \tag{3.9}$$

and R_c follows the definition in (3.6).

The recursive definition for the constant case (3.4) is adapted accordingly to (3.10).

$$X_{t+1} = X_t - S_{X_t}^{(t)} + R_{X_t} \tag{3.10}$$

As each nurse is healthy on day t with probability $\tilde{\pi}_h(t)$, as determined in (3.3), the number of healthy nurses C_t is Binomial distributed according to (3.11).

$$C_t \sim Binomial(n, \tilde{\pi}_h(t)) \tag{3.11}$$

with an expected number of

$$\mathbb{E}[C_t] = n \cdot \tilde{\pi}_h(t)$$

available nurses.

Theorem 3.4.1. *Let $X_0 \sim Binomial(n, \rho_0)$ denote the number of healthy nurses at time $t = 0$, where each nurse has an independent probability ρ_0 of being healthy. For instance, $X_0 = n \sim Binomial(n, 1)$ assuming all nurses are healthy at the beginning of the pandemic. Then, for all $t \in \mathbb{N}$, the number of healthy nurses X_t follows the Binomial distribution*

$$X_t \sim Binomial(n, \rho_t)$$

with $\rho_t = \frac{q}{p+q}$ for constant infection risks and $\rho_t = \tilde{\pi}_h(t)$, as defined in (3.3), for fluctuating infection risks.

3.4.3 Predicting Available Nurses

We model the nursing capacity as a Markov chain with time epochs T as follows. We define the state space X that represents the number of nurses available, with a total workforce of n nurses.

$$T = \{0, 1, 2, ...\}$$
$$X = \{0, 1, 2, ..., n\}$$

Constant Infection Risk

Based on the recursive definition in (3.4), the transition probability to go from one to the next state is calculated by

$$\mathbb{P}(x_{t+1}|x_t) = \sum_{i=0}^{x_t} \mathbb{P}(S_{x_t} = i) \cdot \mathbb{P}(R_{x_t} = x_t - x_{t+1} + i) \qquad \forall x_t, x_{t+1} \in X,$$

which results in a $n \times n$ transition probability matrix P, with x_t referring to the row and x_{t+1} to the column.

The probability distribution of X_{t+1} given $X_t = x$, which we denote as $\mathbb{P}(X_{t+1} = i|X_t = x)$, corresponds to the x^{th} row of P, denoted as $P_{x,\cdot}$. Note that the first row is indexed by 0 (denoted $P_{0,\cdot}$), since it corresponds to transitioning from 0 present nurses. Similarly, we characterize the probability distribution over multiple time steps. Assuming we want to know the probability distribution of X_{t+s} given $X_t = x$, we need to look at $P^s_{x,\cdot}$, where P is taken to the power of s. The probability distribution of X_{t+s} given $X_t = x$ is therefore given by (3.12).

$$\mathbb{P}(X_{t+s} = i|X_t = x) = P^s_{x,i} \tag{3.12}$$

The previous example is extended with the theory above in Example 3.2.

Example 3.2

Considering the same probabilities as in Example 3.1, we now observe the behavior of the system based on a given nursing capacity on day t. Assume that today 90 nurses are present, and we want to know the probability distribution of the nursing capacity tomorrow. Formally, we compute

$$\mathbb{P}(X_{t+s} = i|X_t = 90) = P_{90,i},$$

which is shown in Figure 3.5. The expected number of nurses for day $t + 1$ is 86.5 with a standard variation of 2.33.

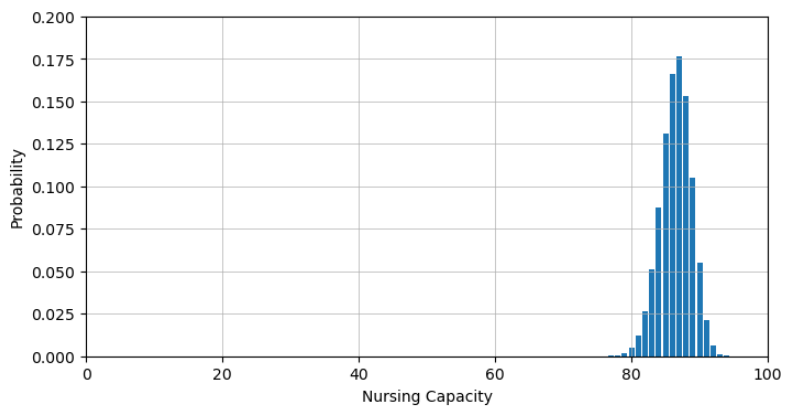


FIGURE 3.5: Probability distribution of healthy nurses for tomorrow given 90 nurses today in example scenario ($n = 100$, $p = 0.05$, $q = 0.1$).

Analyzing the probability distribution 5-days ahead, we compute

$$\mathbb{P}(X_{t+s} = i|X_t = 90) = P^5_{90,i},$$

which can be seen in Figure 3.6. The expected number of nurses for day $t + 5$ is 77.02 with a standard variation of 3.99.

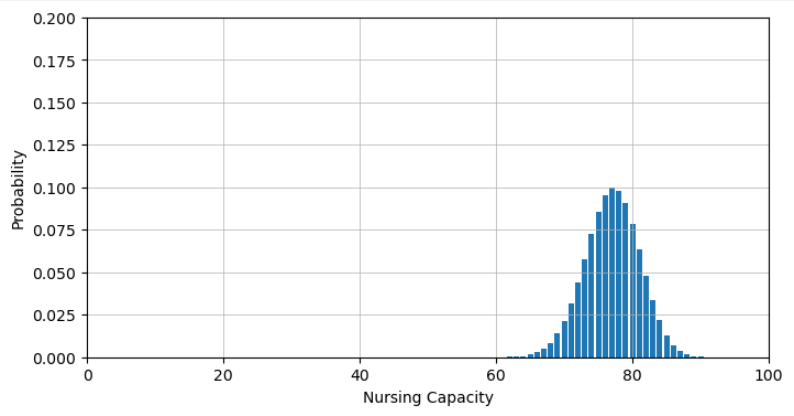


FIGURE 3.6: Probability distribution of healthy nurses in five days given 90 nurses today in example scenario ($n = 100$, $p = 0.05$, $q = 0.1$).

Comparing the two distributions, we observe a probability mass function with less variance for the nursing capacity tomorrow. This indicates that knowing today's number of present nurses provides informative insight into tomorrow's nursing capacity. Given the observed decrease in expected nursing capacity, the current probabilities suggest that more nurses are likely to become sick than to recover in the upcoming days.
Next, we observe the convergence of the expected value of nursing capacities to the expected value of the stationary distribution as given in (3.8). Figure 3.7 shows the convergence from different initial capacities 100, 80, and 20.

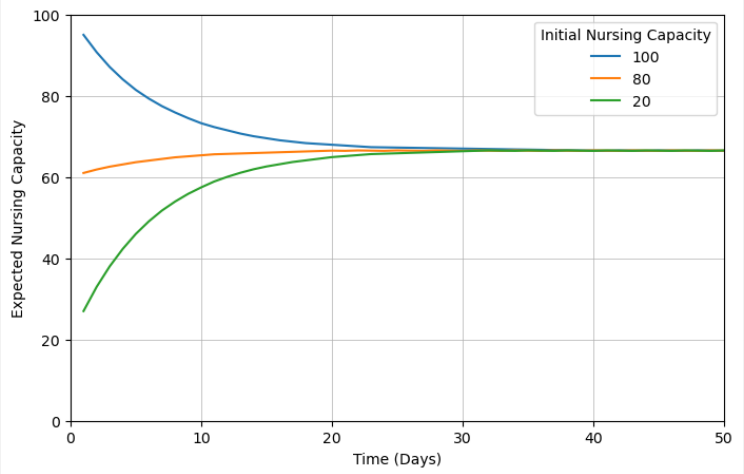


FIGURE 3.7: Convergence of the expected number of healthy nurses to steady state from starting capacities 100, 80, and 20 in example scenario ($n = 100$, $p = 0.05$, $q = 0.1$).

Even for largely different starting capacities, the expected nursing capacity converges to the stationary one in around 30 days.

Fluctuating Infection Risk

Similarly to the case with constant infection probabilities, we define a transition probability matrix $P^{(t)}$ for each day dependent on the corresponding infection risk p_t in (3.13) based on the recursive definition in (3.10).

$$\mathbb{P}(x_{t+1}|x_t) = \sum_{i=0}^{x_t} \mathbb{P}(S_{x_t}^{(t)} = i) \cdot \mathbb{P}(R_{x_t} = x_t - x_{t+1} + i), \tag{3.13}$$

where R_c follows the definition in (3.6) and $S_c^{(t)}$ the one in (3.9). This results in an $n \times n$ transition matrix $P^{(t)}$, where the ${x_t}^{\text{th}}$ row and ${x_{t+1}}^{\text{th}}$ column corresponds to transitioning from x_t to x_{t+1}.

The probability distribution of a future nursing capacity on day $t+s$ based on a known capacity on day t ($X_t = x$) can be computed by multiplying transition probability matrices $P^{(t)}$ to $P^{(t+s-1)}$. Taking the x^{th} row of the resulting matrix results in the probability distribution of the nursing capacity on day $t+s$. Formally,

$$\mathbb{P}(X_{t+s} = i|X_t = x) = P_{x,i}^{(t,t+s)} \tag{3.14}$$

with

$$P^{(t,t+s)} = P^{(t)} \cdot P^{(t+1)} \cdot \ldots \cdot P^{(t+s-1)},$$

where $X \cdot Y$ stands for the dot product of the two matrices X and Y.

To illustrate this concept, consider the following Example 3.3.

Example 3.3

Let us consider an example with varying infection probabilities, which we assume to follow a sine function, representing pandemic waves. To calculate the probability distribution of available nurses, we model the infection risk of nurses $p_t = p(t)$ on day $t \in \mathbb{N}$ as

$$p(t) = 0.1 \cdot sin(\frac{t}{5}) + 0.1 \qquad t = 1, 2, 3, \ldots$$

such that the infection risk varies between 0 and 0.2. For 50 days, values of p_t can be seen in Figure 3.8.

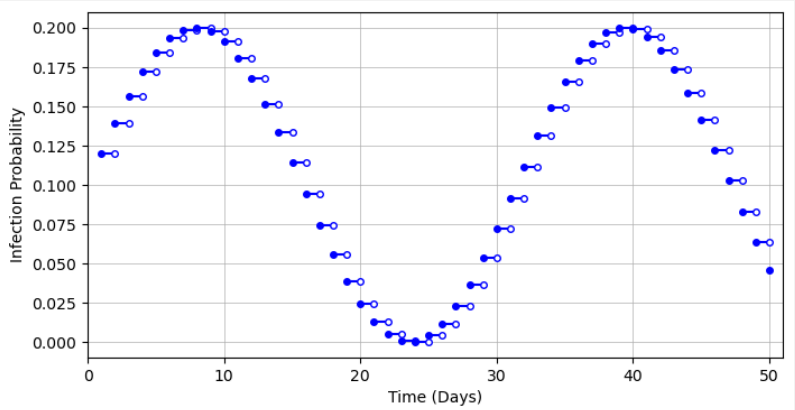
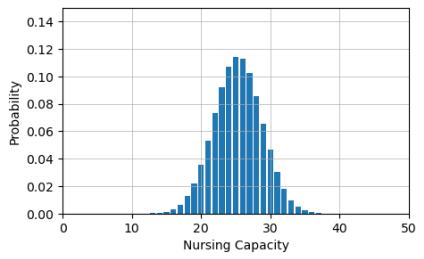
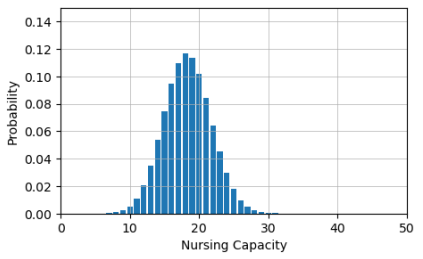


FIGURE 3.8: Daily nurse infection risk p_t used for example scenario.

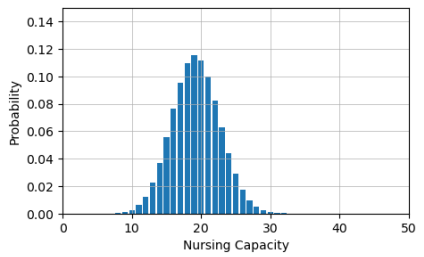
To illustrate the effect of fluctuating infection risks on the nursing capacity, we computed the probability distributions for 50 days ahead, assuming the infection risks of Figure 3.8. We assumed $n = 50$ total nurses, an unchanged constant recovery probability of $q = 0.1$, and an initial capacity of 40 nurses. The probability distribution for every 5 days can be seen in Figure 3.9.



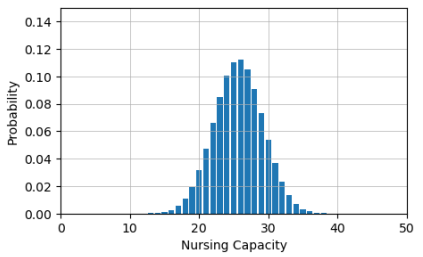
(A) Day 5.



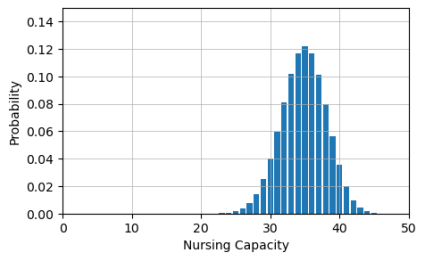
(B) Day 10.



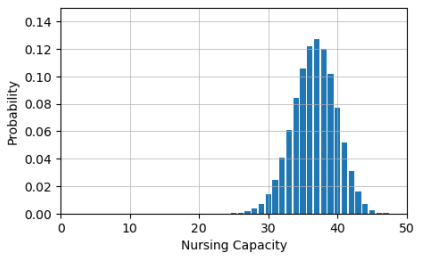
(C) Day 15.



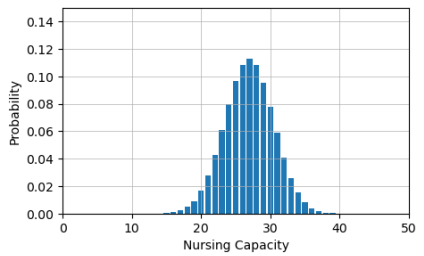
(D) Day 20.



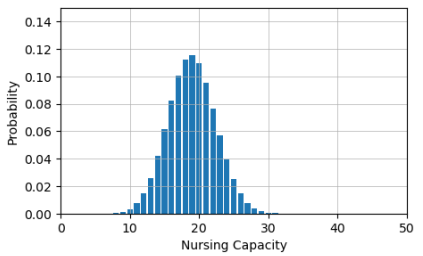
(E) Day 25.



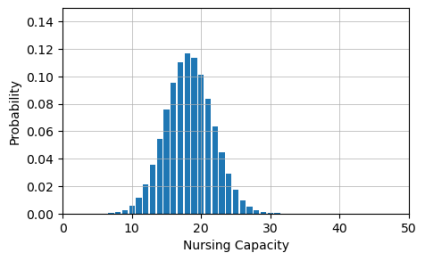
(F) Day 30.



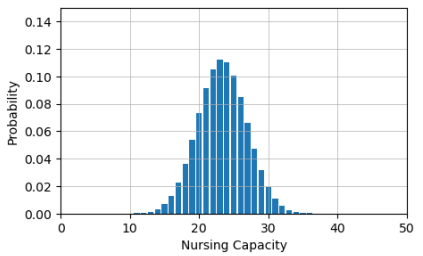
(G) Day 35.



(H) Day 40.



(I) Day 45.



(J) Day 50.

FIGURE 3.9: Probability distributions of the number of healthy nurses with varying infection risks p_t over time in example scenario ($n = 50$, $q = 0.1$).

The analysis reveals that nursing capacity tends to decrease as infection risk for nurses increases and the other way around. However, this relationship is not instantaneous. There is a noticeable time lag between the change in infection risk and its impact on available nursing capacity. For example, although the infection risk begins to rise again after day 24, the expected nursing capacity is higher on day 30 than on day 25. Conversely, after the infection risk peaks around day 40 and decreases again, nursing capacity continues to decline, reaching an even lower expected level by day 45. Regarding the width of the probability distributions, they stay fairly similar over time with a slightly narrower distribution for high nursing staff availabilities. This results from the lower variance in the underlying distributions of nurses getting sick and recovering.

3.5 Forecasting Nursing Capacity

Elective patients must be scheduled in advance for their surgery. Therefore, it is necessary to forecast available resources, as outlined in the following Sections 3.5.1 translates ab to 3.5.3.

3.5.1 Lost Nursing Time due to Absence

When a nurse is sick on a given day, h nursing hours are lost. In reality, nurses typically work four days per week and are off the remaining three, meaning that a sick nurse may result in either roughly 9 or 0 hours lost, depending on whether they were scheduled to work. However, over the long term and across a large workforce, this variation averages out, and it is reasonable to model the loss as h hours per absent nurse per day.

To forecast the number of absent nurses for day $t+s$ as seen from day t, denoted $N_a^{(t,t+s)}$, we use the framework developed in Section 3.4 and take the expectation of the probability distribution of the nursing capacity on day $t+s$. Formally,

$$\begin{aligned} N_a^{(t,t+s)} &= \mathbb{E}[X_{t+s} | X_t = x] \\ &= \sum_{i=1}^{n} \mathbb{P}(X_{t+s} = i | X_t = x) \cdot i \end{aligned}$$

with $\mathbb{P}(X_{t+s} = i | X_t = x)$ as defined in (3.12) and (3.14) for constant and fluctuating infection probabilities, respectively.

Translating this into lost nursing hours due to absence, we get

$$H_a^{(t,t+s)} = h \cdot N_a^{(t,t+s)}. \tag{3.15}$$

3.5.2 Demand of Pandemic Patients

Recall that we treat the prediction of pandemic bed occupancy as an input variable for our model. Let $N_p^{(t,t+s)}$ denote the predicted number of occupied beds on day $t+s$, based on a forecast made on day t. To translate the number of occupied beds into nursing time demand for the respective day, denoted $H_p^{(t,t+s)}$, we multiply by the average NPR for pandemic patients NPR_p.

$$H_p^{(t,t+s)} = \text{NPR}_p \cdot N_p^{(t,t+s)}.$$

Recall that a nurse assigned to pandemic care cannot simultaneously work in a non-pandemic ward. Therefore, the demand of nursing hours of pandemic patients must be rounded up to the next higher multiple of h, as shown in (3.16).

$$\hat{H}_p^{(t,t+s)} = \left\lceil \frac{H_p^{(t,t+s)}}{h} \right\rceil \cdot h, \tag{3.16}$$

where $\lceil \cdot \rceil$ denotes the ceiling function.

3.5.3 Remaining Nursing Time for Elective Care

The prediction for the available nursing capacity for elective care on day $t+s$, made on day t, is then defined as

$$H_e^{(t,t+s)} = h \cdot n - H_a^{(t,t+s)} - \hat{H}_p^{(t,t+s)}, \tag{3.17}$$

where $H_a^{(t,t+s)}$ and $\hat{H}_p^{(t,t+s)}$ are defined in (3.15) and (3.16), respectively. The expected nursing times lost due to absence and demanded by the pandemic ward are subtracted from the amount of nursing hours of the total nursing workforce $h \cdot n$ per day.

We write the realized number of remaining nursing hours on day t as

$$h_e^{(t)} = h \cdot n - h_a^{(t)} - \hat{h}_p^{(t)},$$

where $h_a^{(t)}$ and $\hat{h}_p^{(t)}$ are the realized resources lost in absence or going into pandemic care, respectively.

3.6 Policies for Admitting Elective Care

Let $E^{(t)}$ denote the number of elective patients scheduled for admission on day t, requiring a total of $h(E^{(t)})$ nursing hours, calculated based on the individual NPRs of the scheduled patients. The actual remaining nursing capacity available for elective patients on day t is denoted by $h_e^{(t)}$. Differences between the scheduled demand $h(E^{(t)})$ and the actual remaining capacity for elective patients $h_e^{(t)}$ can lead to cancellations of already planned elective surgeries. Two cases can be distinguished.

1. $h(E^{(t)}) \leq h_e^{(t)}$: The scheduled elective patients can be fully treated, as their total required workload in nursing hours does not exceed the available capacity. The difference $h_e^{(t)} - h(E^{(t)})$ represents unused capacities.

2. $h(E^{(t)}) > h_e^{(t)}$: The scheduled workload exceeds the available capacity, so some of the scheduled patients must be canceled and rescheduled to a later date. The number of cancellations corresponds to the smallest number of patients whose removal brings the total amount of scheduled workload below or equal to $h_e^{(t)}$. This case results in used capacity close to $h_e^{(t)}$, thus high resource utilization.

Based on the predicted available nursing time capacity, the hospital must decide how far in advance and how many elective patients to admit. We focus on the trade-off between efficient resource utilization and number of cancellations, and present results thereof in Chapter 5.

3.6.1 Time of Admission

Assume that patients arriving on day t are planned s days in advance. Thus, admission decision is relies on the forecast $H_e^{(t-s,t)}$, as defined in (3.17). A smaller s generally comes with a more precise forecast, as the probability distribution has less variance with a shorter forecast horizon (e.g., see Example 3.1). However, a certain lead time is needed to carry out pre-operative examinations, and it is generally preferred by patients to be informed about their surgery date ahead of time. Thus, there is a trade-off between the accuracy of capacity forecasts and the practical advantage of admitting patients earlier.

3.6.2 Number of Admissions

Given the forecast $H_e^{(t-s,t)}$ for day t, the hospital must decide on the number of elective patients to admit. It may choose to schedule a patient volume that just fits into the expected available resources. Alternatively, it may decide to schedule more conservatively by leaving a buffer between the forecast capacity and the total demand from scheduled patients, in order to reduce the risk of cancellations. On the other hand, to increase resource utilization, it may opt to overbook, i.e., scheduling more patients than are expected to be treatable. This approach accepts a higher risk of cancellations, but may lead to more efficient use of available resources. The appropriate balance between utilization and cancellations should be evaluated based on the hospital's or health system's priorities.

Chapter 4

Impact of Fluctuating Resources on Backlog

Fluctuating resource availability is a defining feature of health emergencies. This chapter investigates how such variability impacts resource utilization and, consequently, the backlog of care. We focus on the isolated effect of stochastic staff capacity, excluding the influence of pandemic patient demand and its fluctuations. Section 4.1 presents additional assumptions we make in this chapter for analytical tractability. We model our problem as a multi-server queueing process, detailed in Section 4.2. Section 4.3 outlines capacity configurations analyzed in this study, which involve different utilization levels and server availability distributions. To analyze queue length and sojourn time distributions in settings that do not allow for analytical analysis, we conduct numerical experiments. Section 4.4 describes the simulation setup. Section 4.5 explores the defined queueing system in heavy traffic via analytical methods and highlights challenges when dealing with varying server numbers. Section 4.6 reports findings from Monte Carlo simulations regarding queue length and sojourn time and Section 4.7 summarizes the most important findings of this chapter.

4.1 Additional Assumptions

To make the queueing system analytically tractable, we introduce a set of additional assumptions, which only hold for this chapter. First, we assume that NTD arrives according to a homogeneous Poisson process. This replaces patient-level batch arrivals of nursing hours with the arrival of individual units of nursing workload. Similarly, we treat hours of nursing capacity as individually distributed, that is infection and recovery behavior happens at the level of individual hours rather than full shifts of individual nurses. All other assumptions made in Section 3.3 hold.

Admission decisions and operational constraints are not considered in this part of the analysis. Instead, we study the pure effect of fluctuating resources. Capacity is assumed to be used optimally. As long as the waitlist is non-empty, all available available hours of nursing capacity are used. Consequently, service of a patient, which consists of multiple hours of NTD, can be split across days. For example, some nursing hours may be worked off on the current day, and the remainder on subsequent ones.

4.2 Mathematical Queueing Model

We model the delivery of hospital care to patients as a discrete-time queueing system. Time is divided into fixed-length time-intervals, called *slots*, defined as $\{[0,1),[1,2),[2,3),\ldots\}$, where each slot represents one day. We refer to the time interval $[t, t+1)$ as day t.

Workload units arrive at the input of the system stochastically, and wait in a buffer with infinite capacity. Moreover, our modeled hospital operates as a *late arrival system* (also known as *delayed access*), where workload arriving to the waiting list on day t can only start service on day $t+1$. This timing scheme is illustrated in Figure 4.1. Here, the arrival happens at the end of day $t-1$ just after service is completed and patients depart. Service of the next item then starts at the beginning of the next day t.

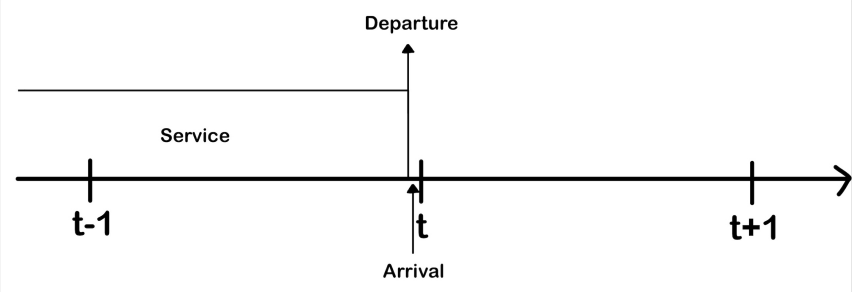


FIGURE 4.1: Time scheme of the late arrival model showing patient arrivals and departures.

Each hour of nursing capacity available on a given day is represented by a single server. Thus, the number of servers, denoted by c directly reflects the total available nursing capacity for that day. We assume deterministic service times of exactly one time slot (i.e., one day), after which workload leaves the system, eliminating the complexity of mid-service capacity fluctuations. Available servers cannot drop during an ongoing service period by definition, ensuring service without interruptions once it has begun.

We define the system contents by (4.1)

$$Q_{t+1} = Q_t - D_t + A_t, \tag{4.1}$$

where Q_t is the system content (workload hours in the queue and in service) at time t, and D_t and A_t denote the departures and arrivals on day t, respectively. D_t is constrained to not exceed the system contents Q_t. In the following, we use the terms *system contents* and *queue length* interchangeably, describing the contents in the queue and in service.

The arrival process is modeled as follows. Workload arrives to the waitlist according to a general uncorrelated arrival process, i.e., the number of arrivals in consecutive slots are independent and identically distributed (i.i.d.) random variables. Each slot, the number of arrivals A_t is drawn from a Poisson counting process with common rate λ, see (4.2).

$$A_t \sim Poisson(\lambda) \tag{4.2}$$

The queue, which consists of nursing hours of elective workload, is served according to a FIFO principle. The service time consists of one slot, and can start and end at slot boundaries only. This also implies that the service of a nursing hour which arrives to an empty system cannot begin until the beginning of the subsequent slot. Consequently, the number of departures D_t is given by the minimum of the system contents and number of

servers, as defined in (4.3).

$$D_t = \begin{cases} Q_t, & \text{if } Q_t < c \\ c, & \text{if } Q_t \geq c \end{cases} \tag{4.3}$$

The described queue is the discrete-time analogue of the continuous *M/1/c* queue, which we denote as *Poisson/1/c*.

We define the utilization ρ of a queueing system as the long-run fraction of time during which the servers are busy processing jobs. Let λ denote the average arrival rate and $\mu = 1$ the average service rate of a single server. For a system with a fixed number of c identical servers, the utilization is defined as

$$\rho = \frac{\lambda}{c \cdot \mu}.$$

A system is considered stable if $\rho < 1$, meaning that on average, the system can handle the incoming workload. If $\rho \geq 1$, the queue length tends to grow unbounded over time, indicating an unstable system.

4.3 Configurations

We consider the system under various demand and capacity circumstances, which are represented as different queue configurations. Section 4.3.1 presents the different server availability patterns used to model fluctuating capacity. Section 4.3.2 outlines the different system loads we examine, representing levels of demand relative to available resources.

4.3.1 Server Distribution

We consider different patterns of server availability. First, we consider a deterministic number of 20 servers. Then, we introduce variability by drawing the number of servers from a Binomial distribution with an expected value of 20. Lastly, we compare to the setting, in which server availability evolves based on Binomial infection and recovery processes based on the nurse absenteeism model introduced in Section 3.4. We describe the latter two serve configurations below.

Binomial Distribution

Each day, the number of available servers is determined by an i.i.d. draw from a Binomial distribution. For comparability with the deterministic setting of 20 servers, the expected number of available servers stays fixed at 20 across all the considered cases. We explore results with maximum possible numbers of 22, 25, 28, and 100 servers, which constitute the support n of the Binomial distribution. The success probabilities r, referring to the likelihood that an individual server is available on a given day, are chosen such that the expected number of servers

$$\mathbb{E}[c] = n \cdot r = 20.$$

We consider four different server distributions

$$c \sim Binomial(n, r)$$

with parameter values shown in Table 4.1.

n	r
22	$10/11$
25	$4/5$
28	$5/7$
100	$1/5$

TABLE 4.1: Parameters for Binomial distributed nursing capacity.

Binomial Infection and Recovery Evolution

Here, server availability evolves as a process of Binomial infection and recovery, as introduced in Section 3.4. The maximum capacity consists of n total servers, each of which *drop out* (get sick) on a given day with probability p and recover with probability q. Again, we consider values of $n \in \{22, 25, 28, 100\}$ with an expected number and a starting capacity of 20 available servers. Setting $q = 0.2$, we determine p such that the expectation of the available servers satisfies

$$\frac{q}{p+q} \cdot n = 20.$$

The values of p, q, and the corresponding steady state success probability $\frac{q}{p+q}$ are summarized in Table 4.2.

n	$\frac{q}{p+q}$	p	q
22	$10/11$	$1/50$	$1/5$
25	$4/5$	$1/20$	$1/5$
28	$5/7$	$2/25$	$1/5$
100	$1/5$	$4/5$	$1/5$

TABLE 4.2: Parameters for nursing capacity following Binomial infection and recovery behavior.

This differs from i.i.d. Binomial variables described previously. Since the availability of nurses on a given day depends on the one of the day before, the number of available servers varies less from day to day. In other words, the availability of servers is temporally autocorrelated.

4.3.2 Utilization

The variability in the number of available servers can affect backlog dynamics differently depending on system utilization. To explore this, we consider several traffic scenarios by adjusting the Poisson arrival rate λ accordingly.

Heavy Traffic ($\rho = 0.995$)

Health systems are typically designed to operate in heavy traffic, where utilization levels approach 100% without exceeding it. Resources are scarce and tightly allocated, yet the system remains stable. This leads to efficient resource usage, with most capacity being utilized.

Moderate Traffic ($\rho = 0.8$)

To understand how server variability impacts systems under less strain, we also examine a moderate traffic scenario. Here, the arrival rate is adjusted to achieve a utilization of $\rho = 0.8$. In steady state, we expect shorter queue lengths than in the heavy traffic case.

Hypercritical Traffic ($\rho = 1$, $\rho = 1.2$)

During health emergencies, demand may exceed available capacity for certain periods. We analyze system behavior under overload conditions, where arrivals overshoot service capacity. These scenarios are simulated for utilization levels of $\rho = 1.0$ and $\rho = 1.2$. Since the queue is unstable, it does not converge to a limiting distribution and instead grows unbounded over time.

4.4 Experimental Setup

For each combination of the server and utilization settings described above, the queueing system is analyzed. We perform 100 simulations of 1,000,000 time steps, where we record system contents and sojourn times. The first 10% of the simulation period are regarded as a *start-up* phase and not used for analyses. The rest is aggregated to report distributions and corresponding 95%-confidence intervals (CIs).

We fit an exponential curve to empirical results by taking the mean queue length as the rate parameter. We validate our simulation by comparing empirical results for $c = 20$ to the corresponding distribution we derive analytically in Section 4.5.

Experiments were run on a computer with an AMD Ryzen 5 5500U processor with 6 cores and 16 GB RAM, running Windows 11. Analyses were performed in Python version 3.8.8 using the web-based interface Jupyter Notebooks.

4.5 Analytical Results

This section explores the queueing system via analytical methods. In Section 4.5.1, we exploit the properties of probability generating functions (PGFs) to derive the system content PGF for the *Poisson/1/c* and *Poisson/1/1* queue following the approach used in [52]. We relate the former to the latter system in heavy traffic and derive the limiting distribution, which matches known results for continuous-time queues. Section 4.5.2 examines the challenges of extending this approach to the *Poisson/1/Binomial* queue, where the number of servers varies according to a Binomial distribution. We show why the analogy to continuous-time queues with varying i.i.d. service times breaks down in this case.

4.5.1 *Poisson/1/c* Queue in Heavy Traffic

Following Gao et al. [52], we define the PGF of arrivals $G_A(z) = \mathbb{E}[z^A]$, where $A_t \sim Poisson(\lambda)$, in (4.4).

$$G_A(z) = \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} z^i = e^{-\lambda(1-z)} \tag{4.4}$$

Since the number of departures are deterministic given the system contents and the number of servers c, we work with the direct definition of D_t in (4.3) instead of defining the PGF as Gao et al. did.

The PGF of the system contents Q_t is defined based on its recursive definition (see (4.1)) as

$$\begin{aligned} G_{Q_{t+1}}(z) &= \mathbb{E}[z^{Q_{t+1}}] \\ &= \mathbb{E}[z^{Q_t - D_t + A_t}] \\ &= \mathbb{E}[z^{A_t} z^{Q_t - D_t}] \\ &= G_A(z)\mathbb{E}[z^{Q_t - D_t}]. \end{aligned} \tag{4.5}$$

Further, we split up the sum of the expectation dependent on whether or not Q_t is less than the number of servers c. For the case of $Q_t \geq c$, Q_t and D_t are independent variables, for $Q_t < c$, they are equivalent.

$$\begin{aligned} \mathbb{E}[z^{Q_t - D_t}] &= \sum_{i=0}^{\infty} \mathbb{P}(Q_t = i)\mathbb{E}[z^{Q_t - D_t} | Q_t = i] \\ &= \sum_{i=0}^{c-1} \mathbb{P}(Q_t = i)\mathbb{E}[z^{Q_t - D_t} | Q_t = i] + \sum_{i=c}^{\infty} \mathbb{P}(Q_t = i)\mathbb{E}[z^{Q_t - D_t} | Q_t = i] \\ &= \sum_{i=0}^{c-1} \mathbb{P}(Q_t = i)\mathbb{E}[z^{i-i} | Q_t = i] + \sum_{i=c}^{\infty} \mathbb{P}(Q_t = i)\mathbb{E}[z^{i-c} | Q_t = i] \\ &= \sum_{i=0}^{c-1} \mathbb{P}(Q_t = i) + \sum_{i=c}^{\infty} \mathbb{P}(Q_t = i) z^{i-c} \end{aligned}$$

Considering the second sum

$$\begin{aligned} \sum_{i=c}^{\infty} \mathbb{P}(Q_t = i) z^{i-c} &= \sum_{i=0}^{\infty} \mathbb{P}(Q_t = i) z^{i-c} - \sum_{i=0}^{c-1} \mathbb{P}(Q_t = i) z^{i-c} \\ &= z^{-c} \sum_{i=c}^{\infty} \mathbb{P}(Q_t = i) z^{i} - \sum_{i=0}^{c-1} \mathbb{P}(Q_t = i) z^{i-c} \\ &= G_{Q_t}(z) z^{-c} - \sum_{i=0}^{c-1} \mathbb{P}(Q_t = i) z^{i-c}. \end{aligned}$$

Inserting into (4.5) yields

$$\begin{aligned} G_{Q_{t+1}}(z) &= G_A(z)\mathbb{E}[z^{Q_t - D_t}] \\ &= G_A(z) \left\{ \left[\sum_{i=0}^{c-1} \mathbb{P}(Q_t = i) \right] + G_{Q_t}(z) z^{-c} - \left[\sum_{i=0}^{c-1} \mathbb{P}(Q_t = i) z^{i-c} \right] \right\} \\ &= G_A(z) \left\{ G_{Q_t}(z) z^{-c} + \sum_{i=0}^{c-1} \mathbb{P}(Q_t = i) z^{i} \left[z^{-i} - z^{-c} \right] \right\}. \end{aligned}$$

Assuming the steady state exists, then the PGF of the system content converges to its equilibrium version as $t \to \infty$. Thus, the PGF of the system contents in steady state $G_Q(z)$ is given by

$$G_Q(z) = G_A(z) \left\{ G_Q(z) z^{-c} + \sum_{i=0}^{c-1} Q(i) z^{i} \left[z^{-i} - z^{-c} \right] \right\},$$

where $Q(i)$ is the equilibrium probability of having i customers in the system. Rearranging the terms leads to

$$G_Q(z)\left[1 - G_A(z)z^{-c}\right] = G_A(z)\left\{\sum_{i=0}^{c-1} Q(i)z^i\left[z^{-i} - z^{-c}\right]\right\}$$

and

$$\begin{aligned} G_Q(z) &= \frac{G_A(z)\sum_{i=0}^{c-1} Q(i)z^i\left[z^{-i} - z^{-c}\right]}{1 - G_A(z)z^{-c}} \\ &= \frac{G_A(z)\sum_{i=0}^{c-1} Q(i)\left[z^c - z^i\right]}{z^c - G_A(z)} \qquad (4.6) \\ &= \frac{e^{-\lambda(1-z)}\sum_{i=0}^{c-1} Q(i)\left[z^c - z^i\right]}{z^c - e^{-\lambda(1-z)}}. \qquad (4.7) \end{aligned}$$

Note that there are still c unknown constants to be determined, which are the probabilities $Q(i)$ of a system content of i for $i = 0, ..., c-1$. While this is possible by analyzing $G_Q(z)$ inside the unit disk of the complex z-plane and using the normalizing condition, we constrain our setting to a heavy traffic setting, where the constants are not required for analysis.

As the system load approaches capacity ($\rho \to 1$), all servers are almost always occupied, making the cases of $Q(i)$ for $i < c$ are negligible. This means that the system behaves as if it processes jobs at a constant rate. We therefore translate the multi-server queue to a single-server queue, where the server operates at a faster speed. The service times are adjusted so that the same number of jobs per time interval are completed.

In the following, we derive results for the *Poisson/1/1* queue with Poisson arrival rate $\lambda \to 1$ resulting in system contents Q and sojourn times W. By Little's Law [53], $Q = \lambda W = W$.

Remark 4.5.1. *Consider a Poisson/1/c queue that we relate to the Poisson/1/1 queue in heavy traffic, with an arrival rate $\lambda \to 1$. The arrival rate of the multi-server queue is then $\tilde{\lambda} = c\lambda \to c$, and service times in the single-server representation are $\frac{1}{c}$. Since the utilization is the same, the queue length $\tilde{Q}$ remains similar to Q. Due to this faster service, sojourn times are c-times less, resulting in $\tilde{W} = \frac{W}{c}$. This goes in line with Little's Law, as $\tilde{Q} = \tilde{\lambda}\tilde{W} = c\lambda\frac{w}{c} = \lambda W = Q$.*

Poisson/1/1 Queue

In the single server setting, D_t simplifies to

$$D_t = \begin{cases} 0, & \text{if } Q_t = 0 \\ 1, & \text{otherwise.} \end{cases}$$

Assuming $\lambda < 1$, such that $\rho < 1$ and the stationary distribution exists, and taking the limit $t \to \infty$, we insert into (4.6), and get

$$\begin{aligned} G_Q(z) &= \frac{G_A(z)\sum_{i=0}^{0} Q(i)[z^1 - z^i]}{z^1 - G_A(z)} \\ &= \frac{G_A(z)Q(0)(z-1)}{z - G_A(z)} \\ &= \frac{e^{-\lambda(1-z)}Q(0)(z-1)}{z - e^{-\lambda(1-z)}} \qquad (4.8) \end{aligned}$$

We now only need to determine one unknown constant $Q(0)$. Since the utilization $\rho = \lambda$ is defined as the fraction of time the server is busy processing jobs, $Q(0) = 1 - \lambda$. As a sanity check we derive it analytically as well. By the properties of PGFs, we know $\lim_{z\to 1} G_Q(z) = 1$, and together with l'Hôpital's rule we get

$$\begin{aligned}
1 &= \lim_{z\to 1} G_Q(z) \\
&= \lim_{z\to 1} \frac{e^{-\lambda(1-z)}Q(0)(z-1)}{z - e^{-\lambda(1-z)}} \\
&= \frac{\lim_{z\to 1} \frac{d}{dz} e^{-\lambda(1-z)}Q(0)(z-1)}{\lim_{z\to 1} \frac{d}{dz} z - e^{-\lambda(1-z)}} \\
&= \frac{Q(0)\lim_{z\to 1} e^{-\lambda(1-z)}(\lambda z - \lambda + 1)}{\lim_{z\to 1} 1 - \lambda e^{-\lambda(1-z)}} \\
&= \frac{Q(0)}{1-\lambda},
\end{aligned}$$

which gives the desired result

$$Q(0) = 1 - \lambda.$$

Inserting into (4.8), the stationary distribution is given by the PGF

$$G_Q(z) = \frac{e^{-\lambda(1-z)}Q(0)(z-1)}{z - e^{-\lambda(1-z)}} \tag{4.9}$$

$$= \frac{e^{-\lambda(1-z)}(1-\lambda)(z-1)}{z - e^{-\lambda(1-z)}}. \tag{4.10}$$

Heavy Traffic Limit

Provided $\lambda < 1$, we have the equilibrium queue length probabilities giving by the generating function (4.10). We follow the approach used in [54], which leverages the properties of characteristic functions (CFs) and derives the heavy traffic limit for the $M_D/G_D/1$ model, where in each time slot either one or zero customers arrive and service times are i.i.d. random variables. Therefore, we bring the PGF into the form (4.11), which was used by Subba Rao [54].

$$\begin{aligned}
G_Q(z) &= \frac{h(z)(1-\lambda)(z-1)}{z - h(z)} \\
&= \frac{(1-\lambda)(1-z)h(z)}{h(z) - z}
\end{aligned} \tag{4.11}$$

with

$$h(z) = e^{-\lambda(1-z)} = G_A(z)$$

We now wish to obtain the distribution of Q in heavy traffic. Since that means considering the case $(1-\lambda) \to 0$, we derive the limiting distribution of $(1-\lambda)Q$.

Following Subba Rao, let $\varepsilon = 1 - \lambda$ and let $\phi_Q(\theta) = \mathbb{E}[e^{i\theta Q}]$ be the CF of the queue length, where i denotes the imaginary unit. We then have

$$\begin{aligned}
\phi_Q(\theta) &= G_Q(e^{i\theta}) \\
&= \frac{(1-\lambda)(1-e^{i\theta})h(e^{i\theta})}{h(e^{i\theta}) - e^{i\theta}}
\end{aligned}$$

We define $\hat{Q} = \varepsilon Q$, leading to the CF

$$\begin{aligned}\phi_{\hat{Q}}(\theta) &= \mathbb{E}[e^{i\theta\hat{Q}}] \\ &= \mathbb{E}[e^{i\theta\varepsilon Q}] \\ &= \phi_Q(\varepsilon\theta) \\ &= \frac{\varepsilon(1-e^{i\varepsilon\theta})h(e^{i\varepsilon\theta})}{h(e^{i\varepsilon\theta})-e^{i\varepsilon\theta}}\end{aligned} \tag{4.12}$$

For $\varepsilon \to 0$, we have the following Taylor expansion of $h(e^{i\varepsilon\theta})$, keeping all terms up to the second order.

$$\begin{aligned}h(e^{i\varepsilon\theta}) &= h(1) + h'(1)(e^{i\varepsilon\theta}-1) + \frac{h''(1)}{2}(e^{i\varepsilon\theta}-1)^2 + o(\varepsilon^2) \\ &= 1 + \lambda(e^{i\varepsilon\theta}-1) + \frac{\lambda^2}{2}(e^{i\varepsilon\theta}-1)^2 + o(\varepsilon^2) \\ &= 1 + \lambda(i\varepsilon\theta - \frac{\varepsilon^2\theta^2}{2}) + \frac{\lambda^2}{2}(i\varepsilon\theta - \frac{\varepsilon^2\theta^2}{2})^2 + o(\varepsilon^2) \\ &= 1 + \lambda(i\varepsilon\theta - \frac{\varepsilon^2\theta^2}{2}) + \frac{\lambda^2}{2}(-\varepsilon^2\theta^2) + o(\varepsilon^2) \\ &= 1 + \lambda i\varepsilon\theta - \frac{\varepsilon^2\theta^2}{2}(\lambda^2+\lambda) + o(\varepsilon^2) \\ &= 1 + (1-\varepsilon)i\varepsilon\theta - \frac{\varepsilon^2\theta^2}{2}((1-\varepsilon)^2 + (1-\varepsilon)) + o(\varepsilon^2)\end{aligned}$$

Inserting into the CF (4.12) gives

$$\phi_{\hat{Q}}(\theta) = \frac{\varepsilon(1-e^{i\varepsilon\theta})[1+(1-\varepsilon)i\varepsilon\theta - \frac{\varepsilon^2\theta^2}{2}((1-\varepsilon)^2+(1-\varepsilon))]}{1+(1-\varepsilon)i\varepsilon\theta - \frac{\varepsilon^2\theta^2}{2}((1-\varepsilon)^2+(1-\varepsilon)) - e^{i\varepsilon\theta}} + o(\varepsilon^2)$$

Now consider the nominator and denominator separately to determine the dominant terms.

Nominator:

$$\begin{aligned}&\varepsilon(1-e^{i\varepsilon\theta})[1+(1-\varepsilon)i\varepsilon\theta - \frac{\varepsilon^2\theta^2}{2}((1-\varepsilon)^2+(1-\varepsilon))] \\ &= \varepsilon(-i\varepsilon\theta + \frac{\varepsilon^2\theta^2}{2} + o(\varepsilon^2))[1+(1-\varepsilon)i\varepsilon\theta - \frac{\varepsilon^2\theta^2}{2}((1-\varepsilon)^2+(1-\varepsilon))] \\ &= -i\varepsilon^2\theta + o(\varepsilon^2)\end{aligned}$$

Denominator:

$$\begin{aligned}&1+(1-\varepsilon)i\varepsilon\theta - \frac{\varepsilon^2\theta^2}{2}((1-\varepsilon)^2+(1-\varepsilon)) - e^{i\varepsilon\theta} \\ &= 1+(1-\varepsilon)i\varepsilon\theta - \frac{\varepsilon^2\theta^2}{2}((1-\varepsilon)^2+(1-\varepsilon)) - (1+i\varepsilon\theta - \frac{\varepsilon^2\theta^2}{2} + o(\varepsilon^2)) \\ &= i\varepsilon\theta((1-\varepsilon)-1) - \frac{\varepsilon^2\theta^2}{2}((1-\varepsilon)^2+(1-\varepsilon)-1) + o(\varepsilon^2) \\ &= -i\varepsilon^2\theta - \frac{\varepsilon^2\theta^2}{2} + o(\varepsilon^2)\end{aligned} \tag{4.13}$$

Here, (4.13) follows from applying a Taylor expansion on $e^{i\varepsilon\theta}$. The limit of the CF for $\varepsilon \to 0$ is therefore given as

$$\begin{aligned}\lim_{\varepsilon\to 0} \phi_{\hat{Q}}(\theta) &= \lim_{\varepsilon\to 0} \frac{-i\varepsilon^2\theta}{-i\varepsilon^2\theta - \frac{\varepsilon^2\theta^2}{2}} \\ &= \frac{2}{2 - i\theta}.\end{aligned}$$

The Lévy's convergence theorem [55] states that if a series of CF converges to some limit, the probability density function (PDF) converges to the PDF corresponding to the limiting CF. Thus, we need to apply an inverse Fourier transform to get the queue length distribution of εQ in heavy traffic.

We recognize the CF of an exponentially distributed variable $X \sim Exp(\alpha)$

$$\phi_X(\theta) = \frac{\alpha}{\alpha - i\theta}.$$

Thus,

$$\lim_{\varepsilon\to 0} \hat{Q} \sim Exp(2) \tag{4.14}$$

and

$$\lim_{\varepsilon\to 0} \mathbb{P}(\varepsilon Q > x) = e^{-2x}.$$

Due to the deterministic service time of one time slot, the sojourn time $\hat{W} = \varepsilon W$ for the single-server setting follows the same distribution.

$$\lim_{\varepsilon\to 0} \hat{W} \sim Exp(2)$$

$$\lim_{\varepsilon\to 0} \mathbb{P}(\varepsilon W > x) = e^{-2x}$$

As stated in Remark 4.5.1, in a translation from this single-server queue to a sped-up single-server queue representing a multi-server queue with c servers, sojourn times $\hat{W}_c$ must scaled down by factor c, resulting in

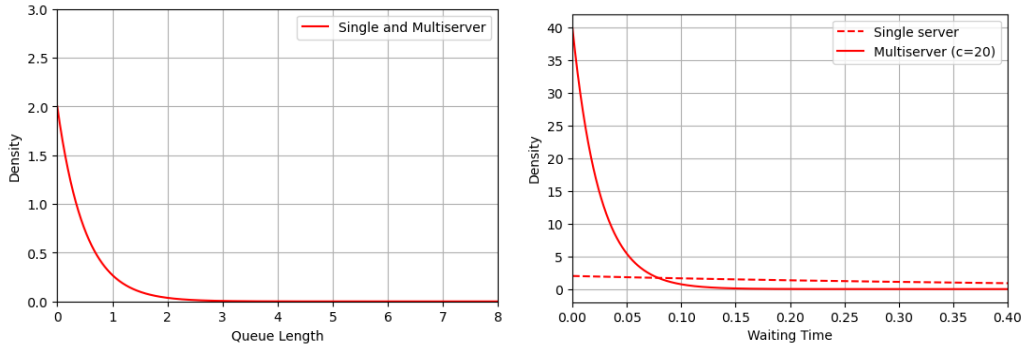
$$\lim_{\varepsilon\to 0} \frac{1}{c}\hat{W}_c \sim Exp(2),$$

which equals

$$\lim_{\varepsilon\to 0} \hat{W}_c \sim Exp(2c). \tag{4.15}$$

Queue lengths $\hat{Q}_c$ for the multi-server case follow the same distribution as the single-server case, given in(4.14).

Figure 4.2 illustrates the derived distributions and highlights the difference in sojourn times between the single-server and multi-server setting.



(A) QL distr. for c constant. (B) ST distr. for $c = 1$ and $c = 20$.

FIGURE 4.2: Heavy traffic limiting distributions for single- and multi-server ($c = 20$) queues with a constant number of servers. QL - queue length, ST - sojourn time.

These results correspond to the heavy traffic limits for the continuous equivalent in [56]. They determined the heavy traffic limits for the $M/G/1$ queue to have the following distributions.

$$\lim_{\rho \to 1} \hat{Q} \sim Exp(\frac{2b^2}{\sigma^2 + b^2}) \tag{4.16}$$

$$\lim_{\rho \to 1} \hat{W} \sim Exp(\frac{2b}{\sigma^2 + b^2}) \tag{4.17}$$

Here, the variable b describes the expected service time, and σ^2 the variance of the service times. For our deterministic setting ($b = 1/c, \sigma^2 = 0$), the results correspond to ours.

The analogy between discrete-time and continuous-time queueing systems has been observed in prior research. As shown by Meisling [57], results for continuous-time systems can be derived as a limiting case of their discrete-time counterparts. By letting the length of time intervals approach zero ($\Delta t \to 0$), Meisling showed that the arrival and service time distributions converge to continuous distributions with the same mean rates. While in his work he considered only one or no arrival at a given time mark, arguing that this limit leads to a Poisson distribution in continuous time, we believe his argumentation still holds in our setting, where the number of arrivals at each time step follows a Poisson counting process. This is because the sum of independent Poisson random variables is itself Poisson distributed.

4.5.2 *Poisson/1/Binomial* Queue in Heavy Traffic

Based on the relation of the *Poisson/1/c* queue to the continuous $M/D/1$ as a special case of the $M/G/1$ queue with service times $\frac{1}{c}$, we explore whether the same relationship exists for the *Poisson/1/Binomial* queue.

We relate our discrete-time case to (4.16) and (4.17), by defining service times $\frac{1}{C}$ for C following a Binomial distribution. The mean and variance of the service time are then computed as $\mathbb{E}[\frac{1}{C}]$ and $Var[\frac{1}{C}]$. Due to the variance term in the denominator of the exponential rate of the continuous-time results, we expect longer queue lengths and sojourn times for higher variability of servers.

These expressions are not straightforward to compute for two reasons. Note that $\mathbb{E}[\frac{1}{C}] \neq \frac{1}{\mathbb{E}[C]}$ and $Var[\frac{1}{C}] \neq \frac{1}{Var[C]}$. Due to the Binomial distribution having a non-zero probability

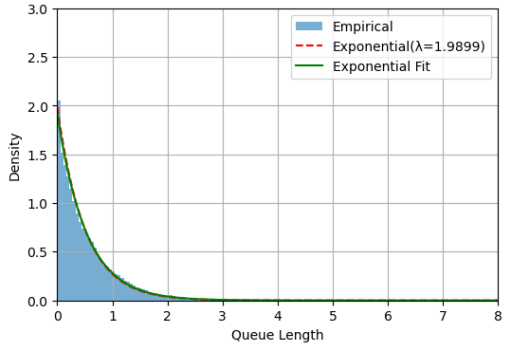
of $C = 0$, these terms are not defined unless we condition on $C > 0$. Even then, the first and second moment of the reciprocals of C, which are needed to derive $\mathbb{E}[\frac{1}{C}]$ and $Var[\frac{1}{C}]$ are cumbersome to compute. Since in our considered distributions, the probability $\mathbb{P}[C = 0]$ is negligibly small, we chose to compute these values numerically.

We translated a random draw of 1 million i.i.d. server capacities $c \in C$ to service times $\frac{1}{c}$. The value 0 did not occur in the sample. We then calculated the mean and variance of these values. The procedure was conducted in Python version 3.8.8. Table 4.3 presents resulting values and the corresponding queue length and sojourn time exponential rate β and ω derived from (4.16) and (4.17), respectively.

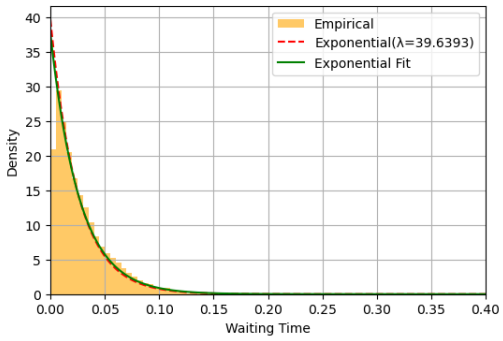
Server Distribution	b	$\sigma^2 \cdot 10^{-5}$	β	ω
$Bin(22, \frac{10}{11})$	0.0502	1.28	1.9899	39.6393
$Bin(25, \frac{4}{5})$	0.0505	2.91	1.9774	39.1572
$Bin(28, \frac{5}{7})$	0.0508	4.24	1.9676	38.7337
$Bin(100, \frac{1}{5})$	0.0522	13.50	1.9056	36.5055

TABLE 4.3: Numerical results for mean and variance of service time, and resulting queue length and sojourn time exponential rates β and ω.

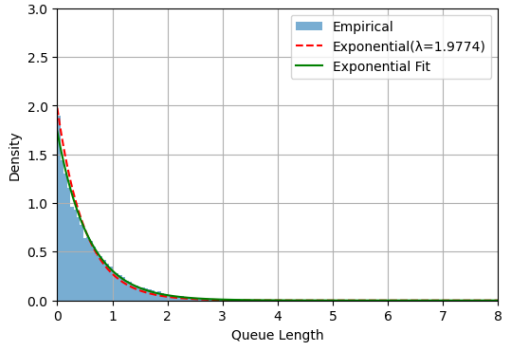
Comparing simulation results from the *Poisson/1/Binomial* queue to the adjusted continuous-time results, we observe significant differences in the heavy traffic limiting distribution. Figure 4.3 compares empirical results and an exponential fit to them (green line) to the analytical continuous results (red dotted line).



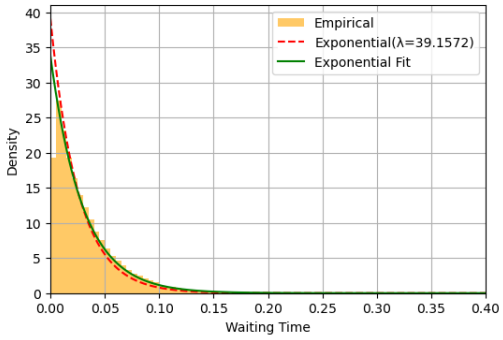
(A) QL distr., $c \sim Bin(22, \frac{10}{11})$



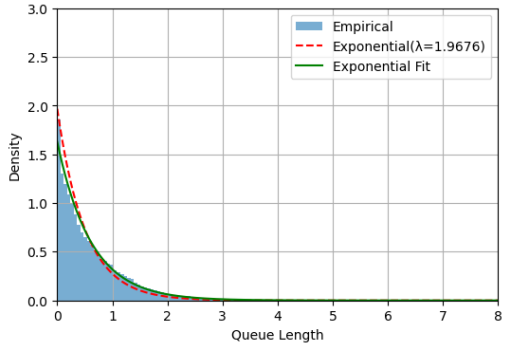
(B) ST distr., $c \sim Bin(22, \frac{10}{11})$



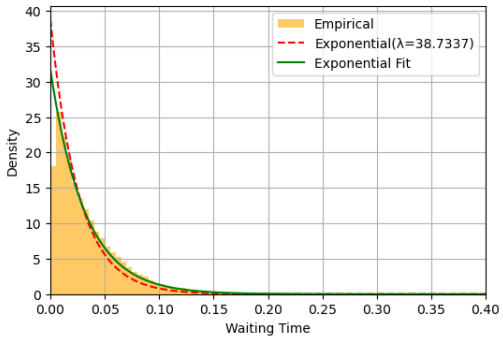
(C) QL distr., $c \sim Bin(25, \frac{4}{5})$



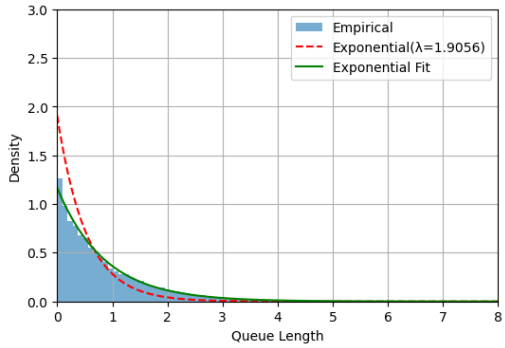
(D) ST distr., $c \sim Bin(25, \frac{4}{5})$



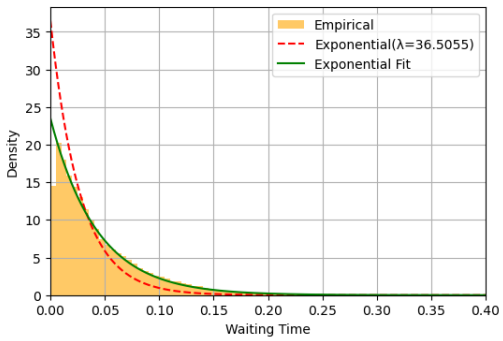
(E) QL distr., $c \sim Bin(28, \frac{5}{7})$



(F) ST distr., $c \sim Bin(28, \frac{5}{7})$



(G) QL distr., $c \sim Bin(100, \frac{1}{5})$



(H) ST distr., $c \sim Bin(100, \frac{1}{5})$

FIGURE 4.3: Heavy traffic limiting distributions of a multi-server queue with the number of servers $C \sim Binomial$. Comparison of empirical results to scaled continuous-time queueing results. QL - queue length, ST - sojourn time, distr. - distribution.

We observe that our empirical results deviate from the queue length distribution of the scaled *M/G/1* queue. Both results indicate longer queue lengths and sojourn times than in the constant 20-server case. However, the results of the scaled continuous queue underestimate the empirical increase in both queue length and sojourn time. In other words, the translation to fluctuating service times underestimates the effect of variability in the number of servers. This is due to the fact that, when doing the translation from server numbers to service times, we do not consider the resulting temporal autocorrelation of service times. The assumption of i.i.d. service times in (4.16) and (4.17) is therefore not fulfilled, leading to differing results.

We illustrate the translation and why i.i.d. servers do not translate to i.i.d. service times for an example queue. Consider a queueing model like described in Section 4.2 over three time steps with server numbers of 20, 10, and 30 on day t, $t+1$, and $t+2$, respectively. Figure 4.4 shows a timeline of service completions, resulting from translating the system into a continuous-time single-server queue. Service completions are indicated by the ticks. Additionally, there is a service completion at the end of each day. Since we assume a non-empty queue, the number of service completions matches the number of available servers. So, when looking at one day in isolation, the translation behaves just as it did in the setting with a constant amount of servers (Section 4.5.1), where continuous-time and discrete-time results are equivalent in heavy traffic. However, considering multiple time steps, service times clearly violate the i.i.d.-assumption.

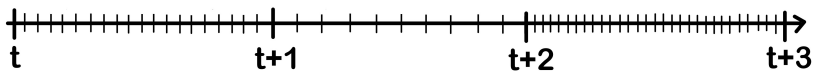


FIGURE 4.4: Timeline of service completions in translation from discrete-time multi-server to continuous-time single-server queue.

Expressions (4.16) and (4.17) with parameter values as given in Table 4.3, however, give results for an *M/G/1* queue with i.i.d. service times. Figure 4.5 illustrates exemplary service completions for the example above.

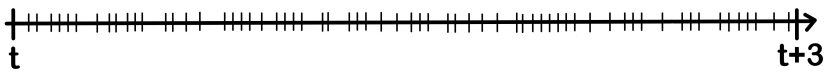


FIGURE 4.5: Timeline of service completions in continuous-time single-server queue with i.i.d. service times.

Note that the mean and variance of service times of these two cases are the same. The queue length and sojourn time distribution in heavy traffic, though, is not. We believe that the temporal autocorrelation of service times, resulting in extended periods of high resource availability increases the likelihood of the queue emptying, which leads to lost resources.

The behavior observed for *Poisson/1/Binomial* queues can be compared to queues in random environments. There, parameters of the queue (e.g., service time) are determined by the state of the queue, which typically follows a Markov chain. Discrete-time queues of this kind are researched in [58], where large discrepancies to standard queueing models are found. Note that in our case transitions of service times happen at a fixed time interval of one day.

In summary, translating a multi-server queue with a varying number of servers to a sped up single-server queue results in temporal autocorrelation of service times. Therefore, (4.16) and (4.17) do not apply, as they assume i.i.d service times.

4.6 Numerical Results

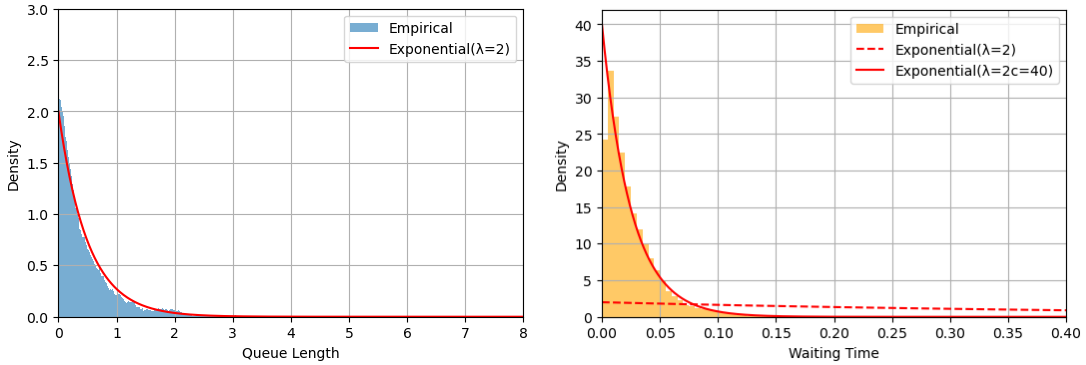
This section reports findings from Monte Carlo simulations regarding queue length and sojourn time. Results are structured according to their traffic scenario. Sections 4.6.1, 4.6.2 and 4.6.3 report results in heavy, moderate and hypercritical traffic, respectively. Within these sections, results are structured as follows. First, we look at a constant number of servers, before investigating a Binomial distribution and an evolution according to Binomial infection and recovery behavior.

4.6.1 Heavy Traffic

This section presents results on the heavy traffic limiting distribution. Before reporting numerical results for a varying number of servers, the distributions for a fixed server number, which we derived analytically, are plotted in Figure 4.6. All figures in this section report queue lengths and sojourn times scaled by the factor $1 - \rho = 0.005$, such that they compare to the analytical results.

Constant Number of Servers

As derived in Section 4.5.1, the queue length and sojourn time for a *Poisson/1/c* queue with a fixed number of servers follow distribution given in (4.14) and (4.15). We validate our simulation by comparing empirical results of the 20-server queue to the analytical ones in Figure 4.6. The mean queue length is 96.75 (95%-CI: [86.18; 108.23]) with a standard deviation (SD) of 98.43 (95%-CI: [82.95; 113.61]).



(A) QL distr. for $c = 20$.

(B) ST distr. for $c = 20$.

FIGURE 4.6: Heavy traffic limiting distributions for multi-server ($c = 20$) queues with a constant number of servers. QL - queue length, ST - sojourn time, distr. - distribution.

The figures show a good fit, successfully validating our simulation procedure. However, we can observe fewer very short sojourn times than indicated in the analytical derivation. This likely results from the LAS discipline, where new arrivals have delayed access to service, which enforces a minimum sojourn time of two time slots, given service times of

one. The small deviation in the queue length distribution may be explained by the fairly wide CIs of simulations, indicating high queue length variability, as shown in Figure 4.7.

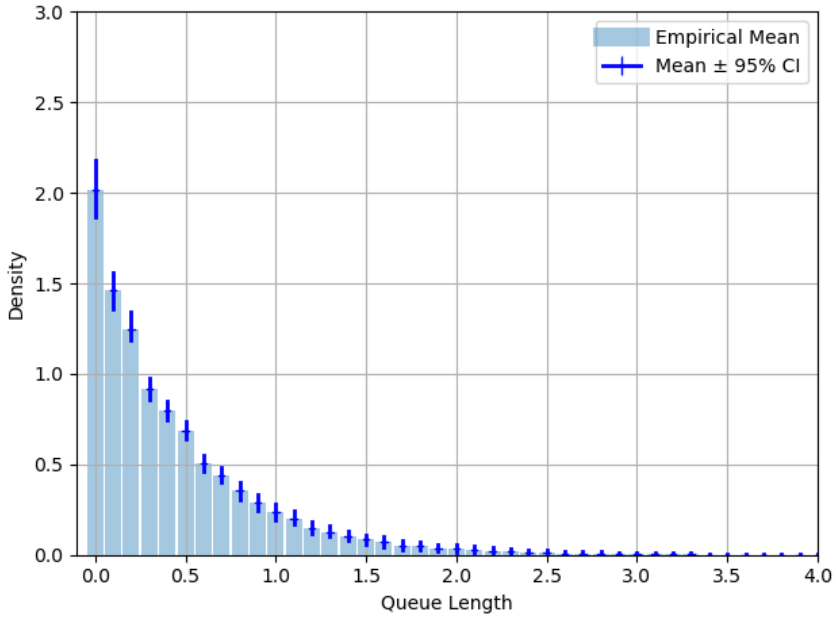


FIGURE 4.7: Confidence intervals of heavy traffic limiting distributions for multi-server ($c = 20$) queues with a constant number of servers.

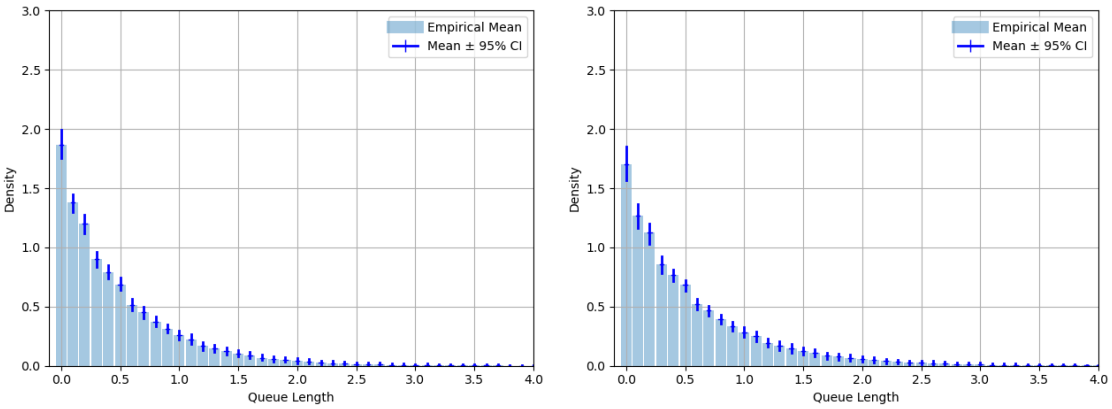
Binomial Distributed Servers

Table 4.4 reports the mean and SD of queue length, including 95% CIs.

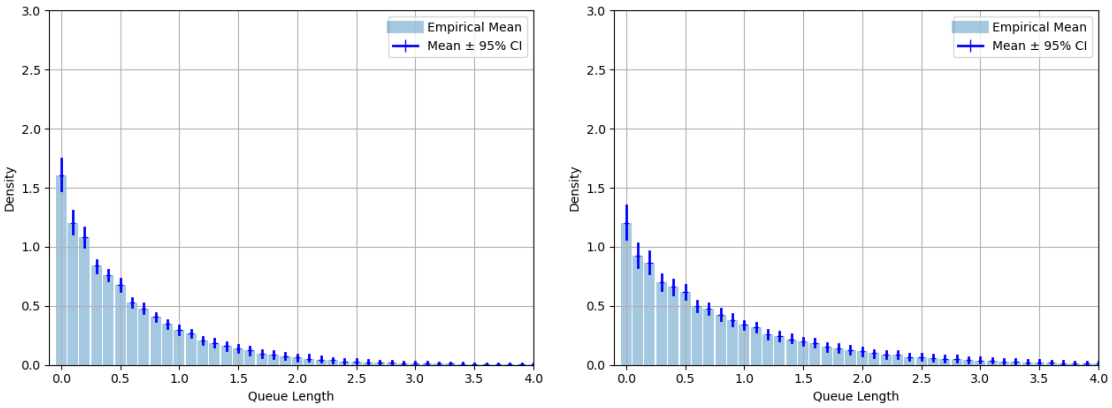
Server Distribution	Mean QL	SD QL
$Bin(22, \frac{10}{11})$	106.28 [91.87; 121.71]	107.07 [89.95; 128.97]
$Bin(25, \frac{4}{5})$	117.12 [104.66; 137.04]	118.36 [97.83; 144.68]
$Bin(28, \frac{5}{8})$	124.61 [107.98; 144.22]	128.27 [102.64; 167.66]
$Bin(100, \frac{1}{5})$	174.14 [151.71; 209.27]	173.03 [142.66; 215.43]

TABLE 4.4: Mean and standard deviation of queue length with Binomial distributed servers in heavy traffic. QL - queue length, SD - standard deviation.

We observer both greater means and SDs for a larger number of total servers. The following Figures 4.8 and 4.9 show queue lengths and sojourn times scaled with the factor $1 - \rho$ to ensure comparability to the analytical results. The first presents the queue length distribution and corresponding CIs, while the latter compares distribution of the different Binomial server settings to the distribution with constant 20 servers in terms of both queue length and sojourn time. Table 4.5 reports rates of the exponential fits to empirical queue length data, which are shown as green lines in Figure 4.9.



(A) $c \sim Bin(22, \frac{10}{11})$ (B) $c \sim Bin(25, \frac{4}{5})$

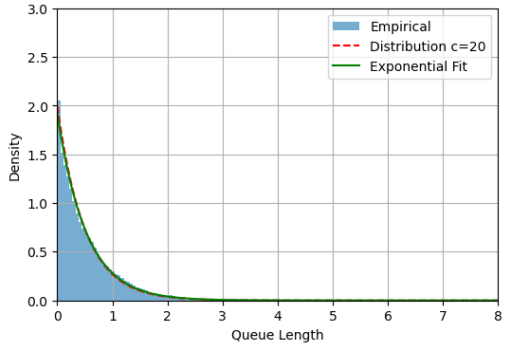


(C) $c \sim Bin(28, \frac{5}{7})$ (D) $c \sim Bin(100, \frac{1}{5})$

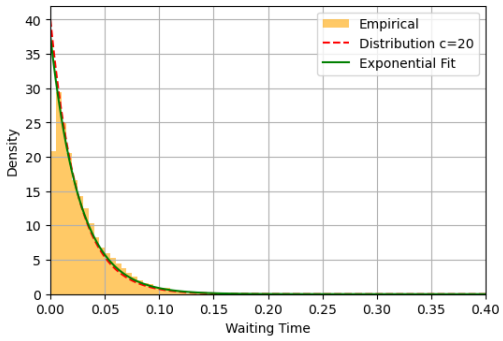
FIGURE 4.8: Confidence intervals of heavy traffic limiting distributions for multi-server queues with the number of servers $C \sim Binomial$.

Server Distribution	Exponential Rate
$Bin(22, \frac{10}{11})$	1.88 [1.64; 2.17]
$Bin(25, \frac{4}{5})$	1.71 [1.46; 1.91]
$Bin(28, \frac{5}{8})$	1.61 [1.39; 1.85]
$Bin(100, \frac{1}{5})$	1.14 [0.96; 1.32]

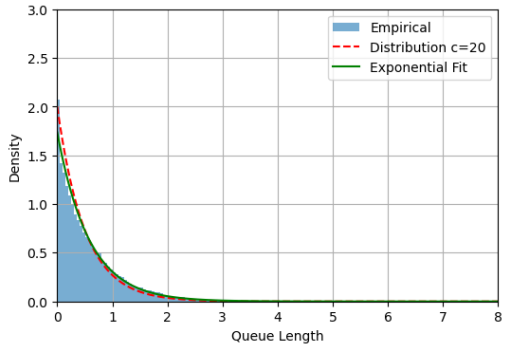
TABLE 4.5: Rates of exponential fit to empirical queue length for Binomial distributed servers.



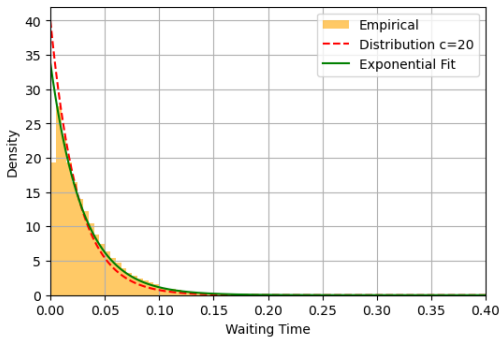
(A) QL distr., $c \sim Bin(22, \frac{10}{11})$



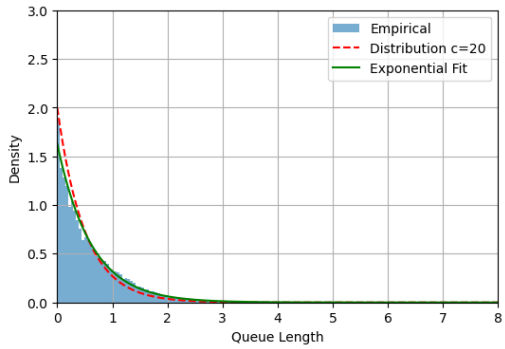
(B) ST distr., $c \sim Bin(22, \frac{10}{11})$



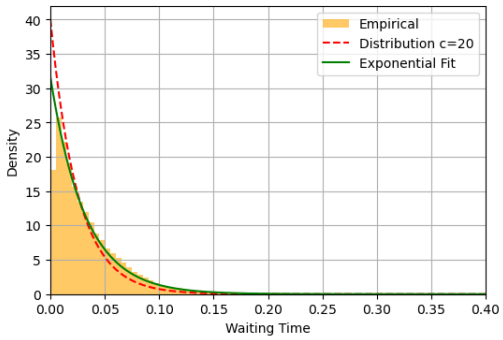
(C) QL distr., $c \sim Bin(25, \frac{4}{5})$



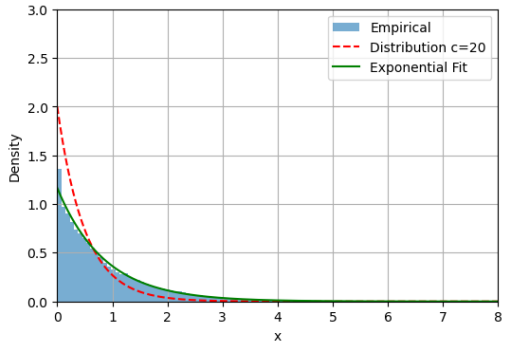
(D) ST distr., $c \sim Bin(25, \frac{4}{5})$



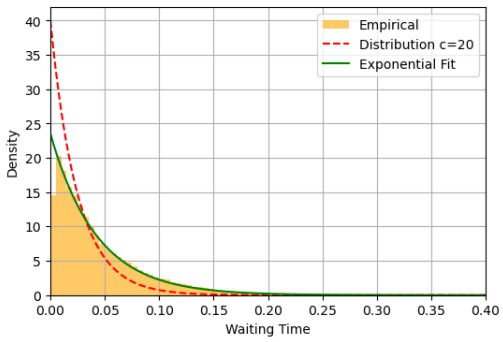
(E) QL distr., $c \sim Bin(28, \frac{5}{7})$



(F) ST distr., $c \sim Bin(28, \frac{5}{7})$



(G) QL distr., $c \sim Bin(100, \frac{1}{5})$



(H) ST distr., $c \sim Bin(100, \frac{1}{5})$

FIGURE 4.9: Heavy traffic limiting distributions of a multi-server queue with the number of servers $C \sim Binomial$. Comparison of empirical results to results for a constant server number $c = 20$. QL - queue length, ST - sojourn time, distr. - distribution.

As can be observed in the plots as well as the exponential rates, stochasticity in the number of servers leads to longer queues and sojourn times. The more variance we observe in the number of servers, the more substantial the increase is (e.g., 28 servers with success probability $\frac{5}{7}$ compared to 22 servers with $\frac{10}{11}$). This behavior also occurs in continuous-time *M/G/1* queues with i.i.d. service times, as suggested by the variance term in the denominator of (4.16) and (4.17).

Servers Following Binomial Infection and Recovery Evolution

Table 4.6 reports the mean and SD of queue length, including 95% CIs.

Server Distribution	Mean QL	SD QL
$n = 22, S_c \sim Bin(c, 0.02)$	167.81 [144.87; 202.63]	174.81 [145.29; 233.52]
$n = 25, S_c \sim Bin(c, 0.05)$	230.76 [191.72; 285.75]	237.51 [185.32; 334.37]
$n = 28, S_c \sim Bin(c, 0.08)$	265.85 [224.44; 317.94]	267.43 [211.17; 334.47]
$n = 100, S_c \sim Bin(c, 0.8)$	175.76 [152.60; 207.21]	178.13 [142.30; 235.22]

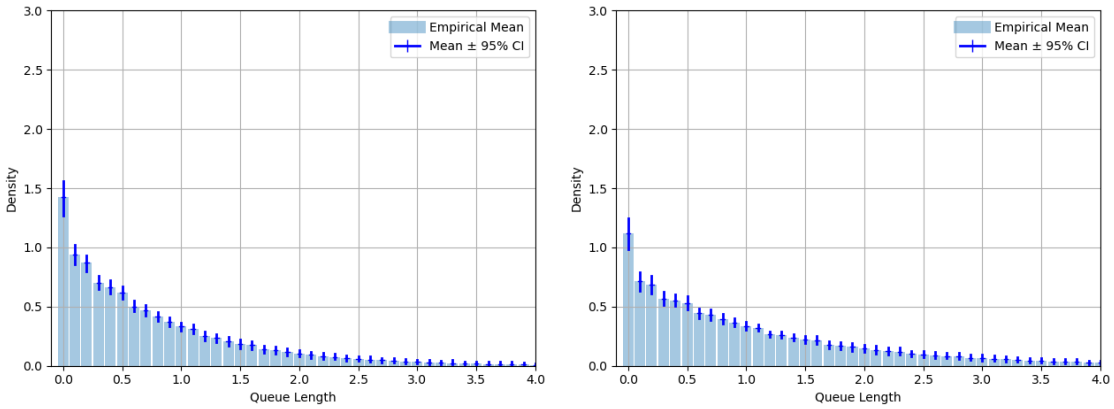
TABLE 4.6: Mean and standard deviation of queue length with servers following Binomial infection and recovery behavior in heavy traffic. QL - queue length, SD - standard deviation.

Up to a total server number of $n = 28$, queue lengths and SDs thereof increase. However, results for $n = 100$ are very similar to $n = 22$, breaking this pattern.

Again, the following Figures 4.10 and 4.11 present scaled results. The first presents the mean distribution and corresponding CIs. The latter compares distributions with servers following different Binomial infection and recovery evolutions to the Binomial server distribution and constant 20 servers. Table 4.7 reports rates of exponential fits to empirical queue length data.

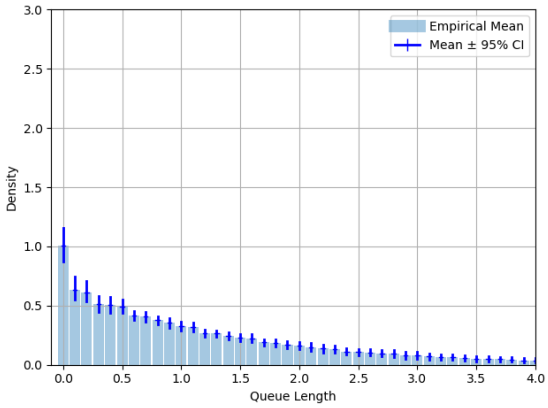
Server Distribution	Exponential Rate
$n = 22, S_c \sim Bin(c, 0.02)$	1.19 [0.99; 1.38]
$n = 25, S_c \sim Bin(c, 0.05)$	0.87 [0.70; 1.04]
$n = 28, S_c \sim Bin(c, 0.08)$	0.75 [0.63; 0.89]
$n = 100, S_c \sim Bin(c, 0.8)$	1.14 [0.97; 1.31]

TABLE 4.7: Rates of exponential fit to empirical queue length for servers following Binomial infection and recovery behavior.

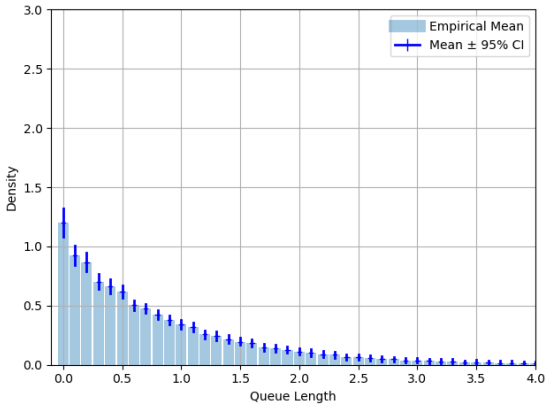


(A) $n = 22, S_c \sim Bin(c, 0.02)$

(B) $n = 25, S_c \sim Bin(c, 0.05)$



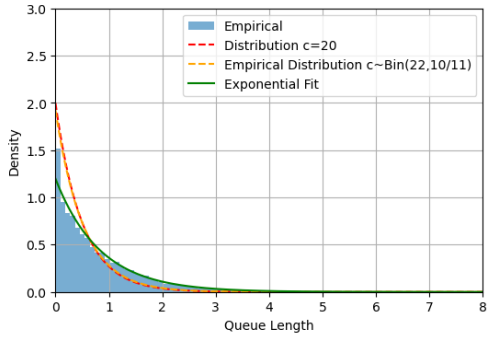
(C) $n = 28, S_c \sim Bin(c, 0.08)$



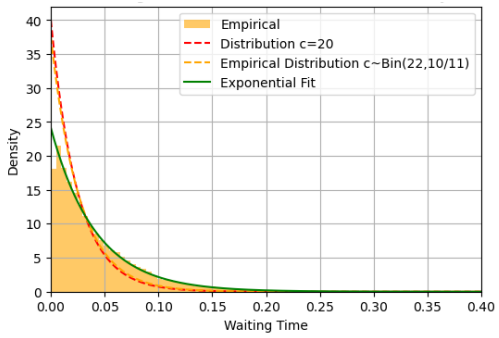
(D) $n = 100, S_c \sim Bin(c, 0.8)$

FIGURE 4.10: Confidence intervals of heavy traffic limiting distributions for multi-server queues with the number of servers following Binomial infection and recovery behavior.

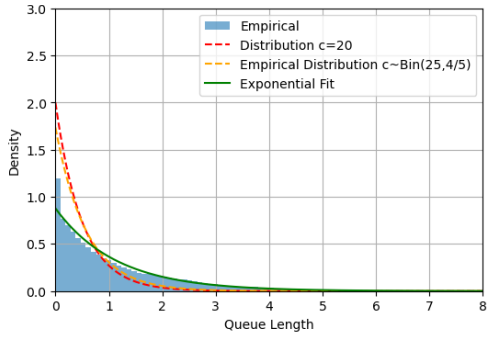
With the number of servers following this infection and recovery process, queues and sojourn times increase compared to the i.i.d. Binomial draw. The queue empties more often due to longer lasting periods of high server availability, leading to more lost resources. This also explains the spike in the histograms of the queue length at zero. However, in the extreme case, when $n = 100$, the difference between the i.i.d. Binomial draw and the infection and recovery behavior is very small. We suspect that due to the high variance in the Binomial distributions of nursing staff getting infected and recovering, the server availabilities on subsequent days vary more than in cases with smaller n. Since this reduces the temporal autocorrelation of available servers, the evolution of infection and recovery is closer to an independent draw from the Binomial distribution.



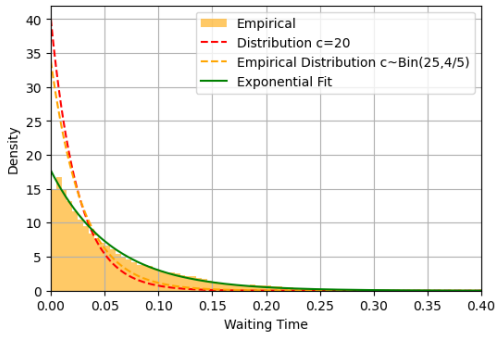
(A) QL distr., $n = 22, p = 0.02$



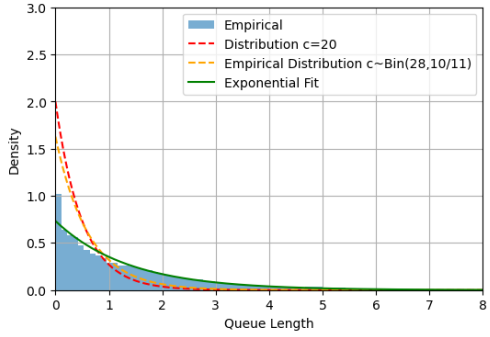
(B) ST distr., $n = 22, p = 0.02$



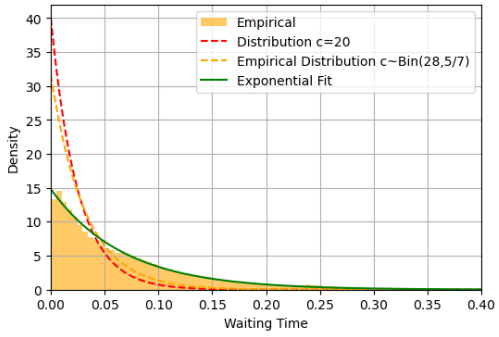
(C) QL distr., $n = 25, p = 0.05$



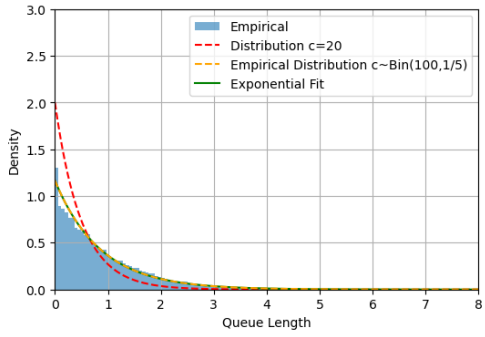
(D) ST distr., $n = 25, p = 0.05$



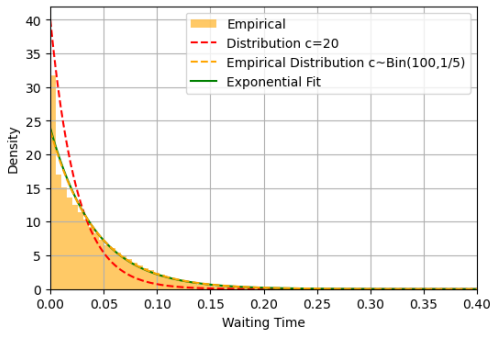
(E) QL distr., $n = 28, p = 0.08$



(F) ST distr., $n = 28, p = 0.08$



(G) QL distr., $n = 100, p = 0.8$



(H) ST distr., $n = 100, p = 0.8$

FIGURE 4.11: Heavy traffic limiting distributions of a multi-server queue with the number of servers following Binomial infection and recovery behavior. Comparison of empirical results to results for $C \sim Binomial$ and constant server number $c = 20$. QL - queue length, ST - sojourn time, distr. - distribution.

4.6.2 Moderate Traffic

For all three types of server configurations, we report means and SDs of queue lengths before illustrating distributions in Figures 4.12, 4.13 and 4.14.

Compared to the heavy traffic setting, the steady state queue length distribution does not follow a geometric distribution. As expected, due to the lower utilization, the queue length remains very short and has far less variance than in heavy traffic.

Constant Number of Servers

For constant 20 servers, the mean queue length is 0.57 (95%-CI: [0.57; 0.58]) with an SD of 1.64 (95%-CI: [1.62; 1.65]).

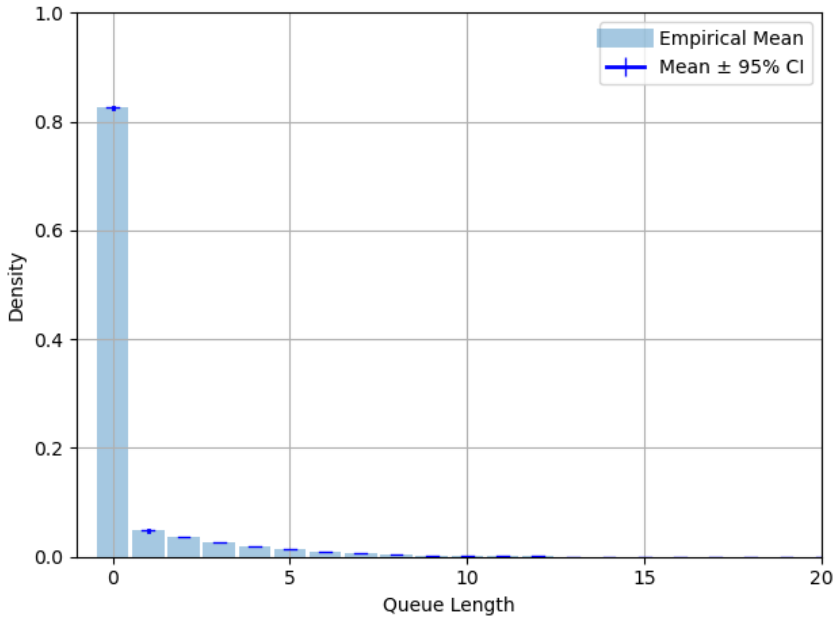


FIGURE 4.12: Confidence intervals of the moderate traffic limiting queue length distribution for multi-server queues ($c = 20$) with a constant number of servers.

Binomial Distributed Servers

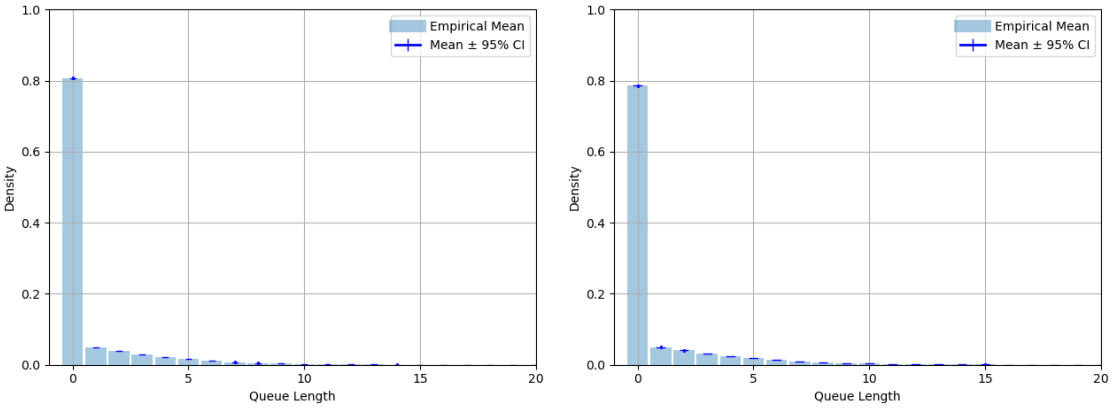
Arithmetic means and SDs of queue lengths can be found in Table 4.8.

Server Distribution	Mean QL	SD QL
$Bin(22, \frac{10}{11})$	0.68 [0.68; 0.69]	1.85 [1.84; 1.87]
$Bin(25, \frac{4}{5})$	0.82 [0.81; 0.82]	2.10 [2.08; 2.11]
$Bin(28, \frac{5}{8})$	0.93 [0.92; 0.93]	2.29 [2.27; 2.31]
$Bin(100, \frac{1}{5})$	1.60 [1.59; 1.61]	3.39 [3.36; 3.42]

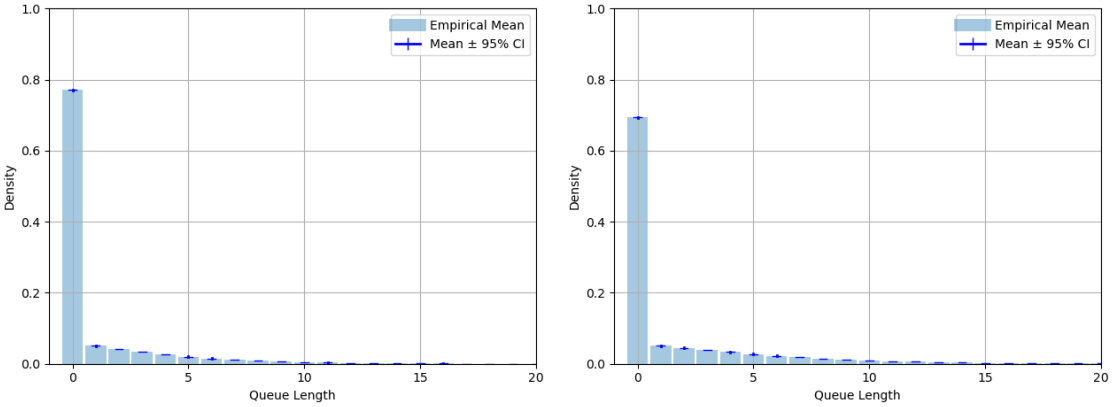
TABLE 4.8: Mean and standard deviation of queue length with Binomial distributed servers in moderate traffic. QL - queue length, SD - standard deviation.

Similarly to the heavy traffic setting, variability in the number of servers leads to an increased queue length and SDs thereof with a more intense effect for higher variability. As can be seen in Figure 4.13 it does not follow an exponential distribution, unlike the

heavy traffic results. Besides, the CIs are much narrower, indicating less variability in the distribution compared to the critical heavy traffic setting.



(A) $c \sim Bin(22, \frac{10}{11})$ (B) $c \sim Bin(25, \frac{4}{5})$



(C) $c \sim Bin(28, \frac{5}{7})$ (D) $c \sim Bin(100, \frac{1}{5})$

FIGURE 4.13: Confidence intervals of the moderate traffic limiting queue length distribution for multi-server queues with the number of servers $C \sim Binomial$.

Servers Following Binomial Infection and Recovery Evolution

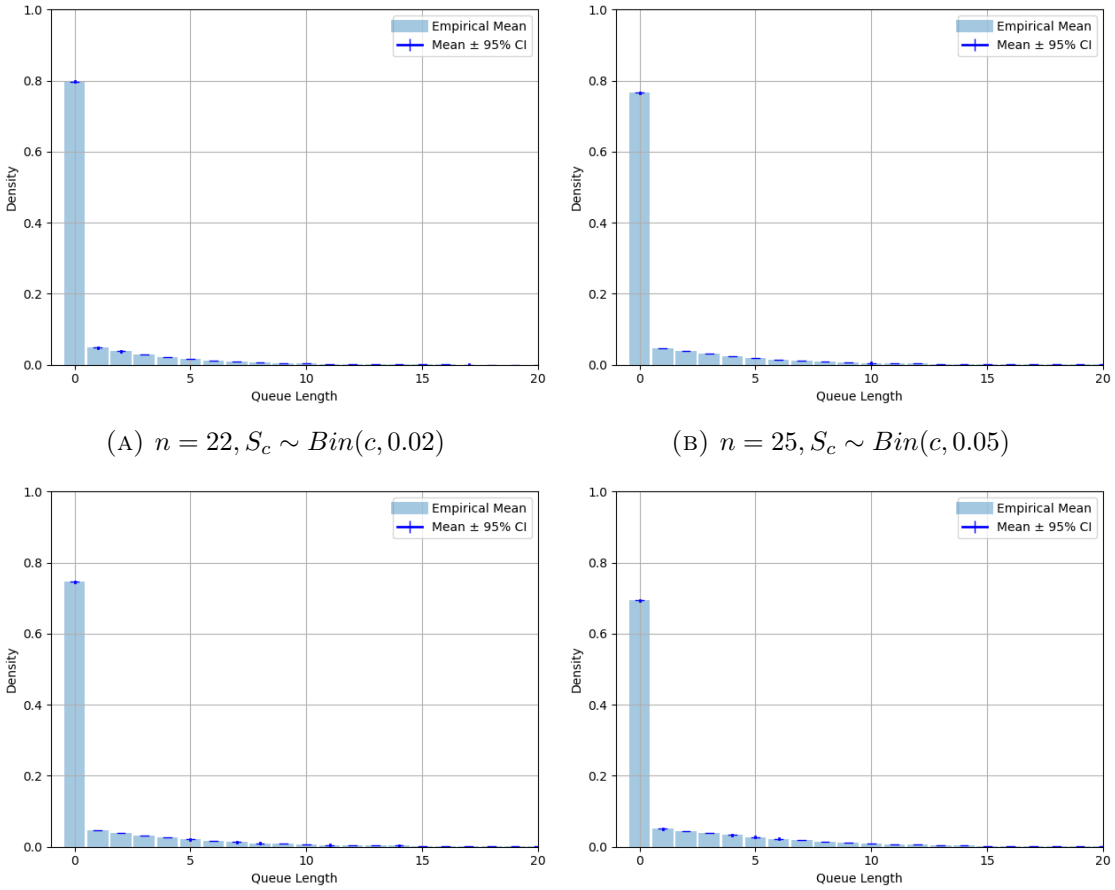
Table 4.9 reports means and SDs of queue lenghts.

Server Distribution	Mean QL	SD QL
$n = 22, S_c \sim Bin(c, 0.02)$	0.82 [0.81; 0.83]	2.28 [2.24; 2.32]
$n = 25, S_c \sim Bin(c, 0.05)$	1.15 [1.14; 1.17]	3.11 [3.06; 3.16]
$n = 28, S_c \sim Bin(c, 0.08)$	1.41 [1.39; 1.43]	3.67 [3.60; 3.75]
$n = 100, S_c \sim Bin(c, 0.8)$	1.60 [1.59; 1.62]	3.39 [3.36; 3.42]

TABLE 4.9: Mean and standard deviation of queue length with servers following Binomial infection and recovery behavior in moderate traffic. QL - queue length, SD - standard deviation.

Again, as already observed under heavy traffic, this autocorrelation of server availabilities leads to longer queues opposed to an i.i.d Binomial draw of the number of servers. On

top, just as in the heavy traffic regime, we observe very similar distributions for the i.i.d. Binomial draw and the infection and recovery evolution for a total amount of 100 servers in Figure 4.13d and 4.14d, respectively.



(A) $n = 22, S_c \sim Bin(c, 0.02)$

(B) $n = 25, S_c \sim Bin(c, 0.05)$

(C) $n = 28, S_c \sim Bin(c, 0.08)$

(D) $n = 100, S_c \sim Bin(c, 0.8)$

FIGURE 4.14: Confidence intervals of the moderate traffic limiting queue length distribution for multi-server queues with the number of servers following Binomial infection and recovery behavior.

4.6.3 Hypercritical Traffic

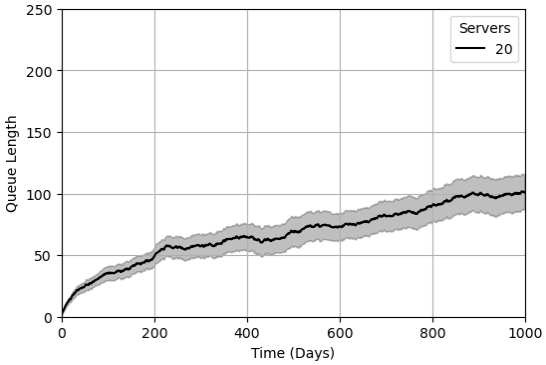
When demand surpasses capacity, queues grow without bound instead of reaching a stable state. We first look at a utilization of 100%, and then of 120%.

100% Utilization

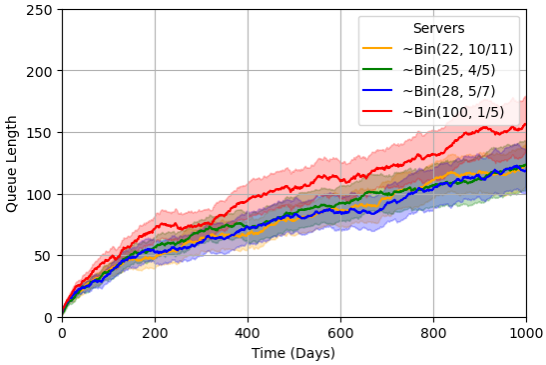
Figure 4.15a shows the queue growth for a constant number of 20 servers. For Binomial availability of servers, queues grow slightly faster, as can be observed in Figure 4.15b. Even though CIs are largely overlapping, we can observe a slight trend of more variability in the number of servers leading to a more intense queue length growth. For Binomial infection and recovery behavior (Figure 4.15c), the queue length grows even faster, with the case of 100 servers being very close to the individual Binomial draw, once again.

Although we observe constant queue growth on average, we attribute the differences between server distributions to emptying queues within single simulation runs. These

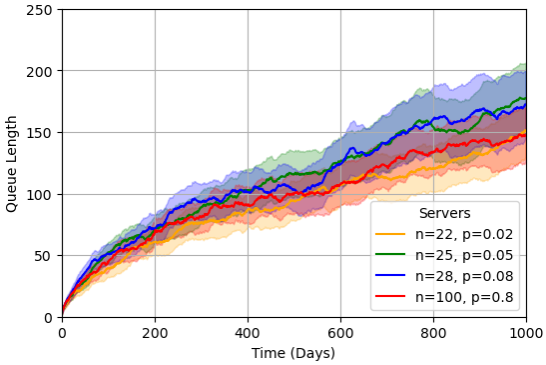
are, just as in the heavy traffic case, more likely with higher variability and temporal autocorrelation of server numbers.



(A) Constant number of servers.



(B) Binomial distribution of servers.

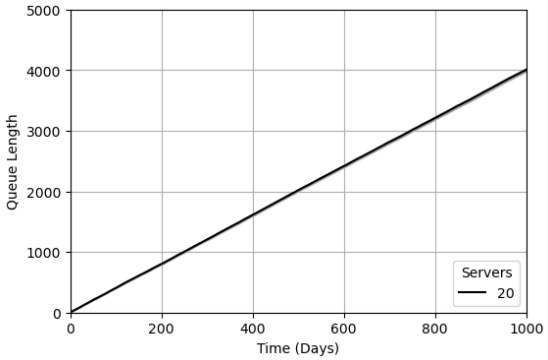


(C) Servers according to Binomial infection and recovery behavior.

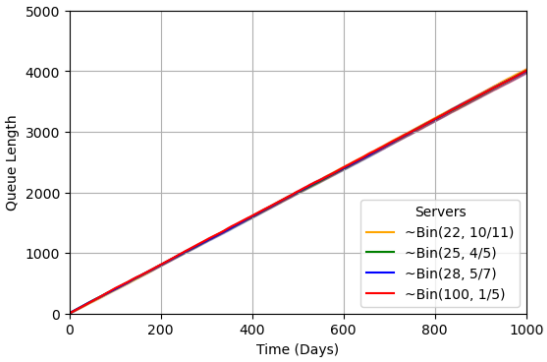
FIGURE 4.15: Queue length evolution with hypercritical utilization $\rho = 1$. Comparison between server numbers $c = 20$, $c \sim \mathit{Binomial}$, and c following Binomial infection and recovery behavior.

120% Utilization

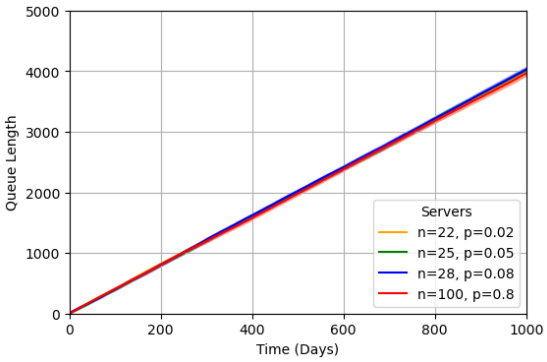
In a highly overloaded setting with 120% utilization, there are no differences in the evolution of the queue length for different server characteristics, as can be seen in Figure 4.16. The queue never empties, meaning that all available servers are always fully occupied. Therefore, also the queues with server variability behave like the constant 20 server queue in the long-term, since the expected number of available servers corresponds to 20.



(A) Constant number of servers.



(B) Binomial distribution of servers.



(C) Servers according to Binomial infection and recovery behavior.

FIGURE 4.16: Queue length evolution with hypercritical utilization $\rho = 1.2$. Comparison between server numbers $c = 20$, $c \sim Binomial$, and c following Binomial infection and recovery behavior.

4.7 Summary

This chapter investigated the impact of variability in nursing resources on the backlog of care. We developed a queueing model and analytically derived the heavy traffic limiting queue length distribution for a constant number of servers. In numerical experiments, we compared to stochastic server scenarios representing the characteristics of varying resources during health emergencies.

We derived the heavy traffic limiting distribution for the *Poisson/1/c* queue representing a ward with a constant number of resources. We related it to a single-server queue and argued why the obtained results correspond to an accordingly scaled continuous single-server queueing system. For the configuration with varying server availabilities, we identified a difference between the discrete multi-server system and scaled continuous single-server queue.

Numerical experiments showed longer queues and sojourn times for the number of servers following a Binomial distribution compared to the constant case in heavy traffic. Additionally, the higher the variance in the number of servers, the bigger this effect. In the setting of server availability evolving according to Binomial infection and recovery behavior, temporal autocorrelation leads to even longer queue lengths due to the higher likelihood of the queue going idle.

We found similar patterns in numerical analyses with lighter traffic, that is an arrival rate corresponding to an utilization of 0.8 in the constant server case. In heavily overloaded settings, however, variability of servers does not impact queue length, as long as the expectation of the number of servers stays unchanged. This is due to the queue never emptying.

The main takeaway is that variability and especially temporal autocorrelation leads to longer steady state queue lengths. So even if the long-time expectation of resources did not change, the characteristics of fluctuating resources during a pandemic leads to a longer queue build up.

Chapter 5

Impact of Admission Policies on Backlog and Cancellations

Policies that determine *when* and *how many* patients are admitted for elective surgery influence how efficiently resources can be used. Especially in pandemic circumstances, where resources are highly uncertain and varying, this is critical. In such settings, scheduling policies can play a decisive role in keeping backlogs manageable or, alternatively, in causing the system to become unstable and overwhelmed.

This chapter investigates the effect of different scheduling strategies in terms of timing and volume of patients. We conduct numerical analyses with different arrival intensities for the scheduling strategies to examine the stability of the system. The evolution of the backlog and cancellations of patients constitute a trade-off, thus we present results for both metrics.

Modeling assumptions for this chapter are the ones outlined in Section 3.3. Section 5.1 motivates and presents parameter choices for numerical experiments. We present results on constant staff infection probabilities in Section 5.2, followed by fluctuating probabilities in Section 5.3. Section 5.4 compactly compares scheduled, realized and canceled workload between all considered scheduling policies. Section 5.5 summarized the most important findings of this chapter.

5.1 Experimental Setup

This section details the key components of our simulation framework. We begin by specifying the parameter values in Section 5.1.1. Next, Section 5.1.2 discusses the modeling of NTD and associated boundary effects. Section 5.1.3 outlines the patient scheduling policies, including timing and volume strategies for scheduling. Finally, Section 5.1.4 presents the computational environment.

5.1.1 Parameter Values

We consider a time horizon of 50 days. The total amount of nursing staff is $n = 25$, and the long-term expected number of present nurses is 20, corresponding to 100 and 125 nursing hours, respectively. Nursing staff absence follows the model introduced in Section 3.4, and we consider two scenarios, one with constant infection probabilities and the other with fluctuating ones. In the constant case, each nurse has a fixed daily probability of becoming unavailable, $p = 0.05$. In the dynamic case, the infection probability varies with time, modeled as (5.1) for $t \in \mathbb{N}$.

$$p(t) = 0.045 \cdot sin(\frac{t}{5}) + 0.05 \tag{5.1}$$

As illustrated in Figure 5.1, this yields values between 0.01 and 0.09, with a mean of 0.05 in the long-term. This allows for comparison to the the case of constant infection probabilities.

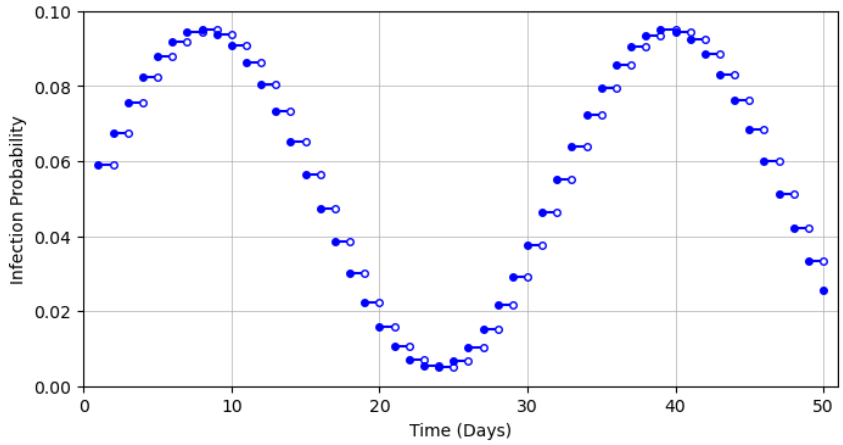


FIGURE 5.1: Daily nurse infection risks p_t used in numerical simulations.

The recovery rate is $q = 0.02$, resulting in fairly short periods of absence. Since we consider absence periods not only due to illness, but also due to care duties and assume that nurses do not have to sit out a long quarantine period, we consider them plausible in reality.

5.1.2 Choice of Nursing Time Demand k

An important modeling aspect is the NTD, denoted by k, which must be met on the admission day of a patient. That is, a patient can only be admitted if their NTD fully fits into the available nursing capacity on that day. This leads to boundary effects, where available capacity remains unused simply because it cannot accommodate any full patient admission due to our modeling choice.

In reality, even though the daily NTD (not multiplied by the patient's LoS) must be accommodated, partial leftover capacity may still be wasted. For example, suppose a nurse is responsible for three patients with an NPR of 1:4. Then a quarter of their capacity may go unused if no additional patient fits the residual time, due to capacity constraints of other shifts or incompatible demands.

While such inefficiencies reflect real-world challenges, choosing a large k could significantly overestimate the impact of this boundary effect. For instance, if $k = 24$ (e.g., corresponding to an NPR 1:4 and LoS of 4 days), the model would often leave capacity unused simply because no other patient fits. Therefore, we set $k = 3$ for our numerical analyses to mitigate the overestimation of lost resources.

5.1.3 Scheduling Policies

We explore various scheduling policies with respect to both the *timing of scheduling* and the *number of patients scheduled*. We explore all combinations of the considered policy

choices for different arrival rates of patients. These arrival rates correspond to 93, 96 and 99 mean nursing hours per day, all allowing for a stable system, if resources are used efficiently.

Timing of Scheduling

Scheduling s days in advance uses the number of nurses present on day $t - s$ and the infection probabilities of days $t - s, \ldots, t - 2, t - 1$ to find the probability distribution of nursing availability on day t. We consider scheduling one, two and three days before admission, and refer to these policies as *1-day-ahead*, *2-day-ahead* and *3-day-ahead* scheduling, respectively.

Scheduling Volume

The probability distribution of available nurses leads to an expected value of nursing capacity for elective care on day t. We explore scheduling according to that expected number, and a conservative and overbooking strategy. In the conservative strategy, the goal is to minimize cancellations. Patients are only scheduled if their admission would still be feasible even if one nurse fewer than expected is present on the day. Conversely, the overbooking strategy accepts a higher risk of cancellations in an effort to use available resources more efficiently. Here, the number of patients scheduled corresponds to the assumption that one nurse more than expected will be available. An additional constraint ensures that the scheduled demand never exceeds the maximum workforce capacity of $n = 25$.

We will further refer to these strategies as *expected-value scheduling*, *conservative scheduling* and *overbooking*, respectively.

5.1.4 Computational Details

Simulations were coded in Python version 3.10.12 and performed on a server at the University of Twente provided via the JupyterLab environment. The CPU server has 64 cores, 128 threads and 1024 GB memory. For each combination of variables (infection probability, arrival rate, timing of scheduling, and scheduling volume), we simulated 50 independent trajectories of 50 days each. For each day, the expected and realized nursing capacity, and the scheduled, canceled and realized nursing time volume was recorded. Simulation results were aggregated to compute daily averages and 95% CIs for these key metrics.

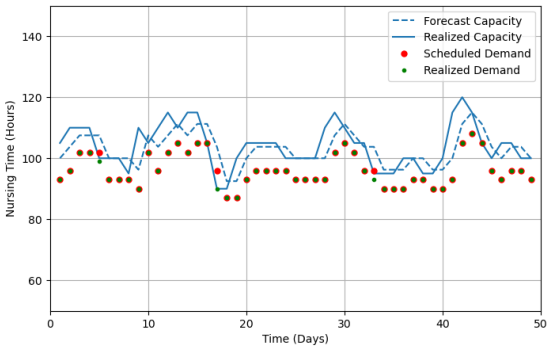
5.2 Results under Constant Infection Risk

This section presents results under a constant staff infection risk. Section 5.2.1 analyzes single simulation runs to gain insight into the dynamics between expected and realized capacities, and corresponding scheduled and realized workload. To clearly show the timing and structure of key dynamics, we present results from a single simulation run rather than aggregated outcomes, which can mask important patterns. Multiple runs showed consistent behavior, so one representative trajectory is sufficient to illustrate the effects of interest without requiring CIs. Section 5.2.2 presents aggregated numerical results on backlog evolution over the whole simulation period. Lastly, we analyze cancellations in Section 5.2.3, again averaged over the trajectories per day for the whole simulation period. As we assume non-empty queues in this analysis, the arrival rate of patients does not influence the cancellation behavior nor the dynamics we analyze in the single trajectory.

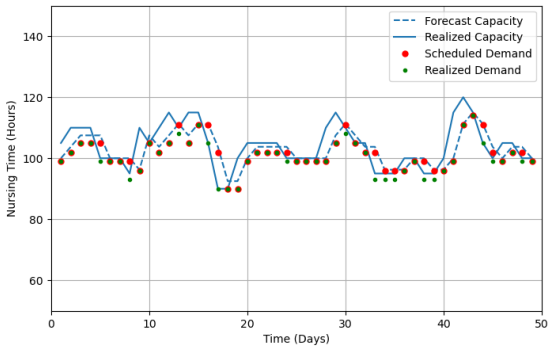
One result for every scheduling policy thus represents all arrival intensities in these two sections.

5.2.1 Individual Trajectories

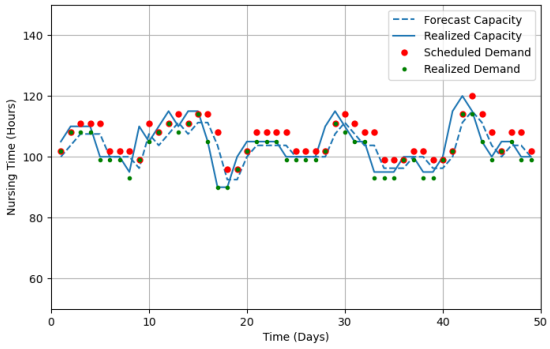
Figures 5.2, 5.3 and 5.4 present randomly chosen trajectories under three volume strategies: *conservative scheduling* (A), *expected-value scheduling* (B) and *overbooking* (C), applied with timing policies of planning one, two and three days in advance, respectively. In each plot, the expected and actual nursing capacities in hours are illustrated by the dashed and solid blue lines, respectively. The workloads that the policy scheduled are represented by red dots. The workloads, which could then actually be realized on the respective day are shown as green dots. Note that realized workload might not coincide with realized capacity, since demand can only be served in blocks of NTD corresponding to full patients. On days, where the green dot does not cover the red one, patients are canceled at short notice and returned to the waitlist.



(A) Conservative scheduling.

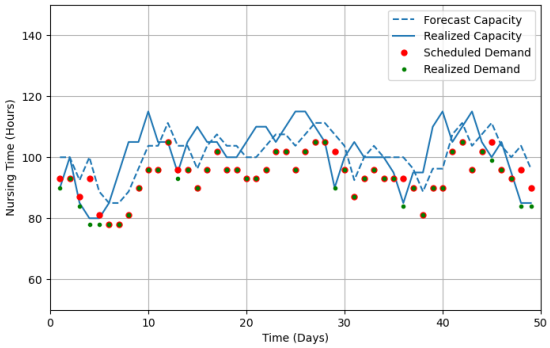


(B) Expected-value scheduling.

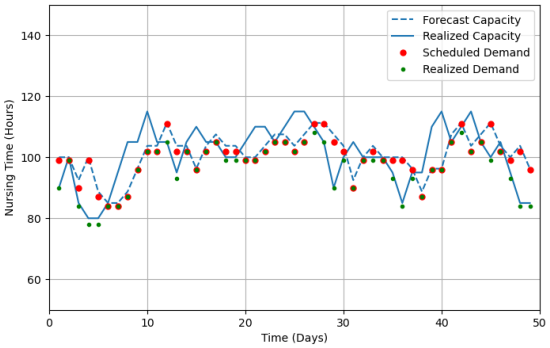


(C) Overbooking.

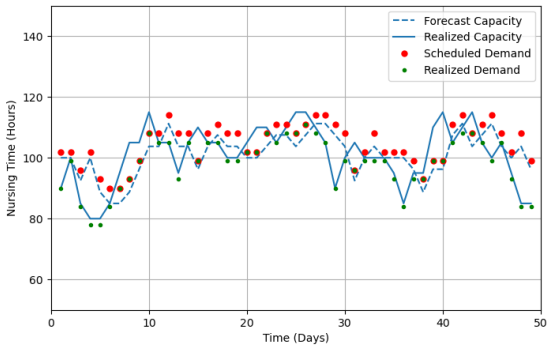
FIGURE 5.2: Example trajectories showing forecast vs. realized capacity and scheduled vs. realized workload under different patient volume strategies. *1-day-ahead* scheduling under constant infection risks.



(A) Conservative scheduling.

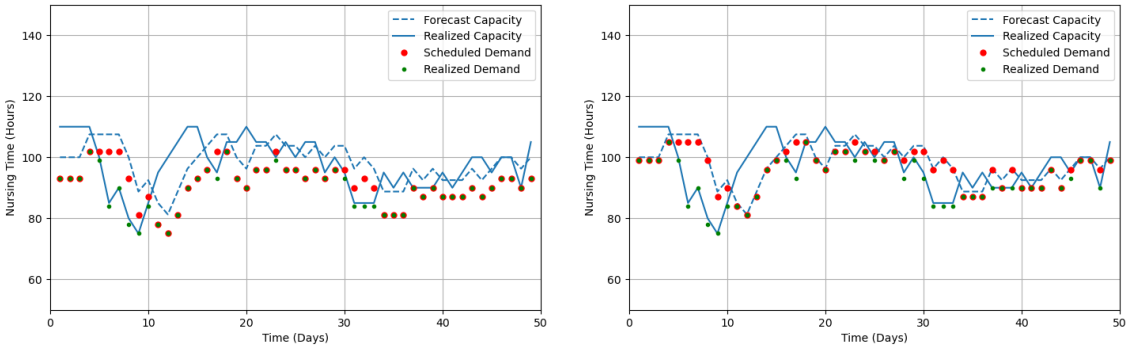


(B) Expected-value scheduling.



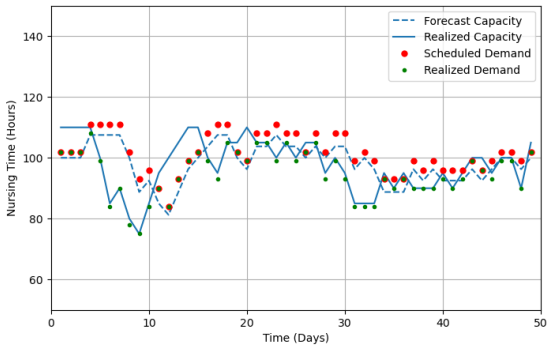
(C) Overbooking.

FIGURE 5.3: Example trajectories showing forecast vs. realized capacity and scheduled vs. realized workload under different patient volume strategies. *2-day-ahead* scheduling under constant infection risks.



(A) Conservative scheduling.

(B) Expected-value scheduling.



(C) Overbooking.

FIGURE 5.4: Example trajectories showing forecast vs. realized capacity and scheduled vs. realized workload under different patient volume strategies. 3-day-ahead scheduling under constant infection risks.

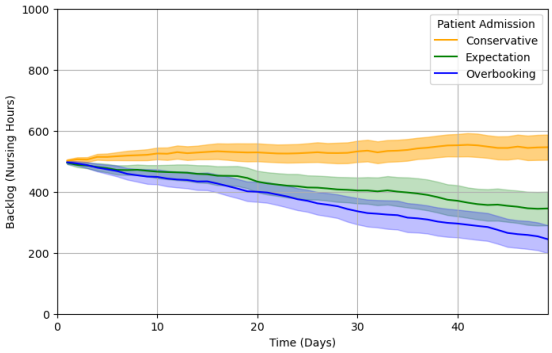
Since the forecast capacity also takes into account the currently available nurses, we observe a time-shift between the behavior of the actual and expected resources. So, for instance, if the actual capacity goes up like on day 2 in Figure 5.2, the expectation, hence also the amount of scheduled patients, goes up only on day 3. This delay corresponds to the scheduling horizon used in the respective policy.

Regarding the volume strategies, scheduling more patients brings realized demand closer to actual capacity, enhancing resource utilization. However, *overbooking* consistently leads to excess scheduled workload and hence results in high cancellation numbers. This highlights a clear trade-off between maximizing capacity utilization and avoiding patient cancellations due to uncertainty.

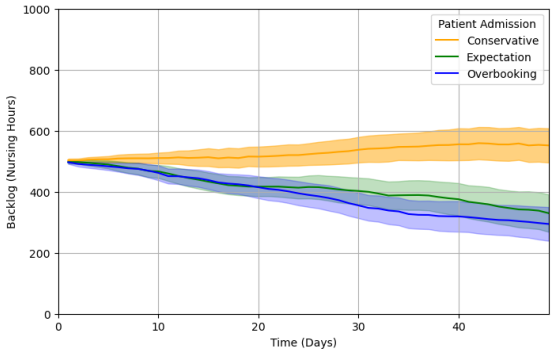
5.2.2 Backlog Evolution

Figures 5.5, 5.6 and 5.7 present numerical results on backlog evolution for different arrival rates of 93, 96 and 99 nursing hours on average, respectively. Subfigures (A), (B) and (C) correspond to the scheduling horizons of one, two and three days, respectively, and compare the impact of patient volume strategies. For each scheduling policy, the mean backlog trajectory is represented as a line and the associated 95% CI indicated by a surrounding shaded area.

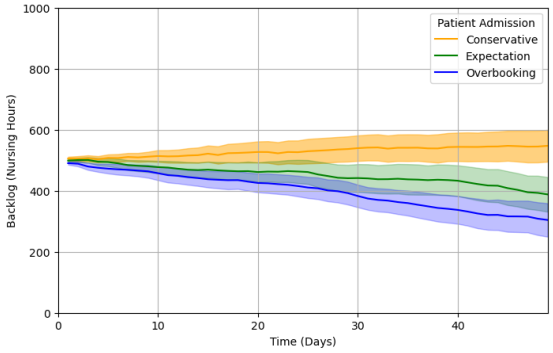
Arrival Rate 93



(A) 1-day-in-advance.



(B) 2-days-in-advance.

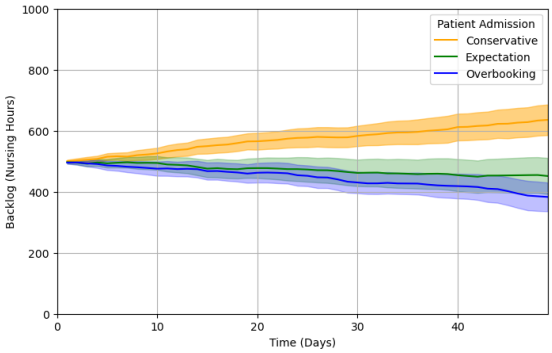


(C) 3-days-in-advance.

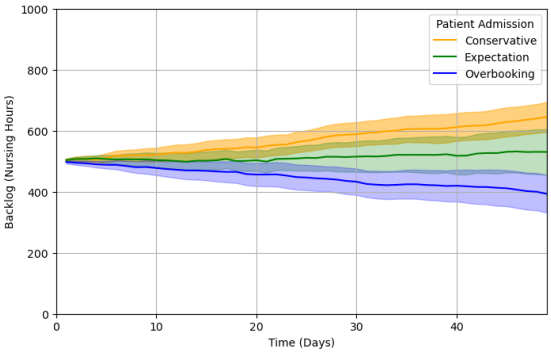
FIGURE 5.5: Backlog evolution under constant infection risks with arrival rate 93. Mean and 95% CI for different timing and patient volume scheduling strategies.

Under conditions of low strain on the healthcare system, all investigated strategies are capable of maintaining system stability or even reducing the backlog. For *conservative scheduling*, the timing policy appears to have little to no effect, with backlog levels remaining relatively constant across all planning horizons. In contrast, both *expected-value scheduling* and *overbooking* lead to a reduction in backlog. However, their efficiency in reducing backlog slightly decreases as the scheduling horizon increases. That is, the further in advance patients are scheduled, the less effective these strategies appear to be.

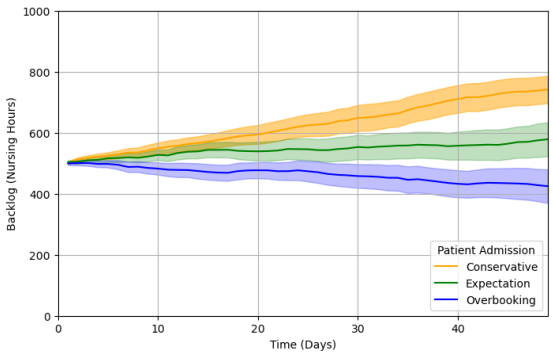
Arrival Rate 96



(A) 1-day-in-advance.



(B) 2-days-in-advance.

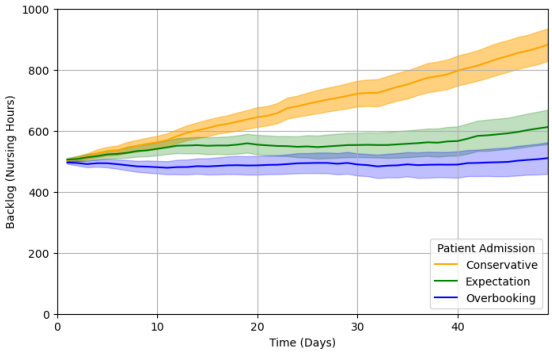


(C) 3-days-in-advance.

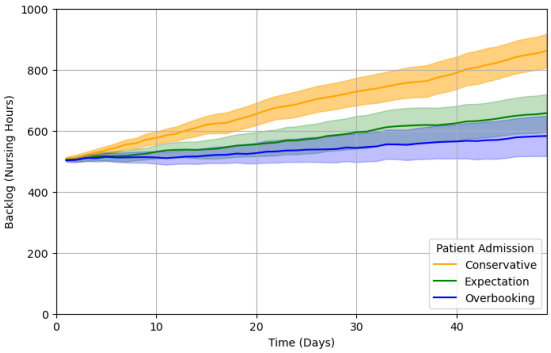
FIGURE 5.6: Backlog evolution under constant infection risks with arrival rate 96. Mean and 95% CI for different timing and patient volume scheduling strategies.

As arrival rates increase, the choice of volume strategy becomes decisive in determining whether the system remains stable or experiences backlog growth. *Conservative scheduling* leads to steadily increasing backlogs, whereas *overbooking* slightly reduces them. *expected-value scheduling* results in a backlog trajectory that hovers near the stability threshold. With a short scheduling horizon of one day in advance, backlogs are most likely to remain stable. However, as the scheduling horizon extends, the likelihood of maintaining stability decreases, and a slight growth over time is to be expected.

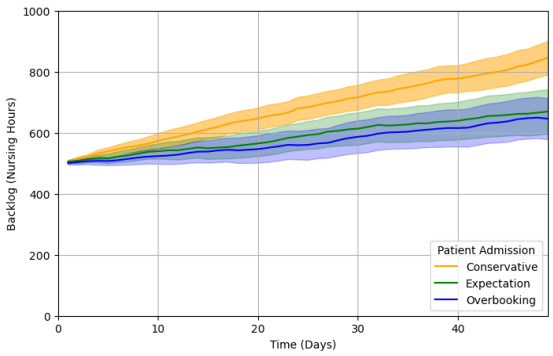
Arrival Rate 99



(A) 1-day-in-advance.



(B) 2-days-in-advance.



(C) 3-days-in-advance.

FIGURE 5.7: Backlog evolution under constant infection risks with arrival rate 99. Mean and 95% CI for different timing and patient volume scheduling strategies.

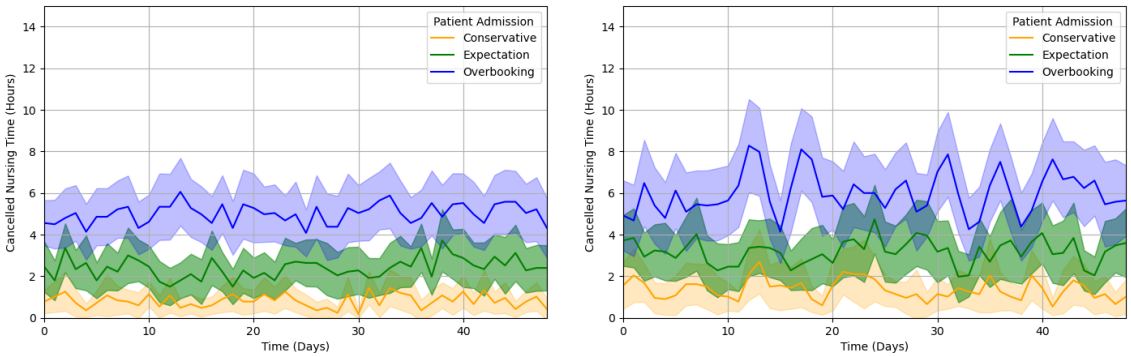
When the arrival rate approaches the operational upper limit, nearly all investigated strategies fail to prevent backlog growth. The only configuration likely to maintain system stability is *overbooking* with a one-day scheduling horizon. *Conservative scheduling* consistently results in rapid backlog accumulation, with the waitlist increasing from 500 to between 800 and 900 nursing hours within the 50-day simulation period, regardless of the timing policy. *Expected-value scheduling* leads to a moderate backlog increase, which is somewhat sensitive to the scheduling horizon. The further in advance patients are planned, the faster the backlog tends to grow. For scheduling horizons longer than one day, even *overbooking* fails to stabilize the system. In such cases, more aggressive overbooking would likely be required to further reduce underutilization of available resources.

In summary, under constant staff infection probabilities, the arrival rate of patients has a strong influence on system stability. At 93 nursing hours arriving on average, the backlog remains stable across all scenarios. In contrast, with 99 hours, the system almost never manages to keep the backlog from increasing. Across most settings, the backlog trajectories under different volume strategies diverge significantly, as reflected in largely non-overlapping CIs. This suggests a strong effect of the chosen volume strategy. Differences between the timing policies (one, two or three days in advance) are relatively small and do not reach statistical significance.

5.2.3 Cancellations

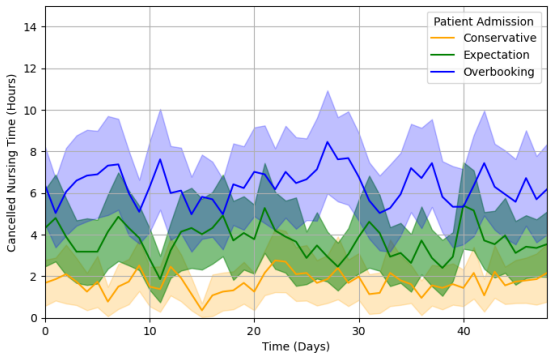
Under the *conservative scheduling* patient volume strategy, the average daily canceled workload amounts to 0.8, 1.4, and 1.7 nursing hours when scheduling one, two, and three days in advance, respectively. When increasing the scheduled volume, these values rise to 2.4, 3.2, and 3.7 nursing hours for *expected-value scheduling*, and to 5.0, 6.0, and 6.4 for *overbooking*.

Figure 5.8 visualized cancellations of all combination of scheduling timing and patient volume policies.



(A) 1-day-in-advance.

(B) 2-days-in-advance.



(C) 3-days-in-advance.

FIGURE 5.8: Cancellations under constant infection risks. Mean and 95% CI for different timing and patient volume scheduling strategies.

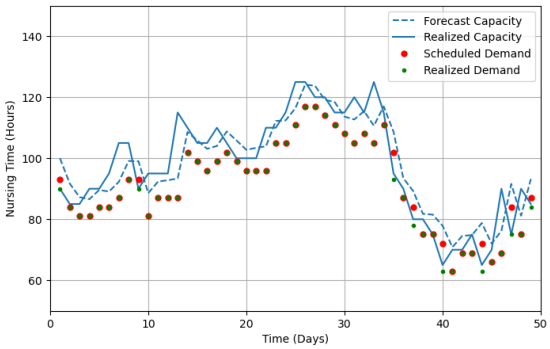
No temporal trend can be observed. With respect to timing policies, cancellations tend to increase slightly as scheduling is done further in advance. However, these differences are not statistically significant, as CIs overlap. We notice, though, that the cancellation trajectories become more volatile with longer planning horizons, indicating occasional larger mismatches between expected and actual capacity. The most notable differences result from the patient volume strategy. As already seen in the single simulation trajectories in Section 5.2.1, scheduling a larger patient volume leads to more cancellations. While under *conservative scheduling*, cancellations hardly exceed 2 nursing hours per day, this value increases to over 6 hours for *overbooking* quite often.

5.3 Results under Fluctuating Infection Risk

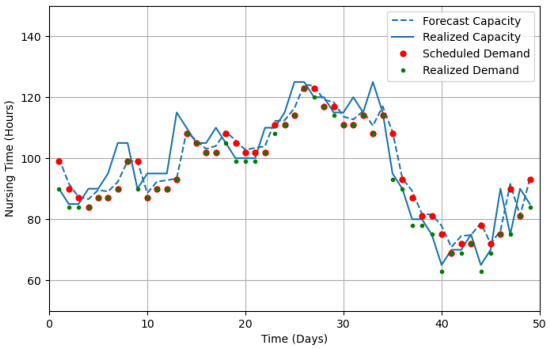
This section presents results under a fluctuating staff infection risk following (5.1). Section 5.3.2 analyzes single simulation runs to gain insight into the dynamics between expected and realized capacities, and corresponding scheduled and realized workload. Again, we chose to display individual trajectories, to clearly show the timing and structure of key dynamics, and checked for consistent behavior in multiple runs. Section 5.3.2 presents aggregated numerical results on backlog evolution over the whole simulation period. Lastly, we analyze cancellations in Section 5.3.3. The arrival rate does not influence the cancellation behavior nor the dynamics analyzed in the single trajectory. Therefore, one plot per scheduling policy represents all arrival intensities in the respective sections.

5.3.1 Individual Trajectories

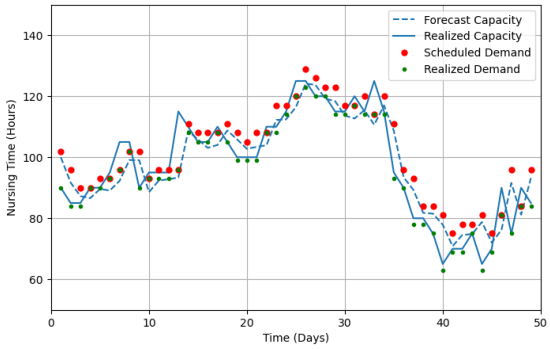
Figures 5.9, 5.10, and 5.11 illustrate randomly selected trajectories under *conservative scheduling* (A), *expected-value scheduling* (B), and *overbooking* (C), each for timing policies of one, two, and three days in advance, respectively.



(A) Conservative scheduling.

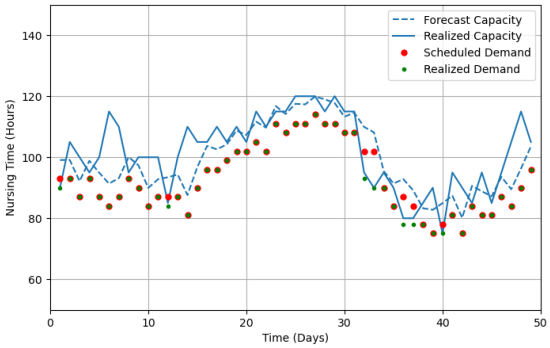


(B) Expected-value scheduling.

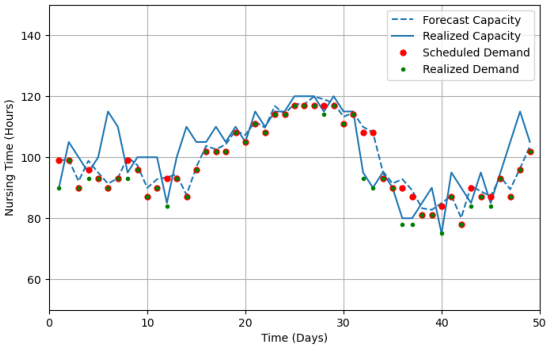


(C) Overbooking.

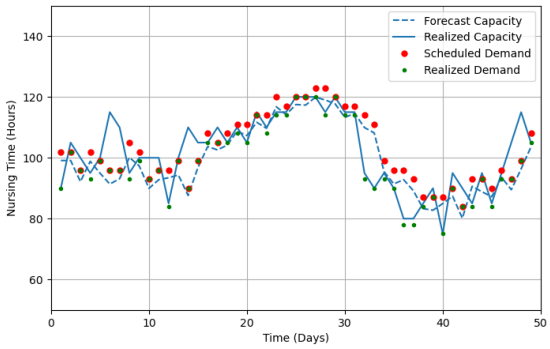
FIGURE 5.9: Example trajectories showing forecast vs. realized capacity and scheduled vs. realized workload under different patient volume strategies. *1-day-ahead* scheduling under fluctuating infection risks.



(A) Conservative scheduling.

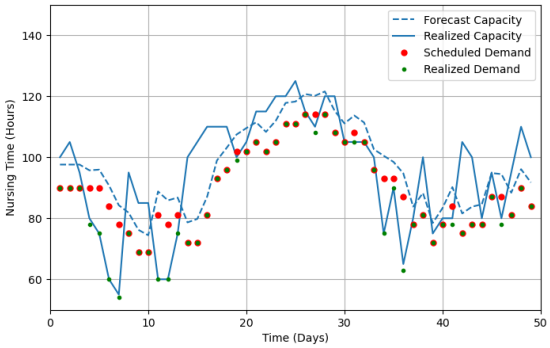


(B) Expected-value scheduling.

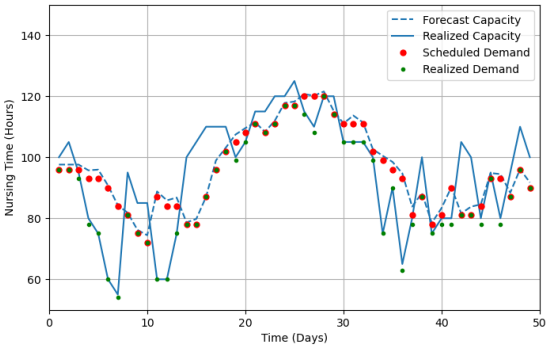


(C) Overbooking.

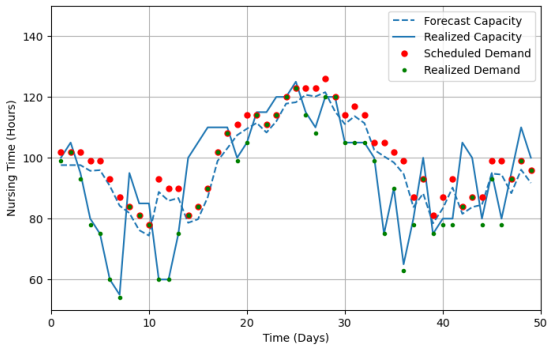
FIGURE 5.10: Example trajectories showing forecast vs. realized capacity and scheduled vs. realized workload under different patient volume strategies. *2-day-ahead* scheduling under fluctuating infection risks.



(A) Conservative scheduling.



(B) Expected-value scheduling.



(C) Overbooking.

FIGURE 5.11: Example trajectories showing forecast vs. realized capacity and scheduled vs. realized workload under different patient volume strategies. *3-day-ahead* scheduling under fluctuating infection risks.

The evolution of capacity over time reflects the influence of fluctuating staff infection probabilities. As discussed in Example 3.4.3, where we analyzed probability distributions of staff availability based fluctuating infection probabilities following a similar sine wave pattern, a time lag becomes apparent. In all trajectories, the maximum capacity is reached around day 30, even though infection rates hit their minimum around day 24, see Figure 5.1.

Similar to the scenario with constant infection probabilities, scheduling larger patient volumes leads to more efficient use of resources but also increases the likelihood of cancellations. Again, a time lag emerges between actual and expected capacity, determined by how long in advance patients are scheduled. For example, when a surge in available capacity occurs, such as on day 13 in Figure 5.9, this is reflected in the expected capacity (and scheduling decisions) the day after, as a scheduling horizon of one day was used.

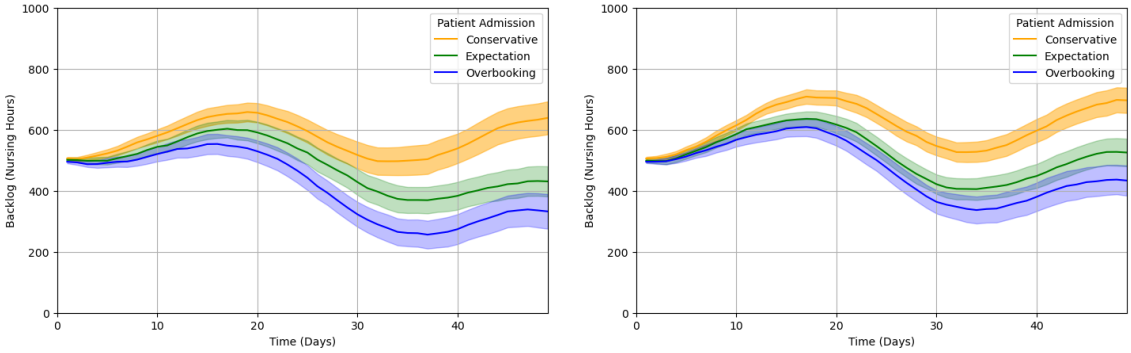
We observe the nature of fluctuating probabilities in future capacity expectations. Unlike the constant case, equal realized capacities no longer imply equal future expectations. This is evident in Figure 5.11, days 16 to 18, where realized capacities remain stable. Nevertheless, expected capacities for days 19 to 21 vary, as they incorporate the trend of decreasing infection probabilities.

5.3.2 Backlog Evolution

Figures 5.12, 5.13 and 5.14 illustrate backlog evolutions for scheduling one (A), two (B) and three (C) days ahead under arrival rates corresponding to an average number of 93, 96 and 99 nursing hours, respectively.

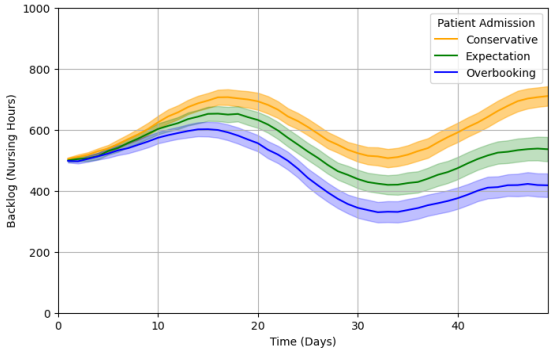
Note that we do not simulate for a multiple time of the cycle length of infection probabilities, which correspond to the phase length of (5.1). To draw conclusions about the trend of the backlog in the long-term, we should therefore compare the backlog at two points in time, which are one phase apart. Based on a a phase length of $10\pi \approx 31$ days, the trend can be observed by comparing day 19 to day 49.

Arrival Rate 93



(A) 1-day-in-advance.

(B) 2-days-in-advance.



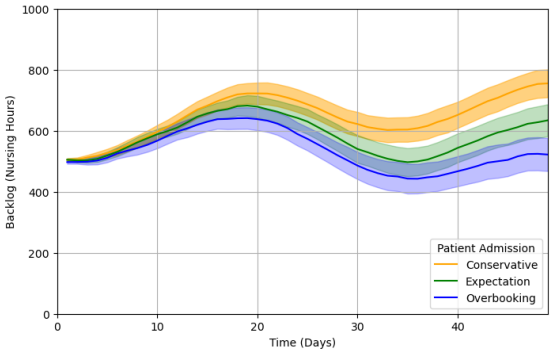
(C) 3-days-in-advance.

FIGURE 5.12: Backlog evolution under fluctuating infection risks with arrival rate 93. Mean and 95% CI for different timing and patient volume scheduling strategies.

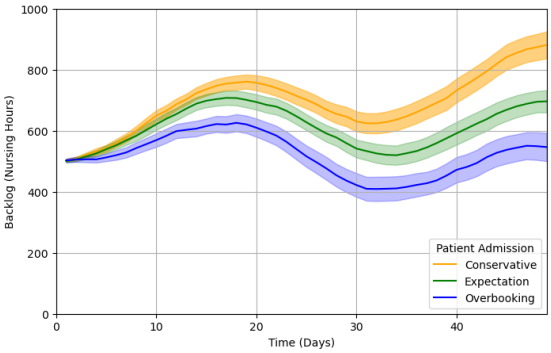
With an arrival rate corresponding to 93 nursing hours on average, there is little strain on the system. The system can definitely be kept stable and backlog decreases for the *overbooking* and *expected-value scheduling* strategy under all timing policies. For *conservative scheduling* of patient volume, backlog probably can be kept stable on average with some uncertainty if taking the worst case estimate of the CIs.

Regarding the width of the CIs, we notice that CIs become narrower as patients are scheduling horizons increase. The same can be observed under higher arrival rates below. The reason for this behavior is not immediately clear and requires further investigation.

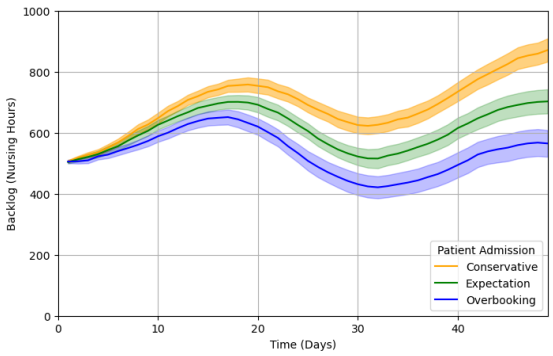
Arrival Rate 96



(A) 1-day-in-advance.



(B) 2-days-in-advance.

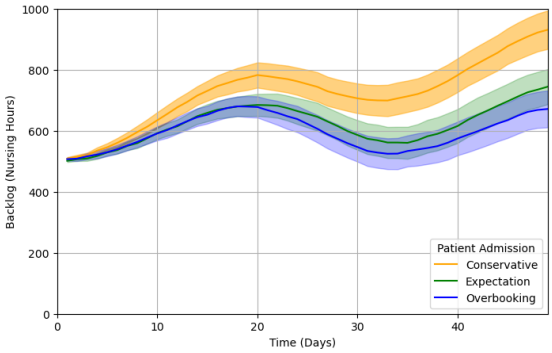


(C) 3-days-in-advance.

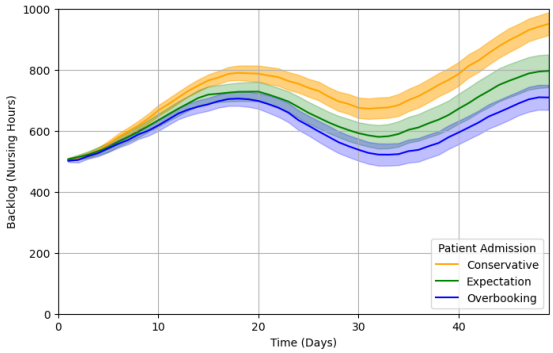
FIGURE 5.13: Backlog evolution under fluctuating infection risks with arrival rate 96. Mean and 95% CI for different timing and patient volume scheduling strategies.

Using a *conservative scheduling* strategy, backlogs do not return to their initial levels following the first wave of high infection probabilities, resulting in an upward trend of the length of waitlist. In contrast, *overbooking* produces a downward trend, as it successfully reduces the backlog after one phase. For *expected-value scheduling*, the stability of the backlog is highly likely for one-day-ahead scheduling. For the other two timing policies, there is some uncertainty about stability considering the worst case of the CIs.

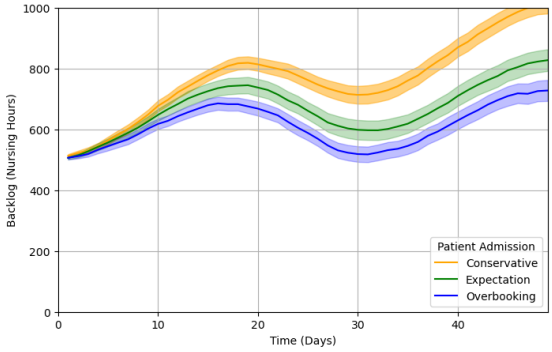
Arrival Rate 99



(A) 1-day-in-advance.



(B) 2-days-in-advance.



(C) 3-days-in-advance.

FIGURE 5.14: Backlog evolution under fluctuating infection risks with arrival rate 99. Mean and 95% CI for different timing and patient volume scheduling strategies.

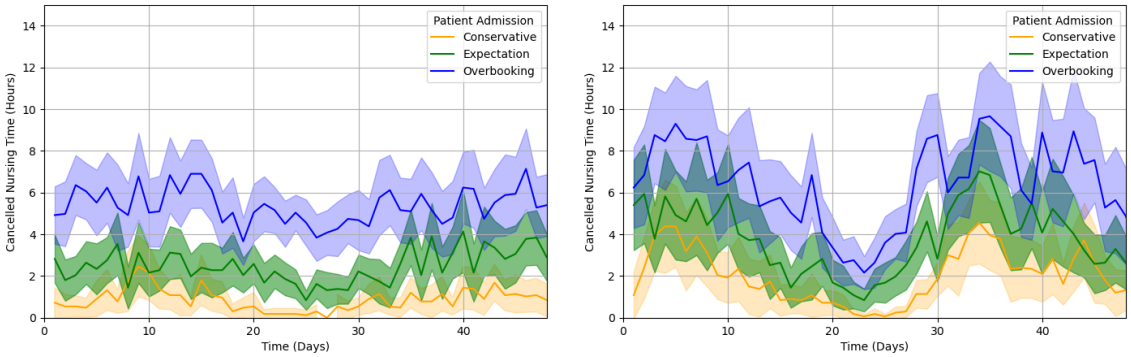
Under conditions of high system strain, at least the *overbooking* strategy is necessary to prevent rapid backlog growth. Even then, scheduling with planning horizons of only one or two days ahead appear to be the only approach likely to maintain system stability. All other policies result in rising backlogs, since buildup during times of little resources cannot be worked of during phases of low nurse infection rates. The lower the patient volume scheduled and the larger the scheduling time horizon, the faster the length of the waitlist increases.

Summarizing, under fluctuating staff infection probabilities, backlog accumulates during periods of high infection rates and (partially) recovers during times of high staff availability. Again, arrival intensity and patient volume strategy are the main drivers of backlog dynamics, while the timing of scheduling has a comparatively minor effect.

5.3.3 Cancellations

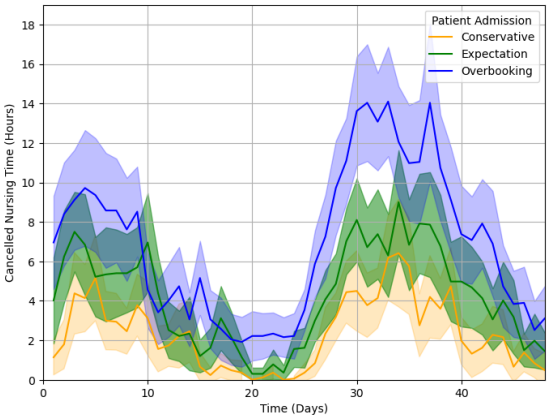
Under *conservative scheduling*, the average canceled workload is 0.7, 1.8, and 2.3 nursing hours for scheduling one, two, and three days ahead, respectively. When patient volume increases, cancellations rise to 2.5, 3.6, and 4.1 hours for *expected-value scheduling*, and further to 5.2, 6.1, and 7.2 hours for *overbooking*.

Figure 5.15 shows cancellations and corresponding 95%-CIs for *1-*, *2-* and *3-day-ahead* scheduling comparing the three considered patient volume strategies.



(A) 1-day-in-advance.

(B) 2-days-in-advance.



(C) 3-days-in-advance.

FIGURE 5.15: Cancellations under fluctuating infection risks. Mean and 95% CI for different timing and patient volume scheduling strategies.

Figure 5.15a shows that when scheduling only one day ahead, cancellations hardly reflect changes in infection probability. However, planning longer in advance reveals clear patterns of high and low cancellations over time. While timing has little effect when infection probabilities are constant, it creates large fluctuations in cancellations when infection probabilities vary, and these fluctuations grow with an increasing planning horizon. For example, cancellations for *3-days-ahead overbooking* fluctuate between 2 and 14 nursing hours over the simulation period. In contrast, the same strategy produces a steady amount of around 7 hours of cancellations under constant infection probabilities, see Figure 5.8.

The longer scheduled in advance, the higher the cancellation numbers during peak times and the lower during low phases. The maximum is reached shortly after day 30 and the minimum shortly after day 20. In the infection risk illustrated in Figure 5.1, these points in time correspond to the minimum infection risk and a rising infection risk, respectively. In other words, when infection probabilities are at their lowest and nurse availability continues to rise, the least cancellations happen. On the other hand, shortly after the point of the maximum nursing resources when availability starts to drop again, a lot of cancellations occur. Looking at the graphs of Section 5.3.2, this is the time where backlog is recovering and hits its minimum of a phase, before rising again.

5.4 Trade-off Between Backlog and Cancellations

Scheduling higher volumes of elective patients reportedly can improve utilization of available resources. However, it comes at a cost — the cost of increased short-notice cancellations of elective patients. To quantify this trade-off across different scheduling policies and under constant and varying staff infection risks, we computed descriptive statistics based on numerical simulations. Table 5.1 reports the average daily scheduled, realized and canceled elective workload in nursing hours, as well as the proportion of cancellations relative to the scheduled workload. Results for constant infection probabilities are aggregated over the full simulation period. For fluctuating probabilities, averages are computed over one infection phase length, corresponding to approximately 31 days. Arithmetic means of nursing hours are presented with corresponding 95%-CIs in square brackets.

Scheduling Policy	Scheduled	Realized	Canceled	CxRt
Constant Infection Probabilities				
1-day conservative	92.4 [90.2; 94.5]	91.6 [89.3; 93.8]	0.8 [0.2; 1.4]	0.009
1-day expected	99.0 [97.0; 101.1]	96.6 [94.3; 98.9]	2.4 [1.4; 3.5]	0.024
1-day overbooking	104.0 [101.9; 106.1]	99.0 [96.4; 101.6]	5.0 [3.6; 6.4]	0.048
2-day conservative	93.1 [90.9; 95.3]	91.7 [89.4; 94.0]	1.4 [0.4; 2.3]	0.015
2-day expected	98.7 [96.5; 100.8]	95.5 [93.1; 97.9]	3.2 [1.8; 4.6]	0.032
2-day overbooking	103.0 [100.9; 105.2]	97.0 [94.5; 99.5]	6.0 [4.2; 7.8]	0.058
3-day conservative	93.5 [91.5; 95.7]	91.8 [89.8; 93.9]	1.7 [0.6; 2.8]	0.018
3-day expected	98.8 [96.6; 101.0]	95.1 [92.7; 97.6]	3.7 [2.1; 5.3]	0.037
3-day overbooking	103.0 [100.9; 105.2]	96.6 [94.1; 99.2]	6.4 [4.3; 8.4]	0.062
Fluctuating Infection Probabilities				
1-day conservative	94.6 [92.6; 96.5]	93.8 [91.8; 95.9]	0.7 [0.1; 1.3]	0.007
1-day expected	98.8 [96.7; 100.9]	96.3 [93.9; 98.7]	2.5 [1.4; 3.6]	0.025
1-day overbooking	104.8 [102.8; 106.8]	99.7 [97.1; 102.2]	5.2 [3.7; 6.6]	0.050
2-day conservative	95.1 [93.8; 96.5]	93.3 [91.6; 95.0]	1.8 [0.8; 2.9]	0.019
2-day expected	100.0 [98.5; 101.5]	96.4 [94.4; 98.4]	3.6 [2.1; 5.1]	0.036
2-day overbooking	104.6 [103.2; 106.0]	98.5 [96.3; 100.8]	6.1 [4.1; 8.0]	0.058
3-day conservative	95.1 [94.0; 96.2]	92.7 [91.1; 94.4]	2.3 [1.2; 3.8]	0.024
3-day expected	100.2 [99.1; 101.3]	96.1 [94.3; 97.9]	4.1 [2.5; 5.7]	0.041
3-day overbooking	104.7 [103.6; 105.8]	97.5 [95.3; 99.6]	7.2 [5.1; 9.3]	0.069

TABLE 5.1: Daily scheduled, realized and canceled elective workload and corresponding cancellation rate. Comparison between different timing and patient volume scheduling policies for constant and fluctuating staff infection risks. CxRt - Cancellation Rate.

Under constant staff infection probabilities, *conservative scheduling* results in quite similar realized workloads across all timing policies. However, as the scheduling horizon increases, the scheduled workload, and thus cancellations, slightly rise to maintain equal realized amounts. In other words, even though the actual realized workload remains stable, scheduling further ahead leads to more expected resources being booked. The other strategies, which involve scheduling a higher patient volume, schedule roughly the same total workload regardless of the planning horizon. Yet, scheduling longer in advance leads to slightly more cancellations in nursing hours, reducing realized elective care. For example, in *overbooking*, an average of 5, 6, and 6.4 nursing hours are canceled daily when scheduling

one, two, or three days ahead, respectively. These cancellations correspond to 4.8%, 5.8%, and 6.2% of the scheduled workload.

Under fluctuating staff infection probabilities, we observe similar patterns, but the planning horizon has a stronger effect on realized elective care. Within the same patient volume strategy, the scheduled workload either remains stable or increases slightly with longer horizons. However, the realized workload tends to drop, leading to more cancellations. This gap between what is scheduled and what is delivered becomes larger the longer in advance scheduling is done.

Comparing constant and fluctuating infection risks, we find slightly higher scheduled and realized workloads under fluctuating conditions, but also observe more cancellations. This suggests that nursing availability is overestimated more frequently, leading to higher scheduled volumes and higher cancellation rates. However, the overestimation that happens has a larger magnitude than the increased cancellations, also leading to more realized workload under fluctuating than under constant infection probabilities. This may be caused by the overestimation of capacity that typically happens after periods of low infection rates just when the amount of resources starts to drop again. The individual trajectories in Section 5.3.1 highlighted this pattern.

5.5 Summary

This chapter examined how different scheduling strategies for elective patients affect backlog and cancellations. We analyzed various combinations of scheduling timing and patient volume policies under both constant and fluctuating staff infection probabilities.

We observed a time lag between expected and actual resource availability, tied to the scheduling horizon. Scheduling a higher patient volume improves resource utilization but also raises the risk of cancellations. Under fluctuating staff infection probabilities, resources are typically underestimated when availability starts rising, and overestimated when availability begins to drop after periods of low and high staff availability, respectively.

When studying backlog stability under varying demand levels, patient volume strategy proved more influential than timing. In high-demand scenarios, even slightly overbooking expected capacity, such as assuming one more nurse is available, can be the difference between keeping the backlog stable or not. In contrast, different timing policies have relatively minor impact on realized workload.

While the scheduling horizon has limited effect on realized workload, both the scheduling horizon and patient volume influence the cancellation patterns. Longer horizons lead to more cancellations and more volatility in cancellation rates. For constant infection probabilities, cancellations are stable over time. For fluctuating probabilities, the infection trend is reflected in the amount of canceled workload, especially when scheduling is high in volume and far in advance.

Longer planning horizons increase both scheduled and canceled workload but tend to reduce realized workload, especially under highly fluctuating staff availability, which is a key characteristic of health emergencies. From the perspective of backlog control and minimizing cancellations, shorter planning horizons are generally preferable. However, the choice of patient volume strategy should reflect the trade-off between maximizing realized care and limiting cancellations. System strain can guide this choice. For example, during critical periods when backlog control is essential, accepting more cancellations may be justified to push through as much elective care as possible. In more stable periods, it may be better to reduce scheduled patient volume to avoid excessive cancellations.

Chapter 6

Case Study: First COVID-19 Wave in ZGT Almelo

This chapter applies the developed framework to real infection data and pandemic bed census of a mid-sized teaching hospital in the Netherlands. We compare four different scheduling policies introduced and analyzed in the previous Chapter 5.

We explore the interplay between pandemic demand and nurse absence and how it affects the amount of elective workload that can be realized. In that regard, we investigate the daily difference of backlog, i.e., the addition or removal of workload to/from the waitlist. Additionally, we explore the amount of canceled workload over the course of the study period.

Section 6.1 gives medical background on the pandemic and hospital considered. In Section 6.2, we describe the data sources for predicted and actual pandemic demand as well as nurse absence and how we pre-processed these. Section 6.3 described parameter settings for simulations, scheduling policies considered and gives computational details. Section 6.4 presents results and Section 6.5 summarizes the most important findings.

6.1 Medical Background

We conduct a case study using data from ZGT (*Ziekenhuis Groep Twente*, Dutch for "Hospital Group Twente"), which is a *topklienisch ziekenhuis* (Dutch for "top-clinical hospital") located in the east of the Netherlands. ZGT provides care to around 390,000 residents. Top-clinical hospitals offer specialized care and medical training, but are not affiliated with a university. We focus on the pandemic ward of ZGT Almelo, which was active during all waves of the COVID-19 pandemic.

We examine the first COVID-19 wave in the time frame from 06-04-2020 to 14-06-2020, which covers 10 weeks. The progression of the pandemic during this study period, based on bed census data from the pandemic ward, can be seen in Figure 6.1.

Assumptions are as described in Section 3.3. We use the model to forecast nursing capacity for elective care, which was introduced in Section 3.5. Note that this model assumes that nurses scheduled in the pandemic ward cannot treat elective patients on the same day.

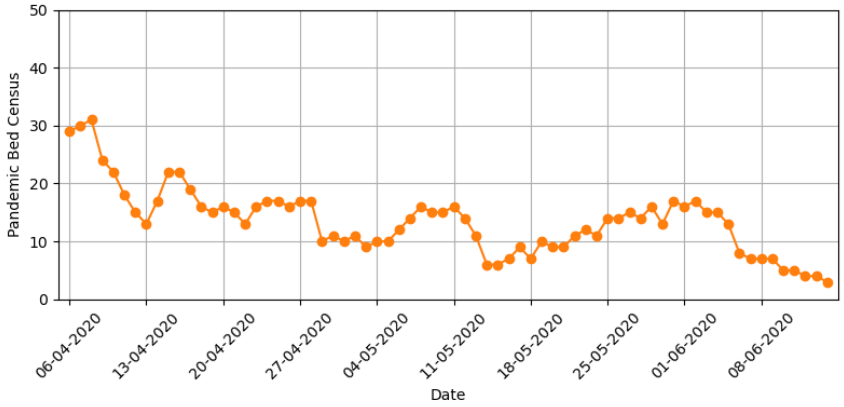


FIGURE 6.1: Pandemic bed census during the first COVID-19 wave in the ward of ZGT Almelo.

6.2 Data

The case study is based on a mix of real data, including pandemic forecasts and realizations, and informed approximations where necessary. Regional daily reported cases give an estimate of the infection risk for nurses and the chance they are needed for care responsibilities outside the hospital, for example within their families. The ward size in terms of the number of nurses and the corresponding NPR for pandemic care is an informed guess. For the remaining parameters, we made practical assumptions. We model one type of elective patient with a constant nursing time demand and Poisson arrival rate.

6.2.1 Estimating Nurse Absence Rates

We lack direct data on nurse absences or infection rates during the first wave of the COVID-19 pandemic, and we do not know whether or how hospitals forecast these in real time. Therefore, we approximate the infection risk for nurses using publicly available data on daily confirmed COVID-19 cases at the regional level.

The RIVM provides detailed datasets on the COVID-19 pandemic via its open data platform. We use the dataset *COVID-19 aantallen gemeente per dag*[1], which reports the number of newly confirmed positive cases per municipality per publication date. Recordings begin on 27-02-2020, corresponding to the first reported case in the Netherlands. We consider the reports of the municipalities supplied by ZGT Almelo (Almelo, Hengelo, Dinkelland, Rijssen-Holten, Twenterand, Tubbergen, Borne, Hof van Twente, Hellendoorn and Wierden [59]) to derive an infection indicator for the region. Figure 6.2 illustrates the number of cases by publication date.

[1] Available at https://data.rivm.nl/covid-19/COVID-19_aantallen_gemeente_per_dag.csv. Accessed 11-04-2025.

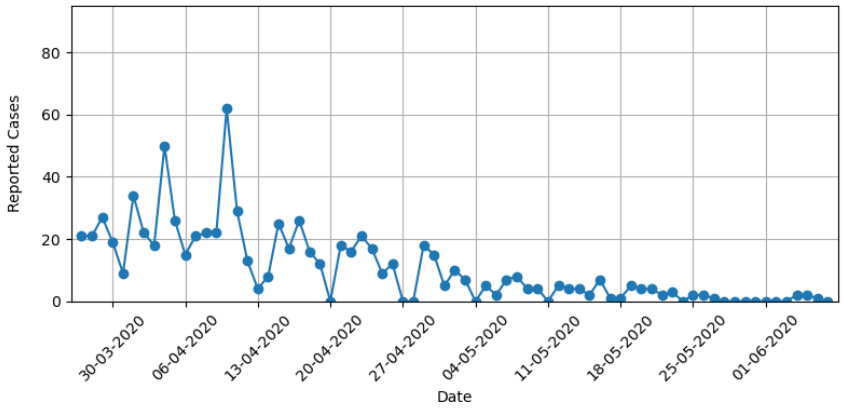


FIGURE 6.2: Positive COVID-19 cases by publication date in the region supplied by ZGT Almelo.

To derive an infection risk, we first divide the number of reported cases by the population of the region (390,000) and create a smoother trend by applying a specific form of a rolling average. Since the data is based on the publishing date, there are weekly patterns (e.g., hardly any published cases on Mondays due to very limited testing on weekends) and there is a time lag between infection, testing and publication. Thus, we take the average over the following week to estimate the current infection rate. That is, to determine the infection risk on day t, we average the population-based infection ratio over days $[t, t+6]$.

Due to limited testing capacities during the first wave, testing was mainly reserved for individuals in high-risk groups and symptomatic healthcare workers caring for vulnerable populations. As a result, the true number of infections likely substantially exceeded the reported values. So, while not giving an accurate estimate in terms of absolute numbers, we assume it reflects relative changes in infection dynamics over time, as no major changes to the national testing strategy were implemented during our study period [60].

Therefore, we scale this infection risk by a constant factor. This factor is chosen based on early published data on healthcare worker infection rates. A study conducted across nine hospitals in the south of the Netherlands found infection rates ranging from 0% and 9.5% per hospital between March 6 and March 8, 2020 [45]. Based on this, we assume a maximum absenteeism rate of 18%, compromising 9% due to infections and 9% due to private care duties or quarantine. We then scale the calculated infection ratios by a factor of 500 to match this value. Within the beginning of the recordings and our study period, the lowest expected nurse census occurs on the days from 04-04-2020 to 10-04-2020. Figure 6.3 shows the resulting evolution of the infection risk.

This scaled infection rate is used for both forecasting future nurse availability and drawing the realized number from the corresponding distribution, which serves as the ground truth in our simulation. In other words, a probable forecasting error in infection risk is not considered.

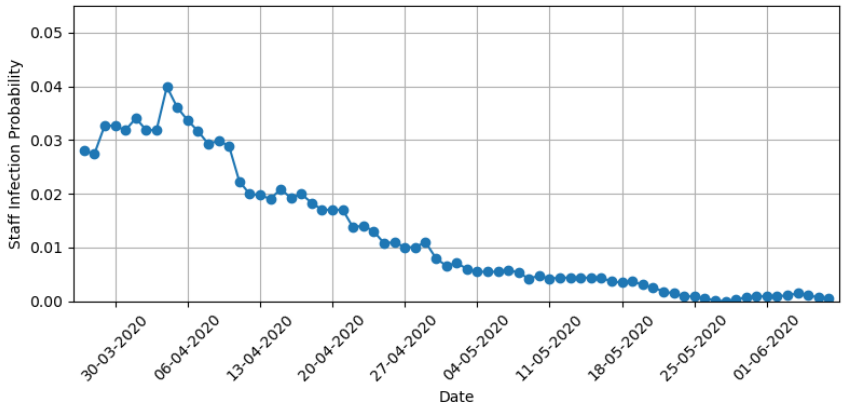


FIGURE 6.3: Staff infection rates derived from reported case data, after smoothing and scaling.

6.2.2 Pandemic Demand

We use pandemic ward bed census forecasts one and three days ahead, provided by the authors of [33], along with the corresponding realized bed census. ZGT Almelo used these numbers to prepare for expected pandemic demand, particularly in deciding when to open or expand COVID-19 wards. The forecasts represent the expected future bed occupancy, generated via simulation using a Poisson Arrival Location Model (PALM). The PALM extends the concept of a Poisson arrival process by introducing a location component. Patients arrive according to a Poisson process, after which they are assigned to a location (ICU or ward) according to a predefined probability distribution. Transfers between locations, as well as departures from the system (due to discharge or death), are also determined by stochastic processes.

Figure 6.4 presents the realized pandemic bed census alongside forecasts with a forecasting horizon of one and three days.

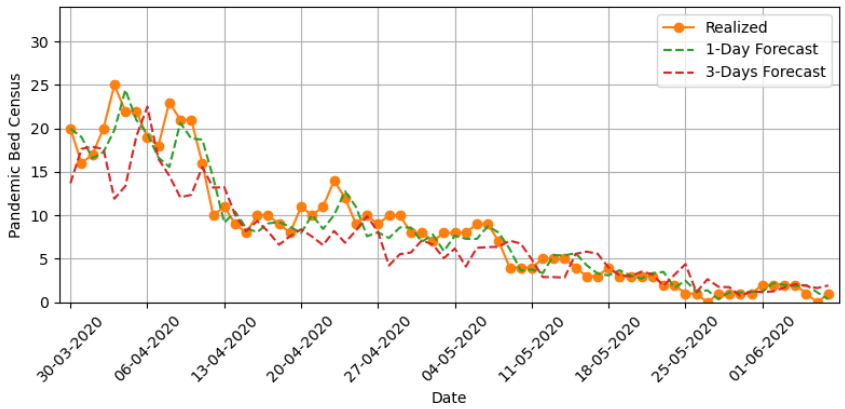


FIGURE 6.4: Forecast (1-day-ahead and 3-day-ahead) and realized pandemic bed census in the ward of ZGT Almelo.

Following the notation introduced in Section 3.5, the forecast data corresponds to $N_p^{(t,t+s)}$, where s denotes the forecast horizon in days. By multiplying this forecast patient count by the pandemic NPR and applying a ceiling division to account for full nursing

shifts, we obtain the expected pandemic workload $\hat{H}_p^{(t,t+s)}$. The actual realized workload is calculated analogously from the realized bed census and denoted $\hat{h}_p^{(t)}$.

Data pre-processing of pandemic demand was performed in R version 4.5.0 on a computer with an AMD Ryzen 5 5500U processor with 6 cores and 16 GB RAM, running Windows 11.

6.3 Experimental Setup

This section presents modeling choices for numerical simulations. We report choices of parameters in Section 6.3.1 and scheduling policies to be investigated in Section 6.3.2. Section 6.3.3 provides computational details.

6.3.1 Parameter Values

During the first COVID-19 wave in the Netherlands, hospital ICUs came under an immense pressure within a few days, demanding to increase capacities. While resources in the ward (non-critical care) were reported to be strained, pandemic demand did not exceed available capacities, mostly due to suspension of elective care [61]. Based on this information, we estimate the total number of nurses as 40 full-time equivalents, resulting in a maximum nursing capacity of 200 hours per day. The maximum pandemic occupation during the study period takes up around 75% of the nursing resources, which we consider a plausible number.

To initialize the dynamic nursing capacity from which forecasts are made s days before the simulation start date (30-03-2020), we first determine the initial available capacity. This is done by applying the approximated nurse infection rates from the beginning of the recordings up to the start of the study period. The ground truth capacity s days prior to 30-03-2020 is then set to this derived value, which corresponds to 35 of 40 nurses.

We set the pandemic nurse-to-patient ratio $\text{NPR}_p = \frac{1}{4}$, meaning that one nurse simultaneously cares for four pandemic patients. In terms of nursing effort, one pandemic patient thus requires 6 hours of nursing time for every day they occupy a ward bed. Elective patients arrive following a Poisson arrival process, with a fixed NTD of 6 hours per patient, which incorporates NPR and LoS. The arrival rate is set to 28 patients per day, which corresponds to 84% of the total available capacity assuming no nursing absences or pandemic demand. In other words, a mean of 168 nursing hours join the waitlist daily. The 84%-value reflects a realistic and operationally stable level under normal conditions, accounting for typical absenteeism, conservative planning buffers, and boundary effects. While close to the operational limit, this rate is considered manageable in practice.

6.3.2 Scheduling Policies

We compare four different selected scheduling policies in terms of timing and patient volume. The first two apply *3-day-ahead* forecasts using *expected-value scheduling* and *overbooking*, respectively. The others use a more reactive *1-day-ahead* scheduling policy, again with an *expected-value scheduling* and *overbooking* approach, where the letter aims to make maximal use of any remaining capacity.

We assume that throughout the simulation period, there is always enough backlog to schedule according to the expected available resources. Translated to the queueing framework, this means the queue never empties, resulting in no *lost resources* due to empty waitlists.

6.3.3 Computational Details

Simulations are coded in Python version 3.10.12 and performed on a server at the University of Twente provided via the JupyterLab environment. The CPU server has 64 cores, 128 threads and 1024 GB memory.

For each policy, we simulate 25 trajectories over the study period. We are limited to this number due to computational reasons. For each day, we record the forecast and realized nursing capacities, forecast and realized pandemic workload, scheduled elective workload, elective backlog additions, and canceled elective workload. To illustrate the relationship between forecast error and scheduling outcomes, we also show a single simulation trajectory displaying forecast and realized values.

All results are averaged across the simulations, and 95% CIs are computed. Note that the only randomness across simulations comes from the nurse capacity forecasts and realizations, since the pandemic data are provided by the hospital. Additionally, we compute a benchmark result, assuming no forecasting error and no boundary effects occur, which serves as a theoretical upper bound on achievable performance.

6.4 Results

In the following, we present and interpret results of the case study. Section 6.4.1 gives metrics on available resources for elective care and provides an upper bound on how much nursing workload can be realized. Section 6.4.2 illustrates a single trajectory for each policy, showing the dynamics between forecast and actual resources, realized workload, and canceled workload. Finally, simulation results of 25 trajectories for each policy are presented in Section 6.4.3. We report realized and canceled elective workload and its effect on the backlog evolution, showcasing the trade-off between backlog and cancellations.

6.4.1 Spare Resources

To evaluate realized workload under different scheduling policies, we first determine the mean spare resources across simulations that can be used for elective care. We also compute an upper bound on the realized elective workload, assuming perfect accuracy of pandemic and nurse absence forecasts.

Figure 6.5 shows pandemic demand and a sample simulation of nursing capacity measured in nursing hours. The maximum nursing capacity is indicated by a black dotted line at 200 nursing hours. Gray areas represent *lost* resources due to pandemic demand and nurse absenteeism. The white area between these two indicates capacity available for elective care. Note that even with perfect foresight of future spare resources, 100% utilization of these resources is not always possible. This is because workload can only be served in chunks of $NTD = 6$ nursing hours, which have to completely fit into the available resources on the respective day. We refer to this phenomenon as *boundary effects*.

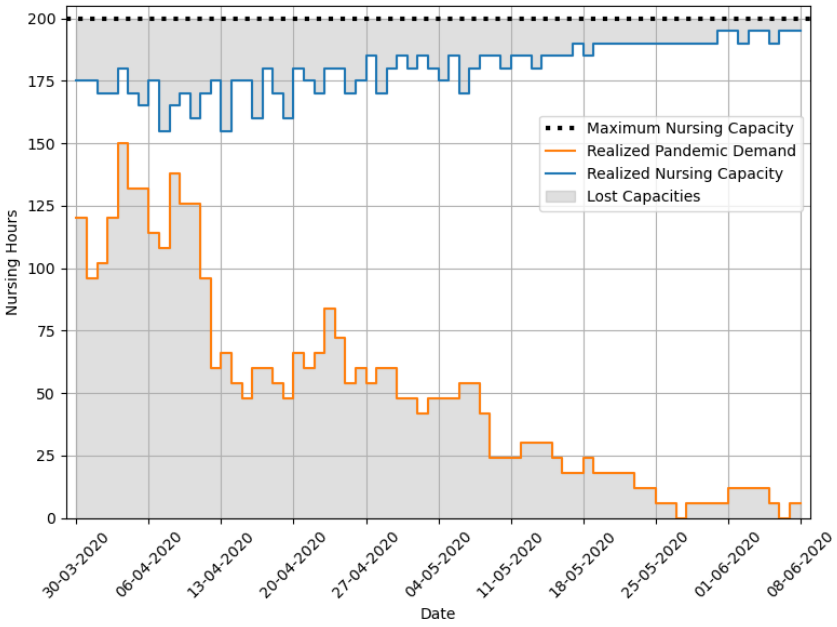


FIGURE 6.5: Spare resources for elective care between pandemic demand (orange) and example trajectory of nursing resources (blue) per day.

We calculated the following values based on all performed simulations, i.e., a total of 100 trajectories for four different scheduling policies. Note that we can just aggregate these results for this purpose, since these upper bounds only depend on the pandemic demand, which is deterministic, and the nursing availability, which is not influenced by scheduling.

Across the study period, there were 9,050 (95%-CI: [8,981; 9,118]) hours of nursing capacity available for elective care. Considering boundary effects, scheduling under perfect knowledge of future resources would result in realized elective workload of 8,861 (95%-CI: [8,793; 8,928]) nursing hours. This number serves as an upper bound of possible realized workload and helps us to assess how well a certain scheduling policy performs.

6.4.2 Example Trajectories

The following Figures 6.6, 6.7, 6.8 and 6.9 show forecast and realized resources, and corresponding realized and canceled elective patient workload for *3-day-ahead expected-value scheduling*, *3-day-ahead overbooking*, *1-day-ahead expected-value scheduling* and *1-day-ahead overbooking*, respectively. To highlight the timing and structure of key dynamics, we present results from individual simulation runs instead of aggregated summaries, which can obscure patterns. We examined multiple trajectories and observed similar behavior across all of them. Given this consistency, we concluded that one representative run adequately illustrates the behavior of interest, and CIs were not necessary for this purpose.

The plots should be interpreted as follows. Maximum possible nursing resources per day amount to 200 hours, indicated by the black dotted line. Some of these resources are *lost* due to nurse absenteeism, which is represented by the gray area defined by the realized nursing capacity (blue line). An additional amount of resources is needed for pandemic care, illustrated by the gray area in the bottom defined by the realized pandemic demand (orange line). When scheduling patients in advance, though, these values are not known yet, and scheduling decision must be made based on forecasts. These are represented by the dashed lines in the respective color of the resource. The combination of scheduled workload

and actual available capacities determine how much workload is realized on a given day, shown as the green area. If scheduled workload exceeded the resources available, e.g., because pandemic demand got underestimated, patients are canceled, represented by the red area.

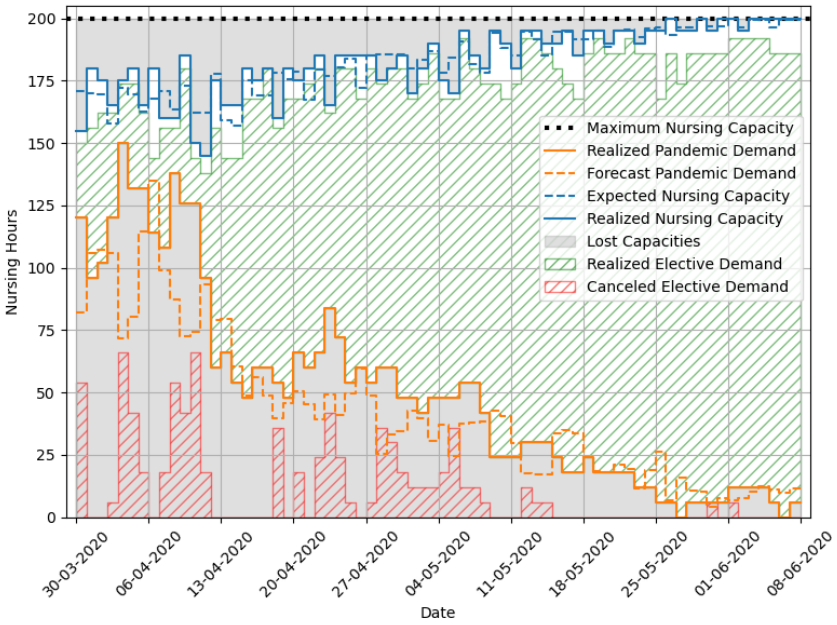


FIGURE 6.6: Forecast and realized resources, and corresponding realized and canceled elective patient workload per day. Example trajectory under *3-day-ahead expected-value scheduling*.

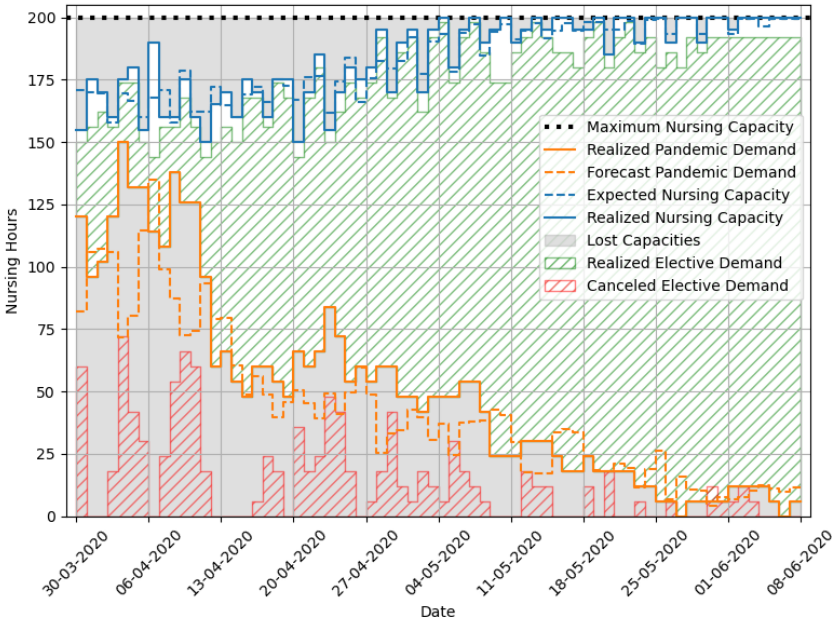


FIGURE 6.7: Forecast and realized resources, and corresponding realized and canceled elective patient workload per day. Example trajectory under *3-day-ahead overbooking*.

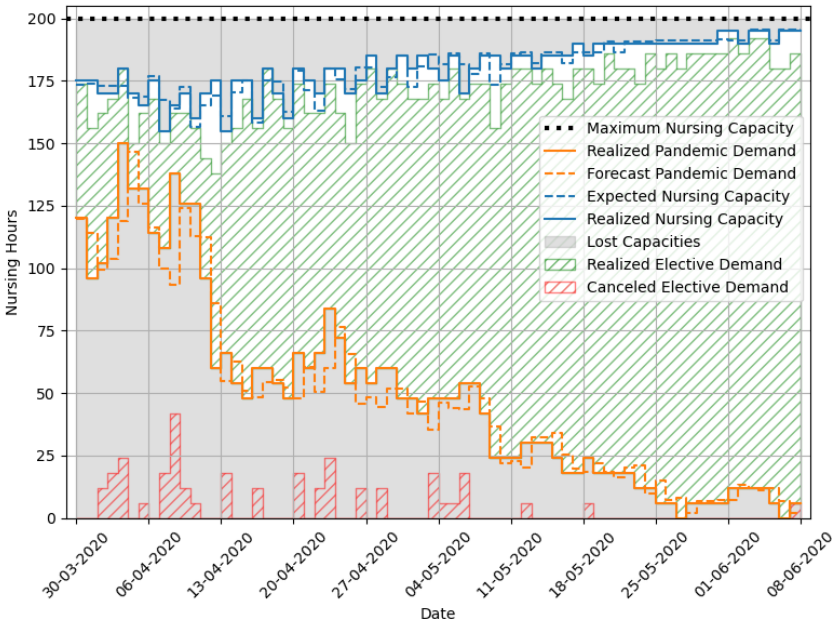


FIGURE 6.8: Forecast and realized resources, and corresponding realized and canceled elective patient workload per day. Example trajectory under *1-day-ahead expected-value scheduling*.

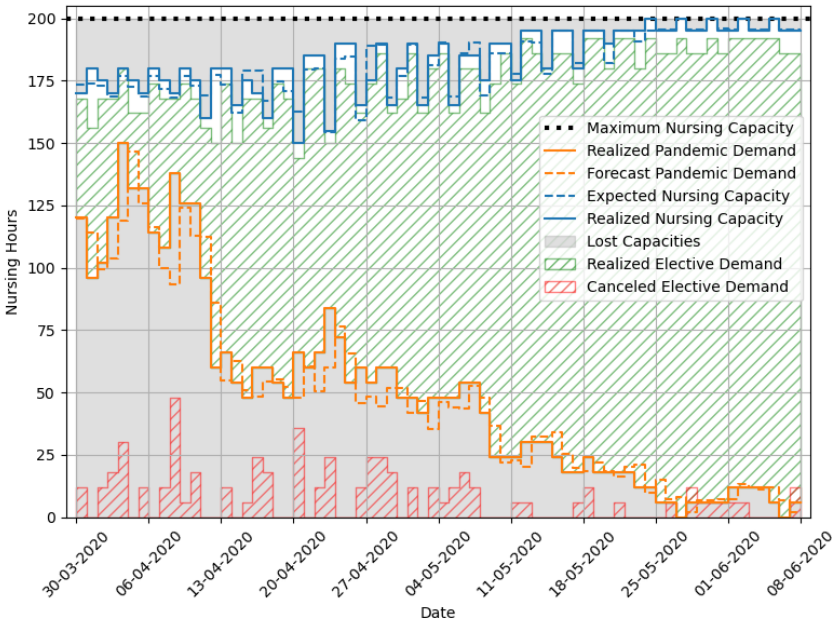


FIGURE 6.9: Forecast and realized resources, and corresponding realized and canceled elective patient workload per day. Example trajectory under *1-day-ahead overbooking*.

We observe an interaction between pandemic demand and nursing capacity. When pandemic demand was high, nursing capacity tended to be low. This intensified pressure on resources during infection surges of the pandemic.

Comparing *3-day-ahead* and *1-day-ahead* scheduling, the longer horizon systematically

underestimated pandemic demand. This led to a an overestimation of available resources. In contrast, the 1-day forecast predicted pandemic demand more accurately with a slight one-day time lag in reacting to changes of actual pandemic demand. Forecasting errors in nursing capacity occurred in both cases but are minor, both because we assumed no prediction error in infection rates and because fluctuations in nursing availability were smaller compared to pandemic demand.

The persistent overestimation of resources in the *3-day-ahead* scheduling led to a high number of cancellations. These occurred exactly on days when forecast pandemic demand (dashed orange lines) fell far below realized demand (solid orange lines), and they align with visible spikes in cancellations (red bars). This pattern is dominant for both scheduling practice in *3-day-ahead* scheduling, but also appears, to a lesser extent, in *1-day-ahead* scheduling. The same relationship exists between overestimating nursing capacity and cancellations, though the effect is minor in this case study due to only slight prediction errors in nursing capacity.

Regarding realized workload, cancellations led to full utilization of the remaining available capacity. Conversely, underestimating available resources resulted in wasted capacities, which are visible as the white parts above the green area of realized workload. For instance, this occurred in the days leading up to 18-05 in *3-day-ahead* scheduling, see Figure 6.6 and 6.7, where less pandemic demand arises than forecast.

Finally, we compare *expected-value scheduling* and *overbooking* within the respective timing policies. Across both scheduling horizons, *overbooking* caused more cancellations. On days, where even *expected-value scheduling* resulted in canceled workload, *overbooking* led to more cancellations, which did not improve utilization. On the other hand, *overbooking* seems to have utilized spare resources a bit more effectively, as we notice less lost capacities (white areas) in Figure 6.7 and 6.9 compared to their *expected-value scheduling* counterparts in Figure 6.6 and 6.8, respectively.

6.4.3 Trade-off Between Backlog and Cancellations

The following Figures 6.10, 6.11, 6.12 and 6.13 show the evolution of backlog changes and cancellations over time for the four investigated scheduling policies. In each figure, the first plot shows the daily change of backlog, i.e., the net addition or removal of workload to/from the waitlist, in nursing hours. The dotted line at 0 therefore indicates no change, meaning that fluctuations around this line imply a stable backlog. Cancellations in the lower graph are also expressed in nursing hours. Lines represent means across simulation runs, with shaded areas indicating 95%-CIs.

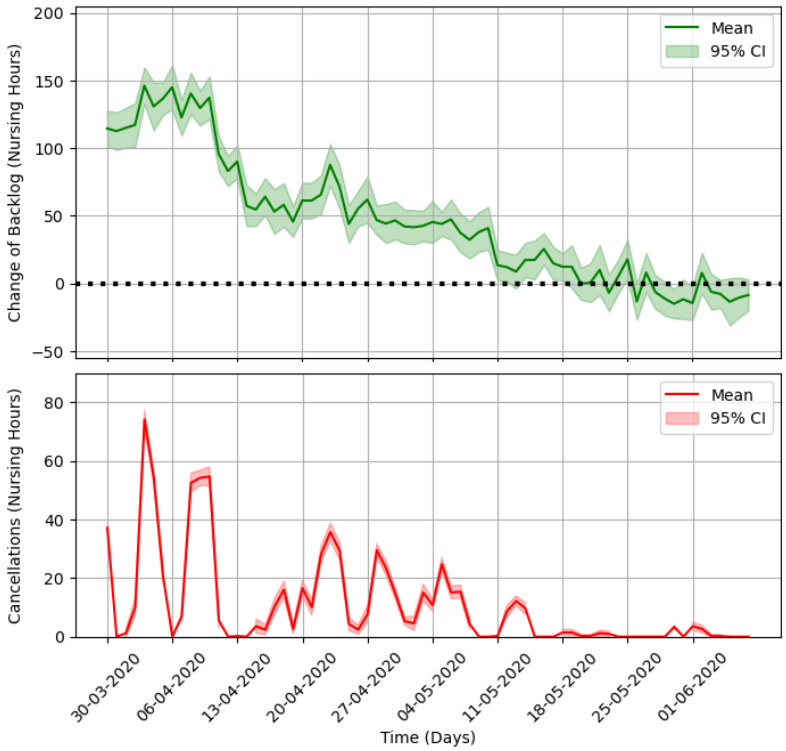


FIGURE 6.10: Change of backlog and canceled patient workload under *3-day-ahead expected-value scheduling*.

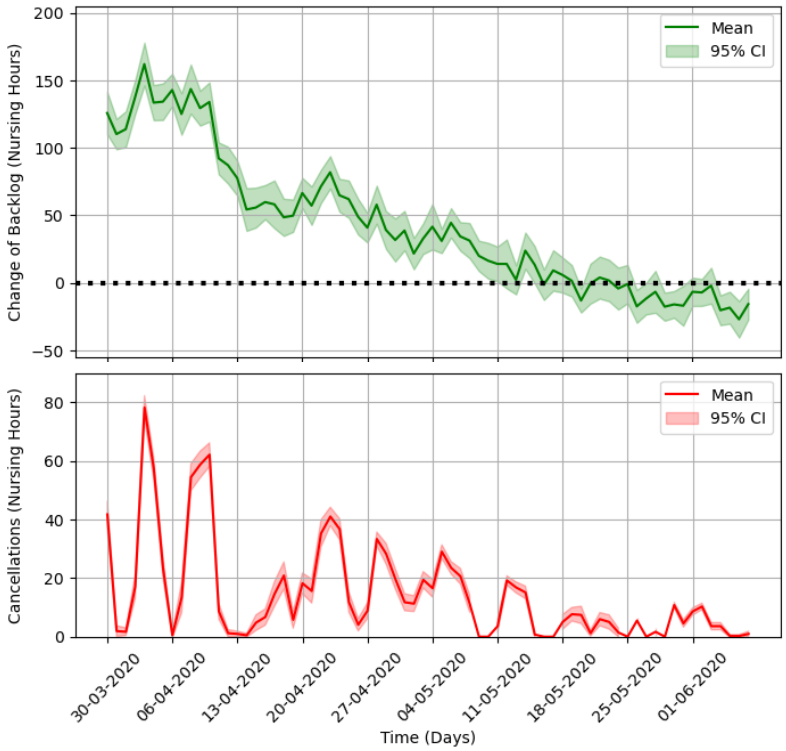


FIGURE 6.11: Change of backlog and canceled patient workload under *3-day-ahead overbooking*.

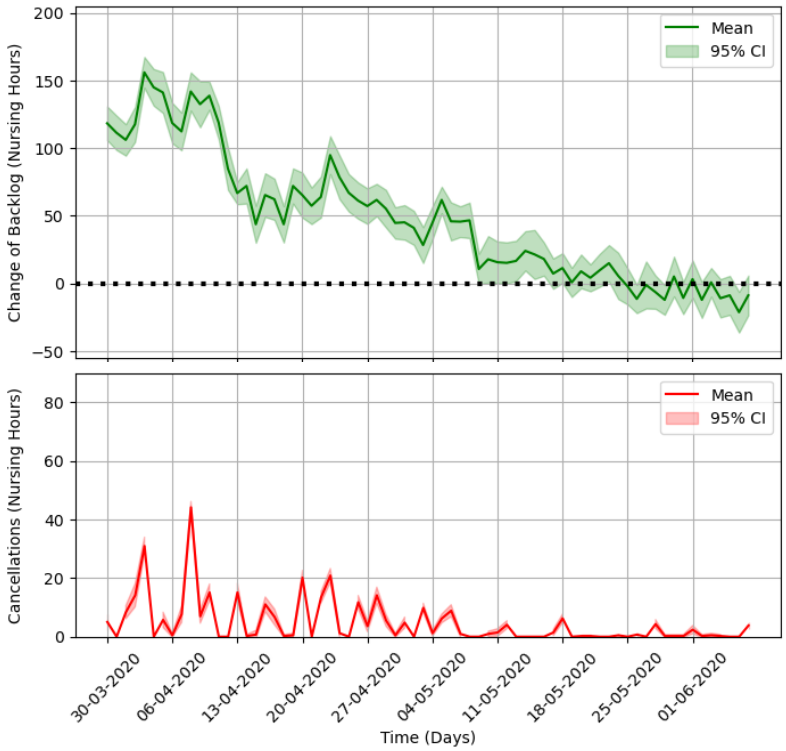


FIGURE 6.12: Change of backlog and canceled patient workload under *1-Day Ahead expected-value scheduling*.

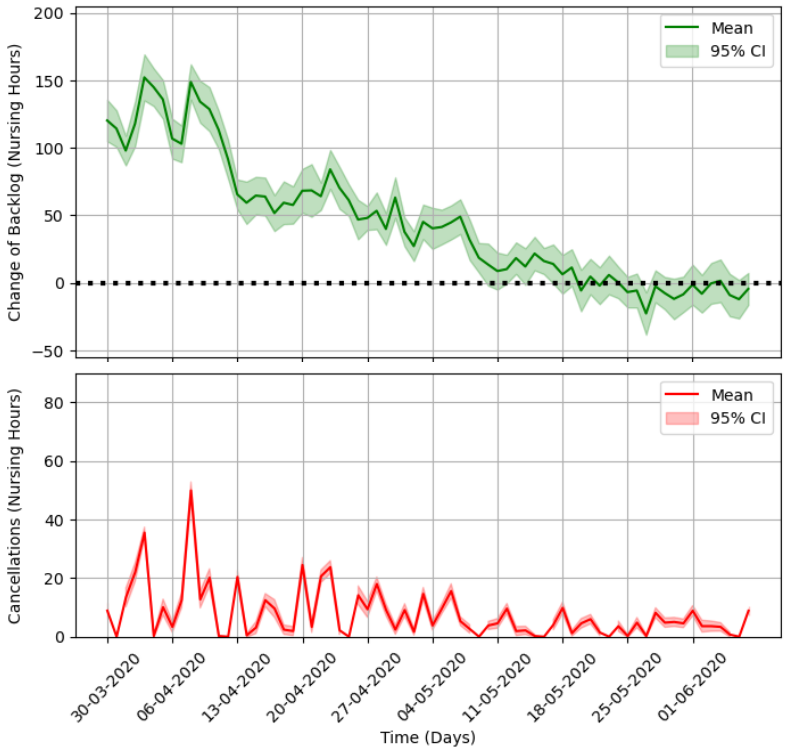


FIGURE 6.13: Change of backlog and canceled patient workload under *1-Day Ahead overbooking*.

We observe a clear relationship between pandemic-related demand and backlog addition. During the early stages of the first COVID-19 wave, when many pandemic patients required care and infection rates among nurses were high, between 100 and 150 nursing hours were added to the backlog per day. As both the pandemic bed occupancy and the number of infected nurses decreased, the backlog additions also declined. Only around 18-05, when the infection risk among nurses dropped below 0.5% (see Figure 6.3) and pandemic care demand fell below 25 nursing hours (see Figure 6.5), a stable backlog could be maintained. Toward the end of the study period, when the pandemic had little remaining impact on available resources, the backlog was even partially reduced. This general pattern holds across all scheduling policies.

Differences in backlog development between the scheduling policies are minor and do not allow for conclusions regarding superior resource utilization based on merely the graphs. However, as expected from previous observations in Section 6.4.2, *overbooking* tended to result in slightly lower backlog additions than *expected-value scheduling* within the same forecasting horizon.

Cancellation numbers show very narrow CIs across simulations, indicating that the primary drivers of cancellations were pandemic-related demand and its predictions, both of which were modeled deterministically. As already stated before, this was likely due to the higher amount of lost resources due to pandemic demand than nursing absence, and our modeling choice to not introduce a forecasting error to infection probabilities of nursing staff.

The days with the highest cancellation rates coincide with underestimations of pandemic demand, which caused an overestimation of spare resources. These instances are identifiable in Figures 6.6 and 6.8 by orange dashed lines (forecast pandemic demand) lying below the solid lines (realized demand) for *3-day-ahead* and *1-day-ahead* scheduling, respectively. Due to larger prediction errors, cancellations were more frequent and severe under *3-day-ahead* scheduling. Within each scheduling horizon, *overbooking* occasionally led to higher spikes in cancellations than *expected-value scheduling*, and introduced additional cancellations on days when the more conservative policy would not have scheduled excessive workload.

In Figure 6.13, despite *overbooking*, it seems like less cancellations occurred than in Figure 6.10, where *expected-value scheduling* was used. This highlights the impact of underestimation of pandemic demand when scheduling three days in advance. Even with more conservative patient volume strategies, *3-day-ahead* scheduling resulted in implicit overbooking. Quantitatively, over the full study period, the *3-day-ahead* policy predicted 3,402 nursing hours of pandemic-related workload, compared to 3,000 hours under the *1-day-ahead* policy. This difference directly translated to higher booked elective workload under the longer forecast horizon.

To evaluate overall performance of policies, Table 6.1 reports the total realized workload, backlog addition and canceled workload of the policies across the study period. Results are shown as average nursing hours over all simulation runs, with 95%-CIs given in square brackets.

Scheduling Policy	Realized Workload	Backlog Addition	Canceled Workload
3-day expected-value	8,431 [8,332; 8,529]	3,261 [3,145; 3,377]	750 [733; 767]
3-day overbooking	8,765 [8,696; 8,834]	2,991 [2,864; 3,118]	1,011 [979; 1,042]
1-Day expected-value	8,483 [8,396; 8,570]	3,337 [3,207; 3,467]	324 [312; 337]
1-Day overbooking	8,592 [8,504; 8,681]	3,167 [3,041; 3,293]	539 [523; 555]

TABLE 6.1: Total backlog addition and canceled workload in nursing hours under different scheduling policies.

Comparing actual workload to the benchmark result of possibly realizing 8,861 nursing hours (see Section 6.4.1), all policies realized over 95% of the maximum possible elective care workload. *3-day-ahead overbooking* significantly performed best, reaching a total of 8,765 nursing hours, corresponding to almost 99% of the upper bound. It also resulted in the lowest addition to the backlog. *1-day-ahead overbooking* achieved the second-highest realized workload, followed by *1-day-ahead expected-value scheduling* and *3-day-ahead expected-value scheduling*. We expected the same ranking (reversed order) for backlog addition, however, notice an inconsistency between the two *expected-value* strategies. Despite realizing more nursing hours of elective care, *1-day-ahead* scheduling resulted in a higher backlog addition. Since simulations were not based on common random numbers, we attribute this to variability in the Poisson-distributed arrival of elective workload and the limited number of simulation runs. Hence, we consider realized workload a more reliable and meaningful performance measure in the remainder of this chapter.

Regarding cancellations, *overbooking* led to significantly more cancellations than *expected-value scheduling* within the same timing policy. Scheduling patients three days in advance resulted in more cancellations than scheduling one day ahead. As already suggested in Section 6.4.2, *1-day-ahead overbooking*, actually led to less canceled workload than *3-day-ahead expected-value scheduling*. While previous results in Chapter 5 also suggested higher cancellations when scheduling further in advance, we believe that this particularly large difference observed here is also a result of systematic underestimation in the 3-day pandemic demand forecast.

Similarly, we would usually expect better performance in terms of realized workload with short-term scheduling. The case study, however, showed the opposite. Once again, we attribute this to the overestimation of resources in *3-day-ahead scheduling*, which mimicked the effect of overbooking.

Assessing the trade-off between realized workload and cancellations, we observe a tendency of an inverse relationship. The lowest realized workload corresponds to the highest canceled workload (*3-day-ahead overbooking*). For stakeholders aiming to minimize cancellations, *1-day-ahead* scheduling is preferable. Overall *1-day-ahead overbooking* appears to achieve the best balance between the two objectives of minimizing backlog growth and cancellations. This policy results in the second best values in both categories.

6.5 Summary

This chapter analyzed how pandemic demand and nurse absenteeism affected nursing resources for elective care during the first wave of COVID-19. It modeled a hospital based on pandemic data from ZGT Almelo and infection numbers in the region. We performed numerical experiments to evaluate the performance of different scheduling policies in terms of realized workload, which determine backlog evolution, and short-notice cancellations.

The case study reveals a relationship between prediction and scheduling accuracy, workload realization, and cancellations. Longer-term (*3-day-ahead*) scheduling consistently overestimated available resources due to underestimating pandemic demand, causing higher cancellation rates. *Overbooking* strategies increased realized workload but also led to more cancellations compared to *expected-value scheduling*.

Surprisingly, the *3-day-ahead overbooking policy* achieved the highest realized workload and lowest backlog growth, essentially benefiting from overestimating capacity, which acted like implicit overbooking. However, this came at the cost of more cancellations. In contrast, *1-day-ahead* scheduling is more accurate, resulting in significantly fewer cancellations but somewhat lower workload realization. The best trade-off was achieved under *1-day-ahead overbooking*, balancing backlog growth and cancellations and performing well in both categories.

Chapter 7

Discussion

This chapter discusses the study. Section 7.1 highlights key findings and interprets them in the given context. Section 7.2 outlines limitations related to model assumptions and the scope of the results. Section 7.3 explains how our findings can support both pandemic preparedness and decision-making during ongoing crises. Finally, Section 7.4 provides recommendations for future research.

7.1 Discussion of Results

Under the model developed in Chapter 3, which assumes independent infection behavior, staff availability follows a Binomial distribution across both constant and fluctuating infection rates. When predicting future nursing capacities, realized numbers of available nurses are considered, resulting in a distribution of future staff availability that is no longer Binomial distributed. Instead, it is characterized by less variance, meaning that the current capacity adds valuable information when predicting near-future capacity.

Lower productivity during pandemics is not solely due to reduced capacity from staff absence or increased pandemic demand. As shown in Chapter 4, a queueing system demonstrated that even if total nursing capacity over some period is equal, the temporal distribution of these resources significantly affects how much workload can be realized. The more variance and the higher the temporal autocorrelation of capacities, the less efficient those resources can be used for utilization below 1. Since wasted resources in these results are attributed to idle queue periods, one could argue that elective care waitlists are never empty, and thus such idle time is unrealistic. However, in practice, hospitals allocate resources to specific specialties and subspecialties. For instance, in the Netherlands, such tactical planning is typically done on a quarterly basis. When looking at the waitlists for these more narrowly defined patient groups, they are far more likely to run empty.

Theoretically, we expected that a discrete-time queueing system with a varying number of servers could be approximated by a correspondingly scaled single-server system. This assumption did not hold. While we successfully derived the heavy-traffic limit of a multi-server system using a translation to single-server queues, the same approach failed under variable server capacity drawn i.i.d. from a Binomial distribution. Empirical results showed that this translated system underestimates queue lengths. In fact, the translation results in a continuous queue with temporally correlated service times, for which standard queueing theory results, which assume i.i.d. service times, do not hold.

Chapter 5 characterized the trade-off between minimizing backlog and minimizing cancellations. To maximize resource utilization, the system needs to be overbooked even if short scheduling horizons are used. This is the best way to prevent backlog growth, but it

also leads to higher cancellation rates. While short-notice cancellations likely only cause negligible costs in hospitals (e.g., due to already performed preoperative examination), they are problematic for patients. Elective surgeries are often long-awaited and require logistic preparation and coordination (e.g., with caregivers). Canceling shortly before the planned procedure signals unreliability, and can strain confidence in the healthcare system. Patients who experience repeated disruptions may become less likely to seek care or follow through with treatment.

To mitigate overly high cancellation rates, shorter scheduling horizons proved themselves as an effective strategy, and also slightly further improve utilization. They allow hospitals to adapt better to sudden changes in the pandemic through more accurate predictions of resources and infection dynamics. However, too spontaneous planning is unfavorable for both patients and staff. Patients need time to prepare (e.g., arrange transport and care, take leave) and staff generally prefers stable and predictable work schedules.

The case study in Chapter 6 confirms the hypothesis that staff absence intensifies pressure on resources during infection surges of the pandemic. Most of the earlier findings were reflected in the case study, with one notable exception: scheduling three days in advance led to more realized workload than scheduling one day before. This unexpected result was due to consistent underestimation of pandemic resources under the longer forecasting horizon. As a result, the system implicitly overbooked capacity, increasing throughput at the cost of exceptionally high cancellation rates. This highlights how forecast errors of capacities heavily impact scheduling effectiveness, sometimes more than the chosen scheduling method itself. Thus, accurate pandemic demand predictions and a solid understanding of nurse infection patterns are the foundation for effective scheduling using these policies.

7.2 Limitations of the Study

One limitation of this study is the lack of validation for the nurse absenteeism model. Due to missing data on staff presence, we could not verify whether our assumptions correctly reflect the link between infection rates in the general population and nursing staff absences. The literature offers no consensus, and competing theories exist about when healthcare workers are most susceptible to infection. We assumed independent infection of nurses according to a Bernoulli process. However, reports during the COVID-19 pandemic indicated clusters of infected staff [62], implying dependent infection behavior of nurses. The extent to which such clusters impacted capacity and contributed to overall absence remains unclear.

Another limitation is how we handled forecasting of nursing resources. We used the infection probability input to determine both forecast and actual nursing capacity, introducing no forecasting error when calculating the probability distribution of available nursing staff. In reality, future infection probabilities must be predicted. As it is typical to make larger errors for longer forecasting horizons, we may assume less accurate prediction of nursing capacity in reality than observed in this study. However, as seen in Chapter 6, prediction errors do not automatically result in worse resource utilization. Assuming that prediction errors result in both over- and underestimation of nursing capacity in the long-term, the implicit overbooking cannot outweigh underestimation of available resources, which is connected to wasted capacities. This is due to the fact that the effect of overbooking is capped by actual available resources, while the range of possible resource waste due to underestimation is a lot larger. Therefore, we may expect more backlog growth in practice than we saw in our model.

For analytical tractability, we made several simplifications. The most significant is

likely the omission of elective patients' LoSs. We compressed their total nursing demand into a single scalar value. As already commented on in Section 5.1.2, this made our model prone to boundary effects. Specifically, requiring the full NTD to fit into the resources of a single day makes it more likely that a patient cannot be scheduled, even if their daily nursing demand would have fit when spread over multiple days. We mitigated this by choosing a low total nursing workload in our experiments, which led to boundary effects we considered plausible in real-world scenarios.

This same simplification implies that scheduling or canceling patients only affects capacity on a single day. In reality, elective patients often stay multiple days. Scheduling a patient on day t also consumes capacity on days $t+1$, $t+2$, etc., depending on the (stochastic) LoS. Similarly, canceling a patient frees up capacity on multiple days. In our model, the only consequence of a cancellation was the short-notice disruption. In practice, it can also lead to lost future capacity. For example, if patients are scheduled three days in advance and a cancellation occurs on day t, a patient with a 3-day LoS might leave some available capacities on days $t+1$ and $t+2$ unused, because no replacement can be scheduled on such short notice. This waste might be reduced by prioritizing cancellations of short-LoS patients.

We assumed a single patient type with a deterministic LoS and modeled one unified waitlist. In reality, hospitals divide capacity across multiple specialties, each with its own waitlist and patient characteristics. Even within a specialty, LoS varies and is influenced by medical complexity, recovery speed, and patient-specific factors. While this simplification likely doesn't affect our main findings on scheduling policy performance, it does limit the model's applicability. The capacity forecasting component can estimate spare capacity under pandemic pressure and staff absence, but does not account for downstream capacity usage from elective patients already admitted and recovering after surgery.

Finally, nursing capacity in the ward alone does not justify admitting a patient. Most elective cases require surgery, which also demands OR availability and surgical staff (e.g., anesthetist, surgeon, assistants). While it may be rare, there is still a chance that a patient requires specialized treatment in the ICU after their procedure. Therefore, capacity in all these units (ward, OR, and ICU) must be considered when scheduling elective care.

7.3 Practical Relevance and Implications

We present a flexible and broadly applicable capacity forecasting model that can be deployed with minimal data requirements. To apply the forecasting model in any hospital, the following data inputs are sufficient:

- Daily forecasts of pandemic patient load
- Daily forecasts of staff infection rates
- Realized staff availability (current or past value)

The scheduling policy insights derived from this study are intended to support pandemic preparedness. One major shortcoming during the COVID-19 crisis was the lack of clear guidelines for elective care admissions. To avoid repeating this, policymakers should proactively define scheduling protocols for future pandemics. Based on our findings on planning horizon and patient volume, different scheduling rules can be tailored to different pandemic phases or severity levels. If hospitals want to test specific policies using their own situation, the model only needs the data mentioned above and observed values for pandemic demand and staff absence.

Choosing between policies not only involves the presented trade-off between backlog and cancellation minimization, but also requires consideration of what is practically possible in a given hospital. The nature of the pandemic itself must also factor into these decisions. For instance, if the evolution of a pandemic is relatively accurately predictable for longer horizons, no spontaneous scheduling policy may be needed. Conversely, in a more volatile situation with rapidly changing trends, greater flexibility in scheduling might be essential to keep backlogs under control.

Although our study centered on nursing capacity in general wards, the model structure extends to other hospital areas. Thinking of the capacity requirements in different hospital areas based on which elective care can be admitted, as stated in Section 7.2, our model can also be used. ICU spare capacity can be forecast similarly, considering pandemic demand of resources (e.g., ventilators, nurses) and similar infection behavior of specialized ICU nurses. A comparable approach applies to surgical departments, where staff availability likely follows similar infection dynamics. If there is no pandemic-related demand, this component can simply be omitted. However, as seen during the COVID-19 pandemic, some ORs were repurposed as ICUs, which would then represent the pandemic demand of OR capacity.

7.4 Recommendations for Future Work

A follow-up study should aim for a better understanding staff absence dynamics during a pandemic. Due to the lack of public datasets, we couldn't validate our model. However, accurate staff availability forecasts are essential for predicting spare capacity, which is a requirement for reliable scheduling of elective care. With access to real data, the model could be extended to include more realistic infection patterns among nurses, such as incorporating time lags between general population infection rates and hospital staff absences, or correlations with pandemic patient load.

While we analyzed policies that schedule or cancel patients at a fixed point in time, more dynamic approaches should be explored. For example, hospitals could schedule patients in advance but leave room to add additional patients later if capacity turns out higher than expected. This would allow longer planning horizons for most patients while still reacting to short-term changes. Some elective procedures likely require less preparation time and could be added at short notice if extra capacity becomes available. Similarly, if forecasts strongly suggest a drop in available capacity, cancellations could be made earlier than on the day of surgery, giving patients and staff more time to adjust. This kind of dynamic decision-making could improve both resource use and cancellations.

To turn this model into a proper decision-making tool for elective care admissions, it should include OR and ICU capacity, as mentioned in Section 7.3. This could be done by creating separate models for each area and using the most limiting resource to make admission decisions. Before implementing such an approach, it should first be developed and tested using historical data, validated against real outcomes, and then simulated under various pandemic scenarios.

Another important area that was beyond the scope of this study is prioritization within elective care. Instead of relying on a basic FIFO approach, admissions should factor in both the expected LoS and the impact of delaying treatment for each patient. Some patients are more affected by postponement than others. QALY-based models, which were used during the COVID-19 pandemic to estimate population-level health loss (see Section 2.1.3), provide a solid starting point. These models can be extended to reflect how delay length translates into individual QALY loss. This would allow hospitals to prioritize patients

based not only on how much capacity they're expected to require, but also on how badly a delay would harm their health. In doing so, resources could be used more effectively while minimizing overall health loss during a crisis.

Chapter 8

Conclusion

This thesis explored how pandemic-related capacity constraints affect the continuation of elective care. We were, to the best of our knowledge, the first to model staff absenteeism linked to the course of a pandemic and combine it with pandemic patient demand.

On the theoretical side, we successfully related a multi-server discrete-time queueing network to a single-server system in heavy traffic and derived the limiting distribution. Results correspond to the distribution of continuous-time queues. We identified challenges when applying the same approach to systems with time-varying server availability, which would better reflect real-world resource fluctuations during a pandemic. Numerical results, however, showed how varying staff availability significantly reduces the effective capacity for elective care.

We tested various elective care scheduling policies in simulated pandemic scenarios, and characterized a trade-off between minimizing the backlog and minimizing the number of canceled patients. There is no single optimal admission strategy, but decisions must be adapted to the specific situation at hand. This includes target throughput, acceptable cancellation levels, pandemic volatility, and operational planning feasibility. We identified the need for accurate forecasts for pandemic demand to allow for scheduling effectiveness. This highlights the importance to further improve existing pandemic demand prediction approaches and extend and validate our proposed nurse absenteeism model. While we highlighted key dynamics of patient volume and scheduling horizon of a simplified policy regime, future research should look into more dynamic planning approaches, which can potentially improve the discussed trade-off.

Our nurse absenteeism model and capacity forecasting approach are data-efficient and easily adaptable for hospital use. They allow planners to estimate spare capacity based on two key inputs: infection-driven staff absence and pandemic demand. To provide a complete decision framework for admitting elective care, model extensions are needed. This includes integrating capacity dynamics from other critical areas such as the OR and ICU. Our model's modular design generally allows for representing these capacities, since they have a pandemic demand component and a staff availability component. To be practically applicable, the interactions between these areas still need to be explicitly modeled and validated.

In summary, the model provides a better understanding of how healthcare capacity and elective care backlogs interact during pandemics. It delivers a forecasting tool and policy evaluation framework that helps identify how to best use limited capacity. These insights give decision-makers and capacity planners a solid foundation for preparing for the next health emergency, to make more effective choices under pressure and minimize population health loss.

Bibliography

[1] COVIDSurg Collaborative. Elective surgery cancellations due to the COVID-19 pandemic: global predictive modelling to inform surgical recovery plans. *Journal of British Surgery*, 107(11):1440–1449, 2020.

[2] Michelle R De Graaff, Rianne NM Hogenbirk, Yester F Janssen, Arthur KE Elfrink, Ronald SL Liem, Simon W Nienhuijs, Jean-Paul PM de Vries, Jan-Willem Elshof, Emiel Verdaasdonk, Jarno Melenhorst, Mark GH Besselink, Jelle P Ruurda, Mark I van Berge Henegouwen, Joost M Klaase, Marcel den Dulk, Mark van Heijl, Johannes H Hegeman, Jerry Braun, Daan M Voten, (...), Schelto Kruijff, and Dutch CovidSurg Collaborative Study Group. Impact of the COVID-19 pandemic on surgical care in the Netherlands. *British Journal of Surgery*, 109(12):1282–1292, 2022.

[3] Marije Oosterhoff, Lisanne HJA Kouwenberg, Adriënne H Rotteveel, Ella D Van Vliet, Niek Stadhouders, G Ardine de Wit, and Anoukh van Giessen. Estimating the health impact of delayed elective care during the COVID-19 pandemic in the Netherlands. *Social Science & Medicine*, 320:115658, 2023.

[4] Sue J Fu, Elizabeth L George, Paul M Maggio, Mary Hawn, and Rahim Nazerali. The consequences of delaying elective surgery: surgical perspective. *Annals of surgery*, 272(2):e79–e80, 2020.

[5] Maroeska M Rovers, Stan RW Wijn, Janneke PC Grutters, Sanne JJPM Metsemakers, Robin J Vermeulen, Ron Van Der Pennen, Bart JJM Berden, Hein G Gooszen, Mirre Scholte, and Tim M Govers. Development of a decision analytical framework to prioritise operating room capacity: lessons learnt from an empirical example on delayed elective surgeries during the COVID-19 pandemic in a hospital in the Netherlands. *BMJ open*, 12(4):e054110, 2022.

[6] Byron Breedlove. Emerging Pathogens Pose Inevitable Surprises. *Emerging Infectious Diseases*, 29(2):462–463, 2023.

[7] Thomas R Frieden, Marine Buissonnière, and Amanda McClelland. The world must prepare now for the next pandemic. *BMJ Global Health*, 6(3):e005184, 2021.

[8] Darren P Mareiniss, Jon M Hirshon, and Bryan C Thibodeau. Disaster planning: potential effects of an influenza pandemic on community healthcare resources. *American journal of disaster medicine*, 4(3):163, 2009.

[9] Michael T Osterholm. Preparing for the next pandemic. In *Global health*, pages 225–238. Routledge, 2017.

[10] Patrick Stewart. When the System Fails. COVID-19 and the Costs of Global Dysfunction. *Foreign Affairs*, 99(4):40–50, 2020.

[11] Milton C Weinstein and William B Stason. Foundations of cost-effectiveness analysis for health and medical practices. *New England journal of medicine*, 296(13):716–721, 1977.

[12] Sarah J Whitehead and Shehzad Ali. Health outcomes in economic evaluation: the QALY and utilities. *British medical bulletin*, 96(1):5–21, 2010.

[13] Rosalind Rabin and Frank de Charro. EQ-SD: a measure of health status from the EuroQol Group. *Annals of medicine*, 33(5):337–343, 2001.

[14] Leslee L Subak and Aaron B Caughey. Measuring cost-effectiveness of surgical procedures. *Clinical obstetrics and gynecology*, 43(3):551–560, 2000.

[15] Jmarchn. Graphical comparison of two projected quality-adjusted life years (QALYs), 2018. URL: https://commons.wikimedia.org/w/index.php?curid=67001576. This work is licensed under the CC BY-SA 3.0 International License. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/3.0/.

[16] Gary C Brown, Melissa M Brown, and Brandon G Busbee. Cost-utility analysis of cataract surgery in the United States for the year 2018. *Journal of Cataract & Refractive Surgery*, 45(7):927–938, 2019.

[17] Richard Fordham, Jane Skinner, Xia Wang, John Nolan, and Exeter Primary Outcome Study Group. The economic benefit of hip replacement: a 5-year follow-up of costs and outcomes in the Exeter Primary Outcomes Study. *BMJ open*, 2(3):e000752, 2012.

[18] Justin D Postma and Marius A Kemler. The effect of carpal tunnel release on health-related quality of life of 2346 patients over a 5-year period. *Journal of Hand Surgery (European Volume)*, 47(4):347–352, 2022.

[19] Roberto de la Plaza Llamas, Lorena Ortega Azor, Marina Hernández Yuste, Ludovica Gorini, Raquel Aránzazu Latorre-Fragua, Daniel Alejandro Díaz Candelas, Farah Al Shwely Abduljabar, and Ignacio Antonio Gemio Del Rey. Quality-adjusted life years and surgical waiting list: Systematic review of the literature. *World Journal of Gastrointestinal Surgery*, 16(4):1155, 2024.

[20] Ardine de Wit, Marije Oosterhoff, Lisanne HJA Kouwenberg, Adriënne H Rotteveel, Ella D Van Vliet, K Janssen, Marieje Stoelinga, Koen Visscher, and Anoukh van Giessen. De gezondheidsgevolgen van uitgestelde operaties tijdens de coronapandemie, 2022.

[21] Benjamin Gravesteijn, Eline Krijkamp, Jan Busschbach, Geert Geleijnse, Isabel Retel Helmrich, Sophie Bruinsma, Céline van Lint, Ernest van Veen, Ewout Steyerberg, Kees Verhoef, Jan van Saase, Hester Lingsma, and Rob Baatenburg de Jong. Minimizing population health loss in times of scarce surgical capacity during the coronavirus disease 2019 crisis and beyond: a modeling study. *Value in Health*, 24(5):648–657, 2021.

[22] Jashvant Poeran, Haoyan Zhong, Lauren Wilson, Jiabin Liu, and Stavros G Memtsoudis. Cancellation of elective surgery and intensive care unit capacity in New York State: a retrospective cohort analysis. *Anesthesia & Analgesia*, 131(5):1337–1341, 2020.

[23] Salman Alsafran, Dalia Albloushi, Danah Quttaineh, Abdullah A Alfawaz, Ahmed Alkhamis, Ali Alkhayat, Maha Alsejari, and Salman Alsabah. The impact of the COVID-19 pandemic on surgeons' and surgical residents' caseload, surgical skills, and mental health in Kuwait. *Medical Principles and Practice*, 31(3):224–230, 2022.

[24] Hanbin Luo, Jiajing Liu, Chengqian Li, Ke Chen, and Ming Zhang. Ultra-rapid delivery of specialty field hospitals to combat COVID-19: Lessons learned from the Leishenshan Hospital project in Wuhan. *Automation in construction*, 119:103345, 2020.

[25] Prateek Behera, Zainab Ahmad, Amol Dubepuria, Nitu Mishra, Anirban Chatterjee, John A Santoshi, Rehan Ul Haq, and Jai Prakash Sharma. Repurposing surgical wards in pandemics–An appraisal of outcomes of COVID-19 patients treated in Orthopaedic wards. *Journal of Family Medicine and Primary Care*, 13(5):1868–1874, 2024.

[26] Juliane Winkelmann, Erin Webb, Gemma A Williams, Cristina Hernández-Quevedo, Claudia B Maier, and Dimitra Panteli. European countries' responses in ensuring sufficient physical infrastructure and workforce capacity during the first COVID-19 wave. *Health Policy*, 126(5):362–372, 2022.

[27] Chen Shu-Ching, Lai Yeur-Hur, and Tsay Shiow-Luan. Nursing perspectives on the impacts of COVID-19. *Journal of Nursing Research*, 28(3):e85, 2020.

[28] Jean-Paul Chretien, Dylan George, Jeffrey Shaman, Rohit A Chitale, and F Ellis McKenzie. Influenza forecasting in human populations: a scoping review. *PloS one*, 9(4):e94130, 2014.

[29] Elaine O Nsoesie, John S Brownstein, Naren Ramakrishnan, and Madhav V Marathe. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and other respiratory viruses*, 8(3):309–316, 2014.

[30] Stef Baas, Sander Dijkstra, Aleida Braaksma, Plom van Rooij, Fieke J Snijders, Lars Tiemessen, and Richard J Boucherie. Real-time forecasting of COVID-19 bed occupancy in wards and Intensive Care Units. *Health care management science*, 24:402–419, 2021.

[31] Mieke Deschepper, Kristof Eeckloo, Simon Malfait, Dominique Benoit, Steven Callens, and Stijn Vansteelandt. Prediction of hospital bed capacity during the COVID-19 pandemic. *BMC health services research*, 21(1):468, 2021.

[32] René Bekker, Michiel uit het Broek, and Ger Koole. Modeling COVID-19 hospital admissions and occupancy in the Netherlands. *European journal of operational research*, 304(1):207–218, 2023.

[33] Sander Dijkstra, Stef Baas, Aleida Braaksma, and Richard J Boucherie. Dynamic fair balancing of COVID-19 patients over hospitals based on forecasts of bed occupancy. *Omega*, 116:102801, 2023.

[34] Stef Baas, Sander Dijkstra, Richard J. Boucherie, and Anne Zander. A stochastic programming approach for dynamic allocation of bed capacity and assignment of patients to collaborating hospitals during pandemic outbreaks. *arXiv preprint*, arXiv: 2311.15898, 2023.

[35] Ruth McCabe, Nora Schmit, Paula Christen, Josh C D'Aeth, Alessandra Løchen, Dheeya Rizmie, Shevanthi Nayagam, Marisa Miraldo, Paul Aylin, Alex Bottle, et al. Adapting hospital capacity to meet changing demands during the COVID-19 pandemic. *BMC medicine*, 18:1–12, 2020.

[36] Vari M Drennan and Fiona Ross. Global nurse shortages—the facts, the impact and action for change. *British medical bulletin*, 130(1):25–37, 2019.

[37] Niels Holthof and Markus M Luedi. Considerations for acute care staffing during a pandemic. *Best Practice & Research Clinical Anaesthesiology*, 35(3):389–404, 2021.

[38] Sharon Dezzani Martin. Nurses' ability and willingness to work during pandemic flu. *Journal of Nursing Management*, 19(1):98–108, 2011.

[39] Monique MA Penturij-Kloks, Simon T de Gans, Mandy van Liempt, Esther de Vries, Fedde Scheele, and Carolina JPW Keijsers. Pandemic Lessons for Future Nursing Shortage: A Prospective Cohort Study of Nurses' Work Engagement before and during 16 Months of COVID-19. *Journal of Nursing Management*, 2023(1):6576550, 2023.

[40] Violeta Lopez, Judith Anderson, Sancia West, and Michelle Cleary. Does the COVID-19 pandemic further impact nursing shortages? *Issues in Mental Health Nursing*, 43(3):293–295, 2022.

[41] Shujin Jiang, Nan Kong, Kathleen Abrahamson, Mingyang Li, and Yuehwern Yih. Optimal nursing home service scheduling under COVID-19 related probabilistic staff shortage: A two-stage stochastic programming approach. In *2021 Industrial and Systems Engineering Research Conference, year=2021.*

[42] Mark N Abramovich, Eric S Toner, and Jason Matheny. Panalysis: a new spreadsheet-based tool for pandemic planning. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 6(1):78–92, 2008.

[43] World Health Organization. Prevention, identification and management of health worker infection in the context of COVID-19: interim guidance, 30 October 2020. Technical report, World Health Organization, 2020.

[44] Reina S Sikkema, Suzan D Pas, David F Nieuwenhuijse, Áine O'Toole, Jaco Verweij, Anne van Der Linden, Irina Chestakova, Claudia Schapendonk, Mark Pronk, Pascal Lexmond, Theo Bestebroer, Ronald J Overmars, Stefan van Nieuwkoop, Wouter van den Bijllaardt, Robbert G Bentvelsen, Miranda ML van Rijen, Anton GM Buitin, Anne JG van Oudheusden, (...), Marjolein FQ Kluytmans van den Bergh, and Marion PG Koopmans. COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study. *The Lancet Infectious Diseases*, 20(11):1273–1280, 2020.

[45] Chantal B Reusken, Anton Buiting, Chantal Bleeker-Rovers, Bram Diederen, Mariëtte Hooiveld, Ingrid Friesema, Marion Koopmans, Titia Kortbeek, Suzanne PM Lutgens, Adam Meijer, Jean-Luc Murk, Ilse Overdevest, Thera Trienekens, Wouter Timen, Aura ad van den Bijllaardt, Jaap van Dissel, Arianne van Gageldonk-Lafeber, Dewi van der Vegt, (...), and Jan Kluytmans. Rapid assessment of regional SARS-CoV-2 community transmission through a convenience sample of healthcare workers, the Netherlands, March 2020. *Eurosurveillance*, 25(12):2000334, 2020.

[46] Klaudyna Grzelakowska and Jacek Kryś. The impact of COVID-19 on healthcare workers' absenteeism: infections, quarantines, sick leave—a database analysis of the antoni jurasz university hospital No. 1. in bydgoszcz, Poland. *Medical Research Journal*, 6(1):47–52, 2021.

[47] Julia A Bielicki, Xavier Duval, Nina Gobat, Herman Goossens, Marion Koopmans, Evelina Tacconelli, and Sylvie van der Werf. Monitoring approaches for health-care workers during the COVID-19 pandemic. *The Lancet infectious diseases*, 20(10):e261–e267, 2020.

[48] Edward J Mascha, Patrick Schober, Joerg C Schefold, Frank Stueber, and Markus M Luedi. Staffing with disease-based epidemiologic indices may reduce shortage of intensive care unit staff during the COVID-19 pandemic. *Anesthesia & Analgesia*, 131(1):24–30, 2020.

[49] Mart L Stein, James W Rudge, Richard Coker, Charlie van Der Weijden, Ralf Krumkamp, Piya Hanvoravongchai, Irwin Chavez, Weerasak Putthasri, Bounlay Phommasack, Wiku Adisasmito, Sok Touch, Le M Sat, Yu-Chen Hsu, Mirjam Kretzschmar, and Aura Timen. Development of a resource modelling tool to support decision makers in pandemic influenza preparedness: The AsiaFluCap Simulator. *BMC public health*, 12:1–14, 2012.

[50] Helena C Maltezou, Caterina Ledda, and Nikolaos V Sipsas. Absenteeism of healthcare personnel in the COVID-19 era: a systematic review of the literature and implications for the post-pandemic seasons. In *Healthcare, volume=11, number=22, pages=2950, year=2023, organization=MDPI.*

[51] Kirstin Khonyongwa, Surabhi K Taori, Ana Soares, Nergish Desai, Malur Sudhanva, Will Bernal, Silke Schelenz, and Lisa A Curran. Incidence and outcomes of healthcare-associated COVID-19 infections: significance of delayed diagnosis and correlation with staff absence. *Journal of Hospital Infection*, 106(4):663–672, 2020.

[52] Peixia Gao, Sabine Wittevrongel, and Herwig Bruneel. Discrete-time multiserver queues with geometric service times. *Computers & Operations Research*, 31(1):81–99, 2004.

[53] John DC Little. A proof for the queuing formula: L= λ w. *Operations research*, 9(3):383–387, 1961.

[54] S Subba Rao. Heavy Traffic Analysis for Discrete Time Queues. *Research Reports from the Department of Operationsn*, 234, 1973. URL: https://commons.case.edu/wsom-ops-reports/234. Accessed: 02-02-2025.

[55] Paul Lévy. Variables aléatoires. *Paris. Gauthier-Villars*, 1937.

[56] Boris V Gnedenko and Igor N Novalenko. *Introduction to Queuing Theory, 2nd ed.* Birkhäuser Boston, MA, 1989.

[57] Torben Meisling. Discrete-Time Queuing Theory. *Operations Research*, 6(1):96–105, 1958.

[58] Rein Nobel and Annette Rondaij. A Discrete-Time Queueing Model in a Random Environment. In *Queueing Theory and Network Applications: 14th International Conference, QTNA 2019, Ghent, Belgium, August 27–29, 2019, Proceedings 14, pages=330–348, year=2019, organization=Springer.*

[59] ZGT. Over ZGT, 2025. URL: https://www.zgt.nl/over-zgt/. Accessed: 04-06-2025.

[60] Rijksinstituut voor Volksgezondheid en Milieu (RIVM). Testing policy for COVID-19, 2020. URL: https://www.rivm.nl/en/novel-coronavirus-covid-19/testing-policy. Accessed: 04-06-2025.

[61] Enrico R de Koning, Mark J Boogers, Saskia LMA Beeres, Iwona D Kramer, Wouter J Dannenberg, and Martin J Schalij. Managing hospital capacity: achievements and lessons from the COVID-19 pandemic. *Prehospital and Disaster Medicine*, 37(5):600–608, 2022.

[62] Michael Klompas, Meghan A Baker, Chanu Rhee, Robert Tucker, Karen Fiumara, Diane Griesbach, Carin Bennett-Rizzo, Hojjat Salmasian, Rui Wang, Noah Wheeler, Glen R Gallagher, Andrew S Lang, Timelia Fink, Stephanie Baez, Sandra Smole, Larry Madoff, Eric Goralnick, Andrew Resnick, Madelyn Pearson, (...), and Charles A Morris. A SARS-CoV-2 cluster in an acute care hospital. *Annals of Internal Medicine*, 174(6):794–802, 2021.

Disclosure of AI Use

During the preparation of this work, the author used OpenAI's ChatGPT in order to assist with phrasing and formulation. The author also used Overleaf's built-in language check feature solely to assist with correcting syntax. All research ideas, models, interpretations, and scientific content originate entirely from the author, with guidance and input from the supervisor. After using this service, the author reviewed and edited the content as needed and takes full responsibility for the content of this work.