UNIVERSITY OF ICELAND

# Pattern Scheduling

A Practical Approach to Preventing Surgery Cancellations Due to Uncertainty in Surgery Times, Bed Availability, and Arrivals of Semi-Acute Elective Patients

Ásgeir Örn Sigurpálsson

2025

FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

# Pattern Scheduling

A Practical Approach to Preventing Surgery Cancellations
Due to Uncertainty in Surgery Times, Bed Availability, and Arrivals
of Semi-Acute Elective Patients

Ásgeir Örn Sigurpálsson

Dissertation submitted in partial fulfillment of a
*Philosophiae Doctor degree in Industrial Engineering*

Supervisors
Dr. Thomas Philip Rúnarsson
Dr. Rögnvaldur Jóhann Sæmundsson

Doctoral Committee
Dr. Edmund Kieran Burke
Dr. Rögnvaldur Jóhann Sæmundsson
Dr. Thomas Philip Rúnarsson

Opponents
Dr. Jaideep J Pandit
Dr. Michael O'Sullivan

Faculty of Industrial Engineering, Mechanical Engineering and
Computer Science
School of Engineering and Natural Sciences
University of Iceland
Reykjavik, January 2025

Pattern Scheduling: A Practical Approach to Preventing Surgery Cancellations Due to Uncertainty in Surgery Times, Bed Availability, and Arrivals of Semi-Acute Elective Patients
(Pattern Scheduling)

Dissertation submitted in partial fulfillment of a *Philosophiae Doctor* degree in Industrial Engineering

Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
School of Engineering and Natural Sciences
University of Iceland
Dunhagi 5
107 Reykjavik
Iceland

Telephone: 525-4000

Bibliographic information:
Ásgeir Örn Sigurpálsson (2025). *Pattern Scheduling: A Practical Approach to Preventing Surgery Cancellations Due to Uncertainty in Surgery Times, Bed Availability, and Arrivals of Semi-Acute Elective Patients*, PhD dissertation, Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, 143 pp.

ISBN 978-9935-9807-7-9

Copyright © 2025 Ásgeir Örn Sigurpálsson
All rights reserved

Printing: Háskólaprent, Fálkagata 2, 107 Reykjavík
Reykjavik, Iceland, January 2025

# Abstract

Demographic change, increasing cost of care, and shortage of hospital workers pose challenges in hospital management. As a response, hospitals maintain high utilisation of their existing resources by maximising patient throughput to minimise waiting times. However, maintaining a high resource utilisation on a continuous basis is likely to result in last-minute cancellations or rescheduling events due to multiple sources of uncertainty. In this context, surgery scheduling has received central attention from researchers and healthcare officials globally.

This research aims to increase our understanding of how to maintain a high throughput of elective patients under limited resources while minimising the combined risk of last-minute cancellations and rescheduling events. The objective is to develop mathematical models that address the problem practically and statistically accurately while considering several sources of uncertainty frequently resulting in last-minute cancellations and rescheduling events. A single surgical speciality, General Surgery at Landspitali Hospital, was selected for the computational experiments, and the results were compared to the actual scheduling data.

The results show that using chance constraints makes it possible to reduce the risk of last-minute cancellations due to uncertainty in surgery times and length of stay. However, utilising such constraints makes the problem computationally intractable. Therefore, Pattern Scheduling is proposed to overcome the computational challenges by specifying practical rules. Further results show that leaving 20% of the weekly operating room capacity unreserved makes it possible to reduce the need for rescheduling to accommodate unpredictable arrivals of semi-acute elective patients while maintaining high utilisation.

# Útdráttur

Hækkandi meðalaldur, aukinn umönnunarkostnaður og skortur á heilbrigðisstarfsfólki valda sífellt fleiri áskorunum í rekstri sjúkrahúsa. Því er eðlilegt að sjúkrahús hafi það að markmiði sínu að hámarka nýtingu auðlinda með því að ná fram sem mestu flæði sjúklinga og lágmarka biðtíma eftir veittri þjónustu. Það reynist þó þrautin þyngri því ýmsir óvissuþættir leiða oft til frestana og endurröðunar. Af þessum sökum hefur röðun skurðaðgerða fangað athygli rannsakenda og heilbrigðisstarfsmanna víða um heim.

Tilgangur þessarar rannsóknar er að auka skilning okkar á því hvernig hámarka megi flæði sjúklinga í valaðgerðum í umhverfi þar sem auðlindir eru takmarkaðar. Samhliða því þarf að lágmarka tilfærslur á borð við frestanir og endurröðun. Markmiðið er að þróa stærðfræðileg líkön sem leysa verkefnið á hagnýtan en nákvæman hátt og taka tillit til óvissu af ýmsum toga sem leiðir til fyrrnefndra tilfærslna. Líkönin verða þróuð með þarfir almennra skurðlækninga á Landspítalanum í huga og verða niðurstöður þeirra bornar saman við raungögn.

Helstu niðurstöður sýna fram á að með því að nota líkindaskorður í stærðfræðilegum líkönum megi á nákvæman hátt lágmarka frestanir með skömmum fyrirvara sem skapast af óvissu í skurðtímum og legulengd. Með notkun slíkra skorða eykst hins vegar reikniþungi verkefnanna með aukinni stærð. Því var þróuð aðferð sem kallast mynsturröðun, en með henni má koma í veg fyrir reiknifræðilegar áskoranir með því að tilgreina hagnýtar reglur. Frekari niðurstöður sýna fram á að með því að skilja 20% af heildarskurðtíma eftir auðan í hverri viku má lágmarka endurröðun, sem skapast af ófyrirsjáanlegum komum sjúklinga með háan forgang, en þó þannig að nýting auðlinda sé hámörkuð.

# Contents

**Abstract** iii

**Útdráttur** v

**Contents** vii

**List of Figures** xi

**List of Tables** xiv

**List of Original Publications** xv

**Abbreviations** xvii

**Nomenclature** xix

**Acknowledgments** xxv

**1 Introduction** 1
- 1.1 Motivation and Background . . . 1
- 1.2 Research Objectives . . . 6
- 1.3 Contributions . . . 6
- 1.4 Outline . . . 8

**2 Literature Review** 11
- 2.1 Scheduling Levels . . . 11
- 2.2 Patient Pathways and Uncertainty at the Operational Level . . . 14
  - 2.2.1 Uncertainty in Surgery Times . . . 15
  - 2.2.2 Uncertainty in Length of Stay . . . 19
  - 2.2.3 Uncertainty in Arrivals . . . 21
- 2.3 Summary . . . 24

**3 Methodology** 27
- 3.1 Model Development Process . . . 27
- 3.2 Data Collection and Analysis . . . 28
- 3.3 Problem Description . . . 31
- 3.4 Experimental Setup . . . 34

3.4.1 Uncertainty in Surgery Times . . . 35
3.4.2 Uncertainty in Length of Stay . . . 35
3.4.3 Uncertainty in Semi-Acute Elective Arrivals . . . 36
3.5 Summary . . . 36

**4 Uncertainty in Surgery Times 37**
4.1 Model Development . . . 39
4.1.1 Pattern Generation . . . 40
4.1.2 Optimisation with Limited Downstream Ward Beds . . . 44
4.2 Experimental Study . . . 48
4.2.1 Instance Generation . . . 49
4.2.2 Parameter Settings . . . 49
4.2.3 Results . . . 50
4.3 Conclusions . . . 53
4.4 Summary . . . 57

**5 Uncertainty in Length of Stay 59**
5.1 Model Development . . . 60
5.1.1 Pattern Generation . . . 61
5.1.2 Ward Combinations Optimisation . . . 65
5.1.3 Robust Ward Optimisation . . . 69
5.2 Experimental Study . . . 70
5.2.1 Parameter Settings . . . 71
5.2.2 Comparison . . . 73
5.2.3 Parameter Analysis . . . 76
5.3 Conclusions . . . 82
5.4 Summary . . . 83

**6 Uncertainty in Semi-Acute Elective Arrivals 85**
6.1 Planning Gaps in the Surgery Program . . . 87
6.1.1 Gap-reserving Strategies . . . 87
6.2 Model Development . . . 90
6.2.1 Base Scheduling Model . . . 92
6.2.2 Rescheduling Model . . . 96
6.3 Computational Experiments . . . 98
6.3.1 Comparison of the Gap-reserving Strategies . . . 100
6.4 Conclusions . . . 106
6.5 Summary . . . 107

**7 Discussion and Conclusions 109**
7.1 Practical Implications . . . 111
7.2 Future Research . . . 111

**8 Contributions 113**

**Appendices 126**

**A Supplementary Data: Statistical Analysis of Surgery Times** **126**

**B Practical Heuristics** **127**
B.1 Pattern Scheduling . . . . . 127
B.1.1 Pattern Generation . . . . . 127
B.1.2 Scheduling . . . . . 128
B.2 Ward Combinations . . . . . 129
B.2.1 Generation of Ward Combinations . . . . . 129
B.2.2 Discretisation . . . . . 131
B.2.3 Ward Combination Optimisation . . . . . 131

**C Scheduling and Rescheduling Models** **135**
C.1 Front Load . . . . . 135
C.2 Daily Limit . . . . . 136
C.3 Balance ORs . . . . . 137

**D Supplementary Results: Uncertainty in Semi-Acute Elective Arrivals** **139**

# List of Figures


2.1 Common pathways of patients after surgery. . . . . 15
3.1 Distribution of historical surgery times for five frequently performed operations at General Surgery. . . . . 30
3.2 Distribution of historical surgery times for a specific type of operation performed by each surgeon at General Surgery. . . . . 31
3.3 The probability of being in the ward on a given day following surgery for two types of operations. . . . . 32
4.1 Discretisation of an empirical distribution of daily ward probabilities using three scenarios, each with an associated probability. . . . . 44
4.2 Example of a single pattern consisting of two patients requiring ward admission following surgery. . . . . 45
5.1 Discretisation of an empirical distribution of daily ward probabilities into five scenarios, each with an associated probability. . . . . 65
5.2 Approximation of an empirical distribution of daily ward probabilities with the worst-case scenario as illustrated by the dashed line using certainty $\omega = 0.25$. . . . . 71
5.3 Visualisation of the actual surgeries for each day and OR and the corresponding ward occupancy. . . . . 75
5.4 Visualisation of the optimal solution for each day and OR for the Ward Combination Optimisation and the corresponding ward occupancy. . . 77
5.5 Visualisation of the optimal solution from the robust ward optimisation for each day and OR and the corresponding ward occupancy. . . . . 78
6.1 The actual schedule of patients of the GS speciality during regular working hours by the first two weeks of a month for each OR (1 to 3) and day (1 to 12). . . . . 88
6.2 The actual schedule of patients of the GS speciality during regular working hours by the second two weeks of a month for each OR (1 to 3) and day (15 to 26). . . . . 89
6.3 Distribution of the number of patients scheduled per week across the planning horizon for the Base Schedule and the Final Schedule, divided by the three gap-reserving strategies and actual outcome. . . . . 103
6.4 Distribution of the number of open ORs (blocks used) per week across the planning horizon for the Base Schedule and the Final Schedule, divided by the three gap-reserving strategies and actual outcome. . . . 104
6.5 Distribution of the days between scheduled and rescheduled appointments across the three gap-reserving strategies (SR time). . . . . 106

6.6 Distribution of the number of days between scheduled and rescheduled appointments (SR time) across the three gap-reserving strategies and surgeons. . . . . 107
B.1 Example of how ward combinations are generated and validated towards the risk of exceeding a given bed capacity. . . . . 130
B.2 Linking an actual ward combination of patients to a pre-generated ward combination using indexing and bed availability to determine if the actual combination is valid. . . . . 133

# List of Tables


3.1 Summary statistics for the distributions of surgery times† and length of stay in the ward for each surgeon at the GS speciality. Throughput is shown for the year 2019. . . . . 29
3.2 Summary statistics distributions of surgery times and length of stay in the ward for each surgeon at the GS speciality for semi-acute elective arrivals. Throughput is shown for the year 2019. . . . . 30
3.3 Master Surgery Schedule for General Surgery surgeons. . . . . 32
4.1 Summary of statistics for the most frequently performed type of operations by each surgeon at General Surgery . . . . . 49
4.2 Overall comparison of the total number of patients scheduled and the risk of exceeding the ward beds using an actual Master Surgical Schedule (MSS) and an optimised MSS. . . . . 51
4.3 Comparison of the actual Master Surgery Schedule (MSS) and an optimised MSS. The table depicts the allocation of surgeons (1-9) to operating rooms ($r_1$ and $r_2$) for each day along with the ratio of inpatients in the parenthesis. . . . . 52
4.4 Comparison of the patient throughput under different parameter settings of planning horizon lengths and waiting list sizes for each surgeon. . 53
4.5 Optimal Master Surgery Schedule for a planning horizon length of 14 days with 30 patients on the waiting list of each surgeon. . . . . 54
4.6 Sub-Optimal Master Surgery Schedule for a planning horizon length of 28 days with 30 patients on the waiting list of each surgeon. . . . . 55
4.7 Comparison of the historical inpatient ratio to the scheduled inpatient ratio for each surgeon under different planning horizon lengths and waiting list sizes. . . . . 56
5.1 Parameter settings used to generate the feasible patterns. . . . . 72
5.2 Overall comparison between optimal solutions of the ward combinations optimisation (WCO), the robust ward optimisation (RWO) and the actual schedule for the planning horizon of one month. . . . . 74
5.3 Quality of solutions and computational requirements for different configurations of ($M^A$, $\Omega$, $|K|$) for the ward combination optimisation. . . 79
5.4 Quality of solutions and computational requirements for different configurations of ($M^A$, $\omega$) for the robust ward optimisation. . . . . 81
6.1 Parameters settings used to generate the feasible patterns. . . . . 100
6.2 Parameters used for the scheduling and rescheduling models. . . . . 100

6.3 Weights used for the objective function and are selected to ensure the order of priority. . . . . 101
6.4 Comparison of the experiment results for three gap-reserving strategies. Results are shown for the base model (Base Schedule) and after rescheduling (Final Schedule). . . . . 102
6.5 Summary statistics for the five most common types of operations that were rescheduled under each of the three gap-reserving strategies. . . 105
A.1 A pairwise comparison between the median surgery time of five frequent types of operations using Wilcoxon's rank sum test. . . . . 126
A.2 A pairwise comparison between the median surgery time of the operators for the same type of operation using Wilcoxon's rank sum test. . . . . 126
B.1 Example of two specific types of operation where their empirical daily empirical values have been discretised into five possible scenarios. . . 132
D.1 Summary statistics for the distributions of surgery times and length of stay (LOS) in the ward for each at General Surgery speciality. Throughput is shown for the year 2019. . . . . 139
D.2 Results for the actual scheduling data for each month. Results are shown for the base model (Base Schedule) and after rescheduling (Final Schedule). . . . . 140
D.3 Computational results for each month for the Front load strategy. . . . 141
D.4 Computational results for each month for the Daily limit strategy. . . . 142
D.5 Computational results for each month for the Balance ORs strategy. . 143

# List of Original Publications

**Paper I:** T.P. Runarsson and **A.O. Sigurpalsson**, 2019. Towards an evolutionary guided exact solution to elective surgery scheduling under uncertainty and ward restrictions, IEEE Congress on Evolutionary Computation (CEC), 2019, pp. 419-425, doi: 10.1109/CEC.2019.8790174.

**Paper II:** **A.O. Sigurpalsson**, T.P. Runarsson, R.J. Saemundsson, 2020. Stochastic Master Surgical Scheduling Under Ward Uncertainty. In: Bélanger, V., Lahrichi, N., Lanzarone, E., Yalçındağ, S. (eds) Health Care Systems Engineering. ICHCSE 2019. Springer Proceedings in Mathematics Statistics, vol 316. Springer, Cham, doi:10.1007/978–3–030–39694–7_13

**Paper III:** **A.O. Sigurpalsson**, T.P. Runarsson and R.J. Saemundsson, 2022, Bounding the Likelihood of Exceeding Ward Capacity in Stochastic Surgery Scheduling, Applied Sciences 12, no. 17: 8577. https://doi.org/10.3390/app12178577

**Paper IV:** **A.O. Sigurpalsson**, R.J. Saemundsson, and T.P. Runarsson, 2025. Mind the Gap: Strategies for Semi-Acute Patient Scheduling in Elective Surgery. Submitted.

# Abbreviations

**AASD** Accumulated Average Surgery Duration

**C-G** Column Generation

**DRO** Distributionally Robust Optimisation

**GS** General Surgery

**ICU** Intensive Care Unit

**Landspitali** National University Hospital of Iceland

**LOS** Length of Stay

**MIP** Mixed Integer Programming

**MSS** Master Surgery Schedule

**OR** Operating Room

**PACU** Post Anaesthesia Care Unit

**Pattern** An unordered list of patients that can be assigned to a single block.

**RO** Robust Optimisation

**RWO** Robust Ward Optimisation

**SAA** Sample Average Approximation

**Semi-acute** Elective patients of a high medical priority

**SP** Stochastic Programming

**S-R** Time between scheduled and rescheduled appointments.

**WC** Ward Combination

**WCO** Ward CSombination Optimisation

# Nomenclature

## Sets and Indices

| | |
|---|---|
| $o \in O$ | Set of surgeons. |
| $d \in D$ | Days in the planning horizon. |
| $v \in V$ | Weeks in the planning horizon. |
| $r \in R$ | Available operating rooms. |
| $a \in A$ | Available beds in the ward. |
| $t \in T$ | Cycle lengths in the planning horizon. |
| $i \in I$ | All patients on the waiting list for surgery. |
| $i \in I_o$ | Patients of surgeon $o$. |
| $i \in I_p$ | Patients included in pattern $p$. |
| $p \in P$ | List of all patterns $p$. |
| $p \in P_o$ | Patterns including surgeon $o$ where $P_o \subseteq P$. |
| $p \in P_s$ | Patterns including patient $s$ where $P_s \subseteq P$. |
| $(d, p, r) \in DPR$ | Pattern $p$ for days $d$ and rooms $r$ for which patients and surgeons are available. |
| $l \in L$ | Indexes for the ward combinations.. |
| $k \in K$ | List of all ward scenarios. |
| $k \in K'$ | List of all ward scenarios, excluding the first and the last. |
| $j \in J$ | Days after surgery starting from 0 (same day as the surgery). |

## Parameters

| | |
|---|---|
| $N^s$ | Soft upper bound on the number of patients assigned to each pattern. |
| $M^P$ | Upper bound on the maximum number of patients assigned to each pattern. |
| $\bar{M}^a$ | Upper bound on the maximum number of semi-acute elective patients assigned to a pattern. |
| $M^{ICU}$ | Upper bound on the number of patients assigned to a pattern and requiring ICU admission. |
| $M^{\phi}$ | Upper bound on the number of patients assigned to a pattern and admission to resource $\phi \in \Phi$. |
| $\bar{M}^{ICU}$ | Upper bound on the number of patients requiring ICU admission in a pattern. |
| $M^W$ | Upper bound on the number of days a patient stays in the ward. |
| $M^I$ | Upper bound on the number of days a patient stays in the ICU. |
| $M^A$ | Upper bound on the number of staffed ward beds. |
| $M^A_{ICU}$ | Upper bound on the number of staffed ICU beds. |
| $C_{d,r}$ | Opening hours of a block allocated to day $d$ and room $r$. |
| $C_p$ | Number of patients assigned to pattern $p$. |
| $h_o$ | Desired ratio of inpatients from the throughput for surgeon $o$. |
| $F_a^{50}$ | Upper bound on the number maximum number of patients with 50% chance of being in the ward on a day with $a$ number of staffed beds available. |
| $F_{l,a}$ | Feasibility of ward combination $l$ when there are $a$ beds available in the ward. |
| $g_i^{Ward}$ | A binary parameter with the value 1 if the patient $i$ requires ward admission after surgery and 0 otherwise. |
| $g_i^{ICU}$ | A binary parameter taking the value 1 if patient $i$ requires ICU admission after surgery, otherwise 0. |

| | |
|---|---|
| $g_i^{\phi}$ | A binary parameter taking the value 1 if patient $i$ requires admission to resource $\phi \in \Phi$ after surgery. |
| $G_p$ | Number of in-patients for pattern $p$. |
| $n_p^{ICU}$ | Number of patients in pattern $p$ that require ICU admission following surgery. |
| $\bar{n}_d^{\omega}$ | Number of patients with certainty $\omega$ in the ward from previous plan on day $d$. |
| $h_G$ | Required balance (ratio) between in-patients and out-patients in the planning horizon |
| $\rho_k$ | Ward probabilities associated with scenario $k \in K$. |
| $n_k$ | Number of patients that belong to scenario $k$. |
| $\rho'_{i,d}$ | Probability of a patient $i$ being in the ward on the day $d$. |
| $Q_{j,p}^{50}$ | Number of patients with a 50% chance of being in the ward on the day $j$ for pattern $p$. |
| $Q_{j,p}^{100}$ | Number of patients with a 100% chance of being in the ward on the day $j$ after surgery and belonging to pattern $p$ |
| $Q_{j,k,p}$ | Number of patients on day $j$ after surgery belonging to scenario $k$ and pattern $p$. |
| $Q_{j,p}^{ICU}$ | Number of patients in ICU on day $j$ after surgery belonging pattern $p$. |
| $\bar{n}_{d,k}$ | Number of patients operated in the previous planning period that occupy the ward on day $d$ and belong to scenario $k$. |
| $\bar{n}_d^{ICU}$ | Number of patients operated in the ICU on day $d$ from previous planning horizon. |
| $\bar{n}_d^{w}$ | Number of patients in the ward on day $d$ from previous planning horizon. |
| $\delta$ | Limit on the probability that a pattern exceeds $C_{d,r}$. |
| $\Omega$ | Limit on the likelihood that a ward combination exceeds the number of available beds in the ward. |
| $\Upsilon_{o,p}$ | Parameter containing the ratio of surgeries from the total number of surgeries for surgeon $o$ in patients $p$. |
| $\gamma_i$ | The day on which the patient $i$ is scheduled to in the current schedule. |

| | |
|---|---|
| $M$ | A large number. |
| $S^P$ | Set of patients with a high medical priority. |
| $\omega$ | Level of certainty that the number of beds occupied in the ward are kept below $M^A$. |
| $D^v$ | A parameter specifying the week of day $d$ |
| $\mathbb{1}_{\omega \geq \rho'_{i,j}}$ | Robust parameter taking the value 1 if $\omega \leq \rho'_{i,j}$. |
| $\delta_p$ | Probability that the sum of surgery duration for pattern $p$ surpasses $C_{d,r}$. |
| $\delta'$ | Accepted risk of entering overtime. |
| $\Delta_{d,r}$ | Threshold for extended overtime. |
| $\delta_p^\Delta$ | Probability that the sum of surgery duration for pattern $p$ surpasses $C_{d,r} + \Delta_{d,r}$. |
| $\delta^{\Delta'}$ | Accepted risk of entering extended overtime. |
| $w$ | Factor weighing the relative contribution of regular and extended overtime in the objective function. |
| $\lambda^w$ | Penalty cost for the maximum peak in ward occupancy in the planning horizon. |
| $\lambda^{ov}$ | Penalty cost for selecting pattern with either regular or extended overtime. |
| $\lambda^{FL}$ | Penalty cost for the Front load gap-reserving strategy. |
| $\lambda^{DL}$ | Penalty cost for selecting a pattern with more than $N^S$ patients assigned for the Daily limit gap-reserving strategy. |
| $\lambda^{BO}$ | Penalty cost used to minimise the maximum number of blocks used each week for the Balance ORs gap-reserving strategy. |
| $\lambda^{RS}$ | Penalty cost for re-scheduling patients. |

# Variables

| | |
|---|---|
| $a_d$ | Number of available ward beds on day $d$. |
| $u_{d,r}$ | 1 if $\delta_p > \delta'$, 0 otherwise. |
| $v_{d,r}$ | 1 if $\delta_p^\Delta > \delta^{\Delta'}$, 0 otherwise. |

| | |
|---|---|
| $n_{d,k}$ | Number of patients that belong to scenario $k$ and occupy the ward on day $d$. |
| $n_d^w$ | Number of patients in ward on day $d$. |
| $\bar{n}_d^{50}$ | Number of patients with 50% chance of being in ward on day $d$. |
| $\bar{n}_d^{100}$ | Number of patients with 100% chance of being in ward on day $d$. |
| $P^{in}$ | Continuous variable denoting the total number of in-patients scheduled across the planning horizon. |
| $\bar{R}_{d_r}$ | A binary indicator variable taking the value 1 if on day $d$ there are more than $N^i$ patients assigned. |
| $V^B$ | Maximum number of blocks utilised each week |
| $w^{max}$ | Maximum ward bed occupancy in the planning horizon. |
| $\Delta_i^+$ | The number of days between the assignment of patient I in the current planning horizon and the assignment in the previous planning horizon. |
| $\Psi_i$ | The date on which the patient $i$ was scheduled in the previous schedule. |

# Decision variables

| | |
|---|---|
| $x_{d,p,r}$ | 1 if pattern $p$ is scheduled to day $d$ and room $r$, 0 otherwise. |
| $\Delta_i$ | 1 if patient $i$ is assigned to a different date in the current planning horizon otherwise 0. |
| $z_{i,p}$ | 1 if patient $i$ is assigned to pattern $p$, otherwise 0. |
| $z_{d,a}$ | 1 if $a$ beds are available in the ward on day $d$, otherwise 0. |
| $y_{d,l}$ | 1 if ward combination $l$ is assigned to day $d$, otherwise 0. |
| $\Xi_{d,a}$ | 1 if on day $d$ there are $a$ beds available. |

## Random variables

| | |
|---|---|
| $S(i)$ | Surgery duration for patient $i$. |
| $W(l)$ | Total number of patients belonging to scenarios $k \in K'$ for ward combination $l$. |
| $B\Big(n_k(l), \rho_k\Big)$ | Binomially distributed random variable for the $n_k(l)$ number of patients in ward combination $l$ belonging to scenario $k$ with probability $\rho_k$. |

# Acknowledgments

This research was funded by the Icelandic Technology Development Fund (grant number 175373-0611). Without their support, this research would not have been possible. Additionally, I received financial support from the PhD student travel grant at the University of Iceland, which assisted me with conference travel costs. For this, I am thankful.

I would like to express my sincere gratitude to my supervisors, Prof. Thomas Philip Rúnarsson and Prof. Rögnvaldur Jóhann Saemundsson, for their constant support over the years. Their encouragement, suggestions and insightful feedback have been invaluable. I also thank Prof. Edmund Kieran Burke for his feedback and valuable suggestions. Lastly, I would like to acknowledge the staff and the managers at Landspitali for providing insights and support for this project.

I also like to thank my family and friends for all their support throughout this journey. My girlfriend, Anna Lísa, has offered me unconditional support and patience. Words cannot describe how thankful I am. Lastly, I want to thank my parents, Hjördís and Sigurpáll, for their unwavering support and faith in me over the years, which has been fundamental on this journey.

Along this journey, I endured two significant losses of dear family members. In 2021, I lost my beloved dog. Nói was my companion during the early stages of this research. Last summer, Anna Lísa's father, Benedikt, passed away. Benedikt always showed a genuine interest in my work and was eagerly looking forward to the completion of this chapter and my defense. I will forever be grateful for their support. May they both rest in peace.

# 1 Introduction

This thesis considers the problem of the elective operating room (OR) planning and scheduling. The main focus will be at the operational level, where actual elective patients on the waiting list are assigned to ORs and days (blocks) one to several weeks in advance. The goal is to develop scheduling models that take multiple sources of uncertainty into account to reduce the risk of last-minute cancellations and rescheduling events while enabling a high throughput of patients. This is achieved by assigning an unordered but feasible combination of patients (termed Pattern) to each block.

The scheduling models are intended for practical use at the General Surgery (GS) speciality at the National University Hospital of Iceland (Landspitali). This speciality is one of the most complex surgical specialities to schedule within any hospital due to the nature of the operations performed. GS has a long list of patients awaiting surgery but limited access to resources. As a result, available resources are continuously utilised at or close to maximum capacity in an attempt to maximise the number of patients scheduled. However, due to uncertainty in surgery times, length of stay (LOS) in the ward and the intensive care unit (ICU) after surgery, as well as the unpredictable arrivals of elective patients of a high medical priority (referred to as semi-acute elective patients from now on), the GS speciality has historically experienced last-minute cancellations and rescheduling events. Therefore, the hospital seeks ways to minimise the number of such occurrences when scheduling elective patients one to several weeks in advance but without reducing the utilisation of its resources and the desired throughput of patients.

## 1.1 Motivation and Background

The practical scheduling needs of Landspitali are the driving force behind the models developed in this thesis. Landspitali is an approximately 650-bed hospital in Reykjavik, Iceland, and is the leading healthcare provider in the country. This thesis focuses on one of the surgical specialities at Landspitali. The GS speciality was selected as it is one of the most challenging surgical specialities at the hospital and has historically experienced disruptions to its schedule. The majority of these disruptions stem from last-minute cancellations, mostly caused by ward bottlenecks and overtime, and rescheduling events due to arrival of semi-acute elective patients.

Demographic change, together with the rising cost of care, are expected to continue posing challenges in the management of Landspitali in the near future, as is true for other hospitals in the world (Eshghali et al., 2024). These challenges are further intensified with a widespread shortage of human resources, such as nurses (Tamata and Mohammadnezhad, 2023), which pressures hospitals to maintain a high level of resource utilisation to maximise the number of patients scheduled. However, due to the inherent variability of healthcare processes, maintaining a high patient throughput requirement on a continuous basis is likely to result in an unbalanced flow of patients, last-minute cancellations and disruptive rescheduling events. Such events are highly undesirable for patients, their families, the hospital and its staff (Ivarsson et al., 2004; Armoeyan et al., 2021; Pattnaik et al., 2022). Thus, healthcare officials are constantly seeking ways to reduce the likelihood of those events but without decreasing the throughput of patients.

In this regard, OR planning and scheduling problem has received particular attention amongst researchers due to their cost (Denton et al., 2007, 2010; Guerriero and Guido, 2011; Neyshabouri and Berg, 2017) and inter-dependency with a variety of resources at hospitals (Cardoen et al., 2010) that constrain the flow of patients. In general, the decisions in the problem are classified into three hierarchical levels. The first level is the strategic level, where long-term decisions are made (at least one year) about overall surgical capacity, i.e. including the ORs (Blake and Carter, 2002) and equipment (Wachtel and Dexter, 2008). At the tactical level, a cyclic timetable of the ORs, termed the Master Surgery Schedule (MSS), is created. In this timetable, blocks (a combination of days and ORs) are allocated to different surgical specialities, matching their total block hours as determined at the prior level. The MSS repeats itself every one or two weeks but is fixed months in advance (Blake and Donald, 2002; Guerriero and Guido, 2011). At the operational level, elective patients are selected from the waiting list and assigned to different blocks one to several weeks in advance, and their sequence is determined within the block. The former, when patients are assigned to blocks, is known as advance scheduling, which is of focus in this thesis, while the latter, when their sequence is determined, is known as allocation scheduling and is not considered (Samudra et al., 2016). The advance scheduling of elective patients has proven challenging due to the multiple sources of uncertainty inherent to the process. These different sources of uncertainty must be considered when elective patients are assigned to blocks up to several weeks in advance, in an attempt to avoid future last-minute cancellations and disruptive rescheduling events while maintaining high utilisation (Shylo et al., 2013).

Numerous sources of uncertainty impact the advance scheduling of elective patients (Van Riet and Demeulemeester, 2015), but three typically cause the most disruptions (Min and Yih, 2010; Van Riet and Demeulemeester, 2015; Riise et al., 2016; Zhang et al., 2019). Firstly, uncertainty in surgery times may cause under- (Kroer et al., 2018) or over-utilisation (Batun et al., 2011) of the blocks, where the latter is associated with last-minute cancellations (Hans et al., 2008). For instance, if a block on a specific day goes into overtime, e.g., due to a case taking longer than expected, the following case, if any, might be cancelled at the last minute as staff may not be available after hours to perform the surgery (Adams, 2019). Secondly, uncertainty in patients' LOS in the ward

and ICU may cause the unavailability of staffed beds (Augusto et al., 2010; Min and Yih, 2010), resulting in last-minute cancellations of other scheduled patients (Min and Yih, 2010; Neyshabouri and Berg, 2017). For example, a patient staying longer than expected can reduce the available bed capacity for future patients which may cause a last-minute cancellation. Finally, the unpredictable arrival of emergency patients who need to be operated on within the same day (Van Riet and Demeulemeester, 2015), as well as the arrival of semi-acute elective patients who typically need surgery within a week or two, may lead to the rescheduling of previously planned patients (Zonderland et al., 2010). The focus is on the latter group in this thesis.

Uncertainty in surgery times is widely addressed in mathematical models in the academic literature (Samudra et al., 2016; Van Riet and Demeulemeester, 2015; Zhu et al., 2019). Surgery times are typically dependent on the type of operation and the surgeon (Wang et al., 2021). However, other factors such as the surgeon's experience (Opit et al., 1991), type of anaesthetic (Dexter et al., 2008), and patient characteristics (Samudra et al., 2016; Najjarbashi and Lim, 2019; Gür et al., 2024) have also been shown to impact the variability. There are two ways possible to hedge against last-minute cancellations due to uncertainty in surgery times when operating under a high throughput requirement. One the one hand, by assuming the worst-case scenario for each patient, namely the worst-case surgery time, it is possible to hedge against such events (Addis et al., 2014, 2016). Inevitably, this approach leaves the hospital's resources underutilised, such as the blocks and, therefore, the surgeons. Moreover, it ignores all distributional information as the worst-case is only considered (Shehadeh and Padman, 2021). On the other hand, given that the time distributions are available for the different types of operations, e.g. from the hospital's historical data, approximations and chance constraints may be used to minimise the risk. With approximations, the time distributions are replaced with deterministic values, e.g. percentiles (Spratt and Kozan, 2021) or planned slacks (Hans et al., 2008; van Oostrum et al., 2008; Pandit and Tavare, 2011). With chance constraints, the different time distributions are fully utilised and can, therefore, effectively limit the risk of exceeding a given capacity to a specified threshold in a statistically accurate way (Shylo et al., 2013; Wang et al., 2014; Landa et al., 2016; Kamran et al., 2018). In other words, when chance constraints are applied they constrain the probability of an event occurring (Adams, 2019) which is often the desired outcome in practise (Kamran et al., 2018).

Suppose a scheduler wants to schedule four cases in a single block of a certain capacity after two weeks. The hospital could use simulation with historical data on surgery times to estimate the probability of exceeding the block's capacity. This means the scheduler could know in advance if this combination of cases has a high risk of a last-minute cancellation based on historical data. In other words, if 1000 surgery times are randomly sampled with a replacement for each patient, then 1000 scenarios of total surgery times could be generated. The hospital could then specify an upper bound on the number of scenarios exceeding the block's capacity. For example, the hospital could permit 300 out of 1000 scenarios to run into overtime. Consequently, if the simulation indicates that 500 scenarios ran into overtime, the scheduler must modify the schedule to avoid last-minute cancellations. Nevertheless, many studies, as well as hospitals, ignore

uncertainty in surgery times and schedule cases by their mean values instead (Lamiri et al., 2008a; Fei et al., 2008; Pandit and Tavare, 2011), thereby increasing the risk of last-minute cancellations (Min and Yih, 2010).

Following surgery, some patients (in-patients) must recover in downstream facilities, like in the ward or ICU, for one or more days (Fügener et al., 2014). Patients' LOS commonly depends on the type of operation and is highly stochastic (Min and Yih, 2010; Wang et al., 2021) similar to surgery times. A lack of staffed downstream beds may lead to last-minute cancellations or early discharge of other patients (Neyshabouri and Berg, 2017), the latter of which has been linked with an increased risk of readmittance (Jonnalagadda et al., 2005; Baker et al., 2009). With a fixed number of staffed beds, the only way to guarantee that last-minute cancellations are avoided is to assume that the bed capacity is tailored towards the worst-case LOS for each patient (Min and Yih, 2010; Jebali and Diabat, 2015; Neyshabouri and Berg, 2017). However, such conservative assumptions will translate to low utilisation of beds, which is undesirable and may reduce the throughput of patients. As a result, Jebali and Diabat (2017) implemented chance constraints in their model to hedge against exceeding the number of ICU beds to overcome this barrier. In their setting, ICU bottlenecks were kept below a specified threshold, reducing the risk of last-minute cancellations.

Hospitals can employ three predictive disruption management policies to account for the unpredictable and disruptive arrivals of semi-acute elective patients (Van Riet and Demeulemeester, 2015). First, hospitals may reserve gaps in some or all of its blocks in a deterministic (Van Riet and Demeulemeester, 2015) or stochastic manner (Lamiri et al., 2008b; Molina-Pariente et al., 2018; Jebali and Diabat, 2017), combining the different flows of scheduled elective patients and new arrivals (Wullink et al., 2007; Westbury et al., 2009). Second, hospitals may set aside one or more blocks for the new arrivals, thereby completely separating the flow of patients (Antognini et al., 2015; Parmar et al., 2022). Lastly, hospitals can use a combination of these two strategies (Zonderland et al., 2010; Van Riet and Demeulemeester, 2015).

Given that hospitals reserve gaps for semi-acute elective arrivals, it can be carried out in the following way. To begin with, an advance schedule is created for one to several weeks based on a list of elective patients awaiting surgery while simultaneously applying a strategy for reserving gaps for semi-acute elective arrivals. Finally, when semi-acute elective patients arrive, they are accommodated in the reserved gaps with minimal disruption to the advance schedule of other elective patients. These gap-reserving strategies can either be a simple heuristic (Jebali et al., 2006; Addis et al., 2016; Kamran et al., 2019; Spratt and Kozan, 2021), which human schedulers at hospitals can apply manually, or more sophisticated approaches where optimisation is used (Min and Yih, 2010; Molina-Pariente et al., 2018). A simple gap-reserving strategy employed by many hospitals is to leave some fraction, e.g. 15%, of each block unreserved for new arrivals and other changes (Jebali et al., 2006; Van Riet and Demeulemeester, 2015). However, Lamiri et al. (2008a) showed that stochastic gaps result in lower cost, e.g., due to overtime, compared to a model using deterministic gaps. Thus, several studies have attempted to propose human-friendly heuristics that reserve stochastic gaps, such

as limiting the number of surgeries a surgeon performs in a day (Kamran et al., 2019) or imposing bounds on the total block utilisation (Addis et al., 2016; Spratt and Kozan, 2021).

Although it is well known that uncertainty in surgery times, uncertainty in LOS in downstream units and unpredictable arrivals cause disruptions (Min and Yih, 2010; Van Riet and Demeulemeester, 2015; Riise et al., 2016; Zhang et al., 2019), they are only considered simultaneously in a handful of studies. Notably, Min and Yih (2010), Jebali and Diabat (2017) and Zhang et al. (2019) account for these different sources in their models. However, Min and Yih (2010) and Jebali and Diabat (2017) focus on reserving gaps for emergency patients but neglect semi-acute elective patients. In addition, none of the these studies addresses the combined risk of last-minute cancellations, such as by using chance constraints to hedge against exceeding the capacity of the blocks and the staffed beds simultaneously. Moreover, rescheduling is avoided. While rescheduling is considered by Addis et al. (2016), Kamran et al. (2019) and Spratt and Kozan (2021), these studies neglect the LOS in downstream resources. Additionally, these studies employ different human-friendly gap-reserving strategies to accommodate future arrivals, whereas Min and Yih (2010) and Jebali and Diabat (2017) optimise the gap size of each block.

Despite the extensive literature on the subject, there is still a severe lack of real-life implementations of the proposed models and methods at hospitals (Harris and Claudio, 2022). It is evident that reducing the combined risk of last-minute cancellations while minimising the risk of rescheduling is required for hospitals and patients. Complex heuristics or approximations are often proposed as the problem size increases, e.g., by accounting for multiple sources of uncertainty and longer planning horizons (van Oostrum et al., 2008; Shylo et al., 2013; Aissaoui et al., 2020; Shehadeh, 2022). It has been reported that tractable data-driven approaches are lacking (Harris and Claudio, 2022; Shehadeh, 2022), and due to the complexity of healthcare problems, such an approach might necessitate an unconventional combination of solution methodologies (Aringhieri et al., 2013).

The literature on the operating room planning and scheduling problem is extensive and highlights that uncertainty in surgery times, LOS and arrivals cause most disruptions and must be accounted for when scheduling elective patients in advance. Uncertainty in surgery times has received central attention among researchers. However, integrating two or more sources of uncertainty has only been done in a handful of studies, although it would significantly reduce the risk of last-minute cancellations and rescheduling in advance. Moreover, an analysis of the performance of different gap-reserving strategies for semi-acute elective patients is needed to understand their impact on rescheduling when other elective patients have been scheduled to blocks in advance. Lastly, practical and statistically accurate approaches remain lacking.

## 1.2 Research Objectives

This research aims to increase our understanding of how to schedule a high throughput of elective patients under a limited amount of resources while minimising the risk of last-minute cancellations and rescheduling in advance. The objective is to develop mathematical models that address the advance scheduling of elective patients in a practical manner. Practical models must consider multiple sources of uncertainty to minimise the combined risk of last-minute cancellations and disruptive rescheduling in advance. Likewise, the models must enable high throughput of patients while accounting for limited downstream resources and unpredictable arrivals of semi-acute elective patients. However, taking the multiple sources of uncertainty into account in a mathematical model is generally computationally intractable when solved in a statistically accurate way, e.g. with simulation using the hospital's historical data, which is impractical. As a result, approximations or other complicated heuristics may be used instead. Given these challenges, the following three modelling objectives are specified:

1. **Uncertainty in Surgery Times.** The objective is to assess how the uncertainty in surgery times may be specified in a practical and statistically accurate way while taking into account the limited number of staffed ward beds and the high throughput requirement of elective patients.
2. **Uncertainty in Length of Stay**. The objective is to determine how last-minute cancellations due to uncertainty in surgery times and LOS in the ward can be minimised when elective patients are scheduled to blocks up to weeks in advance while maintaining existing utilisation.
3. **Uncertainty in Semi-Acute Elective Arrivals.** The objective is to compare different gap-reserving strategies for accommodating semi-acute elective patients into the long-term schedule of other elective patients, but without resorting to excessive overtime. This comparison aims to understand how these strategies impact the need for rescheduling, while reducing the risk of last-minute cancellations, due to uncertainty in surgery times and LOS in the ward, and maintaining existing utilisation.

## 1.3 Contributions

Despite the extensive literature, only a handful of papers take into account multiple sources of uncertainty simultaneously in the advance scheduling of elective patients. It is known that uncertainty in surgery times and uncertainty in LOS may cause highly undesirable last-minute cancellations. Moreover, the arrival of semi-acute elective patients may cause multiple unfortunate rescheduling events if not planned for. At the same time, a widespread lack of human resources in healthcare, which limits the

capacity of each hospital, together with the ongoing demographic changes, continues to pressure hospitals to continuously utilise their resources at maximum capacity.

This thesis aims to address these research gaps by developing models considering multiple sources of uncertainty for the advance scheduling of elective patients. In addition, it explores how hospitals can proactively minimise the risk of last-minute cancellations and rescheduling when operating under a high throughput of elective patients and limited resources. Accounting for multiple sources of uncertainty in a statistically accurate way using simulations is computationally intensive which is impractical. Therefore, another aim is to propose a data-driven and practical approach that uses simulations to solve the problem.

All contributions in this thesis have been published as journal articles, conference proceedings, conference presentations and posters and can be summarised as follows:

1. **Uncertainty in Surgery Times**

   A novel two-step, data-driven, practical approach is proposed, termed *Pattern Scheduling*, for the advance scheduling of elective patients to blocks, proactively minimising the risk of last-minute cancellations due to uncertainty in surgery times. The approach resolves uncertainty in surgery times using Monte-Carlo sampling with historical data in a computationally tractable way. The novelty of the approach stems from generating all feasible patterns, which take place in the first step. A pattern is a pre-generated unordered combination of one or more patients that can be assigned to a single block. Pattern feasibility is determined by practical rules set by the hospital, such as limiting patterns to one ICU patient, and probabilistic restrictions on overtime, such as permitting only patterns with less than a 30% chance of exceeding block capacity. Infeasible patterns, such as those violating the practical rules or restrictions, are eliminated. In the second step, a mixed integer programming (MIP) model is used to assign feasible patterns to blocks given some objective while taking into account ward restrictions. Pattern Scheduling is practical in two ways. First, it is computationally practical since the search space is reduced and does not require using approximate models. In other words, the uncertainty of each pattern is verified using Monte Carlo sampling with historical data, so it is statistically accurate. Patterns with a high chance of overtime are eliminated from the search space. Finally, it is clinically practical, as the generated patterns follow the practical rules of the speciality, further limiting the search space (Sigurpalsson et al., 2018, 2019a; Runarsson and Sigurpalsson, 2019a,b; Sigurpalsson et al., 2019b,c, 2020).

2. **Uncertainty in Length of Stay**

   Chance constraints are employed to limit the risk of exceeding block capacity and the availability of staffed ward beds in the advance scheduling of

elective patients, thereby minimising the risk of last-minute cancellations that can occur on the day of surgery (Sigurpalsson et al., 2019b, 2020, 2022).

3. **Uncertainty in Semi-Acute Elective Arrivals**

   Three gap-reserving strategies for reserving gaps into the long-term schedule of other elective patients to accommodate future semi-acute elective arrivals are compared with two strategies from the literature and one proposed in this thesis. The scheduling and rescheduling models used for this comparison take into account practical rules, downstream availability, equipment availability, and probabilistic restrictions on overtime while reducing the risk of last-minute cancellations in advance (Sigurpalsson et al., 2021, 2025).

## 1.4 Outline

This thesis is structured into the following chapters:

**Chapter 2: Literature Review:** A general overview of the current state of the literature on OR planning and scheduling at the operational level with a focus on uncertainty in surgery times, uncertainty in downstream LOS and new arrivals of patients is provided in this chapter. The chapter concludes with a summary of the main research gaps.

**Chapter 3: Methodology:** This chapter describes the methodology used to develop the models in this thesis along with a brief description of the data collection and analysis. This is followed by a section providing the case study on which the mathematical models are based upon. Finally, the chapter concludes with the experimental setup used for the computational experiments conducted in this research.

**Chapter 4: Uncertainty in Surgery Times:** This chapter introduces a novel approach, termed Pattern Scheduling, which addresses uncertainty in surgery times in a practically and statistically accurate way while addressing the limited amount of staffed ward beds in a probabilistic manner. Computational experiments explore the approach's capability and tractability using various instances generated from Landspitali's historical data. This chapter is based on Sigurpalsson et al. (2020) and Runarsson and Sigurpalsson (2019b).

**Chapter 5: Uncertainty in Length of Stay:** A novel way to minimise the combined risk of last-minute cancellations due to uncertainty in surgery times and LOS in the

downstream wards using the pattern scheduling approach is provided in this chapter. Ward combinations are proposed allowing one bounding the risk of exceeding the staffed ward beds in a statistically accurate way in a MIP model. Experiments are conducted using the pattern scheduling approach with the ward combinations, and a comparison is made with the state-of-the-art robust optimisation and actual data. This chapter is based on Sigurpalsson et al. (2022).

**Chapter 6: Uncertainty in Semi-Acute Elective Arrivals:** A comparison is made between three different gap-reserving strategies, with two from the literature and one proposed in this chapter for accommodating semi-acute elective arrivals in several ways while taking into account restrictions on overtime, staffed ward and ICU beds and equipment availability. Scheduling and rescheduling models are proposed where the different strategies are implemented. This chapter is based on Sigurpalsson et al. (2022, 2025).

**Chapter 7: Discussion and Conclusions:** A summary of the findings, heuristics, and avenues for future research are provided in this chapter.

# 2 Literature Review

This chapter provides an overview of the literature on advance scheduling of elective patients, which takes place at the operational level. However, for the reader, this chapter begins with a brief description of the different hierarchical levels used to classify the decisions of the OR planning and scheduling problem based on a block scheduling approach. This is followed by a literature review of the main sources of uncertainty inherent to the scheduling process at the operational level, which frequently lead to last-minute cancellations and disruptive rescheduling events. These sources must be considered when scheduling patients up to weeks in advance to hedge against the number of such occurrences. Finally, the chapter concludes with a summary identifying the main research gaps.

## 2.1 Scheduling Levels

Planning and scheduling occur at three hierarchical levels under the block scheduling approach, where the output of one level serves as input to the subsequent level (Guerriero and Guido, 2011; Zhu et al., 2019; Akbarzadeh and Maenhout, 2024). Strategic level is at the top of the hierarchy, where long-term decisions, usually years in advance, are made about the overall needs for block capacity and how it is allocated to different surgical specialities (Blake and Donald, 2002). The next level is the tactical level, where a cyclic timetable, or MSS, is created based on the total block time allocated to different specialities (Guerriero and Guido, 2011). As such, each speciality is allocated a certain number of blocks in the MSS. The blocks may also be allocated to specific surgery teams, to specific surgeons or a mix of surgical procedures (Schneider et al., 2020; Beliën et al., 2008; van Oostrum et al., 2008). Finally, at the operational level patients, awaiting surgery, are assigned to blocks (Cappanera et al., 2018; Denton et al., 2007), and their sequence within the block is determined (Kroer et al., 2018). The assignment may be done weeks in advance but needs to be updated to accommodate new patient arrivals, both emergency (Van Riet and Demeulemeester, 2015) and semi-acute (Zonderland et al., 2010). The following paragraphs describe the planning and scheduling of each of these levels. Open scheduling approach, which is not considered in this thesis, focuses exclusively on the operational level where blocks are made available to all specialities (Akbarzadeh and Maenhout, 2024).

### Strategic Level

The decisions made at the strategic level are very important for every hospital, as they constrain the flexibility of the subsequent levels (Akbarzadeh and Maenhout, 2024). Generally, the task at this level is to define the hospital's mission and translate it into actions (Hans et al., 2012). Therefore, decisions made at this level are long-term and made several years in advance.

When hospitals define their mission, they must also specify the resources needed to satisfy the demand years in advance on a highly aggregated level (Hans et al., 2012; Marques et al., 2012; Zhu et al., 2019). This is known as capacity planning (Jacobs et al., 2018). Inadequate capacity results in longer waiting times for patients (Choi and Wilhelm, 2014), and consequently, it must be carefully planned. Other decisions made at this level include upgrading and constructing new facilities (May et al., 2011), which expand the ORs' capacity (Lovejoy and Li, 2002). Finally, decisions on buying new equipment that can modify the hospital's ability (Wachtel and Dexter, 2008), are also made at this level.

Once the capacity planning has been performed, hospitals can start translating their long-term mission into medium-term actions, e.g. up to a year in advance (Blake and Carter, 2002) and is known as capacity allocation. At this level, total block hours are allocated to surgical specialities (Guerriero and Guido, 2011), surgeon groups (Santibáñez et al., 2007; Schneider et al., 2020) or surgery groups (Dexter et al., 2002; Marques et al., 2012; Choi and Wilhelm, 2014; Yahia et al., 2016). The allocation is often performed in conjunction with defining the annual hospital budget (Blake and Carter, 2002), which sets the total number of block hours for the year from the overall capacity.

There are several ways possible to allocate the block hours to the various surgical specialities but commonly it is based on their historical utilisation of the block hours. However, as pointed out by Dexter et al. (2002), any changes made to the allocation, e.g. by changing the allocation of block hours among the surgeons while maintaining the same total hours, can lead to increased costs for the hospital if based on historical utilisation. Consequently, they suggest incorporating the hospital's beds and implants in the allocation process to lower the cost for the hospital. Finally, decisions on offering extended block capacity to maximise the throughput of elective patients may also be taken at this level. Following the COVID-19 pandemic, this became particularly important due to the rapid growth of the already long waiting lists (Oliveira et al., 2023).

### Tactical Level

Once long-term strategic decisions have been made, a cyclic timetable known as the MSS is determined. The MSS commonly repeats every week or two but is fixed months in advance. The available blocks are allocated to surgical specialities, teams,

or individual surgeons at this level based on the total number of block hours allocated at the prior level (Blake and Donald, 2002; Guerriero and Guido, 2011). Nevertheless, MSS may also specify the mix of operations performed in each block (Adan et al., 2009). If so, the MSS can estimate the aggregated resource requirement (Blake and Donald, 2002), e.g., ward and ICU occupancy (van Oostrum et al., 2008; Schneider et al., 2020). An example of how the MSS for General Surgery at Landspitali is shown in Chapter 3 in Table 3.3.

A new MSS is established whenever a change occurs at the strategic level (Marques et al., 2019). For instance, if the number of staffed block hours increases or decreases due to changes in the hospital's budgeting, the MSS must be changed accordingly. Moreover, seasonal fluctuations in demand (Dexter et al., 1999a) and staff availability (e.g., due to holidays) may also require a new MSS to be created (Blake and Donald, 2002). Such regular adjustments are essential to give the surgeons enough time to complete their elective cases in response to changing demands (Dexter et al., 1999a).

### Operational Level

The final level is the operational level, where the scheduling of actual patients takes place. At this level, elective patients are assigned to appropriate blocks, such as those of their surgeon or speciality, and their starting time within the block is determined (Zhu et al., 2019; Otten et al., 2019). The time horizon of this level typically ranges between one to several weeks in advance. Additionally, decisions are made in response to unanticipated events and the process is continuously monitored (Hans et al., 2012; May et al., 2011; Stuart and Kozan, 2012).

According to the framework in Hans et al. (2012), decisions at this level can be classified into *Offline* and *Online* decisions:

- *Offline* decisions are made in advance and are further decomposed into:
  - *Advance Scheduling* refers to assigning actual elective patients awaiting surgery to pre-allocated blocks in the MSS (Samudra et al., 2016). Patients are assigned to these blocks based on their medical priority, which determines the maximum waiting time on the waiting list (Marques and Captivo, 2017). Other objectives include minimising the time patients spend on the waiting list (Adams et al., 2023) and maintaining an appropriate balance between underutilisation and overutilisation of the blocks (McIntosh et al., 2006). Finally, operational constrains, such as the availability of staffed ward and ICU beds, are also taken into account (Min and Yih, 2010).
  - *Allocation Scheduling* determines the starting time of the patients in each block (Samudra et al., 2016). However, some hospitals may solve the advance scheduling and allocation scheduling simultaneously (Kroer et al., 2018).

- *Online* decisions are made in response to unforeseen events (Hans et al., 2012) and also involve monitoring the elective schedule on a continuous basis (Stuart and Kozan, 2012). This category addresses various topics, including rescheduling due to emergency arrivals (Addis et al., 2016; Kamran et al., 2019) to minimise the waiting time for emergency surgeries (Parmar et al., 2022), managing semi-acute elective arrivals (Zonderland et al., 2010), responding to last-minute cancellations caused by the actual realisation of surgery times (Kamran et al., 2019). Additionally, cases may be moved between blocks to prevent last-minute cancellations when a block's schedule is running late (Zhou and Dexter, 1998).

The following section, presents a literature review on the pathways of elective patients and the main sources of uncertainty at the operational level. Particular attention will be given to how uncertainty is taken into account in mathematical models to reduce the risk of last-minute cancellations and rescheduling events, which are the main focus of this thesis. For the vast literature and the recent advances can be found in the comprehensive reviews by Cardoen et al. (2010); Guerriero and Guido (2011); Van Riet and Demeulemeester (2015); Samudra et al. (2016); Soh et al. (2017); Zhu et al. (2019); Wang et al. (2021); Bellini et al. (2024).

## 2.2 Patient Pathways and Uncertainty at the Operational Level

Elective patients are typically selected from the waiting list and assigned to pre-allocated blocks in the MSS based on their medical priority and waiting time (Marques and Captivo, 2017). However, the selection also depends on the availabilities of the hospital's resources to avoid last-minute cancellations and rescheduling events. In general, patients experience anxiety and disappointment as a response to last-minute cancellations or other disruptions (Ivarsson et al., 2004; Eshghali et al., 2024). Consequently, such occurrences should be avoided at all costs. Moreover, these events have been shown to contribute to more costs for patients due to disruptions to their daily lives, lost working days (Pattnaik et al., 2022) and accommodation expenses for those patients residing far from the hospital (Armoeyan et al., 2021). Additionally, they impact the hospital, through lower resource utilisation, and emotionally affect its staff (Armoeyan et al., 2021).

Figure 2.1 illustrates the most frequent pathways a patient follows after surgery (OR) and, consequently, the resources the patient may occupy. The figure shows that a proportion of patients are discharged from the hospital on the same day as their surgery (outpatients). In contrast, other patients (inpatients) are admitted to downstream resources (here, ICU and ward) for one or more days before being discharged. More complicated pathways are still possible than shown in the figure (Fügener et al., 2014).

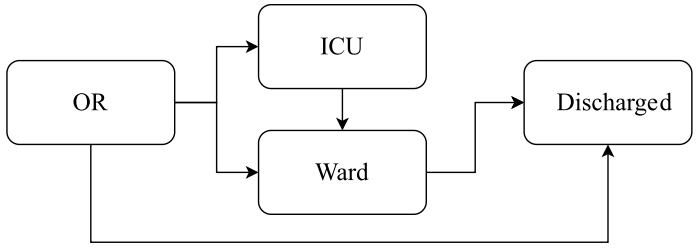



*Figure 2.1. Common pathways of patients after surgery.*

See Fügener et al. (2014) for more complicated flows.

What poses significant challenges when scheduling elective surgeries are the time uncertainties related to each resource as shown in Figure 2.1, and other sources of uncertainties, such as staff absence, which is, however, less predictable (Van Riet and Demeulemeester, 2015; Samudra et al., 2016). Despite the number of different sources of uncertainties, three frequently cause the most disruptions (Min and Yih, 2010; Van Riet and Demeulemeester, 2015; Riise et al., 2016; Zhang et al., 2019). First, surgery times and the time between surgeries have high variability and can lead to overutilisation of blocks (Batun et al., 2011) and potentially result in last-minute cancellations (Hans et al., 2008). Second, the LOS in downstream resources may lead to the unavailability of staffed beds (Augusto et al., 2010), which may also cause last-minute cancellations (Min and Yih, 2010; Neyshabouri and Berg, 2017). Finally, unpredictable arrivals of patients, e.g. emergency or semi-acute elective patients, may cause the rescheduling of previously planned patients (Zonderland et al., 2010; Van Riet and Demeulemeester, 2015; Addis et al., 2016; Kamran et al., 2019; Spratt and Kozan, 2021) unless planned for. The following sections review how different sources of uncertainties can be accounted for in mathematical models to minimise last-minute cancellations and rescheduling events.

### 2.2.1 Uncertainty in Surgery Times

A single surgery consists of three time intervals: pre-incision, incision, and post-incision (Batun et al., 2011; Marques and Captivo, 2017). During the pre-incision interval, patients are positioned on beds in the ORs, and anaesthesia starts (Batun et al., 2011). Next, the incision takes place. Finally, the post-incision phase includes closing the incision (Batun et al., 2011) and preparing the patient for transfer to downstream facilities for recovery (Augusto et al., 2010). After the patient leaves the OR, the cleaning starts (Dexter et al., 2003; Marques and Captivo, 2017; Kroer et al., 2018). Despite the different time intervals, the total surgery time in this thesis is defined as the interval from the start of the pre-incision to the end of the post-incision phase, with cleaning time included.

Predicting surgery times in advance is a complex task (Macario, 2009; Wang et al., 2021; Shehadeh, 2022). Generally, the time required for a single operation is usually dependent on the type of operation (Zhou and Dexter, 1998; Dexter et al., 2008; Min and Yih, 2010; Marques and Captivo, 2017; Pandit, 2020) and the surgeon performing it (Zhou and Dexter, 1998; Strum et al., 2000; Dexter et al., 2008; Marques and Captivo, 2017). However, other factors such as the surgeon's experience (Opit et al., 1991), the type of anaesthetic (Dexter et al., 2008), and the patient's history of illness (Gür et al., 2024) may also contribute to the the variability.

Hospitals frequently assign patients to blocks manually using the mean time, typically based on the type of operation and the surgeon (Opit et al., 1991; Zhou and Dexter, 1998) to predict how long a single operation will take (Dexter et al., 1999c; Batun et al., 2011; Pandit, 2020). In general, these mean values are computed at regular intervals so that the most recent data is used (Dexter and Macario, 1996; Zhou et al., 1999; Macario, 2009) to reflect changes due to seasonal variability (Dexter, 1996) or new surgical techniques (Zhou et al., 1999).

Scheduling by the mean has long been criticised as surgery times have high variability and consequently, uncertainty is neglected (Pandit, 2018, 2020; Gür et al., 2023). In other words, if four cases of a single operation that take, on average, two hours to complete are assigned into a single block of eight hours, it is unlikely that it will take exactly eight hours to complete the cases (Pandit, 2018). Several studies have shown that surgery times follow right-tailed distributions like the log-normal distribution (Zhou and Dexter, 1998; May et al., 2000). Therefore, it raises the risk of underutilising the block as the mean is larger than the median (Kroer et al., 2018). At the same time, the surgery time may also surpass the mean, which will, in turn, overutilise the block (Denton and Gupta, 2003; Batun et al., 2011). In other words, scheduling a three-hour case into a four-hour block leaves the block under-utilised for an hour. Overutilisation is the scheduled time that exceeds the block capacity. Overutilising a block may result in last-minute cancellations (Hans et al., 2008; Min and Yih, 2010) due to overtime restrictions (Wang et al., 2014; Adams, 2019) as staff may not be available after hours. On the other hand, underutilisation does not result in last-minute cancellations but may cause increased waiting time for patients. Therefore, finding an appropriate balance between under- and overutilising each block is important (McIntosh et al., 2006; Pandit and Dexter, 2009), while the risk of cancellations is minimised (Pandit et al., 2007).

The academic literature addressing uncertainty in surgery times within mathematical models is dense. Assigning patients to pre-allocated blocks can be considered as a bin packing problem (Hans et al., 2008) and by assuming that time distributions for surgery times (Kroer et al., 2018) or limited data (Marques and Captivo, 2017) are available for any given type of operation, several different optimisation approaches can be applied to account for uncertainty.

Two-stage Stochastic Programming (SP) is frequently used to take uncertainty into account (Kamran et al., 2018; Shehadeh, 2022; Gür et al., 2023). The objective is typically to minimise the expected value, such as overtime costs (Kroer et al., 2018;

Jebali and Diabat, 2015; Molina-Pariente et al., 2018; Min and Yih, 2010), resulting in risk-neutral solutions (Najjarbashi and Lim, 2019) across the planning horizon. In other words, suppose that three cases are assigned to the same block of a certain capacity, each with a randomly generated distribution of surgery times from historical data. Using the two-stage SP, the optimal solution is a schedule with the least expected overtime cost. This cost is calculated by randomly sampling surgery times across thousands of scenarios, measuring the minutes that exceed the block's capacity in each scenario, averaging the minutes in overtime of all scenarios, and multiplying it with a cost associated with overtime. However, the variance of the time completion of the last case is generally high (Hans et al., 2008). Therefore, the risk of exceeding the block's capacity may be high, and thus, last-minute cancellations are possible with a high probability. Thus, other ways are required to prevent last-minute cancellations in the advance scheduling of elective patients.

One approach to minimise the risk of last-minute cancellations due to overtime restrictions is through the use of chance constraints. As such, the block capacity is satisfied for a predefined fraction of scenarios of surgeries within a block, which is often the desired outcome in practice (Kamran et al., 2018). Referring to the same scheduling problem discussed in the previous paragraph, it is possible to estimate the likelihood of exceeding the block's capacity. This probability can be calculated by counting how many scenarios have total surgery time greater than the block's capacity and dividing that number by the total number of scenarios. If the likelihood is larger than a predefined threshold, selected by the hospital, the combination of the three surgeries can not be assigned to same block due to a risk of last-minute cancellation. Formally, if we let $z_i \in \{0,1\}$ be a binary decision variable indicating if the patient $i$ has surgery on the day $d \in D$ in the room $r \in R$ with capacity $C_{r,d}$ then the probability exceeding the capacity can be bounded by the following chance constraint:

$$Pr[f(z_1,\ldots,z_n) \geq C_{r,d}] \leq \delta \tag{1}$$

where $f(z_1,\ldots,z_n)$ denotes the distribution function for the stochastic sum of all surgeries time, including the time between surgeries and $\delta$ is the maximum allowed probability of exceeding the block's capacity and defined on the interval $0 \leq \delta \leq 1$. For example, if the hospital selects $\delta = 0.25$, it accepts the risk that one out of every four blocks exceeds its capacity. This concept is referred to as **regular overtime** in this thesis.

One of the earliest attempts to implement constraint (1) in an OR planning and scheduling model involved planning a slack to each block to hedge against last-minute cancellations due to uncertainty in the total surgery time. By planning a slack, the risk of overtime was minimised, no surgeries were cancelled, and the block utilisation was improved (Hans et al., 2008). Following that publication, several studies have included the concept of planned slack in their scheduling models (van Oostrum et al., 2008; Schneider et al., 2020; van den Broek d'Obrenan et al., 2020).

An important assumption when utilising the planned slack is that the sum of the surgery times assigned to a block is normally distributed making it possible to use a deterministic

mathematical scheduling model without requiring simulation. Consequently, cases are assigned to blocks by their historical mean values. However, the total surgery time of the block is calculated as the sum of the individual mean values ($\mu$) and a planned slack where the size is determined by the variance of the surgeries in the block ($\sigma^2$). By standardising the normal distribution, it is possible to calculate a Z-score for the schedule, which can be used to determine the probability of exceeding a block capacity (Pandit and Tavare, 2011; Proudlove et al., 2013).

In optimisation models, the Z-score must be selected in advance. This ensures that all blocks in the optimised solution will have a probability lower than $\delta$ of exceeding their block capacity. Nevertheless, Shylo et al. (2013) pointed out that assuming that the sum of surgery times in a block is normally distributed is only valid for a high volume of patients scheduled in a block. Furthermore, Soh et al. (2024) highlighted that chance constraints may fail to capture the severity of the overtime when utilised as a single measure.

Extending previous studies Shylo et al. (2013) implemented a chance constraint bounding the risk of exceeding the maximum allowed overtime (extended block capacity) in a block:

$$Pr[f(z_1,\ldots,z_n) - C_{r,d} > L] \leq \Delta \tag{2}$$

where $L$ is the maximum allowed overtime, and $\Delta$ the maximum allowed probability of exceeding $L$ defined at the interval $0 \leq \Delta \leq 1$. As the chance constraint guarantees a low risk of last-minute cancellations in each block, the author's objective is to minimise the expected undertime to maximise the utilisation of the blocks. The concept of using extended block capacity is also implemented in Pandit and Tavare (2011); Wang et al. (2014); Landa et al. (2016); Kamran et al. (2018) and is referred to as **extended overtime** in this thesis.

Despite the effectiveness of using chance constraints to hedge against last-minute cancellations, they are only implemented in a handful of studies (Kamran et al., 2018). This may be due to the computational issues that arise when utilising these chance constraints, often requiring tailored heuristics or approximations (van Oostrum et al., 2008; Shylo et al., 2013; Shehadeh, 2022).

An alternative to two-stage SP is Robust Optimisation (RO) which addresses uncertainty in surgery times while minimising the risk of exceeding the capacity of each block. RO is generally suitable when data is limited, as the worst-case value is only needed for each scheduled surgery (Addis et al., 2014, 2016; Marques and Captivo, 2017; Wang et al., 2019). This is unlike SP, where it is assumed that whole time distributions are available (Shehadeh and Padman, 2021).

When using RO, uncertainty sets are generated, reflecting uncertainty in surgery times and the optimisation is based on the worst-case outcome. However, it is possible to relax the conservatism of RO by using budget of uncertainty. In that case, the decision maker must make a trade-off between conservatism and the protection level by selecting how many surgeries assigned to each block are assigned their worst-case value while

others are assigned their average values (Bertsimas and Sim, 2004). In general, RO is known for conservative solutions. It neglects distributional characteristics and, thus, cannot capture frequency information of the parameters (Wang et al., 2019). Moreover, it may leave the blocks and other resources underutilised.

Distributionally Robust Optimisation (DRO) has been developed recently to overcome the barriers of RO (Wang et al., 2019; Shehadeh and Padman, 2021; Shehadeh, 2022). Unlike RO, DRO uses distributional information by assuming that time distributions are partially known and that uncertainty resides in ambiguity sets generated based on several descriptive statistics of uncertain parameters (Wang et al., 2019). The optimisation, in this case, selects the worst-case probability distribution within the ambiguity set (Wang et al., 2019; Shehadeh and Padman, 2021). Nonetheless, as noted by Wang et al. (2016), selecting the worst-case distribution may be an unrealistic assumption as it tends to protect against very conservative scenarios while performing poorly in more likely ones.

### 2.2.2 Uncertainty in Length of Stay

Following surgery, a patient is either discharged or admitted to downstream resources for recovery for one or more days, as shown in Figure 2.1. As discussed in the previous section, the patient's LOS in the downstream resources is highly stochastic (Shehadeh and Padman, 2021), as is true for surgery times. LOS is often based on the operation type (Min and Yih, 2010; Schneider et al., 2020) although the patient's characteristics, such as complications (Cohen et al., 2009), may also impact the variability.

Uncertainty in LOS has been attributed to a lack of downstream resources that may cause last-minute cancellations (Jebali and Diabat, 2015), early discharge (Jonnalagadda et al., 2005), or transfers of patients (Neyshabouri and Berg, 2017; Zhang et al., 2019; Shehadeh and Padman, 2021). When a patient gets cancelled at the last minute due to a shortage of downstream bed capacity, it not only delays their treatment but may also cause less block utilisation that day (Wang et al., 2022). Increasing the downstream bed capacity might sound tempting to prevent last-minute cancellations in such occurrences. However, that will increase the number of patients to each nurse which has been associated with worse patient outcomes (Kane et al., 2007; Dall'Ora et al., 2022). In the same vein, van Oostrum et al. (2008); Schneider et al. (2020) point out that such peaks in demand have been shown to cause last-minute cancellations. Moreover, such peaks in occupancy may also cause early discharge of other patients, which has been associated with readmissions, so again, bottlenecks (Jonnalagadda et al., 2005; Neyshabouri and Berg, 2017).

While many studies account for uncertainty in surgery times, as discussed in Section 2.2.1, only a handful account for uncertainty in LOS in downstream resources simultaneously (Shehadeh, 2022; Eshghali et al., 2024). However, by doing so, the combined risk of last-minute cancellations caused by overutilisation of blocks and downstream bed capacity, would be significantly reduced, resulting in more robust

schedules (Min and Yih, 2010). One reason downstream resources may be less considered is that hospitals may use an overflow allowing patients to be admitted to a ward outside their speciality with available capacity to avoid last-minute cancellations caused by downstream bottlenecks (Wang et al., 2022). However, such strategies may first, increase the LOS (Izady and Mohamed, 2021) and, second, raise the risk of readmission (Stowell et al., 2013), which is costly for hospitals.

Min and Yih (2010) provides one of the earliest studies to account for uncertainty in surgery times and LOS simultaneously. The problem is formulated using a two-stage SP model with the objective of minimising expected overtime cost and patient-related costs. In their model, ICU beds are considered as a fixed number each day. An equivalent deterministic MIP model is proposed, ignoring all uncertainty. Results show significant improvements in average block utilisation and a reduction in last-minute cancellations for the same level of throughput compared to the deterministic model, suggesting the importance of considering uncertainty. Later, Jebali and Diabat (2015) extended the work of Min and Yih (2010) by incorporating uncertainty in LOS, both in the ward and ICU, as well as uncertainty in surgery times, into a two-stage SP model aimed at minimising patient costs, and under- and over-time costs. The model was reformulated as an equivalent deterministic MIP model and solved using the Sample Average Approximation (SAA) algorithm. The result showed that taking ward and ICU beds into account when scheduling elective patients will result in fewer last-minute cancellations.

Jebali and Diabat (2017) extended the study of Min and Yih (2010) by implementing a chance constraint of exceeding the limited ICU beds to a two-stage SP model. Consequently, the model bounds the daily probability of exceeding the ICU capacity to a predefined threshold, thereby reducing the risk of last-minute cancellations due to downstream bottlenecks. This is unlike the studies of Min and Yih (2010); Jebali and Diabat (2015) where the capacity had to be tailored for all possible scenarios for the LOS of each patient. Jebali and Diabat (2017) employed the SAA to solve the problem, minimising patient-related costs, expected block utilisation, and penalty cost for exceeding ICU capacity. The results indicate increased robustness of the solutions but come with higher costs and lower block utilisation. Similar problems are studied by Zhang et al. (2019) and Zhang et al. (2020).

Recently RO (Neyshabouri and Berg, 2017; Makboul et al., 2021) and DRO (Shehadeh and Padman, 2021) have been used to solve the problem. Neyshabouri and Berg (2017) propose a two-stage RO model taking into account uncertainty in LOS in ICU and surgery time, assuming the worst-case to minimise patient-related costs and overtime costs. Makboul et al. (2021) uses RO but assumes the worst-case scenario for bed availability each day rather than considering uncertainty in LOS of each patient. Recently, Shehadeh and Padman (2021) formulated the problem using DRO to minimise the cost of postponing surgeries while minimising the worst-case expected overtime cost, idle time cost, and cost of exceeding the ICU capacity. Similar to Neyshabouri and Berg (2017), an adapted column-and-constraint generation method is proposed to solve the problem.

Another approach to account for uncertainty in LOS is to compute the probabilities, based on historical data, that a patient will be in the ward or ICU on a given day following surgery (Adan et al., 2009; van den Broek d'Obrenan et al., 2020), as illustrated in Example 2.1. This approach enables the use of deterministic MIP models while accounting for uncertainty in downstream resources. However, considering uncertainty in LOS in this way can result in last-minute cancellations when operating at the maximum bed capacity. Namely, if we suppose the ward occupancy each day is the sum of many independent and identically distributed Bernoulli random variables, then the sum can be approximated by a normal distribution. For instance, if there are three beds and the expected bed occupancy is three, there can be up to 50% chance of exceeding the beds that day (van den Broek d'Obrenan et al., 2020). Consequently, there is a high risk of last-minute cancellations. As a result, variation in bed occupancy are commonly minimised in an attempt to avoid exceeding the given capacity (Beliën and Demeulemeester, 2007; Fügener et al., 2014; van den Broek d'Obrenan et al., 2020; Schneider et al., 2020).

**Example 2.1** *Suppose that historical data has been collected on the LOS in the ward for a specific type of operation from four patients. The data shows that two patients stayed for one day and two for three days. Using this information, it is possible to compute the ward bed occupancy for a future patient undergoing the same type of operation. Namely, the probability of staying one day is 100% as all four patients stayed at least one day, while the probability of staying 2 or 3 days is 50% , respectively, as two of the four patients stayed at least three days. Note that a patient who stays three days also stays first for two days. Otherwise, the probability is 0% as no patient in the data set has stayed more than three days. Suppose a human scheduler wants to schedule five patients to undergo this surgery next Monday. By summing the individual probabilities of each patient each day, it is possible to estimate the bed occupancy in advance. As a result, there will be 5 beds needed on Monday* $(5 \cdot 100\%)$*, 2.5 beds on Tuesday and Wednesday* $(5 \cdot 50\%)$ *respectively and 0* $(5 \cdot 0\%)$ *for the rest of the week.*

### 2.2.3 Uncertainty in Arrivals

Even when accounting for uncertainty both in surgery times and LOS when assigning elective patients weeks in advance to proactively avoid last-minute cancellations on the day of surgery, there remains a risk of rescheduling previously scheduled patients to accommodate unpredictable new arrivals (Addis et al., 2016). Generally, elective patients can be assigned to blocks in advance, but new arrivals, such as the arrivals of semi-acute elective patients, are unpredictable (Lamiri et al., 2007; Zonderland et al., 2010) and must be planned for.

The literature considering emergency arrivals is extensive, but the literature addressing semi-acute elective arrivals is limited, which is the focus of this thesis. Semi-acute elective patients are fundamentally different from emergency patients as they are elective

patients with a high medical priority and must be accommodated within a week or two into the advance of other elective schedule (Zonderland et al., 2010). Emergency patients, however, must be operated on within a day (Van Riet and Demeulemeester, 2015). Consequently, there is more flexibility in the assignment of the semi-acute elective arrivals to blocks. Furthermore, as pointed out by Epstein and Dexter (2013), there is flexibility in timing between two rescheduled appointments for other elective patients, meaning that rescheduled patients can be assigned to appropriate blocks without increasing variability. Despite the lack of literature addressing semi-acute arrivals, a similar methodology can be used to reserve capacity for these patients (Zonderland et al., 2010).

Van Riet and Demeulemeester (2015) outlined three policies to account for uncertainty in arrivals. First, hospitals can separate the flow by setting aside one or more blocks daily for new arrivals (Antognini et al., 2015; Parmar et al., 2022). In this case, all semi-acute elective patients would be scheduled in dedicated blocks. Second, new arrivals are placed in the existing blocks, combining the flow of previously scheduled elective patients and the new arrivals (Wullink et al., 2007). This means that gaps are reserved in some or all blocks within the long-term schedule of the elective patients. Consequently, when semi-acute patients arrive, they are accommodated into these gaps with the least disruption to other scheduled patients. Finally, hospitals can use a mix of the first two policies (Zonderland et al., 2010; Van Riet and Demeulemeester, 2015). In other words, the hospital sets aside blocks for the semi-acute arrivals and additionally reserves gaps in other blocks within the long-term elective schedule.

Reserving gaps in one or more blocks to accommodate new arrivals is classified as a predictive disruption management policy (Addis et al., 2016; Kamran et al., 2019; Akbarzadeh and Maenhout, 2024). In general, these gaps can be reserved in a deterministic (Van Riet and Demeulemeester, 2015) or stochastic (Lamiri et al., 2008b; Molina-Pariente et al., 2018; Jebali and Diabat, 2017) manner in some or all of the available blocks (Wullink et al., 2007; Westbury et al., 2009) using gap-reserving strategies that humans can apply, or sophisticated approaches, like optimisation.

Lamiri et al. (2008a) proposed a Monte-Carlo optimisation method to assign elective patients under uncertain demand of emergency patients where gaps are reserved assuming stochastic demand. The results demonstrate that patient costs and overtime costs are lower when compared to an equivalent deterministic model assuming average emergency demand. This study is further extended by Lamiri et al. (2008b) and Lamiri et al. (2009). In Lamiri et al. (2009), several heuristics and meta-heuristics are proposed to solve the problem. In Lamiri et al. (2008b), patient assignment to blocks is addressed as the total daily capacity was only considered by Lamiri et al. (2008a). Regardless, an important drawback of both studies is that surgery times are assumed to be deterministic despite their stochastic nature, as discussed in Section 2.2.1.

Some papers consider uncertainty in surgery times and uncertainty in daily emergency demand simultaneously. Lamiri et al. (2007) solves the problem using a Monte-Carlo simulation and Column Generation (C-G) to minimise patients' related costs and ex-

pected overtime costs. Molina-Pariente et al. (2018) considers a similar problem but in contrast the availability of responsible surgeon is considered in their model. The model considers uncertainty in surgery time, the surgeon's capacity, and the arrival of emergency patients. A Monte-Carlo optimisation is proposed to solve the problem with the objective of minimising the expected under- and overtime costs and the cost of exceeding the block capacity. The results indicate that significant cost reductions (e.g., over-time/under-time) are possible by considering the uncertainty compared to a deterministic model. Nonetheless, uncertainty in downstream LOS is neglected in Lamiri et al. (2007); Molina-Pariente et al. (2018), although it would reduce the risk of last-minute cancellations, as discussed in Section 2.2.2.

In practice, hospitals typically apply gap-reserving strategies to reserve gaps for arrivals as they rarely have personnel with optimisation skills (Shehadeh, 2022). As a result, a common approach is to dedicate some fraction, e.g. 85% of the block capacity to the elective program and 15% idle for the unexpected (Jebali et al., 2006; Van Riet and Demeulemeester, 2015). Several papers have proposed gap-reserving strategies that hospitals can implement (Addis et al., 2016; Kamran et al., 2018, 2019; Spratt and Kozan, 2021; Davarian and Behnamian, 2022). Addis et al. (2016) and Davarian and Behnamian (2022) introduced time-dependent bounds on the utilisation of each block in the schedule, gradually increasing the reserved slack towards the end of the planning horizon. A similar strategy is implemented by Kamran et al. (2018). Kamran et al. (2019) restricted the number of surgeries a surgeon can perform daily to reserve slacks while Spratt and Kozan (2021) reserved weekly capacity for each speciality for a specific number of non-elective surgical patients. Dexter et al. (1999b) compared several algorithms for scheduling add-on elective cases into the remaining elective block time at a cut-off time the day before surgery. The goal was to understand which algorithm maximises block utilisation. The result showed that an offline algorithm, where all add-on cases are considered the day before surgery, and assigned from the longest to shortest surgery time to blocks with the best fit, performed the best. Additionally, the algorithm allowed up to 15 minutes of exceeding the regular block capacity in cases where no block had sufficient time for an add-on. However, as pointed out by the authors, the algorithm may not be practical as surgeons and patients will need to wait for the cut-off time to know if or when the surgery takes place.

Rescheduling may become necessary to accommodate new arrivals if the gaps reserved are insufficient (Addis et al., 2016; Akbarzadeh and Maenhout, 2024; Eshghali et al., 2024). Scheduling and rescheduling is considered by Addis et al. (2016); Kamran et al. (2019); Spratt and Kozan (2021); Davarian and Behnamian (2022); Adams et al. (2023); Eshghali et al. (2024). In their work, an optimised long-term schedule is created with gaps reserved to account for possible future arrivals. During the actualisation of the optimised schedule, new arrivals are accommodated to the reserved gaps. If the gaps are insufficient, rescheduling is required. In Kamran et al. (2019); Spratt and Kozan (2021); Eshghali et al. (2024), the schedule is updated daily to account for these changes but every week in Addis et al. (2016); Davarian and Behnamian (2022); Adams et al. (2023). Despite the significance of downstream resources, it is disregarded in some of those studies, increasing the risk of last-minute cancellations.

Downstream resources are considered in Davarian and Behnamian (2022); Eshghali et al. (2024). However, Davarian and Behnamian (2022) neglects time uncertainties, and rescheduling of elective patients is only permitted within the same day. Eshghali et al. (2024) considers uncertainty in surgery times and uncertainty LOS in the PACU but neglects other downstream resources such as the ward and ICU. In addition, both of these papers focus on emergency patient arrivals.

## 2.3 Summary

A large body of literature exists on the OR planning and scheduling problem. The literature highlights that the problem is highly stochastic and uncertainty cannot be overlooked in mathematical models. Moreover, it suggests that uncertainty in surgery times, LOS in downstream resources, and the arrival of new patients cause the most disruptions (Min and Yih, 2010; Van Riet and Demeulemeester, 2015; Riise et al., 2016; Zhang et al., 2019). Uncertainty in surgery times has received central attention among researchers (Gür et al., 2023). However, most studies ignore the combined risk of two or more sources of uncertainty (Eshghali et al., 2024). The following research gaps have been identified:

- Chance constraints are barely implemented despite their effectiveness in hedging against last-minute cancellations
  - A handful of papers consider bounding the risk of exceeding the block capacity, but only the work of Jebali and Diabat (2017) considers bounding the risk of exceeding the downstream ICU beds. Regardless, Jebali and Diabat (2017) ignores bounding the risk of exceeding the block capacity. To best of the author's knowledge, no study has considered this combined setting at the operational level.
- Two-stage SP remains the most popular optimisation technique for addressing uncertainty. However, as the size of the problem increases, particularly with the implementation of chance constraints, tailored heuristics or approximations are often proposed to solve the problem (van Oostrum et al., 2008; Shylo et al., 2013; Aissaoui et al., 2020; Shehadeh, 2022; Gür et al., 2023)
- Practical and statistically accurate optimisation approaches are lacking (Harris and Claudio, 2022; Shehadeh, 2022). They may require an unconventional combination of solution methodologies (Aringhieri et al., 2013).
- Uncertainty in LOS in the ward is often ignored, and the main focus has been on the ICU.
- A comparison of gap-reserving strategies used to reserve gaps for unpredictable semi-acute elective arrivals is needed to understand their impact on the need of

rescheduling when operating under a limited amount of downstream resources and a high throughput of elective patients.

The subsequent chapter outlines the methodology used to develop the models in this work. General Surgery, one of Landspitali's surgical specialities, receives a special attention in the model development of this thesis. The speciality was selected due its scheduling complexities which have frequently led to last-minute cancellations and rescheduling events in the past.

# 3 Methodology

The methodology used in this thesis is outlined in this chapter. The chapter begins by describing the model development process. This is followed by a brief presentation of the data set collected for the study from Landspitali Hospital and analysis of it. The problem description that follows provides a detailed overview of the scheduling practises and challenges at the General Surgery (GS) speciality in detail, which the models of this thesis are based on. This chapter concludes with a brief description of the experimental setup used for the computational experiments performed in subsequent chapters.

## 3.1 Model Development Process

The practical scheduling needs of Landspitali are the driving force behind the models developed in this thesis. The hospital, which is located in Reykjavik, is the largest healthcare provider in Iceland (Ministry of Health, 2021) and operates approximately 650 ward beds. In cooperation with Landspitali, one of its surgical specialities, GS speciality, was selected for this study. First, the GS speciality is one of the most complex specialities to schedule due to the nature of operations performed. Additionally, the speciality has a long list of patients awaiting surgery but a limited amount of resources. Consequently, the speciality maintains a high utilisation of its resources. Running close to full capacities each day, however, is challenging due to multiple sources of uncertainty inherent to the process. These different sources of uncertainty have caused last-minute cancellations and disruptive rescheduling in the past making it hard to build schedules weeks in advance. Finally, by selecting a single speciality for the study, the mathematical models or approaches will be realistic and tailored to the speciality's needs. Therefore, it is hoped that the speciality can use the models and results to help with their scheduling challenges.

To develop a practical surgery scheduling models that hospitals or surgical specialities can use, the models must be flexible and adaptable so that unique needs can be considered each time. The academic literature was reviewed as a starting point for the model development process. The goal was to gather information about the problem at other hospitals, identify the problem's common constraints and objectives, and finally,

see which modelling techniques are frequently used to formulate the problem. Even if the academic literature is extensive, the focus is often on scheduling multiple surgical specialities. As a result, the models may lack details on how individual specialities operate. Therefore, an important and inevitable second part of the model development process was to visit the hospital regularly for three reasons:

- To become familiar with the ORs, the wards, ICU, and the Post Anaesthesia Care Unit (PACU) to understand first-hand the daily operations in those units and their main problems.
- Meet the hospital's staff to get a holistic view of the problem and understand their scheduling challenges and desired outcomes.
- To present the hospital's data for validation and computational results for thorough feedback. The feedback received during this process was used to improve the models, such as by adding constraints. Comparisons to actual scheduling data were also made to understand the difference between human-made and optimised schedules.

This model development process, as described above, enabled close collaboration with the hospital and in models that solve the actual scheduling challenges the GS speciality faces. Although the models of this thesis were not implemented, it is believed that developing models in this way may help implement future models.

## 3.2 Data Collection and Analysis

A large historical dataset was collected from the hospital for this study for GS speciality spanning the years 2009-2019. The data originated from the hospital's database and included the following fields: operation date, registration date, medical priority, type of operation, the OR, the surgeon, and timestamps for the starting and end time of the surgery as well as starting time and end time of the anaesthesia for each anonymously identified patient. Additionally, the data set included information on the ward and ICU departments, along with corresponding LOS for patients who stayed in those units for at least a day following their surgery. Data analysis of the data set was performed in R.

As of 2019, an average of $95 \pm 30$ (mean$\pm$standard deviation) surgeries were performed each month at the GS. Up to 40% of the patients admitted to the ward for one or more days (in-patients) while the remaining portion is discharged from the hospital the same day as their surgery (out-patients). The number of semi-acute elective arrivals contributed to $28 \pm 8\%$ of the total throughput each month, with 30% on average requiring admission. It is assumed that semi-acute surgeries are elective patients who underwent surgery within two weeks after registration to the waiting list. The number

of blocks used each month similarly varied, with $39 \pm 9$ on average being utilised each month out of 48 blocks available (See Table 3.3), with each block having $83 \pm 9\%$ utilisation rate.

There are ten surgeons working in the speciality, each with a unique list of patients and specialisation in certain operations, even if all are capable of performing the standard operations. Tables 3.1 and 3.2 show summary statistics for each surgeon in 2019. The former table summarises all elective patients (including semi-acute elective patients), while the latter focuses specifically on semi-acute elective arrivals. The tables show that the case mix of each surgeon is fundamentally different. For example, surgeons 5 and 7 have a high throughput and a high ratio of semi-acute elective arrivals. However, the average surgery time and the ratio of in-patients are significantly lower for surgeon 7. Surgeon 9 has a patient throughput close to the median but has the highest ratio of semi-acute arrivals. Similar to surgeon 7, the average surgery time and the ratio of in-patients are relatively low. Analysing the surgery times of the surgeons, one can see that they are right-tailed as the mean is larger than the median in all cases. However, the interquartile range (IQR) between the first and the third quartile, indicates that some surgeons have high values (e.g. surgeon 5) and others lower values (e.g. surgeon 7). A higher IQR indicates higher variability but lower values less variation. A similar effect is observed for the LOS in the ward.

*Table 3.1. Summary statistics for the distributions of surgery times[†] and length of stay in the ward for each surgeon at the GS speciality. Throughput is shown for the year 2019.*

| | | **Surgery time (in minutes)** | | | | **Length of stay (in days)** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Surgeon** | **#Pat.** | **Mean** | **Median** | **SD** | **IQR** | **#In-Pat** | **Mean** | **Median** | **SD** | **IQR** |
| 1 | 103 | 143.7 | 112.5 | 40.7 | 13.1 | 30 | 3.6 | 3.3 | 2.1 | 4.2 |
| 2 | 90 | 130.5 | 123.1 | 33.7 | 59.5 | 35 | 4.6 | 4.5 | 1.2 | 1.2 |
| 3 | 109 | 141.8 | 124.8 | 43.9 | 90.4 | 33 | 3.4 | 2.5 | 1.7 | 1.4 |
| 4 | 118 | 177.5 | 173.3 | 36.8 | 108.9 | 66 | 4.7 | 4.2 | 1.3 | 1.2 |
| 5 | 141 | 178.3 | 170.6 | 40.1 | 112.5 | 88 | 4.5 | 4.0 | 1.2 | 0.9 |
| 6 | 91 | 157.6 | 93.9 | 44.6 | 96.0 | 26 | 6.3 | 6.9 | 2.5 | 5.3 |
| 7 | 180 | 87.9 | 78.1 | 26.9 | 8.2 | 15 | 4.7 | 4.0 | 1.8 | 1.9 |
| 8 | 144 | 125.6 | 100.8 | 40.7 | 55.6 | 30 | 3.7 | 3.3 | 2.0 | 1.5 |
| 9 | 108 | 90.4 | 66.5 | 19.3 | 59.1 | 27 | 5.3 | 5.1 | 1.8 | 2.5 |
| 10 | 57 | 109.0 | 112.8 | 25.8 | 13.0 | 12 | 1.2 | 1.1 | 0.2 | 0.2 |

[†]Surgery times of other surgeons used if fewer than 30 data points exist for each type of operation. #Pat is the total number of all elective patients (including semi-acute elective patients) operated on in 2019.
SD=standard deviation. IQR is the interquartile range between the first and the third quartile. #In-Pat is the the number of patients admitted to the ward following surgery.

Figure 3.1 portrays the distribution of surgery times for five common operations performed by the GS speciality. The figure shows that each operation has a right-tailed distribution, which is in line with the literature. In addition, the non-parametric Kruskal-Wallis test indicates that the median surgery time differs significantly across different types of operations, highlighting the importance of distinguishing between them

*Table 3.2. Summary statistics distributions of surgery times and length of stay in the ward for each surgeon at the GS speciality for semi-acute elective arrivals. Throughput is shown for the year 2019.*

| | | Surgery time (in minutes) | | | | Length of stay (in days) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Surgeon** | **#Pat.** | **Mean** | **Median** | **SD** | **IQR** | **#In-Pat** | **Mean** | **Median** | **SD** | **IQR** |
| 1 | 18 | 111.6 | 112.5 | 35.1 | 23.2 | 6 | 2.6 | 2.2 | 1.4 | 1.5 |
| 2 | 21 | 128.7 | 115.4 | 30.0 | 92.2 | 9 | 5.0 | 5.1 | 1.0 | 1.2 |
| 3 | 41 | 135.1 | 96.8 | 41.8 | 43.7 | 6 | 5.1 | 4.8 | 2.7 | 4.3 |
| 4 | 31 | 142.6 | 114.3 | 31.1 | 91.2 | 15 | 4.8 | 5.1 | 1.2 | 1.1 |
| 5 | 48 | 170.9 | 170.6 | 37.7 | 112.5 | 34 | 4.6 | 4.2 | 1.4 | 1.2 |
| 6 | 23 | 162.8 | 93.9 | 48.2 | 138.0 | 7 | 6.4 | 6.9 | 2.7 | 3.5 |
| 7 | 63 | 81.3 | 78.1 | 26.7 | 0.0 | 5 | 6.4 | 7.3 | 1.7 | 1.8 |
| 8 | 15 | 140.6 | 109.3 | 56.8 | 87.7 | 4 | 3.3 | 2.5 | 2.2 | 1.2 |
| 9 | 49 | 67.0 | 55.5 | 15.7 | 0.0 | 7 | 5.6 | 5.1 | 1.9 | 2.8 |
| 10 | 3 | 107.4 | 108.5 | 19.2 | 31.8 | 0 | - | - | - | - |

#Pat is the total number of semi-acute elective patients operated on in 2019. SD is the standard deviation. IQR is the interquartile range between the first and the third quantile. #In-Pat is the the number of patients admitted to the ward following surgery.

when generating time distributions for mathematical models ($p < 0.001$). A pairwise Wilcoxon rank sum test was conducted to further analyse the differences between pairs of operations shown in Figure 3.1. The results revealed that the median time is significantly different between all pairs of operations ($p < 0.001$) except for pairs I and IV, where the difference is not significant (see Table A.1 in Appendix A).

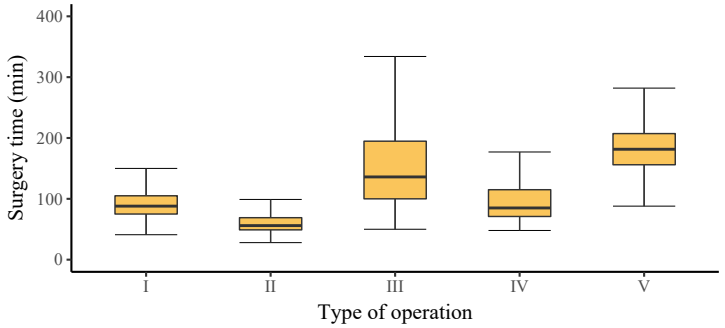



*Figure 3.1. Distribution of historical surgery times for five frequently performed operations at General Surgery.*

Operations labelled with Roman numbers.

In Figure 3.2, the historical surgery times for operation I, previously shown in Figure 3.1, are analysed to understand the impact of the surgeons on the variability. The figure indicates a large variation in the median surgery time between the surgeons where

surgeon 7 has the shortest surgery time, while surgeons 2 and 9 have the longest. The analysis suggests that distinguishing between the surgeons performing individual types of operations is important ($p < 0.001$). A pairwise Wilcoxon rank-sum test revealed that the median time varies between all surgeons performing this specific type of operation. However, no statistical significance was observed between between surgeons 1, 2 and 9 (see Table A.2 in Appendix A).

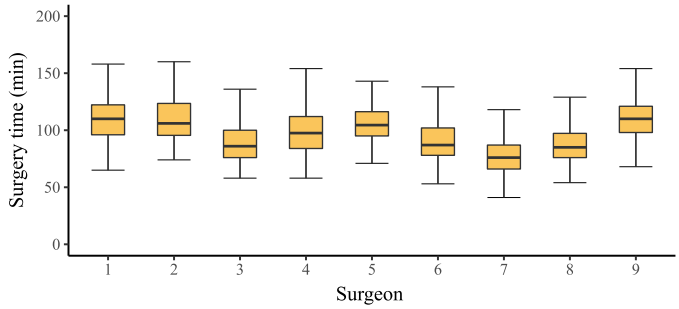



*Figure 3.2. Distribution of historical surgery times for a specific type of operation performed by each surgeon at General Surgery.*

After each surgery, cleaning must be performed taking on average, $22.9 \pm 7.2$ minutes to complete. It was assumed that cleaning times are associated with the period between two surgeries. Under this assumption, long cleaning times, such as those spanning a couple of hours, were identified in the data but removed. Extended cleaning times may still be possible in practise, for instance following a surgery of an infectious patient where the OR must be thoroughly sterilised afterwards.

Figure 3.3 shows the length of stay for two types of operations frequently performed in the speciality. The figure illustrates the probability that a patient is in the ward on a given day following surgery and is calculated using historical data as described in Section 2.2.2 with Example 2.1. The figure shows that the distributions have very different characteristics. For instance, a patient undergoing the operation on the left-hand side has a high probability of staying 13 days or longer in the ward, whereas a patient undergoing the operation shown on the right-hand side has a small probability of being in the ward beyond 7 days. The figure highlights the importance of distinguishing between the different types of operations ($p < 0.001$).

## 3.3 Problem Description

GS speciality performs around 1200 elective surgeries annually, with up to 40% requiring ward admission. The speciality performs upper- and lower-abdomen surgeries. The hospital uses a block scheduling approach. Each year, a cyclic MSS is realised based on

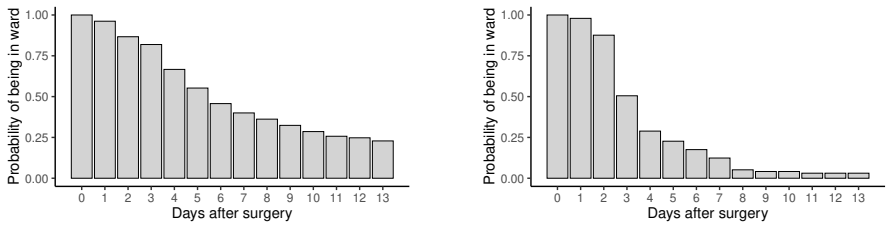



*Figure 3.3. The probability of being in the ward on a given day following surgery for two types of operations.*

the demand for surgical specialities. The current MSS has a cycle of one week, and at least one block is allocated to each surgical speciality in each cycle. The MSS employs two block lengths, 7.5 hours for Mondays-Thursdays and 5.25 hours for Fridays. Each block length spans the entire working day. The GS speciality allocates its blocks to each of its surgeons. An exception is a single surgeon who uses only the open blocks on Fridays or is allocated to one of the existing blocks in the MSS when possible.

The current MSS is shown in Table 3.3. The table shows that two blocks are allocated to the speciality daily with an additional block available on Wednesdays and Fridays respectively. Blocks on Fridays are available to all surgeons requiring additional capacity during the week and to the one surgeon not assigned to any block. On Wednesdays, surgeons 4 and 5 share blocks, and both must be present to perform robotic surgeries. On Thursdays, one or both of the surgeons can be present to perform non-robotic surgeries.

*Table 3.3. Master Surgery Schedule for General Surgery surgeons.*

| | **Monday** | **Tuesday** | **Wednesday** | **Thursday** | **Friday** |
|---|---|---|---|---|---|
| OR 1 | 8 | 6 | 3 | 1 | Open |
| OR 2 | 9 | 2 | 7 | 4/5 | Open |
| OR 3 | - | - | 4/5 (Robot) | - | Open |

When an elective patient is registered to the waiting list, the patient is assigned a medical priority and a type of operation determined by the surgeon. In general, the medical priority determines the urgency and maximum response time for the surgery. Three medical priority codes are available at the speciality: one week, four weeks and three months. However, many elective patients, particularly the ones with a three-month medical priority, exceed their waiting time limits (Embætti landlæknis, 2021).

A human scheduler assigns elective patients to the blocks of their surgeons one week in advance. The selection is based on a combination of medical priority and waiting time and is performed in cooperation with the surgeons. However, the scheduler must also consider the capacities of the different resources that are limited and commonly shared with other specialities. Moreover, equipment availability, such as the robot, may be restricted (see Table 3.3). Exceeding the capacity of certain resources, such as the

ward beds or blocks, may cause last-minute cancellations, which shall be prevented at all costs. In addition, the number of out-patients shall be restricted in a block due to the operating hours of the PACU in an attempt to avoid possible bottlenecks. Each week, the surgeons, OR managers, anaesthesiologists, ward managers and ICU managers meet to review the schedule for the following week. This meeting is used as a venue to foresee possible disruptions in advance that may cause last-minute cancellations. Depending on the outcome of this meeting, it may be necessary for the scheduler to modify the plan.

The GS speciality has developed practical rules gathered from experience over time that the scheduler applies. These rules may be seen as blueprints and aimed to standardise the scheduling process while making it possible to minimize the risk of last-minute cancellations. The scheduler applies the rules in the following way. First, the sum of the mean surgery times of patients in a block should not exceed the block's capacity. The mean surgery time is calculated for each operation. However, going beyond the block capacity may be possible in some blocks as two surgery teams are available after hours for emergency patients. Regardless, it should be kept to a minimum. Finally, to account for limited access to downstream resources and to avoid overtime, the following rules are applied:

- *Outpatient Quota*: A maximum of five outpatients may be assigned per block to allow for sufficient recovery time in PACU during operating hours.
- *Patient Quota*: A maximum of six patients may be assigned per block to hedge against unexpected overtime.
- *ICU Bed Quota*: Only one ICU bed is available to the GS speciality. Therefore, only one patient who requires admission to the ICU after surgery can be assigned each day, and only if the bed is available at the end of the surgery.
- *Ward Bed Quota*: Only six staffed ward beds are available to the GS speciality. Therefore, a maximum of six patients who require admittance to the ward after surgery can be assigned each day, the exact number depends on how many patients from the GS speciality are already in the ward that day.

Typically, it is known with certainty which patients will be admitted to the ward or ICU. The scheduler manually forecasts ward and ICU bed occupancy up to seven days in advance by utilising the mean value associated with each patient's type of operation. In general, it may be possible to temporarily increase the number of ward beds to hedge against last-minute cancellations. However, such peaks in occupancy have historically caused last-minute cancellations, so the scheduler attempts to avoid such peaks. The number of ICU beds cannot be increased and is a hard constraint when scheduling.

The unpredictable arrivals of semi-acute elective patients pose significant challenges for the scheduler. Semi-acute elective patients commonly have a medical priority of one week or four weeks but should preferably undergo surgery within a week or two.

Semi-acute patients and other elective patients share the same blocks at the speciality. Consequently, the scheduler must anticipate how much capacity is necessary to leave unreserved for semi-acute arrivals each week. As a result, the schedule presented at the weekly meeting contains multiple gaps. In general, if the gaps are insufficient, given a high throughput of elective patients, rescheduling of already scheduled patients will occur as the total block time is finite each week. However, if the gap is too large, there is a risk of under-utilising the hospital's resources, which lowers the elective throughput. Both outcomes are undesirable, and scheduling elective patients more than a week in advance is desired by the speciality to give patients with lower priority more time to prepare for their surgery. This has turned out to be a difficult task due to these unpredictable semi-acute arrivals. Therefore, the hospital seeks ways to do so but in a way that maximises the block utilisation while minimising the risk of last-minute cancellations and the number of rescheduling events.

## 3.4 Experimental Setup

The experiments are performed on a Windows desktop machine with 32GB Intel Core i7-7700 3.60 GHz with four cores and solved with Gurobi. All models use the data set described in Section 3.2 and solved as follows:

- Step I: A set of feasible patterns is generated under practical rules and restrictions on overtime using historical data. This step is performed offline prior to the start of the optimisation performed in Step II and requires a few minutes of computational time.
- Step II: The feasible patterns of patients are assigned to pre allocated blocks using a MIP model under some objective and restrictions on downstream resources.

For Step I, 1000 scenarios of surgery times are sampled with replacement from historical data for each patient based on the patient's specific type of operation and performing surgeon using up to the most recent 100 data points for the combination. In cases where fewer than 10 data points of surgery times are available for the surgeon, the data is combined with data from other surgeons. If no data is available, the planned surgery time is used, which applies to approximately 5% of different types of operations in 2019. At the same time, LOS distributions, as shown in Figure 3.3, are generated for each patient requiring ward and ICU admission based on the patient's type of operation and are similarly created offline prior to the start of the optimisation using data of all surgeons and all years. Other parameter settings and data for each model were obtained either from historical data or the hospital and are based on standard practice at GS speciality. They will be presented in more detail in Chapters 4 to 6. The following section briefly presents the unique experimental setup of each chapter of this thesis.

### 3.4.1 Uncertainty in Surgery Times

The purpose of the computational experiments is to determine how the size of each surgeon's waiting list and the length of planning horizon affect the construction of the MSS in terms of allocation of surgeons to blocks, patient throughput and computational tractability. Waiting lists are generated for each surgeon with types of operation sampled at the right historical frequencies for the surgeon from historical data. Additionally, the results of using an optimal MSS are compared with an actual MSS of the GS speciality, where the allocation of surgeons to blocks is fixed.

It is assumed that it is unknown which surgeries are admitted to the ward beforehand, but their admission probability is based on the surgeon's historical inpatient ratio. A simple heuristic is proposed to hedge against exceeding the staffed ward beds by assuming three possible ward scenarios. After optimisation, the risk of exceeding the number of staffed beds is evaluated each day using Monte Carlo sampling with undiscretised LOS distributions to estimate the true probability of exceeding the number of staffed beds.

The model was programmed in Python 3.6 and solved using Gurobi. The time limit for each experiment was set to 6 hours.

### 3.4.2 Uncertainty in Length of Stay

The computational experiments aim to compare the solution quality and computational time between the two models for bounding the risk of exceeding staffed ward beds under various parameter settings. The prior model uses ward combinations, while the latter uses robust optimisation.

The ward combinations are generated offline prior to the start of the optimisation in C and used in the MIP model. Each ward combination is simulated 1000 times using Monte-Carlo sampling, where each combination is verified towards the risk of exceeding the different number of staffed ward beds. The MIP model was programmed using AMPL mathematical programming language and solved using Gurobi. The models schedule the same set of patients as the specialty did for a single month, allowing for a comparison with the actual scheduling data..

Each solution is verified with Monte-Carlo sampling using the complete undiscretised empirical distributions of surgery times and LOS from historical data to estimate the true probability of exceeding the block capacity and the staffed ward beds.

### 3.4.3 Uncertainty in Semi-Acute Elective Arrivals

The aim of computational experiments is to compare three gap-reserving strategies to understand how they impact the need for rescheduling, block utilisation and overtime. The scheduling and rescheduling of semi-acute elective patients will be conducted over twelve four-week planning horizons, and the schedule will be updated with each new semi-acute elective arrival. During each planning horizon, no elective patients are added to the waiting list, only semi-acute elective patients. Therefore, it is possible to compare the optimised results to some extend, using the gap-reserving strategies, to actual scheduling data.

The scheduling and rescheduling MIP models were programmed in AMPL and solved with Gurobi. Each solution is verified with Monte-Carlo sampling using undiscretized empirical distributions of surgery times and LOS from historical data to estimate the true probability of exceeding the block capacity and the staffed ward beds.

## 3.5 Summary

This chapter presented the methodology that will be used in the thesis. It included a description of the model development process used to develop mathematical models inspired by practical challenges of GS speciality at Landspitali. This was followed by a presentation and analysis of the data set collected from the hospital, a problem description of the GS speciality on which the models are based, and finally, a brief description of the experimental setup that will be used in this thesis was given.

The data set analysis revealed the importance of considering uncertainty in surgery times and LOS when scheduling to hedge against last-minute cancellations. Moreover, the analysis highlighted the importance of distinguishing between types of operation when time distributions are generated for surgery times and LOS. However, from the problem description, it became evident that the combined risk of last-minute cancellations, associated with surgery times and LOS, and rescheduling events, due to unpredictable arrivals of semi-acute elective patients, poses significant challenges for the human scheduler, which commonly assigns elective patients to blocks one week in advance. However, the hospital seeks to give patients appointments more than a week in advance while minimising the risk of rescheduling and last-minute cancellations and maximising block utilisation.

In the following chapter, uncertainty in surgery times is considered, with a focus of how it can be specified practically and statistically accurately in a mathematical model while considering a high throughput of elective patients and limited number of staffed ward beds.

# 4 Uncertainty in Surgery Times

This chapter aims to specify uncertainty in surgery times in a practical and statistically accurate way under a high throughput of elective patients while accounting for a limited number of staffed ward beds. The work in this chapter is based on Sigurpalsson et al. (2020), Sigurpalsson et al. (2022), and Runarsson and Sigurpalsson (2019b).

The literature addressing uncertainty in surgery times is extensive, and the importance of accounting for this uncertainty is widely acknowledged, as discussed in Section 2.2.1. By ignoring uncertainty, e.g., by scheduling cases by the mean (Marques and Captivo, 2017), as is often done by hospitals (Batun et al., 2011; Pandit, 2018, 2020) and some studies (Lamiri et al., 2008b), the risk of last-minute cancellations increases. That is to say, if an operation takes longer than its mean time, the following patient might be cancelled at the last minute due to overtime restrictions (Wang et al., 2014). At the same time, uncertainty in LOS in the downstream ward should also be considered, as discussed in Section 2.2.2. This source of uncertainty is known to cause a lack of available beds, which may result in last-minute cancellations or early discharges of other patients (Neyshabouri and Berg, 2017) and, thus, cannot be ignored. Regardless, the combined risk of last-minute cancellations due to uncertainty in surgery times and uncertainty in LOS in the ward is frequently disregarded in advance scheduling models (Shehadeh, 2022).

The two-stage SP model solved with SAA is often employed to take the uncertainty into account, and chance constraints are used to reduce the risk of last-minute cancellations (van Oostrum et al., 2008; Pandit and Tavare, 2011; Shylo et al., 2013; Kamran et al., 2018; Adams et al., 2023). However, as the problem becomes larger, e.g. when more patients are scheduled, tailored heuristics (Aissaoui et al., 2020; Shehadeh, 2022) or column generation (van Oostrum et al., 2008) may be required to solve the problems. Consequently, practical and statistically accurate optimisation approaches are lacking (Harris and Claudio, 2022; Shehadeh, 2022) and such approaches may require an unconventional combination of solution methodologies (Aringhieri et al., 2013).

A novel data-driven and practical approach, termed *Pattern Scheduling* is proposed in this chapter. The approach consists of two steps. In the first step, all feasible patterns of elective patients awaiting surgery are generated for each surgeon. A *Pattern* is defined as a pre-generated feasible, unordered combination of one or more patients who can be assigned to a single block for their surgeon. The feasibility of a pattern is determined by

practical rules set by hospitals (e.g., an upper limit on the number of patients assigned to a block) on the one hand and probabilistic restrictions on overtime on the other. Consequently, patterns that violate practical rules, e.g., by exceeding the upper limit on the number of patients assigned to a block or by violating probabilistic restrictions on overtime, are eliminated. In the second step, the feasible patterns that result in the highest throughput of elective patients in the planning horizon are assigned to blocks using a MIP model. However, the assignment is done in a way that maintains the historical ratio of inpatients from the overall throughput of patients for each surgeon and for the whole speciality. As exceeding the staffed ward beds can potentially result in last-minute cancellations, a ward approximation is proposed, which allows bounding the risk of exceeding the staffed ward beds in a probabilistic manner to a specified threshold in a MIP model. The ward approximation assumes only three possible scenarios for each ward patient. Namely, each ward patient has a 100%, 50% or 0% chance of being in the ward on any given day following surgery, where only one scenario applies to each patient in a day. While having more than three scenarios is possible, it requires a more complicated formulation. Consequently, the ward approximation is generalised in Chapter 5 to allow for a higher number of scenarios.

Computational experiments are performed to build several cyclic MSSs that determine the optimal allocation of surgeons to blocks each week under different cycle lengths and waiting list sizes for each surgeon. The goal is to explore how the parameter selection influences the construction of the MSS in terms of surgeon allocation to blocks, patient throughput, and computational tractability. Additionally, the results of the optimal MSS are compared with actual MSS of General Surgery, where the allocation of the surgeons to blocks is fixed, focusing on throughput and surgeon allocation.

The Pattern Scheduling approach is inspired by the work of van Oostrum et al. (2008), in which all possible patterns are generated with probabilistic restrictions on overtime. In their study, column generation is used to manage the exponential growth of possible patterns. However, using column generation, requires the stochastic values to be approximated with deterministic values. The novelty of the proposed method stems from generating all feasible patterns rather than all possible patterns.

Pattern Scheduling is practical in two ways. On the one hand, it is computationally practical since the search space is reduced and does not require the use of approximate models. In other words, the uncertainty of each pattern is verified using Monte Carlo sampling with historical data, so it is statistically accurate. Patterns with a high chance of overtime are eliminated from the search space. On the other hand, it is clinically practical, as the generated patterns follow the practical rules of the speciality, which limits the search space further.

The chapter is organised as follows. The next section presents model development, where the *Pattern Scheduling* approach is introduced. In Section 4.2, experiments are conducted where the approach is used to construct cyclic MSSs under various parameter settings. A summary of the findings is provided at the end of the chapter.

## 4.1 Model Development

The general problem addressed in this section is finding the optimal allocation of surgeons to blocks for a cyclic MSS, which repeats every five working days, while maximising the throughput of elective patients and accounting for the limited staffed ward beds. It is assumed that each block can only be allocated to a single surgeon and that elective patients (in- and out-patients) most likely to be scheduled for each surgeon can be assigned to their blocks. The patient scheduling must be performed in a way that maintains the historical ratio of inpatients to the throughput for each surgeon and the whole speciality. Additionally, the scheduling must minimise the number of possible last-minute cancellations that can occur on the day of surgery in advance. As a result, uncertainty in surgery times and LOS must be considered by imposing bounds the risk of exceeding block capacity and the staffed ward beds, respectively, in a probabilistic manner.

A novel, two-step approach, termed *Pattern Scheduling*, is proposed to solve the problem and works in the following way:

1. *Pattern Generation*: A pattern is a feasible unordered combination of one or more patients a specific surgeon can operate on in a block on a given day. The first step in generating feasible patterns involves extracting all patients belonging to a specific surgeon. Then, all possible combinations of these patients are generated following the hospital's practical rules. Patterns that do not follow practical rules are not generated. Even if patterns are generated following practical rules, some may carry a high risk of exceeding block capacity, which may result in last-minute cancellations. Therefore, each pattern is validated against the risk of exceeding its block capacity using Monte-Carlo sampling with historical data for surgery times. All patterns where the simulated probability is higher than the threshold $\delta$ are eliminated. This procedure is performed for all surgeons.
2. *Optimisation*: Given the set of feasible patterns generated in the previous step, a deterministic MIP model is used to allocate the surgeons to blocks by assigning their patterns of patients to blocks in a way that maximises the overall throughput of elective patients. The scheduling model also accounts for restrictions on staffed ward beds in a probabilistic manner while maintaining the historical ratio of inpatients to the total throughput. The proposed model is solved using a commercial solver.

Both steps are described in more detail in Section 4.1.1 and Section 4.1.2.

### 4.1.1 Pattern Generation

The hospital has allocated a different number of blocks to the surgical specialities within the planning horizon ($T$) consisting of the days $d \in D$ and determined the opening hours of each block ($C_{d,r}$). Hence, the task of each surgical speciality is to allocate their blocks to their surgeons ($o \in O$) so that each surgeon is assigned one or more blocks each week ($v \in V$). Patients ($i \in I$) on the surgeon's waiting list ($I_o$) can then be assigned to their surgeon's ($o \in O$) block.

For each surgeon, a set of feasible patterns is generated that can be assigned to the surgeon's blocks. It is assumed that the feasibility of a pattern is determined by practical rules specified by hospitals and probabilistic restrictions on overtime. Evidently, generating patterns without specifying the practical rules results in exponential growth of patterns (van Oostrum et al., 2008). As a result, it becomes impossible to evaluate each pattern for the risk of exceeding its block capacity using simulation with historical data, and approximations are needed. However, hospitals have a range of practical rules that determine which patterns are permitted, and applying these rules can reduce the number of possible patterns. This makes it possible to evaluate each pattern for the risk of exceeding its block capacity using simulation and historical data. The application of practical rules is demonstrated in Example 4.1.

**Example 4.1** *Suppose a hospital has a practical rule that restricts the number of patients admitted to ICU in a single pattern. Now, assume there are five patients on the waiting list for a single surgeon, and two of the patients require ICU admission following surgery. Generating all possible patterns results in 31 unique patterns, assuming that each pattern includes at least one patient. Applying the practical rule results in 23 feasible patterns, where patterns containing two ICU patients have been eliminated.*

Practical rules serve as blueprints to standardise the scheduling and are based on expertise and experience gathered over time, as discussed in Section 3.1. For example, these rules may restrict the number of patients in a pattern (Kamran et al., 2018), the number of ICU admissions (Kim and Horowitz, 2002), the types of operations permitted (van Essen et al., 2014), scheduling balance for surgery groups (Banditori et al., 2013; M'Hallah and Visintin, 2019), time limits on overtime (Jebali and Diabat, 2015) and equipment availability (van Essen et al., 2014). In this chapter, two practical rules are applied:

- Upper bound on the number of patients assigned to a pattern.
- Balance between in- and out-patients assigned to a pattern based on historical ratios for each surgeon.

Even if the patterns generated are feasible according to practical rules, they may still violate restrictions on overtime, potentially leading to last-minute cancellations.

Therefore, each pattern must be validated using Monte Carlo sampling with historical data. Patterns where the simulated probability is higher than a predefined overtime threshold are eliminated to hedge against last-minute cancellations on the day of surgery. This process further explained in Example 4.2.

**Example 4.2** *Suppose that 23 patterns feasible according to practical rules have been generated from five patients have been generated, as presented in the example above. Now, assume that each pattern has been evaluated using Monte-Carlo sampling using historical data to estimate the risk of exceeding the block capacity. The simulation shows that 10 out of the 23 patterns surpass the predefined overtime threshold of 25%. Therefore, these patterns are eliminated as they are not feasible towards overtime restrictions and may cause last-minute cancellations if used.*

The pattern generation is mathematically expressed in the following way. Let $z_{i,p}$ be a binary indicator taking the value 1 if patient $i$ is assigned to pattern $p$, otherwise 0. The practical rules can then be implemented as follows:

1. *Patient Quota*: Let $M^p$ be a parameter specifying the maximum number of patients that can be assigned to each pattern $p$. This upper bound can be imposed using the following constraint:

$$\sum_{i \in I_o} z_{i,p} \leq M^P, \quad \forall\, p \in P, \quad o \in O \tag{3}$$

2. *Balance Between In- and Out-patients*: Let the parameter $h_o$ represent the required scheduling balance between in-patients and out-patients for each surgeon $o$, as determined by the surgeon's waiting list. Additionally, let $g_i^{Ward}$ be a binary parameter, taking the value 1 if patient $i$ requires ward admission and 0 otherwise. It is assumed that it is known in advance which patients require ward admission following surgery. Consequently, the balance between in- and out-patients assigned to each pattern can be obtained in the following way:

$$\left| \sum_{i \in I_o} g_i^{Ward} z_{i,p} - \lceil h_o \sum_{i \in I_o} z_{i,p} \rceil \right| \leq 1, \quad \forall o \in O, \quad \forall p \in P \tag{4}$$

   The first part of the constraints sums up the number of inpatients assigned to a pattern. The second part determines the desired number of inpatients in the pattern for each surgeon $o$ by multiplying their historical ratio of inpatients ($h_o$) by the total number of patients assigned to the pattern. The result is then rounded up to the nearest integer. Preferably, the absolute difference between these two parts should be 0. However, to allow flexibility, it is set to 1. For example, suppose a pattern consisting of four patients, with two requiring ward admission, for a single surgeon whose historical ratio of inpatients is 0.5. The pattern is feasible as $|2 - \lceil 0.5 \cdot 4 \rceil| \leq 1$. Notably, due to the flexibility introduced by allowing the absolute difference to be 1, the number of inpatients in this pattern can range from 1 to 3.

3. *Overtime Verification*: As the block lengths ($C_{d,r}$) are finite and the surgery times are stochastic, only a subset of patterns is feasible. That is to say, a pattern is feasible when the probability of exceeding $C_{d,r}$ is no more than $\delta$, namely:

$$\Pr\left[\sum_{i\in I_o} S(i)z_{i,p} \geq C_{d,r}\right] \leq \delta, \quad \forall p \in P, \quad o \in O, \quad r \in R, \quad d \in D \tag{5}$$

where $S(i)$ is a random variable denoting the surgery time of patient $i$. In this case, the surgery time includes pre-incision, incision, post-incision, and cleaning. Patterns that do not meet the overtime restrictions may result in last-minute cancellations and are therefore eliminated at this step.

- *Exception:* An exception is given to patterns that consists of a single patient as some of these patterns may span the whole day and exceed the block capacity with a high probability. Regardless, it is more common to see multiple patients in a pattern.

The output of the pattern generation is a set of patterns that are feasible according to practical rules and probabilistic restrictions on overtime for each surgeon. These patterns are utilised in the scheduling model proposed in Section 4.1.2. For each feasible pattern, a ward bed occupancy is calculated from historical data by summing the individual probabilities of each patient in the pattern requiring ward admission in the pattern each day following surgery. This process is further described in the next section, which outlines how a ward approximation, assuming only three possible ward scenarios, is implemented to bound the risk of exceeding ward bed capacity.

### Ward Approximation

Historical data makes it possible to estimate the probability that a particular patient will be in the ward on any given day following surgery based on the patient's type of operation as illustrated in Example 2.1 in Section 2.2.2. By summing the different probabilities for each patient in the ward on each day after surgery, it is possible to estimate the total number of patients each day and, thus, the beds needed. Nonetheless, if there is a restriction on the number of staffed ward beds each day ($M^A$), this method does not take into account the probability of exceeding the ward capacity (van den Broek d'Obrenan et al., 2020) as discussed in Section 2.2.2. Hence, last-minute cancellations are possible and a different approach is needed to hedge against such occurrences.

It is possible to minimise this risk of last-minute cancellations by approximating the problem as follows. Assume that there are only three possible scenarios for each patient being in the ward on any given day $j$ after surgery, and that only one scenario applies to each patient on a given day. Namely, the patient will either be in the ward on day $j$ with a 100% certainty, a 50% chance or a 0% chance. Based on this approximation, the patients with a 100% chance on day $j$ determine the number of staffed ward beds

available for patients with a 50% chance, as there is no need to consider the patients with a 0% chance. In other words, if there are four patients with a 100% chance on a given day and six ward beds, then two beds would remain available for patients with a 50% chance. Thus, the problem becomes determining how many patients with a 50% chance can be in the ward each day without exceeding the staffed ward bed capacity.

Suppose there exists at least one patient with a 50% chance of being in the ward on any given day following surgery then it is possible to use the quantile function for the Binomial distribution to determine the maximum number of such patients in a day ($F_a^{50}$) when there are $a \in \{0, \ldots, M^A\}$ staffed ward beds available, given a specified confidence level. In other words, the binomial distribution describes the number of successes in independent repetitions when there are only two possible outcomes, each with an associated probability of occurring.

A classic example of Binomial distribution is flipping a fair coin multiple times and counting how many times the head appears. Imagine extending this experiment by flipping five coins simultaneously instead, but keep on counting the number of heads in each flip and arrange the results in an acceding order. In that case, it is possible to use the quantile function to find the value at which a certain percentage of the results are below. Notably, this example is equivalent to having five patients in the ward on a given day, each having a 50% chance of staying and finding the maximum number of beds needed for a given confidence level. By using the confidence level of 0.05, meaning that 95% of the data is below that value, the maximum number of patients with a 50% chance of being in the ward following surgery ($F_a^{50}$) for a given bed capacity $a$ is, for example:

$$(a,\ F_a^{50}) : (0,\ 0), (1,\ 1), (2,\ 2), (3,\ 3), (4,\ 4), (5,\ 4), (6,\ 5)$$

Therefore, for this example, the scheduler would need four beds for the five patients, for the given confidence level.

Assuming that there are only three possible ward scenarios for each patient in the ward each day means that the empirical ward probabilities for each ward patient must be discretised to 0%, 50% and 100%. For example, if the historical data for the operation presented in Example 2.1 in Section 2.2.2 is updated to include a patient who stayed five days in the ward. The updated empirical distribution would be 100% on day 0 (the day of surgery), 60% on days 1 and 2, 20% on days 3 and 4, and 0% afterwards. One way to approximate the empirical distribution is by discretising the probability interval from 0% to 100% into three intervals. The continuous daily ward probabilities within each interval are then discretised to 0%, 50% and 100% in the following way:

- 0-24% discretised to 0%
- 25-74% discretised to 50%
- 75-100% discretised to 100%

As a result, the discretised empirical ward distribution, using the three scenarios, would be 100% on day 0, 50% on days 1 and 2 (discretised from 60%) and 0% on days 3 and 4 (discretised from 20%). This example is portrayed in Figure 4.1.

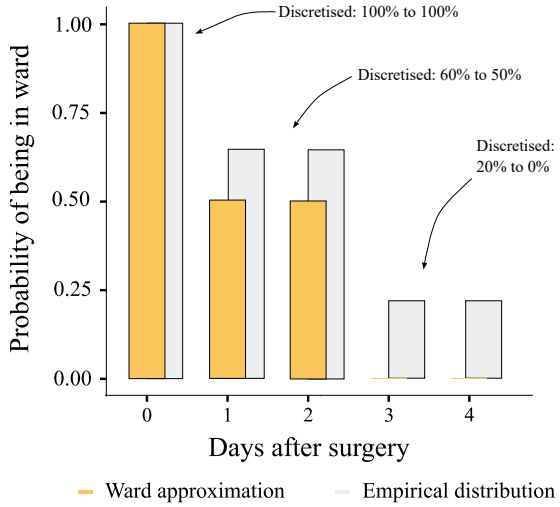



*Figure 4.1. Discretisation of an empirical distribution of daily ward probabilities using three scenarios, each with an associated probability.*

For each feasible pattern, two ward distributions are generated and utilised in the mathematical model presented in the next section. The first distribution counts the number of patients of a pattern $p$ having a 100% chance of being in the ward on day $j$ after surgery ($Q^{100}_{j,p}$). The second distribution counts the number of patients of pattern $p$ having a 50% chance of staying on day $j$ after surgery ($Q^{50}_{j,p}$).

Figure 4.2 illustrates a pattern ($p = 1$) of two patients undergoing the same type of operation, as shown in Figure 4.1, where both patients require ward admission. The distribution for $Q^{100}_{j,p}$ indicates that both patients have 100% of staying on the day of surgery. Meanwhile, $Q^{50}_{j,p}$ shows that on days 1 and 2 after surgery, both patients have a 50% chance of staying. On all other days, they have a 0% chance of staying.

### 4.1.2 Optimisation with Limited Downstream Ward Beds

Once all feasible patterns have been generated that satisfy both practical rules and probabilistic restrictions on overtime, the scheduling problem is reduced to assigning patterns to blocks so that the throughput of elective patients is maximised. This is done while accounting for the limited downstream ward beds in a probabilistic manner using the ward approximation.


| | Day after surgery | | | | | | |
|---|---|---|---|---|---|---|---|
| $j \rightarrow$ | 0 | 1 | 2 | 3 | 4 | $\cdots$ | $M^W$ |
| $Q^{100}_{j,p=1}$ | 2 | 0 | 0 | 0 | 0 | $\cdots$ | 0 |
| $Q^{50}_{j,p=1}$ | 0 | 2 | 2 | 0 | 0 | $\cdots$ | 0 |

*Figure 4.2. Example of a single pattern consisting of two patients requiring ward admission following surgery.*

The example shows the distributions number of patients with a 100% chance ($Q^{100}_{j,p}$) and a 50% chance ($Q^{50}_{j,p}$) of staying on day $j$ to day $M^W$ after surgery for pattern $p = 1$.

Let $x_{d,p,r}$ be a binary decision variable, taking the value 1 if pattern $p$ is assigned to the day $d$ and room $r$. Otherwise, $x_{d,p,r}$ is 0. The assignment of patterns to blocks is usually subject to many restrictions. For example, there may be restrictions on equipment availability, the surgeons and so forth. Consequently, the reduced set $(d,p,r) \in DPR \subseteq D \times P \times R$ is generated, taking such restrictions into account. In this case, the surgeons' availability for a given block is considered, as it depends on the blocks allocated to the speciality each day in the MSS.

Since each pattern spans the entire day, any given block can only have one pattern assigned:

$$\sum_{p \in P,\ r \in R:\ (d,p,r) \in DPR} x_{d,p,r} \leq 1, \quad \forall d \in D \tag{6}$$

Each patient $i$ may be present in multiple patterns, but the patient may only be assigned once:

$$\sum_{(d,p,r) \in DPR:\ i \in I_p} x_{d,p,r} \leq 1, \quad \forall i \in I \tag{7}$$

where $I_p$ are the set of patients included in pattern $p$. Additionally, each surgeon is only permitted to work according to a single pattern per day:

$$\sum_{p \in P_o,\ r \in R:\ (d,p,r) \in DPR} x_{d,p,r} \leq 1, \quad \forall d \in D, \quad o \in O \tag{8}$$

where $P_o \subseteq P$ are the patterns containing the patients of surgeon $o$. Moreover, each surgeon should be assigned to at least one day for a given week $v$ with working days $D^v$:

$$\sum_{p \in P_o,\ d \in D^v,\ r \in R:\ (d,p,r) \in DPR} x_{d,p,r} \geq 1, \quad \forall o \in O,\ v \in V \tag{9}$$

A practical rule was implemented in the generation of the feasible patterns to maintain the correct historical ratio for each surgeon between in- and out-patients assigned to each pattern as expressed with Constraint (4). However, to account for flexibility, the absolute difference between the actual number of inpatients and the desired number of inpatients, which was rounded up to the nearest integer, in a pattern for each surgeon

was allowed to be 1. Therefore, given this flexibility, a bias may have been introduced towards selecting patterns with a higher ratio of outpatients, requiring less surgery time across all surgeons when the objective is to maximise the throughput of patients. Thus, a global ratio $h_G$ of the correct historical balance between inpatients and outpatients for the speciality is imposed in the planning horizon. Otherwise, the optimisation may favour patients with shorter surgery times over those with longer to maximise the throughput.

Let $P^{in}$ be a continuous variable denoting the total number of inpatients scheduled across the planning horizon and determined by the following constraint:

$$P^{in} = \sum_{p \in P, d \in D, r \in R} G_p x_{d,p,r} \tag{10}$$

where $G_p$ denotes the number of inpatients for pattern $p$. Let $C_p$ be a parameter denoting the number of patients assigned to pattern $p$, then the global ratio $h_G$ can be maintained by the following two constraints:

$$h_G \sum_{(d,p,r) \in DPR} C_p x_{d,p,r} - 1 \leq P^{in} \tag{11}$$

$$P^{in} \leq h_G \sum_{(d,p,r) \in DPR} C_p x_{d,p,r} + 1 \tag{12}$$

These constraints provide a limited degree of flexibility in the number of inpatients. For example, if 100 patients are scheduled and the $h_G = 0.5$, then the number of inpatients must be between 49 and 51.

The final step in the model is to accounting for limited the number of staffed ward beds each day to avoid last-minute cancellations that can occur when the capacity is exceeded. As previously discussed, two ward distributions are generated for each feasible pattern $p$. The first distribution denotes the number of patients with a 100% chance of being in the ward on any given day $j$ following surgery (denoted with $Q^{100}_{j,p}$) and the latter for patients with a 50% chance ($Q^{50}_{j,p}$).

Suppose there are $M^A$ staffed ward beds available each day. Then, the number of patients with a 100% chance of staying on a given day can be at most $M^A$. For instance, if there are six beds on a given day, then there can be at most six patients with a 100% chance of staying that day to avoid exceeding the capacity. Next, by subtracting the number of patients with a 100% chance from $M^A$, it is possible to determine how many beds remain available for patients with a 50% chance. For example, if there are five patients with a 100% chance on a single day and six staffed ward beds, then one bed remains available for patients with a 50% chance. The final step is to utilise the parameter $F^{50}_a$, as presented in Section 4.1.1, to determine the number of patients with a 50% chance that can be accommodated when there are $a \in \{0, \ldots, M^A\}$ beds remaining for a given confidence level. For example, if only one bed is available on a given day, then only one patient with a 50% chance can be in the ward that day for a given confidence level. The approximation will now be implemented in the MIP model and requires adding a series of constraints.

First, given that $Q^{100}_{j,p}$ is a parameter denoting the number of patients from pattern $p$ with a 100% chance of being in the ward on the day $j$ after surgery and belonging to pattern $p$, then it is possible to compute the number of such patients each day ($n^{100}_d$) as follows:

$$\bar{n}^{100}_d = \sum_{r\in R, p\in P, j\in\{0,\dots,M^W\}: d-j\in D} Q^{100}_{j,p} x_{d-j,p,r}, \quad \forall d \in D \tag{13}$$

where $M^W$ is the upper bound on LOS in the ward. Note that, the constraint considers all patients with a 100% chance of staying on a given day $d$, also those who had surgeries the days before ($d-j$) and still have a 100% chance of staying. Next, an identical constraint is added, but now for patients with a 50% chance of staying, or:

$$\bar{n}^{50}_d = \sum_{r\in R, p\in P, j\in\{0,\dots,M^W\}: d-j\in D} Q^{50}_{j,p} x_{d-j,p,r}, \quad \forall d \in D \tag{14}$$

Patients with a 100% chance of staying cannot exceed the number of staffed ward beds $M^A$ on any given day, and thus, an upper bound is imposed as follows:

$$\bar{n}^{100}_d \leq M^A, \quad \forall d \in D \tag{15}$$

At this point, it is possible to determine how many beds are remaining for the patients with a 50% chance of being in the ward on any given day using the following constraint:

$$a_d = M^A - \bar{n}^{100}_d, \quad \forall d \in D \tag{16}$$

where $a_d$ is the number of staffed beds available for patients with a 50% chance each day. Finally, the problem becomes determining the maximum number of such patients in the ward in a day $d$ for when there are $a_d$ beds remaining. Consequently, as the last step in bounding the risk of exceeding the staffed ward beds, a connection must be made between $a_d$ to the parameter $F^{50}_a$ which determines the maximum number of the patients as presented in Section 4.1.1.

The first step in connecting $a_d$ to the parameter $F^{50}_a$ is by introducing the binary variable $\Xi_{d,a}$ taking the value 1 if on the day $d \in D$ there are $a \in \{0,\dots,M^A\}$ beds available; otherwise, it is 0. To link $\Xi_{d,a}$ to $a_d$, the following constraint is used:

$$a_d = \sum_{a\in\{0,\dots,M^A\}} a\Xi_{d,a}, \quad \forall\, d \in D \tag{17}$$

In other words, using this constraint the binary variable $\Xi_{d,a}$ only takes the value of 1 when the sum of all possible remaining capacities ($a \in \{0,\dots,M^A\}$) multiplied with the variable equals $a_d$ each day. This means that if there are $a_d = 3$ beds available on a given day and the available bed capacities are $a \in \{0,1,2,3\}$, then the binary variable is $\Xi_{d,a} = 1$ for $a = 3$. It must be noted that $\Xi_{d,a}$ is also 1 for combinations of other capacities where the sum is 3[1]. To overcome this issue, a constraint is added so that

[1]For example, when $a \in \{1,2\}$ then $\Xi_{d,1} = 1$ and $\Xi_{d,2} = 1$ as $1\times 1 + 2\times 1 = 3$.

$\Xi_{d,a}$ can only take one bed availability each day:

$$\sum_{a \in \{0,\ldots,M^A\}} \Xi_{d,a} = 1, \quad \forall d \in D \tag{18}$$

Finally, the variable $\Xi_{d,a}$ can be used as a look-up for the maximum number of patients with a 50% chance of being in the ward on any given day as determined by the parameter $F_a^{50}$, given by look-up index $a$. As a result, it is possible to impose an upper bound on the number of these patients each day so that the chance of exceeding the number of beds is kept below a given confidence. This can be achieved using the following constraint:

$$\bar{n}_d^{50} \leq \sum_{a \in \{0,\ldots,M^A\}} F_a^{50} \Xi_{d,a}, \quad \forall d \in D \tag{19}$$

In other words, as $\Xi_{d,a}$ can only take the value of 1 when it matches a single value for the available bed capacity $a \in \{0,\ldots,M^A\}$ on a given day and thus, the multiplication with $F_a^{50}$ will always be 0 unless $\Xi_{d,a} = 1$. This means that if we have two beds available on day 3, then $\Xi_{3,2} = 1$ (otherwise 0) and the sum can be simplified to $F_3^{50} \times 1 = 3 \times 1 = 3$, which is the upper bound on the number of patients with a 50% chance that day denoted with $\bar{n}_d^{50}$.

As the last step, the objective is defined to maximise the throughput of elective patients:

$$\max \sum_{(d,p,r) \in DPR} C_p x_{d,p,r} \tag{20}$$

where $C_p$ is the number of patients from pattern $p$.

## 4.2 Experimental Study

The purpose of the computational experiments is to determine how the size of each surgeon's waiting list and the length of the planning horizon affect the construction of the MSS in terms of allocation of surgeons to blocks, patient throughput and computational tractability. Additionally, the results of using an optimal MSS are compared with an actual MSS of GS speciality, where the allocation of the surgeons to blocks is fixed, in terms of throughput and surgeon allocation.

The experiments are performed on a 32GB memory Intel Core i7-7700 3.60 GHz with 4 cores. The model is programmed in Python 3.6 using Gurobi version 8.1.0. The time for each experiment is limited to 6 hours.

### 4.2.1 Instance Generation

Different instances are generated from the data set collected and described in Section 3.2. However, only the years 2017 and 2018 are used to estimate the frequencies of each surgeon's different types of operations, assuming that each operation had to be performed at least every other month to be considered. Different waiting list sizes consisting of $I_o \in \{10, 20, 30\}$ operations are sampled for each instance at the frequencies most likely to occur for each surgeon. This means that if a particular surgeon performs, e.g., a gallbladder operation 60% of the time, then in a sampled waiting list of ten patients, there would be approximately six patients awaiting gallbladder operation. Table 4.1 provides summary statistics for each surgeon for the data used in the sampling.

*Table 4.1. Summary of statistics for the most frequently performed type of operations by each surgeon at General Surgery*

| Surgeon | #Type of Operations | Surgery time* | | Ward LOS† | | Historical ratio of inpatients |
|---|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD | |
| 1 | 9 | 184 | 52 | 2.4 | 1.7 | 0.40 |
| 2 | 6 | 137 | 37 | 2.9 | 1.4 | 0.50 |
| 3 | 9 | 188 | 59 | 2.6 | 1.6 | 0.50 |
| 4 | 10 | 167 | 41 | 3.2 | 1.3 | 0.70 |
| 5 | 11 | 162 | 35 | 3.4 | 1.5 | 0.60 |
| 6 | 5 | 214 | 64 | 3.7 | 2.3 | 0.40 |
| 7 | 7 | 90 | 26 | 2.0 | 0.7 | 0.10 |
| 8 | 12 | 143 | 48 | 2.5 | 1.4 | 0.30 |
| 9 | 13 | 84 | 19 | 2.3 | 1.3 | 0.10 |

*Surgery time shown in minutes. †LOS is the length of stay in days in the ward following surgery. # denotes total number. SD is the standard deviation.

### 4.2.2 Parameter Settings

General Surgery is allocated to two blocks (ORs $r \in \{r_1,\ r_2\}$) each working day in the given MSS, where each block has a capacity ($C_{d,r}$) of 450 minutes. Other parameters are selected based on historical data. An upper limit on the number of patients assigned to each pattern is assumed to be $M^P = 6$. Moreover, the historical ratio of inpatients to the throughput of each surgeon ($h_o$) is shown in Table 4.1. The global ratio of inpatients to the total throughput for the speciality is $h_G = 0.38$. Exceeding the block capacity is not necessarily undesirable as two teams are available after hours for emergency surgeries as discussed in Section 3.3, and assumed to be $\delta \approx 0.30$.

Different cycle lengths are considered by using $D = \{7,\ 14,\ 28\}$ days. Nevertheless, blocks are only available during working days, but ward capacities must be considered during the weekends. For the comparison with the actual MSS, it is assumed that the

waiting list of each surgeon is $I_o = 30$ and the cycle length is $D = 7$ days.

#### Solution Verification

Each solution is verified with Monte-Carlo sampling using the complete undiscretised empirical distributions for LOS from historical data described in Section 3.4. This is important due to the nature of the approach, which makes it possible to exceed the number of staffed ward beds. Thus, for each simulated result, the following statistics are reported:

- **Risk of Overflow**: The simulated degree of exceeding the values of a given number of staffed ward beds $M^A$ each day (discretisation error).

### 4.2.3 Results

Table 4.2 summarises the results for the two solutions in Table 4.3. The actual MSS is used for the scheduling in the first solution meaning that the roster days of the surgeons are fixed to specific blocks of the week. The latter solution, the MSS, is optimised, allowing the surgeons' roster to be flexible. Consequently, the optimisation allocates the surgeon to blocks by assigning their patterns of patients resulting in the highest patient throughput. For the comparison, the waiting list of each surgeon $I_o$ consists of 30 patients and the planning horizon is $D = 7$ days. Note that the same set of patterns is used as input for the experiments, but different patterns may be scheduled and, thus, different patients.

The result shows that the number of scheduled patients is equivalent in both solutions (#Patients). However, using the optimised, MSS, the number of inpatients (#Inpatients) is increased by one. Regardless, even if an additional inpatient is scheduled when the optimal MSS is utilised, the overall risk of exceeding the number of staffed ward beds (Risk of Overflow) for all statistics reported is lower compared to the solution using the actual MSS, as shown in the table. However, both solutions are below a specified confidence level of 5%.

The allocation of surgeons to blocks is given in Table 4.3 for the optimal and actual MSS discussed above. The table shows that the surgeons are allocated to different blocks when the optimised MSS is utilised, except for surgeon 7, which is allocated to the same block. As in the actual MSS, the result suggests that surgeon 9 is allocated to two blocks. This could have been predicted, as this surgeon has the lowest average surgery time, average ward probability, and average LOS, as shown in Table 4.1, which is desirable when maximising the throughput of patients given the limited amount of staffed ward beds. Analysing the simulated risk of overflow for specific days reveals that the probability is generally smaller when the optimal MSS is utilised, despite the

*Table 4.2. Overall comparison of the total number of patients scheduled and the risk of exceeding the ward beds using an actual Master Surgical Schedule (MSS) and an optimised MSS.*

| | #Patients | #Inpatients | #Outpatients | Risk of Overflow* Median (%) | Mean (%) | Max (%) |
|---|---|---|---|---|---|---|
| Actual MSS† | 41 | 15 | 26 | 0.0 | 1.0 | 4.0 |
| Optimal MSS† | 41 | 16 | 25 | 0.0 | 0.3 | 1.0 |

Instance: † $(I_o, D) = (30, 7)$. $I_o$ is the number of patients on the waiting list $I_o$ of each surgeon $o$. $D$ is the number of days in the planning horizon * The risk of overflow is the likelihood of exceeding the staffed ward bed capacity.

number of inpatients is higher.

Results of alternating the planning horizon length ($D$) and the waiting list size ($I_o$) for each surgeon to generate a MSS are provided in Table 4.4. The results show that increasing $I_o$ enhances the overall throughput of patients as more scheduling possibilities become available. In addition, having more scheduling possibilities increases the number of patients with short surgery times on the waiting list. One might have expected that 82 patients would be scheduled for a two-week planning horizon ($D = 14$), as 41 patients were scheduled when the planning horizon was a week ($D = 7$), given that $I_o$ remains constant. However, only 78 patients were scheduled, indicating that the optimisation left out cases requiring longer surgery times when the planning horizon was a week to maximise the throughput. A similar effect is observed when four weeks ($D = 28$) are considered. The parameter settings also impact the computational time (CPU time) required to solve the MIP model. First, by alternating $D$ from 7 to 28 but keeping $I_o = 20$ constant the computational time increases from 28 to 948 s (34 times). Lastly, by increasing the number of patients of each surgeon by ten ($I_o = 30$) while alternating $D$ from 7 to 28, the problem becomes computationally challenging to solve and could not be solved optimally within the set time limits of six hours.

Optimal MSSs for the parameter settings $I_o = 30$ and $D = \{14, 28\}$ are shown in Tables 4.5 and 4.6. Similar to what was observed in Table 4.3, surgeon 9 is allocated two blocks each week when $D = 14$. However, by extending the planning horizon to $D = 28$, the surgeon is no longer allocated to two blocks each week. Instead, surgeon 7 is allocated two blocks in weeks I and IV, while surgeons 1 and 9 are allocated to two blocks in weeks II and III, respectively. As predicted, these surgeons, who have the shortest surgery times and LOS (see Table 4.1), complete almost all of their waiting lists within the planning horizon of $D = 28$ when the throughput of patients is maximised. Analysing the simulated ward results, we observe that the risk of exceeding the staffed ward beds surpasses the predefined confidence level on some days (see e.g. Table 4.6 on day 19), while being below the limits on most days.

In Table 4.7, a comparison is made between the historical ratio ($h_o$) and the scheduled ratio of inpatients to the throughput for each surgeon for multiple parameter settings. In addition, a comparison is provided to the global ratio and scheduled global ratio of

*Table 4.3. Comparison of the actual Master Surgery Schedule (MSS) and an optimised MSS. The table depicts the allocation of surgeons (1-9) to operating rooms ($r_1$ and $r_2$) for each day along with the ratio of inpatients in the parenthesis.*

| | **Actual MSS** | | | | | | | **Optimal MSS** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $r_1$ | 8 (2/3) | 6 (3/3) | 5 (1/3) | 4 (2/3) | 7 (1/5) | - | - | 1 (3/4) | 8 (0/3) | 2 (1/4) | 9 (1/6) | 7 (0/5) | - | - |
| $r_2$ | 9 (1/6) | 2 (1/4) | 3 (1/4) | 1 (2/4) | 9 (1/6) | - | - | 6 (3/3) | 5 (3/3) | 9 (1/6) | 3(1/4) | 4 (3/3) | - | - |
| $\Pr[w_d \geq M^A]$ | 0.00 | 0.00 | 0.03 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 |

Instance: $^{\dagger}$ $(I_o,\, D) = (30,\, 7)$. $I_o$ is the number of patients on the waiting list $I_o$ of each surgeon $o$. $d \in D$ are the days in the planning horizon. $Pr[w_d \geq M^A]$ is the simulated probability that the number of patients in ward $w_d$ on a given day $d$ exceeds the number of staffed ward beds $M^A$.

*Table 4.4. Comparison of the patient throughput under different parameter settings of planning horizon lengths and waiting list sizes for each surgeon.*

| $D$ | $I_o$ | #Patients | #Inpatients | #Outpatients | Duality Gap (%) | CPU MIP (sec) | |
|---|---|---|---|---|---|---|---|
| 7 | 10 | 35 | 13 | 22 | 0.00 | 1 | |
| | 20 | 39 | 14 | 25 | 0.00 | 28 | |
| | 30 | 41 | 16 | 25 | 0.00 | 202 | |
| 7 | 30 | 41 | 15 | 26 | 0.00 | 123 | **Actual MSS** |
| 14 | 10 | 60 | 22 | 38 | 0.00 | 4 | |
| | 20 | 72 | 27 | 45 | 0.00 | 320 | |
| | 30 | 78 | 29 | 49 | 0.00 | 986 | |
| 28 | 20 | 124 | 47 | 77 | 0.00 | 948 | |
| | 30 | 136 | 51 | 85 | 0.73 | 21600 | |

$D$ is the number of days in the planning horizon. $I_o$ is the number of patients on the waiting list $I_o$ of each surgeon $o$. # denotes the total number of patients. The duality gap is defined as the gap between the primal objective bound (*Primal*) and the dual objective bound (*Dual*) and calculated as $\frac{|\text{Primal}-\text{Dual}|}{|\text{Primal}|}$ by the mixed integer programming (MIP) solver (Gurobi Optimization, LLC, 2024). CPU is computational time.

inpatients to the total throughput for the same settings. It is evident that the scheduled ratio is often close to or at the historical ratio for each surgeon. However, with short planning horizons, it might be challenging to satisfy this ratio completely. Analysing the global ratio, one can observe that it is satisfied for all settings but with a slight deviation. However, the results suggest that planning beyond one week is required to fulfil the ratio of each surgeon when maximising the throughput. Additional restrictions could be added to better meet each surgeon's ratio and global ratio, such as by considering the types of operations, similar to Banditori et al. (2013); M'Hallah and Visintin (2019). However, it might come at the cost of the flexibility of the provided schedules and the throughput.

## 4.3 Conclusions

The problem considered in this chapter was to specify uncertainty in surgery times in a practical and statistically accurate way under a high throughput of elective patients and a limited number of staffed downstream ward beds. A practical and data-driven approach termed *Pattern Scheduling* was proposed, making it possible. The approach consists of two steps. In the first step, feasible patterns of patients were generated for each surgeon based on practical rules and probabilistic restrictions on overtime using historical data. In the second step, a MIP model was proposed to assign the feasible patterns to blocks so that the throughput was maximised while bounding on the risk of exceeding the staffed ward bed capacity in a probabilistic manner using a ward approximation, which

*Table 4.5. Optimal Master Surgery Schedule for a planning horizon length of* 14 *days with* 30 *patients on the waiting list of each surgeon.*

| | **Week I** | | | | | | | **Week II** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| $r_1$ | 5 (2/3) | 1 (3/4) | 6 (2/3) | 9 (1/6) | 9 (1/6) | - | - | 1 (3/4) | 8 (1/3) | 2 (1/4) | 3 (1/3) | 9 (1/5) | - | - |
| $r_2$ | 8 (1/3) | 3 (1/4) | 4 (2/3) | 7 (0/5) | 2 (1/4) | - | - | 4 (2/3) | 7 (2/4) | 6 (2/3) | 9 (0/5) | 5 (2/3) | - | - |
| $\Pr[w_d \geq M^A]$ | 0.00 | 0.01 | 0.10 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.06 | 0.08 | 0.01 | 0.09 | 0.00 | 0.00 |

Instance:$(I_o,\ D) = (30,\ 14)$. See Table 4.3 for full description.

*Table 4.6. Sub-Optimal Master Surgery Schedule for a planning horizon length of 28 days with 30 patients on the waiting list of each surgeon.*

| | **Week I** | | | | | | | **Week II** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| $r_1$ | 5 (3/3) | 1 (1/3) | 6 (1/3) | 9 (1/5) | 8 (1/3) | - | - | 5 (2/3) | 2 (1/4) | 1 (2/4) | 3 (2/2) | 9 (0/6) | - | - |
| $r_2$ | 7 (0/4) | 4 (2/3) | 2 (2/3) | 7 (0/4) | 3 (1/3) | - | - | 7 (2/4) | 9 (1/5) | 6 (1/3) | 8 (0/3) | 4 (2/3) | - | - |
| $\Pr[w_d \geq M^A]$ | 0.00 | 0.00 | 0.02 | 0.00 | 0.03 | 0.00 | 0.00 | 0.06 | 0.01 | 0.09 | 0.02 | 0.02 | 0.00 | 0.00 |

| | **Week III** | | | | | | | **Week IV** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| $r_1$ | 8 (2/3) | 1 (1/3) | 5 (1/3) | 1 (2/3) | 4 (2/2) | - | - | 8 (1/3) | 6 (1/3) | 4 (2/3) | 9 (1/5) | 5 (1/3) | - | - |
| $r_2$ | 3 (2/3) | 6 (1/3) | 2 (2/2) | 9 (0/5) | 7 (1/4) | - | - | 1 (2/3) | 3 (1/3) | 7 (0/5) | 7 (1/4) | 2 (2/2 | - | - |
| $\Pr[w_d \geq M^A]$ | 0.02 | 0.01 | 0.06 | 0.07 | 0.23 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.05 | 0.01 | 0.00 |

Instance:$(I_o,\ D) = (30,\ 28)$. See Table 4.3 for full description.

*Table 4.7. Comparison of the historical inpatient ratio to the scheduled inpatient ratio for each surgeon under different planning horizon lengths and waiting list sizes.*

| | | | **D = 7** | | | **Actual ↓** | **D = 14** | | | **D = 28** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Surgeon** | **$h_o$** | **$I_o$ →** | **10** | **20** | **30** | **30** | **10** | **20** | **30** | **20** | **30** |
| 1 | 0.40 | | 0.75 (4) | 0.25 (4) | 0.75 (4) | 0.50 (4) | 0.57(7) | 0.86 (7) | 0.75 (8) | 0.58 (12) | 0.50 (16) |
| 2 | 0.50 | | 0.67 (3) | 0.25 (4) | 0.25 (4) | 0.25 (4) | 0.50 (8) | 0.33 (6) | 0.25 (8) | 0.46 (13) | 0.64 (11) |
| 3 | 0.50 | | 0.50 (2) | 0.33 (3) | 0.25 (4) | 0.25 (4) | 0.50 (4) | 0.33 (6) | 0.29 (7) | 0.60 (10) | 0.55 (11) |
| 4 | 0.70 | | 0.50 (2) | 1.00 (3) | 1.00 (3) | 0.67 (3) | 0.50 (4) | 0.67 (6) | 0.67 (6) | 0.70 (10) | 0.73 (11) |
| 5 | 0.60 | | 1.00 (2) | 0.33 (3) | 1.00 (3) | 0.33 (3) | 0.57 (7) | 0.33 (6) | 0.67 (6) | 0.54 (13) | 0.58 (12) |
| 6 | 0.40 | | 0.33 (3) | 0.67 (3) | 1.00 (3) | 1.00 (3) | 0.20 (5) | 0.33 (6) | 0.67 (6) | 0.20 (10) | 0.33 (12) |
| 7 | 0.10 | | 0.00 (5) | 0.50 (4) | 0.00 (5) | 0.20 (5) | 0.00 (9) | 0.25 (12) | 0.22 (9) | 0.20 (20) | 0.16 (25) |
| 8 | 0.30 | | 0.50 (4) | 0.25 (4) | 0.00 (3) | 0.67 (3) | 0.67 (6) | 0.43 (7) | 0.33 (6) | 0.31 (16) | 0.33 (12) |
| 9 | 0.10 | | 0.10 (10) | 0.18 (11) | 0.17 (12) | 0.17 (12) | 0.10 (10) | 0.19 (16) | 0.14 (22) | 0.15 (20) | 0.12 (26) |
| **$h_G$ →** | 0.38 | | 0.37 (35) | 0.36(39) | 0.39(41) | 0.37(41) | 0.37(60) | 0.38(72) | 0.37 (78) | 0.38 (124) | 0.38 (136) |

The number of scheduled inpatients is indicated in the parenthesis. The global historical ratio of inpatients of the speciality is compared to the scheduled global ratio of inpatients for each setting at the bottom. $D$ is the planning horizon length. $h_o$ is the historical ratio of inpatients for each surgeon $o$. $I_o$ is the number of patients on the waiting list of each surgeon $o$. $h_G$ is the historical global ratio of inpatients.

assumed only three possible scenarios.

Several cyclic MSSs were built using the Pattern Scheduling approach, where various planning horizon lengths and different waiting list sizes of each surgeon were considered. The objective was to determine the optimal allocation of surgeons to blocks each week, which maximises the patient throughput while considering limited staffed downstream ward beds in a probabilistic manner. The computational results demonstrated that the allocation of surgeons to blocks, in a planning horizon of one week, was fundamentally different compared to an actual MSS. Although the same throughput of patients was achieved, the number of inpatients was higher using the optimised MSS, and the risk of exceeding the staffed ward beds was lower. Furthermore, the results suggest that considering a longer planning horizon and allowing flexibility in the roster of the surgeons are beneficial to maintain a balanced flow of in- and out-patients. However, longer planning horizons and larger waiting lists increase computational time.

In this chapter, the focus was on building an MSS which maximises patient throughput. However, as demonstrated by the computational results, patients with shorter surgery times and LOS in the ward may be favoured over more challenging patients, particularly for short planning horizons, which is not a desirable outcome. To address this, additional criteria such as patient priority, is necessary.

While the computational results showed promising results in bounding the risk of exceeding the staffed ward beds using the ward approximation, some days exceeded a predefined probability limit. Therefore, a strategy that accepts more than three scenarios for the probability of being in the ward on any given day following surgery is necessary and will make it possible bounding the risk of last-minute cancellations with greater accuracy.

## 4.4 Summary

In this chapter, uncertainty in surgery times was specified in a practical and statistically accurate way. A novel two-step practical approach termed Pattern Scheduling was proposed, making it possible while considering the limited number of staffed ward beds in a probabilistic manner. The result showed that by specifying practical rules, relatively large instances could be solved optimally and that scheduling beyond one week is required as patients with long surgery times and long LOS were left out when a one-week planning horizon was considered. Although the computational results were promising, some days had a high risk of exceeding the staffed ward beds, which may cause last-minute cancellations.

The following chapter aims to address this limitation, reducing the combined risk of last-minute cancellations due to surgery times and LOS in the ward with greater accuracy.

This will enable elective patients to be assigned to blocks weeks in advance while maintaining the existing utilisation.

# 5 Uncertainty in Length of Stay

This chapter considers how to reduce the combined risk of last-minute cancellations due to uncertainty in surgery times and LOS in the ward when operating close to the maximum capacity of operating rooms and staffed ward beds. The objective is to determine how this can be achieved when scheduling elective patients to blocks weeks in advance while maintaining the existing utilisation. This chapter is based on the work published in Sigurpalsson et al. (2022).

In the last chapter, a novel two-step approach, termed *Pattern Scheduling* was proposed. In the first step, feasible patterns, consisting of one or more patients, were generated with feasibility determined by practical rules and probabilistic restrictions on overtime. In the second step, the patterns were assigned to blocks using a MIP model so that the overall throughput of patients was maximised while taking into account the limited number of staffed ward beds. The chapter proposed a ward approximation assuming only three possible scenarios for each ward patient, with only one scenario used each day to hedge against the risk of exceeding the staffed ward beds in the MIP model. The computational results were promising, but some days exceeded the staffed ward beds with a high probability.

This significant limitation is considered in this chapter so that the combined risk of last-minute cancellations associated with exceeding the block capacity and the limited staffed ward beds is further reduced. To address this, Ward Combinations (WCs) are proposed, where the risk of exceeding ward bed capacity is limited to a specified probability but with a lower discretisation error than in the previous chapter, as more than three scenarios can be used. That is to say, empirical distributions for the LOS for each type of operation can be computed from historical data. The probability of exceeding the available staffed ward beds for any given WC of patients with different probabilities of a stay per day can be estimated using Monte Carlo simulation. Therefore, the MIP model will consider every possible patient WC while excluding those with a predetermined risk of exceeding the staffed ward bed capacity. The approach is compared to an equivalent robust formulation, which assumes the worst-case LOS. Consequently, this approach is anticipated to result in less overtime for the same patient throughput. Actual scheduling data for one arbitrarily selected month is compared with the results of models utilising the WCs and another that uses a robust formulation.

As noted in Section 2.2.2, only a handful of studies account for the combined risk

of last-minute cancellations due to uncertainty in surgery times and LOS. The work proposed in this chapter can be seen as an extension to the work of Jebali and Diabat (2017), where chance constraints are implemented to reduce the risk of exceeding the downstream ICU beds. Unlike Jebali and Diabat (2017), the combined risk of exceeding staffed ward beds and exceeding block capacity is considered. Additionally, the risk of exceeding regular (Hans et al., 2008) and extended (Shylo et al., 2013) overtime is considered by minimising the number of blocks selected with either.

The chapter is organised as follows. The following section presents the model development, which includes a description of the problem studied, a model using the WCs and an equivalent robust formulation assuming the worst-case LOS. This is followed by an experimental section that compares results from the different models with actual scheduling data. Furthermore, various parameter settings for the proposed models are also analysed. The chapter concludes with a summary of the main findings.

## 5.1 Model Development

The general problem solved in this chapter is assigning a high throughput of in- and out-patients from the waiting list of each surgeon to blocks weeks in advance under a limited number of staffed ward beds. However, the patients must be scheduled in a way that reduces the combined risk of last-minute cancellations, that can occur on the day of surgery in advance, due to uncertainty in surgery times and uncertainty in LOS while maintaining the existing utilisation. Additionally, the number of blocks resulting in regular and extended overtime must be minimised as well. The following assumptions are made to formulate the problem:

- Medical priorities are formulated as hard constraints. In other words, if a patient has a one-week medical priority, it must be satisfied.
- In contrast to the problem addressed in Chapter 4, it is assumed that the patients to be scheduled for the planning horizon have already been selected by the hospital, and thus, the main objective is to schedule the patients to blocks in a way that reduces the risk of last-minute cancellations. In addition, this assumption allows for a direct comparison with actual scheduling data as the same patients are scheduled for the planning horizon.
- It is assumed that it is known with certainty in advance whether or not a patient requires an ICU and ward admission.

To solve the problem, the two-step *Pattern Scheduling* approach, as presented in Chapter 4, is employed but with some modifications. In the first step, a modified set of practical rules is employed to generate feasible patterns. The setup of this step is

presented in Section 4.1.1. Moreover, overtime statistics are collected for each feasible pattern and are used in the MIP model to determine whether a pattern has regular or extended overtime. In the second step, a MIP model is proposed, where patterns are assigned to blocks while bounding the risk of exceeding the number of staffed ward beds using WCs while minimising the number of patterns selected with regular and extended overtime:

2. *Ward Combination Optimisation*: Given a fixed number of available staffed ward beds, consider all combinations of patient numbers $n_k$ with the discretised probability of stay $p_k$, for $k \in |K|$ where the probability of stay following surgery is discretised into $|K|$ scenarios and dependent on the patient's type of operation. Each such WC is then eliminated if the total patient number exceeds the staffed ward bed limit by a probability $\Omega$. This is computed using Monte Carlo sampling. Given the set of feasible patterns and WCs, a deterministic MIP model is solved using a commercial solver. This is followed by a verification of the solution by Monte Carlo sampling using the complete, undiscretised, empirical distribution for the LOS in the ward.

Sections 5.1.1 and 5.1.2, describe each of these steps in further detail. The second step is reformulated using a robust optimisation described in Section 5.1.3.

### 5.1.1 Pattern Generation

The procedure described in Section 4.1.1 is applied to generate the feasible patterns. However, a modified set of practical rules is employed to determine their feasibility. That is to say, a practical rule maintaining a balance between in- and outpatients within a pattern is removed, and instead, a new rule imposing quota on the ICU admissions per pattern is introduced.

As before, let $z_{i,p}$ be a binary decision variable, taking the value 1 if patient $i$ is assigned to the pattern $p$, otherwise 0. Then, the following set of practical rules is implemented:

1. *Patient Quota*: Let $M^p$ be a parameter specifying the maximum number of patients assigned to each pattern $p$. This upper bound can be imposed using the following constraint:
$$\sum_{i \in I_o} z_{i,p} \leq M^P, \quad \forall\, p \in P, \quad o \in O \tag{3}$$

2. *ICU Quota:* ICU beds are commonly a scarce resource at hospitals. One can reduce the risk of last-minute cancellations by imposing daily ICU quotas (Kim and Horowitz, 2002). Let $g_i^{ICU}$ be a parameter taking the value 1 if patient $i$ require an ICU admission following surgery (0 otherwise) then one can impose

quota on the number of ICU patients ($M^{ICU}$) in a pattern as follows:

$$\sum_{i \in I_o} g_i^{ICU} z_{i,p} \leq M^{ICU}, \quad \forall p \in P, \quad o \in O \tag{21}$$

3. *Overtime Verification:* A pattern $p$ is feasible as long as the risk of exceeding the block's capacity $C_{d,r}$ is no more than $\delta$. The formulation is given with Equation (5) as presented in Section 4.1.1 and shown here below:

$$\Pr\left[\sum_{i \in I_o} S(i) z_{i,p} \geq C_{d,r}\right] \leq \delta, \quad \forall p \in P, \quad o \in O, \quad r \in R, \quad d \in D \tag{5}$$

Monte Carlo sampling is used with historical data to solve this equation. Patterns that are not feasible towards overtime are eliminated from the set $P$.

- *Exception:* As discussed in Section 4.1.1, patterns consisting of a single patient are not eliminated even if the pattern may span the whole day and exceed the limits of $\delta$. This is important to make sure that all patients are assigned to at least one pattern.

At this point, feasible patterns have been generated and can be utilised in the scheduling model to be presented in Section 5.1.2. In addition, output statistics are collected for each feasible pattern, utilised as parameters in the model. The first parameter is the probability of regular overtime $\delta_p$ calculated as follows:

$$\delta_p = \Pr[\sum_{i \in I_o} S(i) z_{i,p} \geq C_{d,r}], \quad \forall p \in P, \quad o \in O, \quad r \in R, \quad d \in D \tag{22}$$

Regular overtime is the simulated probability that a pattern surpasses its regular block capacity as defined in Section 2.2.1. In a similar way, statistics on extended overtime are collected (Shylo et al., 2013) denoted by $\delta_p^{\Delta}$ and calculated as follows:

$$\delta_p^{\Delta} = \Pr[\sum_{i \in I_o} S(i) z_{i,p} \geq C_{d,r} + \Delta_{d,r}], \quad \forall p \in P, \quad o \in O, \quad r \in R, \quad d \in D \tag{23}$$

where $\Delta_{d,r}$ is the time added to extend the block capacity. Thus, extended overtime is defined as the probability that a pattern exceeds its extended block capacity as defined in Section 2.2.1. The model provided in the following section utilises these parameters to determine the number of patterns selected with regular and extended overtime.

For each feasible pattern, a ward bed occupancy is generated from historical data as described in Section 4.1.1, where a ward approximation was applied by assuming only three possible ward scenarios for each ward patient. In the following section, WCs are proposed and applied as a replacement for the ward approximation. In other words, the ward approximation is generalised to consider more scenarios, resulting in lower discretisation error.

### Ward Combinations

From historical data, it is possible to calculate the probability that a patient is in the ward on a given day following surgery, given the patient's type of operation. By summing the individual probabilities of each patient in the ward each day, it is possible to estimate the ward bed occupancy as discussed in Section 4.1.2. However, this approach does not consider bounding the risk of exceeding the staffed ward beds, and thus, the risk of last-minute cancellations is possible. As a result, a ward approximation was proposed to hedge against last-minute cancellations by allowing bounding the risk of exceeding the bed capacity to a specified threshold. Even if the results were promising, some days exceeded the threshold. As a result, this approach is generalised by accounting for more than three scenarios, which minimises the discretisation error and, hence, it allows bounding the risk of last-minute cancellations with greater accuracy.

Let us assume that there are $k \in K$ number of scenarios possible for each patient in the ward on any given day after surgery. Each day, a patient follows one scenario with an associated discrete probability $p_k$, determining the likelihood of staying in the ward that day. Suppose $n_k$ denotes the number of patients with the probability $\rho_k$ of staying in the ward on a given day. In that case, it is possible to approximate the total number of patients on a given day by the sum of independent Binomial distributions. In other words, the total number of patients in the ward on a given day is the sum of the products of $n_k$ and $\rho_k$ for all scenarios $k \in K$.

One way to determine $\rho_k$ is to divide the continuous range of probabilities, from 0 to 1, into $k \in K$ intervals using the following formula:

$$p_k = \frac{k-1}{|K|-1}, \quad \forall k \in K \tag{24}$$

This means that if the number of scenarios is $|K| = 5$, then the possible ward probabilities for each patient in the ward on any given day after surgery are $\rho_1 = 0.00$, $\rho_2 = 0.25$, $\rho_3 = 0.50$, $\rho_4 = 0.75$ and $\rho_5 = 1.00$. As described in Section 4.1.1, patients having a 0% chance of being in the ward on any given day ($\rho_1 = 0.00$) can be excluded. Additionally, patients with a 100% chance of staying on a given day ($\rho_{|K|} = 1.00$) are used to determine the number of staffed ward beds available for the patients with the remaining probabilities given a fixed bed capacity each day. For example, if there are six beds available in the ward on a given day and four patients with a 100% chance of staying that day, then two beds are available for patients with other probabilities. As a result, the problem becomes determining the number of patients $n_k$ with probability $p_k$ of staying in the ward on a given day for the scenarios $k \in K'$ without exceeding the available staffed ward beds to reduce the risk of last-minute cancellations. The set $K' = \{k \in K |\ 2 \leq k \leq |K| - 1\}$ excludes the scenarios where the patients have either a 0% chance or a 100% chance of being in the ward following surgery.

One way to solve this problem is to pre-generate all possible WCs of patients $n_k$ having probability $p_k$ of staying in the ward on a given day for $k \in K'$. This can be mathematically expressed as follows. Suppose $l \in L$ denotes the set of indexes for

the WCs. The number of patients in the ward for each WC is defined as the sum of independent Binomial distributions or:

$$W(l) \sim \sum_{k \in K'} \mathrm{B}\big(n_k(l), \rho_k\big), \quad \forall\, l \in L \tag{25}$$

where $l$ is an index to a particular WC. However, the fundamental challenge is generating appropriate indexing uniquely expressing different numbers of patients $n_k$ and available staffed ward bed capacities $a \in A$ where $A = \{0, \ldots, M^A\}$. One way to address this problem is to use a base-$|A|$ encoding expressed with the following equation:

$$l = \sum_{k \in K'} n_k |A|^{|K'|-k+1} \tag{26}$$

For example, if the number of staffed ward beds $M^A$ is six and $|K| = 5$, then a WC consisting of $(n_2, n_3, n_4) = (4, 1, 3)$ patients is given the index $l = 206$[2].

The number of possible WCs will grow exponentially with the available staffed ward beds $a \in A$ and the number of scenarios $K$. Regardless, given the objective of reducing the number of last-minute cancellations, only some WCs are feasible since many of them have a high probability of exceeding the number of staffed ward beds. In other words, if a WC exceeds the number of available staffed ward beds with a probability higher than $\Omega$, there is a risk of last-minute cancellations. Thus, each WC, indexed with $l \in L$, is verified with Monte-Carlo sampling towards all possible numbers of available beds $a \in A$ where feasible WC must satisfy the following condition:

$$\Pr[W(l) \geq a] \leq \Omega, \quad \forall l \in L,\ a \in A \tag{27}$$

This process results in the binary parameter, $F_{l,a}$, taking the value 1 if WC $l$ is feasible when there are $a$ staffed ward beds available. Otherwise, $F_{l,a}$ is 0. A constraint utilises this parameter in the MIP model, which is presented in Section 5.1.2, and makes it possible to bound the risk of exceeding the staffed ward beds to a specified probability.

To make it possible to use the WCs in the MIP model, $|K-1|$ ward distributions are generated for each feasible pattern, which is then implemented in the scheduling model to determine the number of patients $n_k$ of each scenario $k \in K$ each day. Consequently, each patient's empirical daily ward probabilities in a pattern $p$ must be discretised to the values of $\rho_k$. The discretisation can be performed by multiplying the ward probabilities with the number of scenarios $|K|-1$, rounding the outcome to the nearest integer, and dividing by $|K|-1$. For example, if $|K| = 5$ and the empirical ward probability of a single patient is 0.65 on a given day, the value is discretised to the value $\rho_4 = 0.75$. This process results in the distributions $Q_{j,k,p}$ where $j \in J$ are the days after surgery, $k \in \{2, \ldots, K\}$ are the number of scenarios but excluding the first scenario as these patients have a 0% chance of being in the ward and $p \in P$ patterns.

Figure 5.1, shows how an empirical distribution with continuous probabilities (grey bars) for an arbitrarily selected type of operation has been discretised into $K = |5|$

[2] $4 \cdot 7^{3-2+1} + 1 \cdot 7^{3-3+1} + 3 \cdot 7^{3-4+1} = 196 + 7 + 3 = 206$.

scenarios, as illustrated by the dashed line. This example is similar to what was shown in Figure 4.1 in the previous chapter but applied to actual data. However, in this example, the number of scenarios has been increased from 3 to 5, making the approach more statistically accurate than the ward approximation. In this example, the discretisation assumes that for the day of the surgery and the day after (0 and 1), the patient will be in the ward with probability $\rho_5 = 1$. After the ninth day, the patient has left the ward, and the probability is $\rho_1 = 0$. Thus, the continuous probabilities each day are discretised to the values of the dashed line. In this example, we are only interested in the scenarios between the second and last, $k \in K' = \{2, \ldots, 5\}$. The number of patients in the final level (with $\rho_5 = 1$) determines the daily availability of staffed wards in the MIP model for the remaining levels, based on the overall bed capacity.

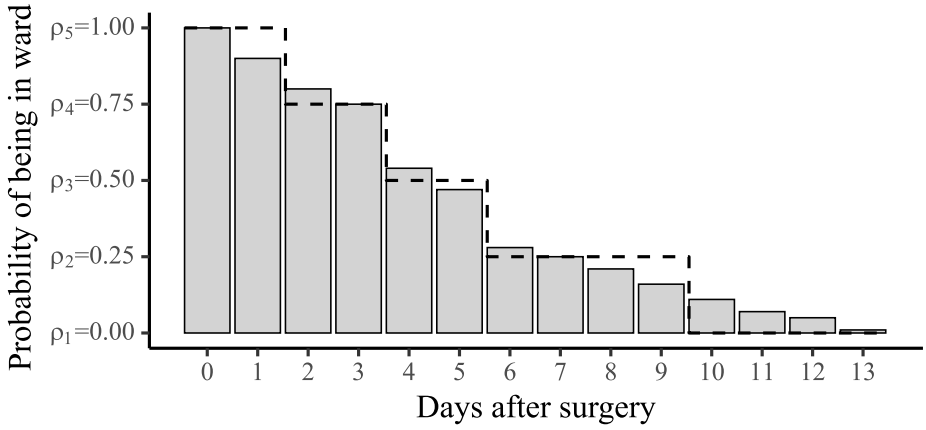



*Figure 5.1. Discretisation of an empirical distribution of daily ward probabilities into five scenarios, each with an associated probability.*

The daily ward probabilities, represented by the bars, are discretised to the values indicated by the dashed line.

### 5.1.2 Ward Combinations Optimisation

The scheduling problem has now been reduced to assigning the feasible patterns to blocks so that the total number of patterns with regular or extended overtime is minimised while bounding the risk of last-minute cancellations due to ward bottlenecks. Thus, the assignment of patterns to blocks is subject to the feasibility of the resulting WC each day. Let $x_{d,p,r}$ be a binary decision variable, taking the value 1 if a pattern $p$ is assigned to the day $d \in D$ and room $r \in R$; 0 otherwise. As pointed out in Section 4.1.2, the assignment of patterns to blocks is subject to many restrictions. For example, this can include the availability of the surgeons and the block's capacity for that day. As a result, the reduced set $(d, p, r) \in DPR \subseteq D \times P \times R$ is generated taking the following restrictions into account:

- Availability of the surgeons for each block.
- Availability of patients and their priorities.
- The feasibility of the patterns for a given block, is dependent on the block's capacity for a given day and room ($C_{d,r}$).

Assuming that each pattern $p$ spans the whole day, only one may be assigned to a block:

$$\sum_{p\in P, r\in R:(d,p,r)\in DPR} x_{d,p,r} \leq 1, \quad \forall d \in D \tag{6}$$

and each surgeon can only be working according to at most one pattern each day:

$$\sum_{p\in P_o, r\in R:(d,p,r)\in DPR:\ P_o\subseteq P} x_{d,p,r} \leq 1, \quad \forall d \in D,\ o \in O \tag{7}$$

Furthermore, each patient $i$ may be present in multiple patterns but can only be scheduled once:

$$\sum_{(d,p,r)\in DPR: i\in I_p} x_{d,p,r} = 1, \quad \forall i \in I \tag{28}$$

where $I_p$ is the set of patients included in pattern $p$. It is assumed that all patients must be scheduled so the throughput is fixed. A daily quota is imposed on the total number of ICU admissions to hedge against last-minute cancellations (Kim and Horowitz, 2002). Therefore, at most $\bar{M}^{ICU}$ can be admitted to the ICU each day:

$$\sum_{p\in P, r\in R:(d,p,r)\in DPR} n_p^{ICU} x_{d,p,r} \leq \bar{M}^{ICU} \quad \forall d \in D \tag{29}$$

where $n_p^{ICU}$ denotes the number of ICU patients in pattern $p$.

The focus now is on bounding the risk of exceeding the limited number of staffed ward beds, which is known to result in last-minute cancellations. For any given schedule, defined by the decision variable $x_{d,p,r}$, it must be ensured that the combinations of patients with the probability $\rho_k$ of being in the ward on a given day are feasible.

This can be accomplished by first linking the actual WCs of patients each day to a corresponding WC, which is pre-generated, as discussed in Section 5.1.1. In other words, if there are three actual ward patients, each with a 75% chance of being in the ward on a given day, then that combination must be matched to a pre-generated WC with three patients with a 75% chance of being in the ward. Lastly, each WC must be linked to the number of available beds each day ($a_d$) to determine if the WC is feasible for a given bed capacity. One can combat this effect by adding several constraints to the model.

As discussed and illustrated in Figure 5.1, the patient's scenarios of interest are the ones between the first and last ($k \in K'$). That is, it is assumed that the patients in the last scenario ($k = |K|$) will be with a 100% certainty ($\rho_{|K|} = 1.00$) in the ward on some day

following surgery while the patients in the first scenario ($k = 1$) have a 0% of being in the ward ($\rho_1 = 0.00$). Assuming the number of staffed ward beds is fixed each day, the number of patients in the last scenario can determine the number of beds available for the patients in other scenarios. At the same time, there is no need to consider the number of patients in the first scenario, as they will have a 0% chance of being in the ward. Formally, for a fixed capacity of staffed ward beds each day ($M^A$) and a schedule defined by $x_{d,p,r}$, the number of available staffed ward beds ($a_d$) for the remaining scenarios ($k \in K'$) can be computed by subtracting the total number of patients with $\rho_{|K|} = 1.00$ from $M^A$ each day. If $Q_{j,\ |K|,\ p}$ is a parameter denoting the number of patients from the pattern $p$ in the ward on the day $j$ after surgery from the scenario $k = |K|$, then the number of staffed ward beds available each day for the remaining scenarios can be calculated as follows:

$$a_d = M^A - \Big(\bar{n}_{d,|K|} + \sum_{\substack{r\in R,\ p\in P,\ j\in\{0,1,\ldots,M^W-1\}:\\ (d-j,\ p,\ r)\in DPR}} Q_{j,\ |K|,\ p} x_{d-j,p,r}\Big), \quad \forall d \in D \tag{30}$$

where $M^W$ denotes the upper bound on LOS in the ward in days for each patient. By multiplying the decision variable $x_{d-j,p,r}$ with the sum of $Q_{j,|K|,p}$ each day, it is possible to find the total number of patients with a 100% certainty of being in the ward on the day $d$. However, some patients from previous planning horizons may still be in the ward during the current one, which impacts the availability of beds. As such, the parameter $\bar{n}_{d,|K|}$ is utilised to determine the number of patients with $\rho_{|K|} = 1.00$ each day from previous planning horizons and estimated using the previous weeks' known schedule. Due to the limitations on the number of staffed ward beds, the following constraint is used to impose an upper bound on the number of available beds each day:

$$a_d \leq M^A, \quad d \in D \tag{31}$$

At this point, the problem is determining if the actual combinations of patients on the day $d$ is feasible towards $a_d$. As the first step, the number of patients $n_{d,k}$ from each scenario $k \in K'$ with probability $\rho_k$ of being in the ward on the day $d$ after surgery must be computed:

$$n_{d,k} = \bar{n}_{d,k} \quad + \sum_{\substack{r\in R,\ p\in P,\ j\in\{0,1,\ldots,M^W-1\}:\\ (d-j,\ p,\ r)\in DPR}} Q_{j,k,p} x_{d-j,p,r}, \quad \forall d \in D, \quad k \in K' \tag{32}$$

where $Q_{j,k,p}$ denotes the number of patients from interval class $k$ with the probability $\rho_k$ of being in the ward on day $j$ after surgery. Similarly, $\bar{n}_{d,k}$ is a parameter specifying the number of patients in each scenario $k$ on the $d$ still in the ward from previous planning horizons.

To connect the actual combination of patients each day to a corresponding WC, the binary decision variable $y_{d,l}$ is introduced, taking the value 1 if WC $l$ is utilised on day $d$; 0 otherwise. To link the pre-generated WC to the actual combinations of patients each day, a base-$|A|$ decoder is constructed, as presented with Equation (26) in the form

of the following constraint:

$$\sum_{l \in L} l y_{d,l} = \sum_{k \in K'} n_{d,k} |A|^{|K'|-k+1}, \quad \forall\, d \in D \tag{33}$$

This constraint may be thought of as searching for a specific row in a table. The right-hand side decodes the combinations of $n_{d,k}$ into a specific row number, corresponding to the settings of WC $l$. This means that if $M^A = 6$, $|K| = 5$, then an actual combination of patients on a given day consisting of $(n_2, n_3, n_4) = (4, 1, 3)$ provides the index 206. Thus, the left-hand side of the constraint can only be satisfied when it takes the same value. However, this constraint does not consider how many WC there are each day. Consequently, the left-hand side is satisfied for any sum that equals the right-hand side. For example, if the left-hand side is 206, then the right-hand side is, e.g. 206 when $l = \{3, 203\}$. As a result, an additional constraint is implemented in the model so that only one WC is utilised each day in the planning horizon:

$$\sum_{l \in L} y_{d,l} = 1, \quad \forall\, d \in D \tag{34}$$

At this point, the actual combinations of patients each day have been connected to a corresponding WC, and thus, the final step is to determine if the WC $l$ assigned to the day $d$ is feasible. Firstly, the WC utilised on day $d$ must be connected to the available beds $a_d$ as computed with Constraint (30). Let $z_{d,a}$ be a binary decision variable, taking the value 1 if on the day $d \in D$ there are $a \in A$ beds available, 0 otherwise. To link $z_{d,a}$ to $a_d$, the following constraint is used:

$$\sum_{a \in A} a z_{d,a} = a_d, \quad \forall\, d \in D \tag{35}$$

The right-hand side of the constraint is satisfied for any sum of ward bed capacities that equals the left-hand side as previously discussed. However, there is only one possible availability of beds each day, and thus, $z_{d,a}$ can only take one value each day:

$$\sum_{a \in A} z_{d,a} = 1, \quad \forall\, d \in D \tag{36}$$

Finally, it is possible to determine if the WC on day $d$ for a given bed availability $a$ is feasible:

$$y_{d,l} \leq \sum_{a \in A} F_{l,a} z_{d,a}, \quad \forall d \in D, \quad l \in L \tag{37}$$

where $F_{l,a}$ is a binary parameter taking the value 1 if a WC $l$ is feasible towards the risk of last-minute cancellations specified by $\Omega$, else 0 for a given number of beds $a$ as presented in Section 5.1.1. To summarise, the purpose of Constraints (32)– (37) is to guide the assignments of the patterns so that the actual combination of patients in the ward each day is feasible.

The objective function will now be presented. As bounds have been placed on exceeding each block's capacity by generating feasible patterns and ward bed capacity with the

WCs, the objective function is to minimise the number of patterns selected with regular and extended overtime and the amount of overtime to strike a balance in the block utilisation.

Let $u_{d,r}$ be a binary indicator variable taking the value 1 if $\delta_p > \delta'$, 0 otherwise, as posed by the following constraint:

$$\sum_{(d,p,r)\in DPR:\delta_p>\delta'} x_{d,p,r} \leq u_{d,r} \tag{38}$$

Similarly, let $v_{d,r}$ be a binary indicator variable taking the value 1 if $\delta_p^{\Delta} > \delta^{\Delta'}$, 0 otherwise, as posed by the following constraint:

$$\sum_{(d,p,r)\in DPR:\delta_p^{\Delta}>\delta^{\Delta'}} x_{d,p,r} \leq v_{d,r} \tag{39}$$

In this context, the variables $u_{d,r}$ and $v_{d,r}$ determine how often the probabilities of the selected patterns surpass the accepted risk of entering regular overtime ($\delta'$) and extended overtime ($\delta^{\Delta'}$) respectively. Therefore, the objective function is to minimise the number of patterns selected that result in regular or extended overtime. Moreover, the degree of surpassing the accepted risk limits is penalised by minimising the squared probabilities $\delta_p$ and $\delta_p^{\Delta}$:

$$\min \quad \sum_{d\in D,r\in R} \left(u_{d,r} + w v_{d,r}\right) \tag{40}$$

$$+ \sum_{(d,p,r)\in DPR} \left([\delta_p]^2 + w[\delta_p^{\Delta}]^2\right) x_{d,p,r} \tag{41}$$

where a weight $w \gg 1$ is posed on extended overtime terms.

### 5.1.3 Robust Ward Optimisation

As discussed in Sections 2.2.1–2.2.2, all distributional information about the LOS is ignored when RO is used. Instead, capacity constraints are added to protect against the worst-case realisation of uncertainty. The remaining challenge is characterising the worst-case outcome, the so-called uncertainty set reflecting each patient's worst-case outcome. In practice, the hospital staff, e.g., surgeons, may be able to determine the worst-case outcome LOS, e.g., if the patient's conditions are known beforehand. Thus, a probabilistic guarantee for feasibility can be made depending on the risk attitude of the hospital.

Suppose the decision maker is very conservative and wants a $\omega$ guarantee that the number of patients in the ward each day does not exceed the number of staffed ward beds $M^A$. Let $\rho'_{i,d}$ be the probability that patient $i$ is in the ward on the day $d$. Hence,

the worst-case realisation each day must satisfy the following constraint,

$$\bar{n}_d^{\omega} + \sum_{\substack{p\in P, j\in\{0,\dots,M^W-1\}, r\in R:\\ ((d-j),p,r)\in DPR}} x_{d-j,p,r}\Big(\sum_{i\in I_p} \mathbb{1}_{\omega\leq\rho'_{i,j}}\Big) \leq M^A, \quad \forall d \in D \tag{42}$$

where $\bar{n}_d^{\omega}$ determines the number of patients with certainty $\omega$ in the ward on day $d$ from the previous planning horizon. In this case, $\mathbb{1}_{\omega\leq\rho'_{i,j}}$ is a binary parameter taking the value 1 when $\omega \leq \rho'_{i,j}$ otherwise 0. In other words, if a patient has a 55% chance of being in the ward on a given day, that value is approximated to 100% if the decision maker wants to be $\omega = 70\%$ certain of not exceeding the ward bed capacity. Thus, if a patient has less than a 30% chance of staying, the value is approximated to 0%. Thus, to formulate the problem using RO, constraints (30)–(37), as described in section 5.1.2, are replaced with a single constraint, namely constraint (42). All other constraints remain the same in the MIP model.

In Figure 5.2, the same example as given in Figure 5.1 in Section 5.1.2 is illustrated. This time, however, the distribution is approximated using the worst-case LOS, as shown by the dashed line with $\omega = 0.25$. As a result, the empirical probabilities on each day are approximated to the values of the dashed line, assuming the worst-case. It is evident from the figure that the approach is highly conservative and may limit scheduling options. For example, it is assumed that the patient is in the ward on days $0-7$ with a 100% chance ($\mathbb{1}_{\omega\leq\rho'_{i,j}} = 1$). From day 8, it is assumed that the patient has been discharged ($\mathbb{1}_{\omega\leq\rho'_{i,j}} = 0$). Even if the worst-case LOS for a single patient is used, there may still be a high probability of exceeding the number of staffed ward beds if there are many patients in the ward, which is similar to what we saw in Chapter 4. For instance, in Neyshabouri and Berg (2017), a slack is added to permit a temporary increase in bed capacity to ensure feasibility. This is not an option, as staffed ward beds are highly restricted. In practice, the suggested strategy would be to decrease $\omega$ values until feasibility is achieved.

## 5.2 Experimental Study

This section compares the solution quality and computational time of Ward CSombination Optimisation (WCO) and Robust Ward Optimisation (RWO) under various parameter settings. Additionally, the solutions of WCO and RWO are compared to actual scheduling data. From the dataset described in Section 3.2, four weeks (from now on referred to as a month) is arbitrarily selected for the computational experiments. During that month, 103 patients were scheduled, with approximately 30% of the patients admitted to the ward following surgery for one or more days and 10 admitted to the ICU.

Several assumptions are made for the computational experiments. First, ten patients had

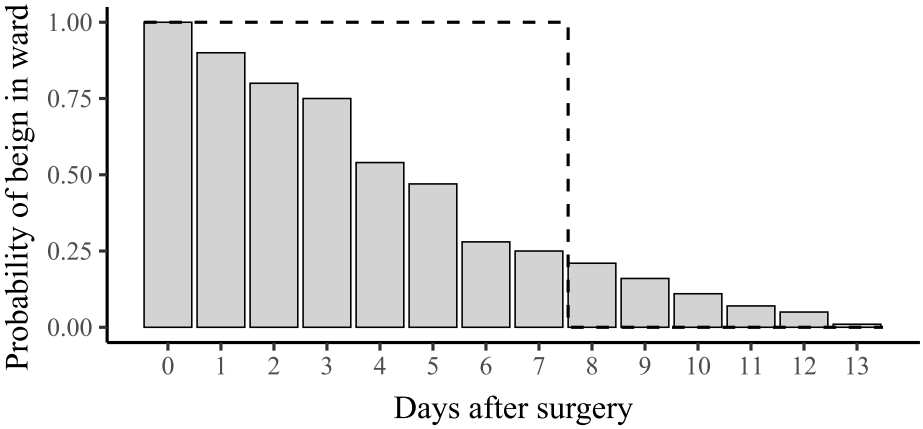



*Figure 5.2. Approximation of an empirical distribution of daily ward probabilities with the worst-case scenario as illustrated by the dashed line using certainty* $\omega = 0.25$.

The daily ward probabilities, represented by the bars, are approximated to the values indicated by the dashed line. In this case, all values greater than or equal to 0.25 are given the value of 1.

a one-week medical priority for the given month. Based on analysis of the historical data, it is assumed that these patients must be scheduled within 14 days. Moreover, it is assumed that other patients within the month have equal priority to avoid discrimination between patients. Next, semi-acute elective patients (32 of 103) were registered on the waiting list during the month. In this case, it was assumed that these patients could only be scheduled no earlier than one day after arrival. Finally, the actual availability of the surgeons for each block is used for the given month, which is, in some cases, not according to the MSS. Note that no elective surgeries are performed on weekends. However, staffed ward beds remain limited on these days and must still be considered.

The pattern and the WC generators were coded in C, while the MIP model was programmed using the AMPL mathematical programming language and solved using the commercial solver Gurobi.

### 5.2.1 Parameter Settings

This section presents the parameter settings used for the computational experiments, as well as the solution verification process necessary to validate the optimised solutions. The patterns and WCs are generated offline prior to the optimisation process. The solution verification process and generation of the patterns and WC require a few minutes of computational time. The selection of all parameters is based on standard practice within the speciality of GS, as presented in Section 3.4.

### Pattern Generation

Table 5.1 provides an overview of the parameter settings used to generate the patterns for the computational experiments.

*Table 5.1. Parameter settings used to generate the feasible patterns.*

| Practical Rules | Overtime verification |
|---|---|
| • Patient Quota: $M^P = 6$ is the maximum number of patients assigned to a pattern. | • Block Capacity: There are two distinct block capacities, thus $C_{d,r} = \{330, 450\}$ minutes. |
| • ICU Quota: $M^{ICU} = 1$ is the upper limit on the number of patients requiring ICU to be assigned to each pattern. | • Overtime Restrictions: Threshold on the probability that a pattern exceeds $C_{d,r}$ is set to $\delta = 0.75$.<br>• Extended Block Capacity: The time added to the extended block capacity to $\Delta_{d,r} = 60$ min. |

### Ward Combinations Optimisation

To generate the WCs, the number of scenarios is varied from $|K| = 4, \ldots, 7$ and the corresponding discretised probabilities are computed using Equation (24). Each WC is simulated 1000 times using Monte-Carlo sampling, with each combination verified for the risk of exceeding the availability of different ward beds. These availabilities $a \in A$ are determined by the maximum number of staffed ward beds $M^A = \{5,\ 6,\ 7\}$ and the values of $\omega = \{1.00,\ 0.75,\ 0.50,\ 0.25,\ 0.15,\ 0.10\}$.

The WCO model requires several parameters to be set. First, it is assumed that the number of staffed ward beds is $M^A = 6$, and next, the maximum number of ICU patients admitted to the ICU each day is $\bar{M}^{ICU} = 1$. Monte-Carlo sampling using historical data three weeks before the beginning of the planning horizon is used to estimate the parameter of $\bar{n}_{d,k}$. Finally, it is assumed that the maximum LOS per patient admitted to the ward is $M^W = 14$ days (7 days in practice).

The threshold for entering regular overtime is set to $\delta' = 0.25$. This value is similar to the values used by van Oostrum et al. (2008); Schneider et al. (2020). Additionally, the threshold for entering extended overtime is $\delta^{\Delta'} = 0.25$. The weight between regular and extended overtime, used in the objective function, is set to $w = 10$ based on their relative importance.

#### Robust Ward Optimisation

The same parameter settings selected for WCO are also employed for the RWO. However, to compare the different approaches, it is assumed that $\Omega$ takes the same values as $\omega$, namely $\omega = \Omega$.

#### Solution Verification

Each solution is verified with Monte-Carlo sampling using the complete undiscretised empirical distributions for surgery times and LOS from historical data as described in Section 3.4. This is important due to the nature of the approach, which allows for the possibility of exceeding the number of staffed ward beds and the block's capacity. Therefore, for each simulated result, the following statistics are collected and reported:

- **Risk of Overtime**: The probability that overtime will occur in each pattern used. Mean, median and maximum values are provided.
- **No. Beds Over**: The number of beds that exceed the specified number of staffed ward beds throughout the planning horizon. Minimum, median, mean, and maximum are provided.
- **Risk of Overflow**: The degree of exceeding the values of $\Omega$ and $\omega$ for a given number of staffed ward beds $M^A$ each day (discretisation error). Median, mean, and maximum values are provided.

#### Practical Setting

Similar parameters must be used to compare the optimised solutions (WCO and RWO) with actual scheduling data. Thus, for this comparison, it is assumed that there are $M^A = 6$ staffed ward beds, as stated in Section 3.1. Moreover, since exceeding the staffed ward beds may cause last-minute cancellations, the threshold must be set at a low value. As a result, $\Omega = \omega = 0.15$. Lastly, the number of scenarios is assumed to be $|K| = 5$.

### 5.2.2 Comparison

Table 5.2 summarises optimal solutions from the WCO, RWO, and also actual scheduling data for a single month where the parameter settings reported in Section 5.2.1 under Practical Setting are used. It is evident that the patterns selected using WCO and RWO are less likely to surpass the accepted risk of regular overtime and extended overtime

compared to actual data. The difference is most significant for extended overtime (up to 71% less), whereas it is smaller for regular overtime (up to 25% less). Additionally, compared to actual data, the optimised solutions have a lower risk of overtime (mean, median, and maximum).

The same effect is observed when ward results are analysed, as the models achieve lower values. This is evident for the number of beds exceeding the staffed ward beds (No. Beds Over) and for the risk of exceeding the beds (Risk of Overflow) in the planning horizon. Comparing WCO to RWO, it is evident that the former produces higher quality solutions in terms of overtime and the number of beds over. However, the risk of overtime and overflow remains identical.

*Table 5.2. Overall comparison between optimal solutions of the ward combinations optimisation (WCO), the robust ward optimisation (RWO) and the actual schedule for the planning horizon of one month.*

| | **Blocks** | | | | | **Ward** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Overtime**[1] | | **Risk of Overtime**[2] | | | **No. Beds over**[3] | | | | **Risk of Overflow**[4] | | |
| **Case** | **Regular** | **Extended** | **Mean** | **Median** | **Max** | **Min** | **Median** | **Mean** | **Max** | **Median** | **Mean** | **Max** |
| Actual † | 8 | 14 | 0.20 | 0.32 | 1.00 | 0 | 9 | 9.66 | 27 | 0.02 | 0.15 | 1.00 |
| WCO * | 6 | 4 | 0.10 | 0.17 | 0.68 | 0 | 1 | 1.79 | 14 | 0.01 | 0.06 | 0.25 |
| RWO ‡ | 7 | 5 | 0.08 | 0.17 | 0.69 | 0 | 1 | 1.47 | 18 | 0.01 | 0.04 | 0.26 |

Configurations: † $M^A = 6$; * $(M^A, \Omega, |K|) = (6, 0.15, 5)$; ‡ $(M^A, \omega) = (6, 0.15)$. $M^A$ is the number of staffed ward beds, $|K|$ is the number of scenarios and $\Omega = \omega$ are the thresholds of exceeding $M^A$. [1]Regular and extended overtime show the number of times the selected patterns surpass the accepted risk for each group. [2]The probability that the selected patterns will surpass their block capacity. [3]Measures how many beds exceed the number of staffed ward beds in the simulation. [4]Likelihood of exceeding the number of staffed ward beds.

A visual presentation of the actual, WCO, and RWO scheduling is provided in Figures 5.3–5.5 illustrating the daily differences between the solutions. These solutions are summarised in Table 5.2. In practice, the scheduler manually creates patterns using the Accumulated Average Surgery Duration (AASD). Consequently, the numbers reflect what the scheduler can see in the planning software.

Figure 5.3 reveals that there is an imbalance in the utilisation of the blocks across the planning horizon. Analysis of the AASD for each block shows that many blocks are utilised at or near full capacity. For instance, on days 3, 9, 15, 17, 18, 22, and 24, the AASD of at least one block exceeds 7.5 hours. However, while many blocks are being utilised near full capacity, many blocks have low utilisation (low block hours). For example, on days 4, 5, 8, 10, 11, 16, 19, 23, and 25, at least one block has a low block utilisation.

The risk of exceeding the block's capacity increases as the AASD approaches its maximum capacity (dashed lines) that day. For example, it was impossible to avoid operating close to or exceeding capacity on certain days. This is evident in the figure for single surgeries spanning the whole day, as observed on days 1, 17, 18, 22, and 23.

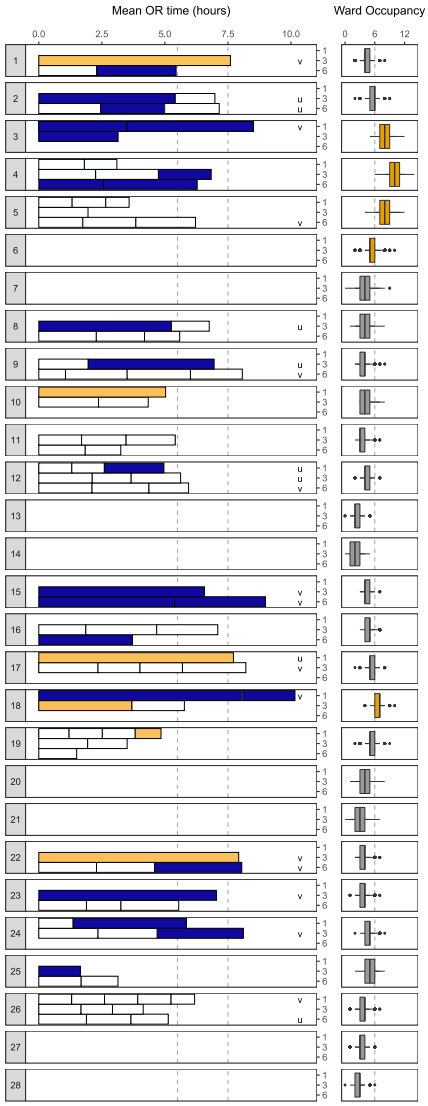



*Figure 5.3. Visualisation of the actual surgeries for each day and OR and the corresponding ward occupancy.*

The numbers from 1 to 28 are the days of the month. The numbers 1,3, and 6 are the numbers of the operating rooms at the hospital. The tags *u* and *v* denote if the threshold for the risk of regular or extended overtime is surpassed, respectively. The blue boxes indicate elective ward patients, while the yellow are semi-acute elective ward patients. Three dashed lines are provided. On the left-hand side, the lines show the opening hours of the blocks, with the lower line for Fridays. On the right-hand side, the dashed line shows the ward capacity.

Figure 5.3 reveals the variability in the staffed ward bed occupancy with the occupancy being the lowest on the weekends (days 6, 7, 13, 14, 20, 21, 27, and 28). There is a high risk of exceeding the staffed ward beds between days 3 and 6 in the first week. In the second week (days 8–14), the ward bed occupancy is lower and evened out. Nevertheless, in the last two weeks (15–28), the ward occupancy variance has returned, with day 18 exceeding the ward bed capacity. However, these variances are less severe than during the initial week. Analysing the number of ward admissions, we see that the highest number of admissions are in weeks 1 and 3, which are also the weeks with days having a chance of exceeding the ward bed capacity.

In Figures 5.4 and 5.5, the optimal solutions for the WCO and RWO schedules are illustrated. The results reveal that only 28 patients for the WCO and 23 patients for the RWO are scheduled for the same day as the actual data, out of a total of 103 patients included in both schedules. This implies that the optimised schedule and the actual schedule are dissimilar. Analysing the data reveals that, for the optimised schedules, the utilisation of both the ward and the blocks in the planning horizon is more evened out. For example, daily ward admissions are relatively balanced, averaging between 1 and 2 per day on most days. In addition, the risk of exceeding the staffed ward beds is lower. However, analysing the total number of ward admissions each week, we observe that the WCO maintains a balance, with 6–8 admissions each week. The actual and the RWO schedule, however, have more variance, with the former ranging from 4-10 and the latter from 6-10 admissions. It can be observed the number of blocks with overtime is lower, resulting in improved and more balanced utilisation. Finally, blocks with a single operation that span the entire day and have a high risk of entering overtime are never combined with other surgeries for optimal solutions. However, this occurs in the actual data (see Figure 5.3 on day 18 in room 1).

When the WCO and RWO schedules are compared, it is evident that they differ in terms of overtime and ward utilisation. There is more overtime in the RWO in the second week but less in the third week. The ward utilisation is lower as more conservative solutions are made by assuming the worst-case scenario for LOS using the RWO. However, both solutions have three days with the risk of exceeding the ward beds but on different days.

### 5.2.3 Parameter Analysis

This section analyses the trade-offs between solution quality and computational time for WCO and RWO for various parameters, as presented in Section 5.2.1. In Table 5.3 and Table 5.4, a summary of the computational results for the various parameter settings used for WCO and RWO, respectively, is provided. It is worth noting that finding feasible solutions for some parameter settings was not possible. Infeasible solutions are indicated by $-$ in the tables and may occur when a low value of $\Omega$ is used for a given staffed ward bed capacity $M^A$.

Changing the parameters $(M^A, \Omega, |K|)$ has a low impact on the number of times regular

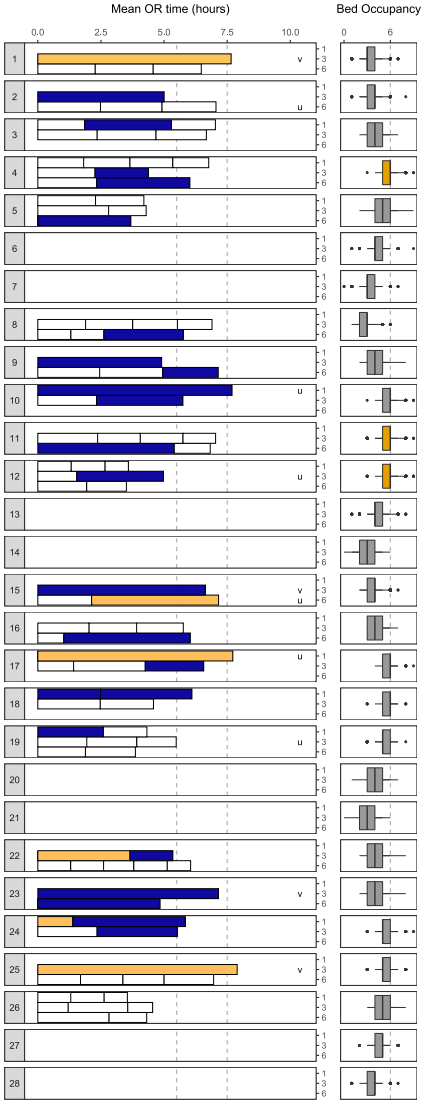



*Figure 5.4. Visualisation of the optimal solution for each day and OR for the Ward Combination Optimisation and the corresponding ward occupancy.*

See Figure 5.3 for a full description.

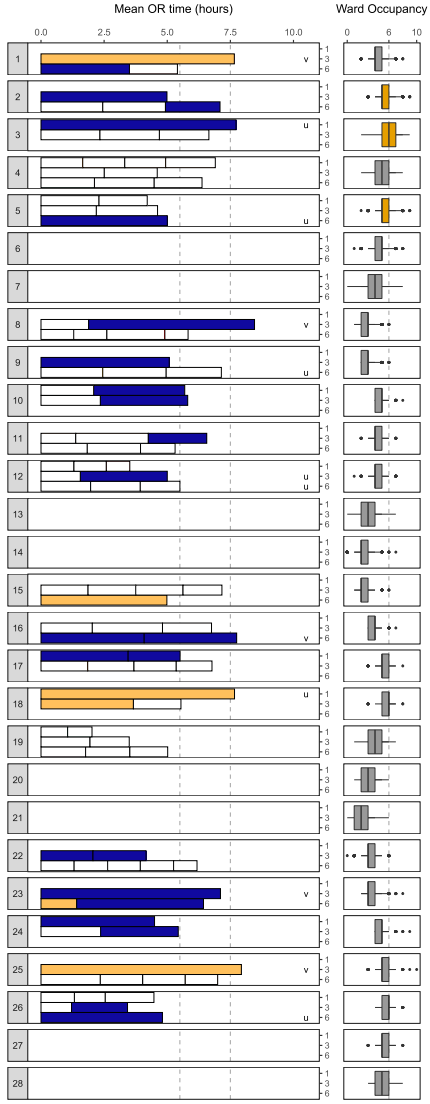



*Figure 5.5. Visualisation of the optimal solution from the robust ward optimisation for each day and OR and the corresponding ward occupancy.*

See Figure 5.3 for a full description.

*Table 5.3. Quality of solutions and computational requirements for different configurations of ($M^A$, $\Omega$, $|K|$) for the ward combination optimisation.*

| Configuration | Block | | Ward | | | | | | | MIP |
|---|---|---|---|---|---|---|---|---|---|---|
| | Overtime [1] | | No. Beds over [2] | | | | Risk of Overflow [3] | | | CPU |
| ($M^A$, $\Omega$, $\|K\|$) | Regular | Extended | Min | Median | Mean | Max | Median | Mean | Max | (s) |
| (5, 1.00, 5) | 6 | 4 | 3 | 16 | 16.18 | 44 | 0.16 | 0.32 | 0.98 | 95 |
| (5, 0.75, 5) | 6 | 4 | 1 | 10 | 11.13 | 46 | 0.15 | 0.28 | 0.89 | 133 |
| (5, 0.50, 5) | - | - | - | - | - | - | - | - | - | - |
| (5, 0.25, 5) | - | - | - | - | - | - | - | - | - | - |
| (5, 0.15, 5) | - | - | - | - | - | - | - | - | - | - |
| (5, 0.10, 5) | - | - | - | - | - | - | - | - | - | - |
| (6, 0.15, 4) | 6 | 4 | 0 | 1 | 2.02 | 20 | 0.02 | 0.06 | 0.31 | 362 |
| (6, 1.00, 5) | 6 | 4 | 0 | 6 | 6.00 | 22 | 0.03 | 0.16 | 0.73 | 121 |
| (6, 0.75, 5) | 6 | 4 | 0 | 5 | 6.07 | 25 | 0.03 | 0.15 | 0.68 | 109 |
| (6, 0.50, 5) | 6 | 4 | 0 | 3 | 3.90 | 19 | 0.05 | 0.11 | 0.56 | 470 |
| (6, 0.25, 5) | 6 | 4 | 0 | 2 | 2.45 | 17 | 0.03 | 0.08 | 0.34 | 790 |
| (6, 0.15, 5) | 6 | 4 | 0 | 1 | 1.79 | 14 | 0.01 | 0.06 | 0.25 | 1080 |
| (6, 0.10, 5) | 6 | 4 | 0 | 1 | 1.58 | 12 | 0.02 | 0.05 | 0.21 | 3365 |
| (6, 0.15, 6) | 6 | 4 | 0 | 1 | 1.61 | 13 | 0.02 | 0.05 | 0.22 | 6505 |
| (6, 0.10, 6) | - | - | - | - | - | - | - | - | - | - |
| (6, 0.15, 7) | - | - | - | - | - | - | - | - | - | - |
| (7, 1.00, 5) | 6 | 4 | 0 | 4 | 4.56 | 22 | 0.00 | 0.09 | 0.88 | 135 |
| (7, 0.75, 5) | 6 | 4 | 0 | 2 | 2.13 | 20 | 0.01 | 0.06 | 0.47 | 170 |
| (7, 0.50, 5) | 6 | 4 | 0 | 1 | 1.88 | 13 | 0.00 | 0.06 | 0.41 | 185 |
| (7, 0.25, 5) | 6 | 4 | 0 | 1 | 1.42 | 12 | 0.01 | 0.04 | 0.28 | 175 |
| (7, 0.10, 5) | 6 | 4 | 0 | 0 | 0.75 | 11 | 0.00 | 0.02 | 0.14 | 325 |
| (7, 0.10, 6) | 6 | 4 | 0 | 0 | 0.21 | 5 | 0.00 | 0.01 | 0.03 | 1500 |

$M^A$ denotes the maximum number of staffed ward beds, $\Omega$ denotes the limit on the likelihood that a ward combination exceeds the number of staffed ward beds and $|K|$ denotes the number of scenarios. CPU is the computational time (CPU) required to solve the mixed integer programming (MIP) model. [1] Regular and extended overtime show how many patterns surpass the accepted risk for each group;[2] Measures how many beds exceed the given number of staffed ward beds ($M^A$); [3] The likelihood of exceeding the number of staffed bed.

or extended overtime occurs, as shown in Table 5.3. As illustrated in Figures 5.4 and 5.5, 50% of the patterns that surpass the risk of entering either regular or extended overtime are composed of single surgeries. Thus, one can not avoid the risk of overtime for the given set of surgeries.

Analysing the results for the simulated ward occupancy shows that changing the parameter settings now affects the results. For example, by lowering the threshold for risk of overflow, $\Omega$, while keeping the same values for $M^A$ and $|K|$, one can see that it lowers all statistics reported for the number of beds over and risk of overflow. However, the most significant difference is in the median and the maximum values. For example, when $M^A = 6$ and $|K| = 5$, the maximum number of beds going over is reduced by 45% and the risk of overflow by 71% when $\Omega$ decreases from 1.00 to 0.10. One can see the same effect by increasing the number of scenarios $|K|$ but keeping the same values for $M^A$ and $\Omega$. For example, when $|K|$ goes from 4 to 6, the maximum number of beds over goes down by 35% when $M^A = 6$ and $\Omega = 0.15$. Similarly, the maximum risk of overflow decreases by 29%.

The parameter settings strongly affect the computational time required to solve WCO. First, as the number of maximum staffed ward beds ($M^A$) increases for a given set of patients, the computational time required decreases. However, as the risk of overflow $\Omega$ threshold decreases, the problem becomes more computationally demanding. When $M^A = 6$ for $|K| = 5$, the computational time increases from 121 to 3365 s (28 times) when $\Omega$ changes from 1.00 to 0.10. Finally, changing the number of scenarios ($|K|$) used to discretise the LOS distributions increases the computational time significantly. In other words, increasing $|K|$ from 4 to 6 increases the computational time from 362 to 6505 s (18 times) when $M^A = 6$ and $\Omega = 0.15$.

In Table 5.4, the computational results for the RWO for the different parameter settings. Changing the $\omega$ values allows one to observe an effect similar to those previously discussed. For most settings, the number of patterns that surpass the accepted risk of regular and extended overtime remains the same. However, for the parameter setting $(6,\ 0.15)$, regular and extended overtime increases by one. The simulated ward results show that by decreasing $\omega$, the median and maximum values for the number of beds over and the risk of ward overflow decreases.

Comparing WCO and RWO, one can see that the results are comparable in terms of the number of beds over and the risk of overflow, although WCO's results are slightly lower. Comparing the overtime, one can see that the same results are achieved apart when $M^A = 6$ and $\Omega = \omega = 0.15$, where it is higher for RWO. One can attain higher quality solutions by utilising the WCO in terms of overtime, number of beds over, and risk of overflow. Moreover, by using the worst-case LOS for RWO, it was impossible to find feasible solutions with $\omega$ greater than 0.15, whereas this was possible for WCO ($M^A = 6$). Analysing the computational time, it is evident that the RWO results in lower computational times for identical parameter settings. For example, by comparing the parameter settings of $(6,\ 0.15,\ 6)$ of WCO to $(6,\ 0.15)$ for RWO, one can see it is 6489 s higher for the WCO.

*Table 5.4. Quality of solutions and computational requirements for different configurations of ($M^A$, $\omega$) for the robust ward optimisation.*

| Configuration | Block | | Ward | | | | | | | MIP |
|---|---|---|---|---|---|---|---|---|---|---|
| | Overtime [1] | | No. Beds over [2] | | | | Risk of Overflow [3] | | | CPU |
| ($M^A$, $\omega$) | Regular | Extended | Min | Median | Mean | Max | Median | Mean | Max | (s) |
| (5, 1.00) | 6 | 4 | 2 | 19 | 19.65 | 47 | 0.17 | 0.28 | 1.00 | 18 |
| (5, 0.75) | 6 | 4 | 0 | 10 | 10.76 | 31 | 0.20 | 0.27 | 0.86 | 27 |
| (5, 0.50) | 6 | 4 | 0 | 7 | 7.44 | 30 | 0.11 | 0.20 | 0.77 | 57 |
| (5, 0.25) | - | - | - | - | - | - | - | - | - | - |
| (5, 0.15) | - | - | - | - | - | - | - | - | - | - |
| (5, 0.10) | - | - | - | - | - | - | - | - | - | - |
| (5, 0.05) | - | - | - | - | - | - | - | - | - | - |
| (6, 1.00) | 6 | 4 | 0 | 9 | 9.37 | 35 | 0.03 | 0.17 | 0.96 | 13 |
| (6, 0.75) | 6 | 4 | 0 | 4 | 4.83 | 22 | 0.03 | 0.12 | 0.72 | 17 |
| (6, 0.50) | 6 | 4 | 0 | 2 | 3.01 | 17 | 0.03 | 0.09 | 0.37 | 21 |
| (6, 0.25) | 6 | 4 | 0 | 2 | 2.12 | 12 | 0.01 | 0.06 | 0.41 | 34 |
| (6, 0.15) | 7 | 5 | 0 | 1 | 1.47 | 18 | 0.01 | 0.04 | 0.26 | 16 |
| (6, 0.10) | - | - | - | - | - | - | - | - | - | - |
| (6, 0.05) | - | - | - | - | - | - | - | - | - | - |
| (7, 1.00) | 6 | 4 | 0 | 5 | 5.69 | 22 | 0.00 | 0.10 | 0.90 | 27 |
| (7, 0.75) | 6 | 4 | 0 | 1 | 1.11 | 12 | 0.00 | 0.03 | 0.36 | 34 |
| (7, 0.50) | 6 | 4 | 0 | 0 | 0.79 | 11 | 0.00 | 0.02 | 0.12 | 18 |
| (7, 0.25) | 6 | 4 | 0 | 0 | 0.72 | 21 | 0.00 | 0.02 | 0.16 | 18 |
| (7, 0.15) | 6 | 4 | 0 | 0 | 0.64 | 8 | 0.00 | 0.02 | 0.15 | 26 |
| (7, 0.10) | 6 | 4 | 0 | 0 | 0.59 | 8 | 0.00 | 0.02 | 0.29 | 136 |
| (7, 0.05) | - | - | - | - | - | - | - | - | - | - |

$M^A$ denotes the maximum number of staffed ward beds and $\omega$ denotes the limit on the likelihood that the robust ward approximation exceeds the number of staffed ward beds. See Table 5.3 for a full description of other parameters.

## 5.3 Conclusions

This chapter considered how to reduce the combined risk of last-minute cancellations due to uncertainty in surgery times and LOS in the ward when operating close to the maximum block capacity and staffed ward beds. The objective was to determine how this can be achieved when scheduling elective patients weeks in advance while maintaining the existing utilisation.

To solve the problem, the *Pattern Scheduling* approach proposed in Chapter 4 was used where uncertainty in surgery times was resolved by generating feasible patterns of patients. Uncertainty in LOS in the ward following surgery was resolved by generating WCs. The optimal assignment of the patterns was determined using a MIP model which allowed bounding the risk of exceeding the number of staffed ward beds using the WCs. The objective was to minimise the number of patterns selected with overtime and the amount of overtime in the planning horizon. The approach was compared with a robust formulation based on the worst-case outcome for LOS for each patient. Both approaches were then compared to actual scheduling data for a single month.

Results demonstrate that both WCO and RWO result in higher-quality solutions when compared to actual scheduling data. First, the number of patterns with a risk of overtime is significantly lower. For the given month and throughput, it was impossible to avoid blocks with overtime as the system was operating near full capacity. Additionally, some patterns included a single operation spanning the whole day and a high risk of overtime. Thus, their selection could not be avoided. Finally, the risk of exceeding the bed capacity is lower. However, WCO maintains a balance in the total number of ward admissions each week, unlike RWO and the actual scheduling data, where more variance is observed. These results suggest that using WCO or RWO can reduce the combined risk of last-minute cancellations.

When analysing the trade-off between solution quality and computational tractability for the WCO and RWO, it was evident that the number of staffed ward beds and their utilisation affect computational tractability. If the utilisation is high, the problem becomes more computationally demanding. Moreover, the tractability is also affected by the number of scenarios used to discretise the empirical LOS distributions for WCO and the specified threshold for the accepted risk of exceeding the ward beds. Increasing the number of scenarios for the WCO while decreasing the threshold for the accepted risk of exceeding the ward beds increases the robustness (e.g., fewer last-minute cancellations) of the solutions but at the computational time cost. By comparing WCO to RWO, it was evident that one can achieve higher quality solutions by using the WCO in terms of block overtime and particularly ward utilisation due to the conservatism of the RWO. However, the RWO's computational times are significantly lower using identical parameter settings.

In this chapter, semi-acute patient arrivals were assumed to be known in advance. However, in practice, their arrivals are unpredictable and can result in disruptive rescheduling

events of previously scheduled patients and, therefore, imbalances in block and ward utilisation. Consequently, it is essential to account for their arrivals in the long-term elective schedule to avoid disruptive rescheduling events.

## 5.4 Summary

In this chapter, the combined risk of last-minute cancellations associated with uncertainty in surgery times and LOS in the ward was addressed when operating under close to full capacity of the block capacity and the staffed ward beds. The objective was to explore how last-minute cancellations could be reduced when elective patients are scheduled weeks in advance while maintaining the existing utilisation. The two-step Pattern Scheduling approach, proposed in Chapter 4, was used to specify uncertainty in surgery time practically and in a statistically accurate way in the first step. In the second step, ward combinations were generated, making it possible bounding the risk of exceeding the staffed ward beds. The approach was compared to an equivalent robust formulation, which assumed the worst-case LOS. The result showed that higher quality solutions are achieved by using the ward combinations in terms of overtime and ward utilisation but at the cost of computational time. Compared to actual scheduling data, both solutions achieved higher quality, suggesting both approaches can reduce the combined risk of last-minute cancellation. In this chapter, it was assumed that semi-acute arrivals were known in advance. However, their arrivals are unpredictable and have historically caused disruptive rescheduling events at GS speciality.

In the next chapter, the unpredictable arrivals of semi-acute elective patients are addressed and how gaps should be reserved to minimise the number of rescheduling events but without resorting to overtime.

# 6 Uncertainty in Semi-Acute Elective Arrivals

This chapter aims to compare three different gap-reserving strategies for reserving gaps in the long-term schedule of elective patients, which are later used to accommodate future semi-acute elective arrivals with minimal rescheduling and without resorting to excessive overtime. The objective is to understand how these strategies impact the number the need for rescheduling when operating under a high throughput of elective patients, limited staffed ward beds, staffed ICU beds, and equipment availability. As a result, the number of last-minute cancellations, associated with uncertainty in surgery times and uncertainty in LOS in the ward following surgery, is considered. This chapter is based on the work in Sigurpalsson et al. (2021, 2025).

The previous chapter focused on building schedules that minimised the combined risk of last-minute cancellations due to uncertainty in surgery times and LOS in the ward when elective patients were scheduled in advance. However, it was assumed that the unpredictable arrivals of semi-acute elective patients, which must be accommodated into the long-term elective schedule with minimal rescheduling, were known in advance. In practice, these arrivals are unpredictable and can cause disruptive rescheduling events of previously planned patients. Consequently, it is essential to account for their arrivals in advance to minimise the number of such occurrences.

As discussed in Section 2.2.3, semi-acute elective patients are defined as elective patients with high medical priority who must undergo surgery within two weeks and share the same block capacity with other elective patients (Zonderland et al., 2010). Unlike emergency patients, who must be operated on within a day (Van Riet and Demeulemeester, 2015), there is greater flexibility in managing the timing of the semi-acute elective arrivals (Epstein and Dexter, 2013). Emergency patients are not considered, as they are operated on during downtime and after hours by dedicated teams, as discussed in Section 3.3.

This chapter considers these unpredictable arrivals by reserving gaps in the elective program using gap-reserving strategies, while minimising the risk of last-minute cancellations due to uncertainty in surgery times and uncertainty in LOS in the ward following surgery, as both events are undesirable outcomes (Pandit, 2018). Alternatively, the maximum peak in the ward bed occupancy is minimised over the planning horizon to avoid last-minute cancellations in the ward (Beliën and Demeulemeester, 2007; Fügener et al.,

2014; van den Broek d'Obrenan et al., 2020; Schneider et al., 2020). With gap-reserving strategies, gaps are reserved in the blocks of the long-term elective schedule, which are later used to accommodate semi-acute elective arrivals with the least rescheduling. It is possible to reserve gaps in a deterministic (Van Riet and Demeulemeester, 2015) or stochastic (Lamiri et al., 2008b; Molina-Pariente et al., 2018; Jebali and Diabat, 2017) manner in some or all of the available blocks. How this is carried out depends on the selected gap-reserving strategy. Generally, gap-reserving strategies can range from simple heuristics, which human schedulers can apply (Jebali et al., 2006; Addis et al., 2016; Kamran et al., 2019; Spratt and Kozan, 2021), to more sophisticated approaches, such as optimisation (Min and Yih, 2010; Molina-Pariente et al., 2018), as discussed in Chapter 2.2.3. This chapter focuses on simple heuristics.

This chapter compares three different gap-reserving strategies with two suggested by the literature (Addis et al., 2016; Kamran et al., 2019) and one proposed in this chapter. Scheduling and rescheduling models are developed to compare these strategies while considering practical aspects of scheduling at the GS speciality and last-minute cancellations. The scheduling and rescheduling of semi-acute elective patients will be performed over multiple four-week planning horizons, with the schedule updated for each new semi-acute elective arrival. During each planning horizon, no elective patients are added to the waiting list, only semi-acute elective patients. As a result, it is possible to compare the optimised results, using the gap-reserving strategies, to actual scheduling data.

The work relates to the studies by Min and Yih (2010); Zhang et al. (2019); Jebali and Diabat (2017), which simultaneously consider uncertainty in surgery times, downstream length of stay, and emergency arrivals, but rescheduling is neglected. Unlike these studies, this chapter focuses on semi-acute elective arrivals, not emergency arrivals, and it includes rescheduling. Similar to the studies by Addis et al. (2016); Kamran et al. (2019); Spratt and Kozan (2021); Adams et al. (2023), scheduling and rescheduling models are proposed in this chapter. However, unlike these studies, it considers the availability of limited downstream resources and equipment availability while minimising the risk of last-minute cancellations due to uncertainty in surgery times and uncertainty in LOS in the ward following surgery. Additionally, the studies by Addis et al. (2016); Kamran et al. (2019); Spratt and Kozan (2021) employ different gap-reserving strategies, but no comparison is made between these strategies or their impact on the need for rescheduling.

The chapter is organised as follows. The next section outlines ways to plan gaps in the surgery program, introducing three gap-reserving strategies, with two suggested by the literature and one proposed in this chapter. This is followed by a section on model development, where scheduling and rescheduling models implementing these different strategies are proposed. Computational Experiments are conducted in Section 6.3, where the aim is to compare the different gap-reserving strategies to understand how they impact the need for rescheduling. Finally, conclusions and a summary of the main findings are provided at the end of the chapter.

## 6.1 Planning Gaps in the Surgery Program

The unpredictable arrivals of semi-acute elective patients pose significant scheduling challenges for the scheduler. That is, when a scheduler plans elective patients up to weeks in advance, gaps must be reserved in the blocks in the program for future semi-acute arrivals. These gaps are then used to accommodate semi-acute elective arrivals, minimising the need to reschedule already scheduled elective patients. The fundamental question for the scheduler is determining in which blocks should have gaps reserved for semi-acute arrivals each week. If the gaps reserved are insufficient, rescheduling becomes inevitable to accommodate semi-acute elective arrivals. Conversely, reserving gaps that are too large leaves the resources underutilised, which may increase the waiting time of elective patients awaiting surgery. Neither of these results are desirable.

Figures 6.1 and 6.2, illustrate the actual scheduling data from the GS speciality for a single month, and the possible disruptions caused by the semi-acute elective arrivals. These figures demonstrate that the number of semi-acute elective patients is highly stochastic. According to the hospital's data for the year 2019, the number of semi-acute elective arrivals was $23.5 \pm 7.3$ (mean$\pm$ standard deviation) each month contributing to $25.4 \pm 6.2\%$ of the total throughput of patients. Although the surgery times for most of these cases are generally short, with each case taking $118.1 \pm 7.0$ minutes on average, Figure 6.2 demonstrates that some may span the whole day, as shown on day 15, posing significant challenges for the scheduler. The figures further reveal that many blocks have a low utilisation (see, e.g. days 3, 10, 16, 19 and 24), suggesting either last-minute cancellations occurred or that the gap reserved in these blocks was too large. In some cases, it was possible to increase the block utilisation by starting the emergency program early (see days 3, 10, 17, 18, 19, 24 and 25). While this may seem desirable, it decreases block utilisation after hours when two surgery teams are available for emergency arrivals, as discussed in Section 3.3. Another possible outcome occurs when the reserved gap is insufficient for the semi-acute arrivals. In such cases, overtime can occur, as observed on days 4, 17, 19, 23, and 26, which may cause last-minute cancellations.

### 6.1.1 Gap-reserving Strategies

The literature on reserving gaps for semi-acute elective patients is limited, with most studies focusing on reserving gaps for emergency arrivals that need to be operated on within the day (Min and Yih, 2010; Samudra et al., 2016). Even fewer studies address the scheduling and rescheduling of patients (Addis et al., 2016; Kamran et al., 2019; Spratt and Kozan, 2021; Adams et al., 2023). In the following paragraphs, two common strategies found in the literature are reviewed, one focusing on block utilisation and the other focusing on avoiding rescheduling. Finally, an alternative strategy that is likely to provide a balance between the other two is proposed.

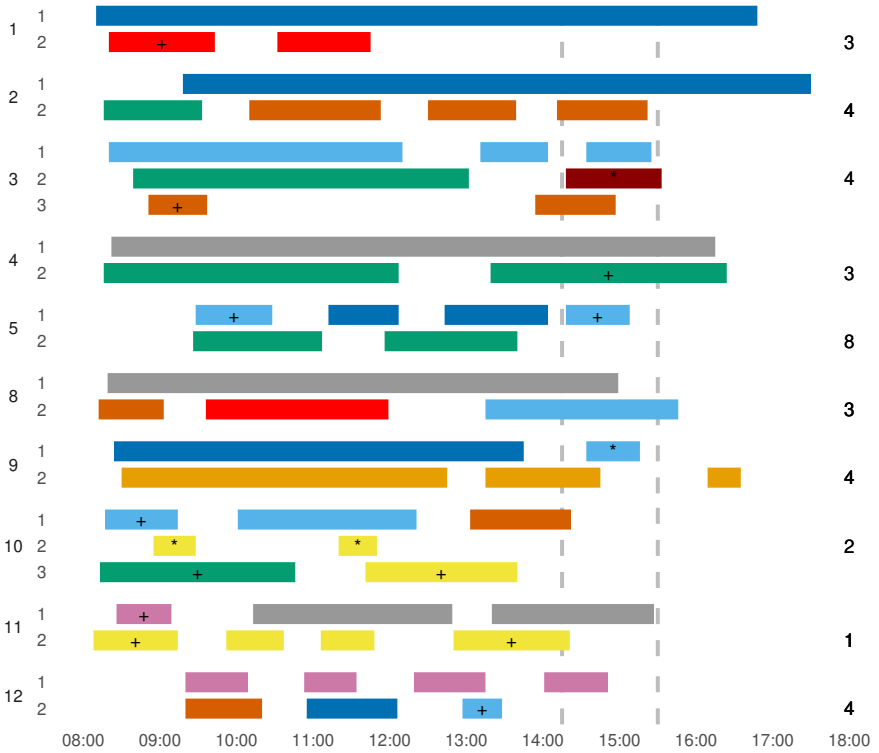



*Figure 6.1. The actual schedule of patients of the GS speciality during regular working hours by the first two weeks of a month for each OR (1 to 3) and day (1 to 12).*

Each patient (box) is assigned a unique colour that represents the patient's surgeon, and the length of the box corresponds to the surgery time. Some boxes are labelled with a $+$ or $*$ to denote semi-acute and emergency patients, respectively. Two dashed lines depict the blocks' capacity with the lower line for Fridays (days 5 and 12). The numbers on the right-hand side show the number of emergency surgeries performed after hours.

## Front Load

The first strategy, *Front load* (sometimes referred to as Forward Scheduling), is a production scheduling technique that schedules tasks as early as possible in the planning horizon. As such, patients are scheduled towards the front of the planning horizon, gradually increasing the reserved gap by each day for semi-acute arrivals or for the already scheduled patients that may need to be rescheduled to accommodate the semi-acute elective arrivals (Addis et al., 2016). It is anticipated that a high block utilisation will be obtained, at least in the short term, but rescheduling and overtime will be required to accommodate semi-acute elective arrivals.

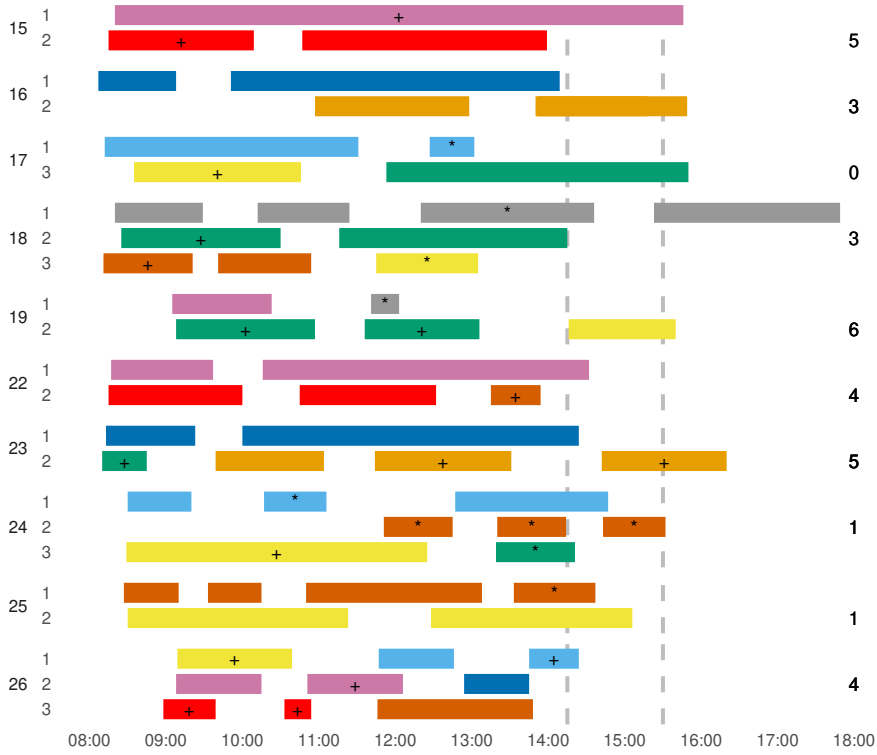



*Figure 6.2. The actual schedule of patients of the GS speciality during regular working hours by the second two weeks of a month for each OR (1 to 3) and day (15 to 26).*

See Figure 6.1 for full description.

### Daily Limit

The second strategy, *Daily limit*, involves posing limits on the number of elective patients assigned to each surgeon on each day (Kamran et al., 2019). With this strategy, stochastic gaps are reserved across the planning horizon in multiple blocks. It is anticipated that this strategy offers a higher flexibility for accommodating semi-acute elective patients compared to *Front load*, but at the cost of requiring more block capacity and carrying a higher risk of lower block utilisation.

### Balance ORs

The third strategy, *Balance ORs*, is an alternative strategy that creates a schedule that minimises the maximum number of blocks per week that can be held constant across the whole planning horizon. This strategy aims to maintain a fixed number of blocks available each week in the planning horizon but minimises that number. It is anticipated that this strategy will strike a balance between the *Front load* and *Daily limit* strategies. In other words, it is expected to require fewer blocks than the *Daily limit* strategy but more than the *Front load* strategy. Additionally, it is expected to result in more rescheduling events than the *Daily limit* strategy but fewer than the *Front load* strategy.

## 6.2 Model Development

This section studies a similar problem to the one considered in Section 5.1. However, it focuses on how gaps are reserved in the long-term elective schedule using gap-reserving strategies. These gaps are later used to accommodate semi-acute elective arrivals so that rescheduling events are minimised and without resorting to excessive overtime. This is done while reducing the risk of last-minute cancellations associated with uncertainty in surgery times and LOS in the downstream ward. In this context, three gap-reserving strategies are compared. The same modelling assumptions presented in Section 5.1 are applied to the problem formulation. Additionally, the following assumptions are made:

- A four-week planning horizon is assumed with no new elective patient arrivals, except semi-acute elective patients. Hence, the throughput of patients is known and fixed for each four-week planning horizon.
- Elective patients and semi-acute elective patients share the same block capacity. However, due to their high medical priority, semi-acute elective patients can be assigned to all blocks.
- Decisions about scheduling semi-acute elective patients and rescheduling of already scheduled patients are made upon each new arrival. Rescheduled patients must be assigned new appointments within the current planning horizon and cannot be postponed to future planning horizons. An extra penalty is incurred for rescheduling a patient with a high medical priority.
- Surgeries requiring specific equipment, such as robots, can only be assigned to specific blocks.
- Unlike the assumption in Section 5.1, this section assumes that staffed ward beds may be temporarily increased. However, peaks in ward bed occupancy must be minimised to reduce the risk of last-minute cancellations.

The two-step Pattern Scheduling approach presented in Section 5.1 is used to solve the problem. However, several modifications are made to account for the gap-reserving strategies and to account for rescheduling:

1. **Base Scheduling Model**: Elective patients, already on the waiting list of each surgeon, are assigned up to four weeks in advance while reserving gaps for future semi-acute elective arrivals. The following modifications are made:
    - **Pattern Generation:** Two new practical rules are implemented. The first rule restricts the number of patients admitted to the ward in each pattern, while the latter restricts the number of out-patients in a pattern.
    - **MIP model:** The model remains mostly unchanged from Section 5.1.2, but the following modifications are made: First, three distinct gap-reserving strategies are employed to reserve gaps in advance scheduling of elective patients, which are later used to accommodate future semi-acute elective arrivals. Next, Constraints (30)–(37) are replaced with a single constraint that determines the maximum peak in ward bed occupancy in the planning horizon. This peak is then minimised to flatten bed occupancy over the planning horizon. The objective is to avoid last-minute cancellations and to leave ward beds available for possible future arrivals of semi-acute elective patients. Finally, an identical constraint is introduced to account for the number of staffed ICU beds. However, the number of staffed ICU beds is assumed to have strict limits unlike the ward beds.
2. **Rescheduling Model**: As the four-week schedule is executed, semi-acute elective patients are added to the waiting list of each surgeon and must be accommodated into the existing long-term schedule with the least rescheduling and without resorting to overtime. The following modifications are made to achieve this:
    - **Pattern Generation**: Due to the priority of semi-acute elective patients, the pattern generation is extended to enable these patients to enter the patterns of other surgeons. The same practical rules still apply as presented above but two additional practical rules are introduced. The first rule limits the number of semi-acute elective patients assigned to each pattern. The second rule ensures that the ratio of elective patients in each pattern is larger than the ratio of semi-acute elective patients when their number in a pattern surpasses a specified threshold.
    - **MIP Model**: The Base Scheduling Model, presented above, is used to accommodate the semi-acute elective patients in to the existing long-term schedule of other elective patients. To account for rescheduling, additional constraints are incorporated into in the model to minimise the number of rescheduling events required to accommodate the semi-acute elective arrivals.

The scheduling and rescheduling models are described in Sections 6.2.1– 6.2.2, respectively.

### 6.2.1 Base Scheduling Model

#### Pattern Generation

The same set of practical rules described in Section 5.1.1 is used to generate the feasible patterns, with the addition of two practical rules. In other words, the first rule imposes an upper bound on the number of outpatients assigned to a pattern, and the second imposes an upper bound on the number of ward patients in a pattern. The first rule is intended to ensure sufficient recovery time in the PACU during operating hours, while the second rule balances out the number of ward admissions, as discussed in Section 3.3. As before, let $z_{i,p}$ be a binary decision variable taking the value 1 if patient $i$ is assigned to a pattern $p$ (0 otherwise) then these practical rules may be implemented as follows:

1. *Patient Quota*: Restricts the number of patients assigned to a pattern to the parameter $M^P$. The formulation is given with equation (3) presented in Section 4.1.1.
2. *Resource Quotas*: As downstream resources are scarce, several practical rules are applied to limit the number of patients admitted to each resource $\phi \in \Phi$ from each pattern. This section specifies the downstream resources as ICU, ward, and PACU. Let $g_i^{\phi}$ be a binary indicator taking the value of 1 if patient $i$ requires resource $\phi \in \Phi$ otherwise 0. Additionally, let $M^{\phi}$ represent the upper limit on the number of patients requiring resource $\phi$. Upper limits for each pattern can be imposed on each specified resource with the following constraint:

$$\sum_{i \in I_o} g_i^{\phi} z_{i,p} \leq M^{\phi}, \forall p \in P, \quad o \in O, \quad \phi \in \Phi \tag{43}$$

   *Overtime Verification*: A pattern $p$ is feasible as long as the risk of exceeding the block's capacity $C_{d,r}$ is no more than $\delta$. The formulation is given with equation (5) as presented in Section 4.1.1. Monte Carlo sampling is used with historical data to solve this equation. Patterns that are not feasible towards overtime are eliminated from the set $P$.

   - *Exception:* An exception is given to patterns that consists of a single patient as some of these patterns may span the whole day and exceed the block capacity with a high probability.

During the verification of the patterns, statistics on regular ($\delta_p$) and extended ($\delta_p^{\Delta}$) overtime were gathered using Equations (22) and (23) from Section 5.1.

#### MIP Model

The scheduling problem has now been reduced to assigning the feasible patterns of elective patients to blocks while reserving gaps for future semi-acute elective arrivals.

As overtime may cause last-minute cancellations, the number of patterns having regular and extended overtime is minimised, as presented in Section 5.1.1. Additionally, the maximum peak in ward bed occupancy is minimised to avoid cancellations.

The restrictions outlined in Section 5.1.2 are applied to determine the assignment of patterns to blocks. However, an additional rule is implemented to restrict surgeries requiring a robot to specific blocks in the MSS. Let $(d,p,r) \in DPR \subseteq D \times P \times R$ represent a reduced set taking these restrictions into account and let $x_{d,p,r}$ be a binary decision variable taking the value 1 if pattern $p$ is assigned to day $d$ and room $r$. The following constraints and objectives are then applied as described in Section 5.1.2:

$$\begin{aligned} \min \quad & \lambda^{ov}\Big( \sum_{d\in D, r\in R} \big(u_{d,r} + \lambda^{ov} v_{d,r}\big) \\ & + \sum_{(d,p,r)\in DPR:} \big(\big[\delta_p\big]^2 + \lambda^{ov}\big[\delta_p^{\Delta}\big]^2\big) x_{d,p,r}\Big) \end{aligned} \tag{44}$$

$$\text{s.t.} \quad \sum_{(d,p,r)\in DPR: i\in I_p} x_{d,p,r} = 1, \quad \forall i \in I \tag{45}$$

$$\sum_{p\in P:(d,p,r)\in DPR} x_{d,p,r} \leq 1, \quad \forall d \in D, \quad r \in R \tag{46}$$

$$\sum_{p\in P_o, r\in R:(d,p,r)\in DPR} x_{d,p,r} \leq 1, \quad \forall d \in D,\ o \in O \tag{47}$$

$$\sum_{(d,p,r)\in DPR:\delta_p \geq \delta'} x_{d,p,r} \leq u_{d,r} \tag{48}$$

$$\sum_{(d,p,r)\in DPR:\delta_p^{\Delta} \geq \delta^{\Delta'}} x_{d,p,r} \leq v_{d,r} \tag{49}$$

The objective function (44) consists of two terms. The first term minimises the number of times patterns with regular ($u_{d,r}$) and extended ($v_{d,r}$) overtime are selected. Those values are determined with Constraints (48)–(49). The second term is added for the degree of surpassing the accepted risk limits by minimising the squared probabilities of $\delta_p$ and $\delta_p^{\Delta}$. A penalty cost of $\lambda^{ov} \gg 1$ is added to the extended overtime for both terms, while the overall objective receives a penalty cost of $\lambda^{ov}$. For a complete description of the objective, see Section 5.1.2. Constraint (45) ensures that each patient is scheduled at most once. Constraint (46) imposes an upper limit on the number of patterns assigned to a single block, as patterns span the entire day, according to the MSS. Constraint (47) limits the number of patterns assigned to each surgeon each day.

There cannot be more than $M_{ICU}^{A}$ patients in the ICU on any given day $d$ following their

surgery due to a limited number of downstream ICU beds. The ICU restrictions are accounted for as follows:

$$\bar{n}_d^{ICU} + \sum_{\substack{r\in R, p\in P, j\in\{0,1,\ldots,M^I-1\}:\\ (d-j,p,r)\in DPR}} Q_{j,p}^{ICU} x_{d-j,p,r} \leq M_{ICU}^A, \quad \forall d \in D \tag{50}$$

In this case, $Q_{j,\,p}^{ICU}$ denotes the number of patients in ICU from pattern $p$ on day $j$ following their surgery, performed on day $(d-j)$, and $M^I$ represents the upper bound on the LOS in ICU for each patient. Some patients may still be in the ICU from previous planning horizons, potentially affecting bed availability in the current planning horizon. As a result, the parameter $\bar{n}_d^{ICU}$ is utilised to denote these numbers and is carried out with simulation using the previous weeks' known schedule. Ward beds are also accounted for in a similar way to Constraint (50). Alternatively, it is assumed that ward beds are infinite, with the peak in the ward bed occupancy being minimised. The following soft constraint is employed to combat this effect:

$$\bar{n}_d^{w} + \sum_{\substack{r\in R, p\in P, j\in\{0,1,\ldots,M^W-1\}:\\ (d-j,p,r)\in DPR}} Q_{j,p}^{w} x_{d-j,p,r} \leq w^{\max} \qquad \forall d \in D \tag{51}$$

where $Q_{j,\,p}^{w}$ denotes the number of patients in the ward from pattern $p$ on day $j$ following their surgery on day $(d-j)$ and $\bar{n}_d$ represents the number of patients remaining in the ward from the previous planning horizon, computed using Monte Carlo sampling. The variable $w^{max}$ determines the maximum ward bed occupancy across the planning horizon. Finally, the maximum peak in the ward bed occupancy is minimised by adding the following term to the objective function:

$$\min \quad \lambda^w w^{max} \tag{52}$$

where $\lambda^w$ is a weight added to the maximum peak in ward bed occupancy. For example, suppose the maximum peak of ward bed occupancy in the planning horizon is three. This means at least three beds are required on one or more days. If $\lambda^w = 10$, then the cost in the objective function would be 30. Thus, by implementing these soft upper limits, the optimisation will balance out the ward occupancy.

Constraints (45)-(49), (50)–(51) and Objectives (44), (52) will be referred to as the *Base Model* from this point onward. Three strategies for reserving gaps for semi-acute elective arrivals are presented in the following paragraphs. Each strategy builds upon the *Base Model* but with additional constraints or objectives implemented.

### Front Load

The Front load strategy is implemented by adding a term to the objective function that gradually increases the daily cost of patient assignments. For instance, if a patient is assigned to a block on day 1, the cost is 1. However, if the patient is assigned to day 19,

the cost increases to 19. As a result, elective patients are naturally scheduled towards the front of the planning horizon to minimise costs:

$$\min \quad \lambda^{FL} \sum_{(d,p,r)\in DPR:P_s\subseteq P} dx_{d,p,r} \tag{53}$$

where $\lambda^{FL}$ is a weight added to the objective.

The strategy consists of Constraints (45)-(49), (50)–(51) and Objectives (44), (52)–(53).

### Daily Limit

To implement the Daily Limit strategy, the binary indicator variable $\bar{R}_{d,r}$ is introduced, taking the value 1 if on day $d$ in room $r$ there are more than $N^s$ patients. The following constraint is added to the *Base Model*, taking this into account:

$$\sum_{p\in P:\ (d,p,r)\in DPR,\ C_p>N^s} x_{d,p,r} \leq \bar{R}_{d,r}, \quad d \in D, \quad r \in R \tag{54}$$

Additionally, the following term is added to the objective function:

$$\min \qquad \lambda^{DL} \sum_{d\in D, r\in R} \bar{R}_{d,r} \tag{55}$$

which minimises the number of patterns selected where the number of surgeries exceeds $N^s$, with a weight $\lambda^{DL}$ applied to the results. In other words, suppose that in the planning horizon, there are four patterns with four patients and two with three patients. The sum of the patterns with more than three patients is four. Thus, if $\lambda^{DL} = 10$, the cost would be 40. As a result, the optimisation will minimise the number of such patterns to minimise the total cost.

The strategy consists of Constraints (45)-(49), (50)–(51), (55) and Objectives (44), (52), (55).

### Balance ORs

A balance in the number of blocks utilised each week can be achieved by minimising the maximum number of blocks used each week in the planning horizon. To combat this effect, the set $V$ is introduced to the model, specifying the weeks of the planning horizon. Additionally, the parameter $D^v$, specifying the week corresponding to day $d$, and, the variable $V^B$, which determines the maximum number of blocks used in any week, are introduced. The variable $V^B$ is defined in the following way:

$$\sum_{(d,p,r)\in DPR:\ D^v=v} x_{d,p,r} \leq V^B, \quad \forall v \in V \tag{56}$$

To minimise the maximum number of blocks used each week, the following term is added to the objective function:

$$\min \quad \lambda^{BO} V^B \tag{57}$$

where the weight $\lambda^{BO}$ is multiplied by the number of blocks that can be kept constant each week. That is, suppose a planning horizon consisting of four weeks, with the number of blocks, utilised each week being 9, 10, 10 and 9, then $V^B$ is 10.

The strategy consists of Constraints (45)-(49), (50)–(51), (56) and Objectives (44), (52), (57).

### 6.2.2 Rescheduling Model

#### Pattern Regeneration

The same practical rules and restrictions on overtime outlined in Section 6.2.1 are applied to generate new patterns and modify the existing ones to accommodate semi-acute elective patients. Additionally, three practical rules are introduced. The first rule allows semi-acute elective patients to enter the patterns of other surgeons, the second rule imposes an upper bound on the number of semi-acute elective patients in a pattern, and the third ensures that the ratio of elective patients to semi-acute patients in a pattern is higher. The first rule is implemented due to the high medical priority of semi-acute elective patients. Rules two and three, however, are added to maintain a balance between these different types of elective patients.

Let $a_i$ be a binary parameter taking the value 1 if patient $i$ is a semi-acute patient; otherwise, 0. If $a_i = 1$, the patient can enter the patterns of other surgeons as long as the practical rules outlined in Section 6.2.1 are satisfied. Additionally, the following rules are applied:

1. *Semi-Acute Quota*: An upper bound is imposed to the number of semi-acute elective patients assigned to each pattern $p$ to $\bar{M}^a$ with the following constraint:

$$\sum_{i \in I} a_i z_{i,p} \leq \bar{M}^a, \quad \forall p \in P \tag{58}$$

2. *Elective Patient Ratio Quota*: A heuristic is introduced to maintain a balance between elective and semi-acute elective patients in a pattern. That is, if the number of semi-acute patients $a_i$ assigned to a pattern exceeds a predefined ratio $\bar{n}^a$, then the ratio of elective patients to semi-acute elective patients in the pattern $p$ must be higher than that of the semi-acute elective patients:

$$\frac{\sum_{i \in I} a_i z_{i,p}}{\sum_{i \in I} z_{i,p} - a_i z_{i,p}} < 1 \quad \forall p \in P \tag{59}$$

For example, suppose the threshold $\bar{n}^a$ is 2. In this case, a pattern of 6 patients becomes infeasible if the number of semi-acute elective patients exceeds 2, as the ratio would be 1 or higher and would not be generated.

### MIP Model

### Scheduling

To assign the semi-acute elective patients to blocks, the *Base model* is used as outlined in Section 6.2.1 with Constraints (45)-(49), (50)–(51) and Objectives (44), (52). However, one modification is made to the reduced set $(d, p, r) \in DPR$ assuming that surgeons can operate on semi-acute elective patients in all blocks. This assumption allows surgeons to share a single pattern. As a result, Constraint (47) must be modified to combat this effect. If one assumes that the patterns selected are likely to fill the entire day due to the optimisation criteria, the following constraint can be used. Let $\Upsilon_{o,p}$ be a parameter specifying the proportion of patients of the surgeon $o$ in the pattern $p$ relative to the total number of patients of the pattern then, the approximation is implemented as follows:

$$\sum_{r \in R,\ p \in P_o : (d,p,r) \in DPR} x_{d,p,r} \Upsilon_{o,p} \leq 1, \quad \forall d \in D, \quad \forall o \in O \tag{60}$$

This constraint is added in an attempt to prevent surgeons from exceeding their total block capacity in a day, as they can now be allocated to multiple patterns due to semi-acute elective arrivals. For instance, suppose a single surgeon is allocated to two blocks in a day where, in one block, the surgeon's proportion of patients from the total patients in each block is 0.4 and 0.5, respectively. In this case, the constraint is satisfied as $0.4 + 0.5 \leq 1$.

### Rescheduling

To account for rescheduling, it is necessary to determine how much the current schedule differs from the previous one. Let $\Psi_i$ be a variable indicating the day patient $i$ is assigned to in the current schedule and computed as follows:

$$\sum_{(d,p,r) \in DPR : i \in I_p} d x_{d,p,r} = \Psi_i, \quad \forall i \in I \tag{61}$$

Let $\gamma_i$ be a parameter denoting the day patient $i$ was assigned to in the previous schedule. Patients can typically be rescheduled both backwards and forwards in time. Nonetheless, rescheduling a patient backwards is prohibited as the patient may not be ready. To prohibit backwards rescheduling, the following constraint is applied:

$$\gamma_i - \Psi_i \leq 0, \quad \forall i \in I \tag{62}$$

For example, if patient $i$ was assigned to day 2 in the previous schedule ($\gamma_i = 2$) and rescheduled to day 1 in the current schedule ($\Psi_i = 1$), it would not be permitted as the difference between the days is higher than 0. Moving a patient forward is permitted and typically done in practice. Thus, let $\Delta_i^+ \geq 0$ be a variable which equals the total number of days patient $i$ has been moved forward:

$$-\gamma_i + \Psi_i \leq \Delta_i^+, \quad \forall i \in I \tag{63}$$

For example, if the patient from the previous example is rescheduled to day 3 ($\Psi_i = 3$), then the variable $\Delta_i^+$ is set to 1. To count the total number of changes between the previous and the current schedule for all patients, the binary indicator variable $\Delta_i$ is introduced, taking the value of 1 if patient $i$ is assigned to a different date in the current planning horizon using the following constraint:

$$M\Delta_i \geq \Delta_i^+, \quad \forall i \in I \tag{64}$$

where $M$ is a large number and is used to ensure that the binary variable takes the value of 1 when $\Delta_i^+ > 0$. The objective is now to minimise the number of changes or:

$$\min \quad \lambda^{RS} \sum_{i \in I} \Delta_i \tag{65}$$

where $\lambda^{RS}$ is a weight for the rescheduling. An additional penalty is applied to rescheduling of patients with a high medical priority (the set $S^p$). These patients should preferably not be rescheduled to a different date. Thus, the following term is added to the objective function:

$$\min \quad \lambda^{RS} \sum_{\substack{i \in I; \\ i \in S^p}} \Delta_i \tag{66}$$

The scheduling and rescheduling consist of the constraints and objectives of the *Base Model* along with the constraints and objectives presented in this section. In other words, the model consists of Constraints (45)-(49), (50)–(51), where Constraint (47) is replaced with Constraint (60) and Objectives (44), (52). In addition, the constraints and objectives for the different gap-reserving strategies are included individually. The rescheduling part of the model consists of Constraints (61)-(64) and (65)-(66).

Full formulations for each of the scheduling and re-scheduling models using the different gap-reserving strategies are provided in Appendix C.1, Appendix C.2 and Appendix C.3.

## 6.3 Computational Experiments

The computational experiments aim to compare the Front Load, Daily limit and Balance ORs gap-reserving strategies to understand how they impact the need of rescheduling,

block utilisation and overtime. Additionally, the results are compared to actual scheduling data of GS speciality for a single year. The pattern generation was programmed in C, whereas the MIP model was programmed in AMPL and solved with Gurobi.

### Data

To compare the optimised results to actual data, 12 four-week intervals were selected from a single year, each consisting of 28 days. For each interval (hereafter month), the same set of patients was scheduled as was done by the speciality.

For the given year, a total of 1141 patients were scheduled, with a median of 109 per surgeon. Of those patients, 362 $(32\%)$ were admitted to the ward after their surgery. In this section, an assumption is made that semi-acute elective patients are patients who were added to the waiting list during the execution of the schedule each month and underwent surgery within the same month. As a result, 312 patients $(27\%)$ were classified as semi-acute elective patients. Of those, 93 there were in-patients $(30\%)$. Finally, it is assumed that elective patients are the patients who were on the waiting list on the Friday preceding the first week of each month.

Tables 3.1 and 3.2 provide summary statistics for each surgeon belonging to the GS speciality. The tables demonstrate that the case mix differs across the surgeons. For example, surgeon 8 has high patient throughput but a low ratio of semi-acute arrivals. Surgeons 5 and 7 have high patient throughput and a high ratio of semi-acute elective arrivals, but the average surgery times and the ratio of in-patients are significantly lower for surgeon 7. Surgeon 9 has patient throughput close to the median but has the highest ratio of semi-acute elective arrivals. Similar to surgeon 7, the average surgery times and the ratio of in-patients are relatively low. These numbers suggest variability in the number of semi-acute arrivals within the speciality and that it may be easier to accommodate semi-acute arrivals when surgery times are shorter and admittance to the ward is not required.

### Parameter Settings

LOS distributions for ward and ICU occupancy are generated for each patient requiring these resources, based on the patient's type of operation, using a maximum of $M^W = 14$ and $M^{ICU} = 14$ days, respectively.

Table 6.1 provides the remaining parameter settings for the pattern generation. To regenerate the patterns, it is assumed that each pattern can accommodate no more than $\bar{M}^a = 4$ semi-acute elective patients. However, if the number of semi-acute patients exceeds $\bar{n}^a = 2$, then the ratio of elective patients must be greater than the ratio of semi-acute patients in the pattern. Table 6.2 provides the parameter settings for the

scheduling and rescheduling models, and Table 6.3 details the weights used by the objective function and their order of priority. The weights are chosen to guarantee that the different objectives are achieved in the specified order.

*Table 6.1. Parameters settings used to generate the feasible patterns.*

| **Practical Rules** | **Overtime verification** |
|---|---|
| • Patient Quota: $M^P = 6$ is the maximum number of patients that can be assigned to a block. | • Block Capacity: There are two distinct block lengths or $C_{d,r} = \{315,\ 450\}$ minutes. |
| • Resource Quotas: Upper bounds are imposed to several downstream resources $\phi \in \Phi$ where $\Phi = \{pacu,\ ward,\ icu\}$, to hedge against exceeding their upper limits $M^\phi$. For the computational experiments, the upper limits to those resources are set to $5, 2$ and 1, respectively. | • Overtime restrictions: Threshold on the probability that a pattern surpasses $C_{d,r}$ is set to $\delta = 0.95$.<br>• Extended block length: The time added to the extended block length to $\Delta_{d,r} = 60$ min. |

*Table 6.2. Parameters used for the scheduling and rescheduling models.*

| **Parameters** | |
|---|---|
| • Regular overtime: $\delta^{\delta'} = 0.25$. | • $M^A_{ICU}$: An upper limit on the number of staffed ICU beds is set to 1. |
| • Extended overtime: $\delta^{\Delta'} = 0.25$. | • $N^S$: The parameter is set to 3 surgeries. |
| • Priority: Patients with a one-week priority have to be scheduled within 14 days (based on historical data), and patients cannot be assigned to blocks before they are registered on the waiting list. | |

### 6.3.1 Comparison of the Gap-reserving Strategies

Table 6.4 summarises the overall results of the experiments for the 12 months (12 four-week planning horizons), but detailed results for each month are provided in Appendix D. The statistics outlined are used to compare the three gap-reserving strategies, which are also compared with the actual outcome.

The Base Schedule reveals that the percentage of total capacity reserved for semi-acute

*Table 6.3. Weights used for the objective function and are selected to ensure the order of priority.*

| Objective | Order of Priority | Cost |
|---|---|---|
| $\lambda^{RS}$ | 1 | 10000 |
| $\lambda^{ov}$ | 2 | 100 |
| $\lambda^{w}$ | 2 | 100 |
| $\lambda^{DL}$ | 3 | 100 |
| $\lambda^{BO}$ | 3 | 100 |
| $\lambda^{FL}$ | 4 | 0.1 |

elective arrivals is around 50% for each strategy and is comparable to the actual outcome. However, there is a large difference in number of blocks in use and the utilisation within each block. Using the *Balance ORs* strategy, 360 of the 576 available blocks are used, leaving 38% of the blocks empty. The numbers are comparable to the numbers for the *Front load* strategy. Using the *Daily limit* strategy, 459 of the 576 available blocks are used, leaving only 20% of the blocks empty. The percentage of empty blocks is 30% for the actual outcome. The number of times overtime is needed and peaks in ward occupancy are similar across the different gap-reserving strategies and, in all cases, lower than the actual outcomes.

By looking at the Final Schedule, it is possible to see that the total capacity utilised ranges from 67% to 69%, being lowest for the actual outcome and highest for the *Front load* strategy. As for the reserved capacity in the base model, the total capacity utilised is similar across the three strategies, but there are significant differences in the number of open blocks and the utilisation within each block. Using the *Front load* strategy, 25% of the available blocks are not used, which is comparable to the 23% that are empty when applying the *Balance ORs* strategy. Again, the application of the *Daily limit* strategy results in the highest number of blocks in use and the lowest number of empty blocks (16%), with the actual outcome in between at 19%.

Looking at the use of overtime in the Final Schedule, there are noticeable differences across the three strategies. When using the *Front load* strategy, surgeons are most likely to require overtime, while they are least likely to do so when using the *Daily limit* strategy. This applies both to regular and extended overtime. The use of overtime is in between when applying the *Balance ORs* strategy but closer to the *Daily limit* strategy. For the actual outcome, the results are closer to the *Front load* strategy for regular overtime and closer to the *Daily limit* strategy for extended overtime.

Analysing the peaks in ward occupancy in the Final Schedule, the results are comparable across the different strategies. The *Front load* strategy has the highest average of ward occupancy and relatively high variance. The average value is similar for the *Daily limit* and the *Balance ORs*, but the use of the *Balance ORs* strategy results in slightly higher variance. The actual outcome has the lowest average value and the lowest variance.

*Table 6.4. Comparison of the experiment results for three gap-reserving strategies. Results are shown for the base model (Base Schedule) and after rescheduling (Final Schedule).*

| | **Base Schedule** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Blocks** | | | | **Overtime** | | **Ward** | | | **SR time** | |
| | **#Open** | **Utilisation** | **Reserved Cap.** | **Empty** | **#Regular** | **#Extended** | **Mean** | **[min,max]** | **#Resch.** | **Median** | **Mean** |
| Actual | 406 | 0.71 | 0.50 | 0.30 | 121 | 65 | 3.30 | [1.20,6.30] | - | - | - |
| Front load | 376 | 0.78 | 0.49 | 0.35 | 75 | 33 | 3.25 | [1.16,5.08] | - | - | - |
| Daily limit | 459 | 0.63 | 0.50 | 0.20 | 72 | 41 | 3.28 | [1.38,5.18] | - | - | - |
| Balance ORs | 360 | 0.81 | 0.49 | 0.38 | 78 | 37 | 3.28 | [1.20,5.32] | - | - | - |
| | **Final Schedule** | | | | | | | | | | |
| Actual | 466 | 0.83 | 0.67 | 0.19 | 223 | 112 | 4.50 | [1.90,7.70] | - | - | |
| Front load | 430 | 0.92 | 0.69 | 0.25 | 231 | 168 | 4.59 | [1.57,7.97] | 89 | 2 | 3.21 |
| Daily limit | 486 | 0.81 | 0.68 | 0.16 | 174 | 98 | 4.53 | [1.88,7.72] | 65 | 1 | 2.51 |
| Balance ORs | 444 | 0.88 | 0.68 | 0.23 | 195 | 129 | 4.52 | [1.56,8.11] | 52 | 1 | 1.96 |

Shown are the number of blocks in use (#Open), the average block utilisation, the reserved capacity from the allocated capacity (Reserved Cap.), and the percentage of empty blocks from the total number of blocks available. Additionally, the number of times surgeons require regular (#Regular) or extended (#Extended) overtime, the average ward bed occupancy (Mean) along with its minimum and maximum [min, max], the number of rescheduling events (#Resch), and the time between scheduled and rescheduled appointments (SR time) are shown.

Finally, looking at the rescheduling in the Final Schedule, a noticeable difference is evident across the three strategies. The use of the *Balance ORs* strategy results in the lowest rescheduling and is around half of the highest number, which is achieved using the *Front load*. The result for the *Daily limit* strategy is in between, closer to the results for the *Balance ORs* strategy. Similarly, using the *Balance ORs* strategy results in the shortest average time (in days) between the scheduled and rescheduled appointments (SR time), and the use of *Front load* leads to the longest average SR time.

Figures 6.3 and 6.4 illustrate the distribution of the number of patients operated on and the number of blocks used, respectively, across the four weeks of the planning horizon. The figures reveal that the number of patients per week and the number of blocks used per week are more variable when the *Front load* strategy is used compared to the other two strategies. The average number decreases for each week, and the variance grows larger. A similar pattern is observed for the actual outcomes in the Base Schedule but less so in the Final Schedule. However, there is not much difference between how the number of patients and the number of open blocks, and their variances, are distributed across the weeks for the *Daily limit* and *Balance ORs* strategies.

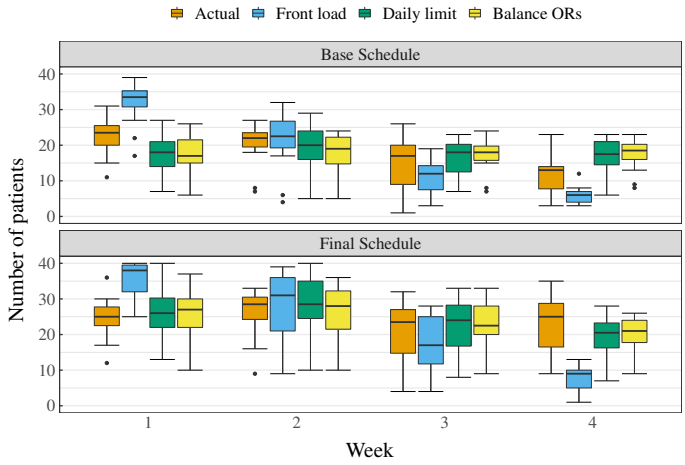



*Figure 6.3. Distribution of the number of patients scheduled per week across the planning horizon for the Base Schedule and the Final Schedule, divided by the three gap-reserving strategies and actual outcome.*

When looking more closely at how the different strategies affect rescheduling, the benefits of the *Balance ORs* strategy become more visible. Figure 6.5 shows the distribution of the SR time in days for each strategy. The figure suggests that the majority of the rescheduled patients are operated on less than a week after their original appointment. Furthermore, using the *Balance ORs* strategy results not only in the fewest number of rescheduling events but also in the shortest average SR time and the lowest variance in SR times. In comparison, using the *Front load* strategy results in the highest number of rescheduling events and the longest SR times with higher variance.

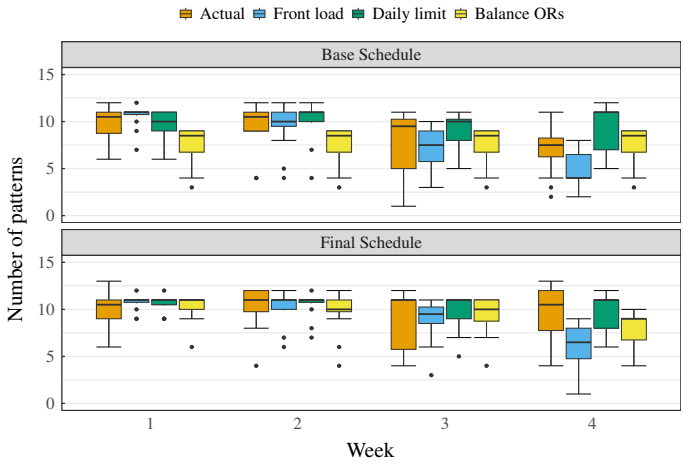



*Figure 6.4. Distribution of the number of open ORs (blocks used) per week across the planning horizon for the Base Schedule and the Final Schedule, divided by the three gap-reserving strategies and actual outcome.*

Similarly, the *Balance ORs* strategy generates the most desirable results when considering what types of operations are commonly rescheduled (see Table 6.5). The table shows that the most frequently rescheduled operation is $s-1$ for all three strategies, but it is least likely to be rescheduled when using the *Balance ORs* strategy. Furthermore, the rescheduled type of operations using the *Balance ORs* strategy have more variations in their average surgery times compared to the other two strategies, providing more flexibility.

Finally, the benefits of the *Balance ORs* strategy become more apparent by looking at the distribution of SR time, as well as the total number of rescheduling events, for each surgeon and strategy (see Figure 6.6). The figure shows, as has been mentioned before, that the use of the *Balance ORs* strategy results in the fewest rescheduled appointments, but in addition, the variation in SR time is low across all surgeons, which is not the case when using the other two strategies. However, when using *Front load* or *Daily limit* strategies, the variation is relatively high for both surgeon 5 and surgeon 7. Surgeon 5 has relatively high average surgery times and relatively many in-patients, whereas the opposite is the case for surgeon 7. The *Balance ORs* strategy seems to accommodate these polar cases more easily in the Base Schedule than the *Front load* or *Daily limit* strategies, resulting in less rescheduling, and the rescheduling is less disruptive for patients because the average SR time is shorter and with less variation.

*Table 6.5. Summary statistics for the five most common types of operations that were rescheduled under each of the three gap-reserving strategies.*

| Front load | | | | Daily limit | | | | Balance ORs | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Surgery | #Resch. | Ratio of Pat. | Avg. Surgery time | Surgery | #Resch. | Ratio of Pat. | Avg. Surgery time | Surgery | #Resch. | Ratio of Pat. | Avg. Surgery time |
| s-1 | 33 | 0.10 | 101.7 | s-1 | 22 | 0.06 | 101.7 | s-1 | 13 | 0.04 | 101.7 |
| s-9 | 4 | 0.10 | 93.9 | s-5 | 9 | 0.21 | 227.6 | s-2 | 8 | 0.21 | 118.2 |
| s-10 | 4 | 0.14 | 143.4 | s-6 | 3 | 0.16 | 151.6 | s-3 | 4 | 0.67 | 62.4 |
| s-11 | 3 | 0.13 | 149.1 | s-7 | 3 | 0.08 | 188.3 | s-4 | 3 | 0.18 | 49.1 |
| s-5 | 3 | 0.07 | 227.6 | s-8 | 2 | 0.25 | 283.3 | s-5 | 3 | 0.07 | 227.6 |

Shown are the number of rescheduling events (#Resch.), the percentage of patients undergoing the given type of operation who were rescheduled (Ratio of Pat.), and the average surgery time (Avg. Surgery time).

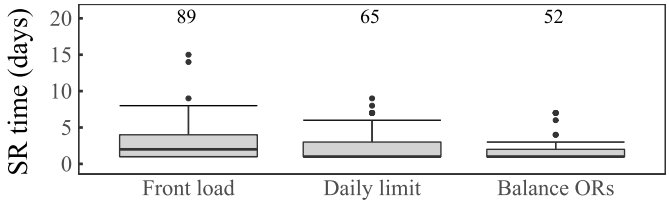



*Figure 6.5. Distribution of the days between scheduled and rescheduled appointments across the three gap-reserving strategies (SR time).*

The numbers above the boxes indicate the total number of rescheduling events.

## 6.4 Conclusions

This chapter focused on how to reserve gaps in a long-term elective schedule to accommodate semi-acute elective patient. Three gap-reserving strategies, two of which have been suggested in the literature (*Front load* and *Daily limit*) and one proposed (*Balance ORs*), were compared to understand how these strategies impact the need for rescheduling, while reducing the risk of last-minute cancellations, due to uncertainty in surgery times and LOS in the ward, and maintaining the existing utilisation. This was done while accounting for limited downstream resources and equipment availability. Scheduling and rescheduling models were proposed using the Pattern Scheduling approach where these gap-reserving strategies are implemented and results were compared for twelve four-week periods using the same set of patients as in the actual scheduling data each period. The use of the same set of patients across each period facilitated a level of comparability between the models and the actual scheduling data. This comparison offered insights into performance efficiency. However, it's worth noting that our understanding of rescheduling dynamics is limited, as the data only presents final schedules and lacks details regarding any rescheduling events that may have occurred. The primary focus is on understanding how different gap-reserving strategies impact the necessity for rescheduling.

The computational results demonstrate that the *Front load* strategy, which reserves gaps late in the planning horizon, promotes high block utilisation at the risk of extensive rescheduling early in the planning horizon. The *Daily limit* strategy reserves gaps by imposing a limit on the number of patients in each block to avoid rescheduling events. However, this comes at the risk of lower block utilisation because the semi-acute elective arrivals are not evenly distributed across the planning horizon. To strike a balance between the benefits of these two strategies the *Balance ORs* strategy is suggested. Using the *Balance ORs* strategy, gaps are reserved evenly across the planning horizon,

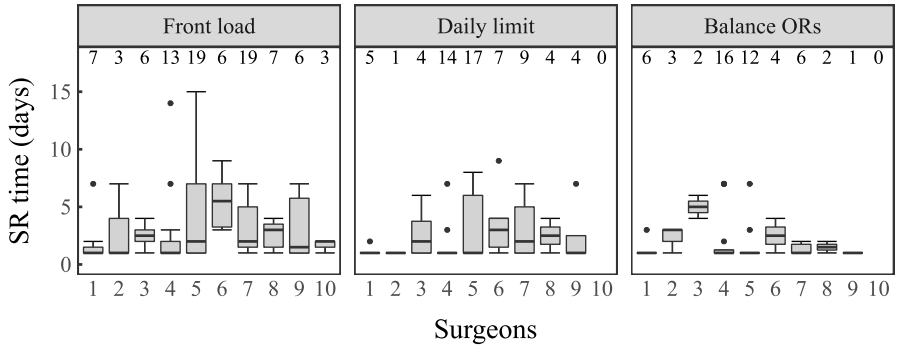



*Figure 6.6. Distribution of the number of days between scheduled and rescheduled appointments (SR time) across the three gap-reserving strategies and surgeons.*

The numbers above the boxes indicate the total number of rescheduling events.

as is done when using the *Daily limit* strategy, but the maximum number of blocks in use is minimised for a given throughput and case-mix to increase block utilisation. However, it is essential that this is not achieved through extensive overtime or by overextending limited downstream resources. The results further suggest that by using the *Balance ORs* strategy, block utilisation is obtained that is similar to what is achieved when using the *Front load* strategy but requires less overtime and produces lower peaks in ward occupancy. Additionally, the time between scheduled and rescheduled appointments is also shorter.

The *Balance ORs* strategy is designed to reduce the maximum number of blocks utilised over a four-week planning period. It achieves this by striking a balance in the number of blocks utilised each week. The outcome of this optimisation reveals a heuristic strategy that schedulers might find useful: filling eight blocks each week but keeping two blocks vacant for rescheduling needs and semi-acute arrivals. Currently, each surgeon is allocated a single block each week. However, implementing a balanced gap-reserving strategy would require surgeons to share blocks in the MSS. This anticipates that their daily plan will likely open up another block and accommodate semi-acute arrivals on-demand. The specific days the empty blocks should be planned for will depend on the availability of the surgeons.

## 6.5 Summary

This chapter explored three gap-reserving strategies that can be used to reserve capacity in the long-term elective program for the unpredictable arrivals of semi-acute elective

patients to understand how they impact the need for rescheduling. The literature suggests two of the strategies, but the third is proposed in this work. Scheduling and rescheduling models were proposed using the two-step Pattern scheduling approach. The gap-reserving strategies were implemented and compared over twelve four-week periods.

The computational results revealed that the proposed strategy, which balances the number of blocks utilised weekly, results in the least number of rescheduling events and the time between scheduled and rescheduled appointments is also the lowest. Compared to a strategy assigning patients early in the planning horizon, a similar block utilisation is achieved, but less overtime is required. Finally, when compared to a strategy that restricts the number of patients in all blocks, a higher block utilisation is achieved. Although the two strategies suggested by the literature may be simpler to implement when scheduling by hand, the proposed strategy suggests a new heuristic: leave 20% of the blocks each week empty for semi-acute elective arrivals. In the following chapter, discussion and conclusions of this thesis are provided.

# 7 Discussion and Conclusions

This research aimed to increase our understanding of how to schedule a high throughput of elective patients under a limited amount of resources while minimising the risk of last-minute cancellations and rescheduling events in advance. The objective was to develop mathematical models that address the advance scheduling of elective patients in a practical manner by accounting for multiple sources of uncertainty to avoid the risk of last-minute cancellations and rescheduling events. However, the practical models must also enable a high throughput of elective patients while considering limited downstream resources.

Three modelling objectives were specified to address three sources of uncertainty, namely: surgery times (Hans et al., 2008; Batun et al., 2011; Kroer et al., 2018), downstream length of stay (Min and Yih, 2010; Jebali and Diabat, 2017), and semi-acute elective arrivals (Zonderland et al., 2010). These sources were selected as they cause the most disruptions for elective patients (Min and Yih, 2010; Van Riet and Demeulemeester, 2015; Riise et al., 2016; Zhang et al., 2019). The first two sources are associated with last-minute cancellations which can occur when an operating room on a given day (block) runs into overtime or when the staffed ward beds are exceeded. The third source of uncertainty is associated with rescheduling of previously planned elective patients, which can occur due to unpredictable semi-acute elective arrivals that must be accommodated into the long-term elective schedule. A single surgical speciality, General Surgery at Landspitali Hospital, was selected for the computational experiments. The results of each experiment were then compared with the speciality's actual scheduling data.

The results demonstrate that uncertainty in surgery times can be specified in a statistically accurate way using chance constraints. These constraints can effectively hedge against last-minute cancellations that can occur when a block capacity is exceeded by bounding the risk to a predefined threshold. However, as the problem size increases, models implementing the chance constraints become computationally intractable (Hans et al., 2008), which is impractical. As a result, approximations or other complex heuristics are often proposed (Shehadeh, 2022). This thesis proposed a novel practical approach called Pattern Scheduling, enabling uncertainty in surgery times to be specified both practically and statistically accurately. A pattern is defined as an unordered combination of one or more patients assigned to a block for a single surgeon. The approach consists of two steps. In the first step, feasible patterns are generated by considering practical rules

and probabilistic restrictions on overtime. In the second step, the feasible patterns are assigned to blocks using a mixed-integer programming (MIP) model that accounts for limited ward capacity. The approach is practical in two ways. First, it is computationally practical since the search space is reduced and does not require using approximate models. Second, it is clinically practical as the generated patterns follow practical rules of the speciality, further limiting the search space. The computational results showed that the approach is tractable up to a certain length of patient list. By specifying practical rules when generating the patterns, it is possible to reduce the problem size, enabling large problems to be solved with an off-the-shelf solver while bounding the risk of exceeding the block capacity. Further results showed that planning beyond one week and flexibility in the roster of the surgeons are essential to ensure the best result in terms of a balanced flow of in- and out-patients.

The risk of last-minute cancellations due to uncertainty in surgery times can be reduced by using Pattern Scheduling. However, with this approach, exceeding the staffed ward beds was still possible, raising the risk of last-minute cancellations. Consequently, Ward combinations (WCs) were proposed, enabling the MIP model to hedge against exceeding the ward bed capacity and thereby reduce the risk of last-minute cancellations. A model utilising the ward combinations was compared to a robust formulation that used the worst case outcome for the length of stay in the ward following surgery. Computational experiments showed that higher quality solutions are achieved using ward combinations optimisation (WCO) compared to robust ward optimisation (RWO) in terms of days in overtime and ward utilisation but at the cost of computational time. However, the computational time is affected by the robustness of the solution, such as the discretisation error of the LOS distributions and the threshold set for exceeding ward bed capacity, and the number of staffed ward beds. Consequently, a trade-off between computational tractability and the robustness of the solutions is observed. Compared to actual scheduling data, both WCO and RWO produce higher quality solutions in terms of overtime and the overall ward numbers, suggesting that both approaches can be utilised to hedge against the combined risk of last-minute cancellations.

Even if the risk of exceeding the block and downstream bed capacity is bounded simultaneously, the potential rescheduling of other elective patients may still occur due to the unpredictable semi-acute elective arrivals that must be accommodated within the existing schedule. However, our knowledge of how gaps should be reserved when elective patients are scheduled up to weeks in advance is limited. As a result, three distinct gap-reserving strategies were compared, two of which have been suggested by the literature (Front load and Daily limit) and one proposed in this thesis (Balance ORs) to understand their impact on rescheduling needs. Scheduling and rescheduling models were proposed to implement these different gap-reserving strategies while accounting for practical rules and limited downstream capacity and without resorting to overtime. The results demonstrate that the Balance ORs strategy, which maintains a balance in the number of blocks used each week, minimises the number of patients rescheduled and the time between two rescheduled appointments. However, the strategy obtains a similar block utilisation compared to the Front load strategy but requires less overtime. Finally, this strategy achieves a higher block utilisation than the Daily limit strategy. While

gap-reserving strategies such as Front load and Daily limit are more straightforward for a human scheduler to apply, the optimisation results may suggest a new gap-reserving strategy that human schedulers can apply. This strategy involves leaving 20% of the blocks empty each week, on average, for semi-acute elective arrivals and rescheduling of other patients. However, this might require the surgeons to operate on days outside their roster.

The overall results of this thesis highlight the importance of using chance constraints in mathematical models to hedge against the combined risk of last-minute cancellations. Pattern scheduling makes it possible to reduce the combined risk of last-minute cancellations using chance constraints and practical rules while avoiding computational issues. By leaving 20% of the blocks in the master surgery schedule empty each week, it is possible to reduce the number of rescheduling events required to accommodate semi-acute elective patients while maintaining a high block utilisation rate.

Even though the results are promising, there are two limitations. First of all, patients were assumed to be available on their scheduled dates. In practice, however, finding elective patients who have waited for months and are willing to come on short notice is often problematic and restricts the scheduling possibilities. Finally, more practical rules may exist beyond those currently implemented in the models of this thesis, which may further restrict the scheduling possibilities. For example, certain combinations of operation types or multiples of the same type of operation may not be assigned together in a block, e.g., due to supply limitations.

## 7.1 Practical Implications

The results of this thesis yield three heuristics that practitioners can adopt. Appendix B.1 provides a heuristic detailing the Pattern Scheduling approach and its implementation. Appendix B.2 follows with a description of how practitioners can generate and implement ward combinations to reduce the risk of last-minute cancellations caused by ward bottlenecks. Finally, it is suggested that 20% of the blocks are left empty each week when scheduling elective patients weeks in advance. These empty blocks can later accommodate semi-acute arrivals and rescheduling of other patients. However, this strategy may require surgeons to operate on days outside their roster.

## 7.2 Future Research

Future research should incorporate the unpredictability of elective patient availability into models. Most models in the academic literature assume that patients are available

for the blocks to which they are assigned to, including those in this thesis. However, this assumption may not be realistic. There are several ways possible to incorporate this into models. One way is to use a stochastic waiting list, thereby reducing the availability of individuals for each schedule. However, data collection is the main challenge when considering stochastic availability. Some patients with high medical priority, for example, may have a restricted time window for their operation, as they must undergo treatment, such as chemotherapy, before their operation and need to be operated on within a week following their treatment. It might be possible to collect data for such restrictions if they are general and can be linked to specific types of operations. Collecting data on the availability of patients with lower medical priority is more problematic, as the waiting list consists of hundreds and may not be feasible to perform. One way to address this problem is to collect logs from the hospital's scheduling system to understand how and when the plans are changed. Another way is to generate theoretical instances of the availability. Given that data can be collected, different scenarios regarding the availability of the patients awaiting surgery can be generated, and optimisation can be performed using the scenarios. In general, this would increase our understanding of how such restrictions impact the overall scheduling process.

Another avenue for future research is to investigate how surgical specialities generate their patterns. Currently, the emphasis in the literature has been on solving large instances as quickly as possible using complex algorithms. Although this work is essential, greater emphasis should be placed on identifying practical constraints to further restrict the problem space. This could involve using machine learning techniques to analyse a large amount of historical data and identify constraints that could be applied. In this regard, simple heuristics could be proposed that human schedulers can apply when scheduling by hand. Similarly, medical priorities should be researched to a greater extent. A general assumption made in the existing literature is to give each patient a priority score based on the product of their days on the waiting list and medical priority. Patients are then selected from the waiting list for days based on those scores. In practice, however, different types of operations may have internal priorities despite being assigned the same medical priority, and thus, a different methodology for patient selection is needed. One could use techniques similar to inverse optimisation to understand the actual values of the medical priorities of type of operations or even individual patients.

Finally, the RWO could be extended further by discretising the worst-case LOS for a single patient into at least three intervals. This would reduce the possibility of exceeding the staffed ward beds when there are many patients in the ward. In doing so, a formulation similar to the WCO would be required.

# 8 Contributions

All contributions in this thesis have been published as journal articles, conference proceedings, conference presentations and posters. The papers published as a part of this thesis are listed below:

**Paper I:** T.P. Runarsson and **A.O. Sigurpalsson**, 2019. Towards an evolutionary guided exact solution to elective surgery scheduling under uncertainty and ward restrictions, IEEE Congress on Evolutionary Computation (CEC), 2019, pp. 419-425, doi: 10.1109/CEC.2019.8790174.

**Paper II:** **A.O. Sigurpalsson**, T.P. Runarsson, R.J. Saemundsson, 2020. Stochastic Master Surgical Scheduling Under Ward Uncertainty. In: Bélanger, V., Lahrichi, N., Lanzarone, E., Yalçındağ, S. (eds) Health Care Systems Engineering. ICHCSE 2019. Springer Proceedings in Mathematics Statistics, vol 316. Springer, Cham, doi:10.1007/978–3–030–39694–7_13

**Paper III:** **A.O. Sigurpalsson**, T.P. Runarsson and R.J. Saemundsson, 2022. Bounding the Likelihood of Exceeding Ward Capacity in Stochastic Surgery Scheduling, Applied Sciences 12, no. 17: 8577. https://doi.org/10.3390/app12178577

**Paper IV:** **A.O. Sigurpalsson**, R.J. Saemundsson and T.P. Runarsson, 2025. Mind the Gap: Strategies for Semi-Acute Patient Scheduling in Elective Surgery. Submitted.

The following author contributions apply to the papers:

A.O.S. and T.P.R. conceptualised the model presented in Paper 1. T.P.R. developed the evolutionary algorithm, A.O.S., and the scheduling model, and both authors verified the algorithm and the model. The models in papers II-IV were conceptualised by A.O.S. and verified by T.P.R. and R.J.S. (Papers II-IV). All authors gave input to the methodology used. A.O.S. conducted the experiments and provided the computational resources for Papers II-IV, but T.P.R. for Paper I. All authors interpreted and validated the results (R.J.S. Papers II-IV). A.O.S. (Papers I-IV) and T.P.R. (Paper I) wrote the original draft

for the papers. A.O.S. made all visualisations. All authors reviewed and edited the paper (R.J.S. Papers II-IV). R.J.S. and T.P.R. did funding acquisition, supervision and project administration.

Conference presentations and posters are listed below:

### Conference Presentations

**I:** **A.O. Sigurpalsson**, T.P. Runarsson, R.J. Saemundsson, P.H. Moller, and V. Hallgrimsdottir, "Practical Surgery Scheduling Under Uncertainty" presented at ORAHS 2018 - Connected Care, Oslo, Norway, Jul. 2018.

**II:** **A.O. Sigurpalsson**, T.P. Runarsson and R.J. Saemundsson, "Elective surgery scheduling at the General Surgery speciality", presented at the Biomedical and Health Sciences Conference at University of Iceland, Reykjavik, Iceland, Jan. 2019.

**III:** T.P. Runarsson and **A.O. Sigurpalsson**, "Towards an evolutionary guided exact solution to elective surgery scheduling under uncertainty and ward restrictions" presented at 2019 IEEE congress on Evolutionary Computation (CEC), Wellington, New Zealand, Jun. 2019.

**IV:** **A.O. Sigurpalsson**, T.P. Runarsson and R.J. Saemundsson, "Stochastic Master Surgical Scheduling Under Ward Uncertainty", presented at HCSE 2019, Montreal, Canada, May 2019.

**V:** **A.O. Sigurpalsson**, T.P. Runarsson and R.J. Saemundsson, "Rescheduling of elective patients upon arrival of new priority patients", presented at the Biomedical and Health Sciences Conference at the University of Iceland, Reykjavik, Iceland, May 2021.

### Posters

**I:** **A.O. Sigurpalsson**, T.P. Runarsson, R.J. Saemundsson, PH Moller and V Hallgrimsdottir, "Better scheduling of surgeries". Researchers' Night, Reykjavik, Iceland, Oct. 2019.

# References


Adams, T., O'Sullivan, M., Walker, C., Wang, K., and Boyle, L. (2023). Application of a risk-averse objective function for scheduling surgeries. *Computers & Operations Research*, 151:106086.

Adams, T. E. (2019). *Data Informed Planning in Healthcare: Practical Models for Rostering and Scheduling under Uncertainty and Risk*. Phd thesis, University of Auckland.

Adan, I., Bekkers, J., Dellaert, N., Vissers, J., and Yu, X. (2009). Patient mix optimisation and stochastic resource requirements: a case study in cardiothoracic surgery planning. *Health Care Manag Sci*, 12(2):129–41.

Addis, B., Carello, G., Grosso, A., and Tànfani, E. (2016). Operating room scheduling and rescheduling: a rolling horizon approach. *Flexible Services and Manufacturing Journal*, 28(1):206–232.

Addis, B., Carello, G., and Tànfani, E. (2014). A robust optimization approach for the operating room planning problem with uncertain surgery duration. In Matta, A., Li, J., Sahin, E., Lanzarone, E., and Fowler, J., editors, *Proceedings of the International Conference on Health Care Systems Engineering*, pages 175–189, Cham. Springer International Publishing.

Aissaoui, N. O., Khlif, H. H., and Zeghal, F. M. (2020). Integrated proactive surgery scheduling in private healthcare facilities. *Computers & Industrial Engineering*, 148:106686.

Akbarzadeh, B. and Maenhout, B. (2024). A study on policy decisions to embed flexibility for reactive recovery in the planning and scheduling process in operating rooms. *Omega*, 126:103061.

Antognini, J. M. O., Antognini, J. F., and Khatri, V. (2015). How many operating rooms are needed to manage non-elective surgical cases? a monte carlo simulation study. *BMC Health Services Research*, 15(1).

Aringhieri, R., Tànfani, E., and Testi, A. (2013). Operations research for health care delivery. *Computers & Operations Research*, 40(9):2165–2166. Operations research for health care delivery.

Armoeyan, M., Aarabi, A., and Akbari, L. (2021). The effects of surgery cancellation on patients, families, and staff: A prospective cross-sectional study. *Journal of PeriAnesthesia Nursing*, 36(6):695–701.e2.

Augusto, V., Xie, X., and Perdomo, V. (2010). Operating theatre scheduling with patient recovery in both operating rooms and recovery beds. *Computers & Industrial Engineering*, 58(2):231–238.

Baker, D. R., Pronovost, P. J., Morlock, L. L., Geocadin, R. G., and Holzmueller, C. G. (2009). Patient flow variability and unplanned readmissions to an intensive care unit. *Crit Care Med*, 37(11):2882–7.

Banditori, C., Cappanera, P., and Visintin, F. (2013). A combined optimization–simulation approach to the master surgical scheduling problem. *IMA Journal of Management Mathematics*, 24(2):155–187.

Batun, S., Denton, B. T., Huschka, T. R., and Schaefer, A. J. (2011). Operating room pooling and parallel surgery processing under uncertainty. *INFORMS Journal on Computing*, 23(2):220–237.

Beliën, J. and Demeulemeester, E. (2007). Building cyclic master surgery schedules with leveled resulting bed occupancy. *European Journal of Operational Research*, 176(2):1185–1204.

Beliën, J., Demeulemeester, E., and Cardoen, B. (2008). A decision support system for cyclic master surgery scheduling with multiple objectives. *Journal of Scheduling*, 12(2):147.

Bellini, V., Russo, M., Domenichetti, T., Panizzi, M., Allai, S., and Bignami, E. G. (2024). Artificial intelligence in operating room management. *Journal of Medical Systems*, 48(1):19.

Bertsimas, D. and Sim, M. (2004). The price of robustness. *Operations Research*, 52(1):35–53.

Blake, J. T. and Carter, M. W. (2002). A goal programming approach to strategic resource allocation in acute care hospitals. *European Journal of Operational Research*, 140(3):541–561.

Blake, J. T. and Donald, J. (2002). Mount sinai hospital uses integer programming to allocate operating room time. *Interfaces*, 32(2):63–73.

Cappanera, P., Visintin, F., and Banditori, C. (2018). Addressing conflicting stakeholders' priorities in surgical scheduling by goal programming. *Flexible Services and Manufacturing Journal*, 30(1-2):252–271.

Cardoen, B., Demeulemeester, E., and Beliën, J. (2010). Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, 201(3):921–932.

Choi, S. and Wilhelm, W. E. (2014). On capacity allocation for operating rooms. *Computers & Operations Research*, 44:174–184.

Cohen, M. E., Bilimoria, K. Y., Ko, C. Y., Richards, K., and Hall, B. L. (2009). Variability in length of stay after colorectal surgery: assessment of 182 hospitals in the national surgical quality improvement program. *Ann Surg*, 250(6):901–7.

Dall'Ora, C., Saville, C., Rubbo, B., Turner, L., Jones, J., and Griffiths, P. (2022). Nurse staffing levels and patient outcomes: A systematic review of longitudinal studies. *International Journal of Nursing Studies*, 134:104311.

Davarian, F. and Behnamian, J. (2022). Robust finite-horizon scheduling/rescheduling of operating rooms with elective and emergency surgeries under resource constraints. *Journal of Scheduling*, 25(6):625–641.

Denton, B. and Gupta, D. (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, 35(11):1003–1016.

Denton, B., Viapiano, J., and Vogl, A. (2007). Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science*, 10(1):13–24.

Denton, B. T., Miller, A. J., Balasubramanian, H. J., and Huschka, T. R. (2010). Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations Research*, 58(4-part-1):802–816.

Dexter, F. (1996). Application of prediction levels to or scheduling. *Aorn j*, 63(3):607–15.

Dexter, F., Abouleish, A. E., Epstein, R. H., Whitten, C. W., and Lubarsky, D. A. (2003). Use of operating room information system data to predict the impact of reducing turnover times on staffing costs. *Anesth Analg*, 97(4):1119–1126.

Dexter, F., Blake, J. T., Penning, D. H., Sloan, B., Chung, P., and Lubarsky, D. A. (2002). Use of Linear Programming to Estimate Impact of Changes in a Hospital's Operating Room Time Allocation on Perioperative Variable Costs. *Anesthesiology*, 96(3):718–724. _eprint: https://pubs.asahq.org/anesthesiology/article-pdf/96/3/718/405415/0000542-200203000-00031.pdf.

Dexter, F., Dexter, E. U., Masursky, D., and Nussmeier, N. A. (2008). Systematic Review of General Thoracic Surgery Articles to Identify Predictors of Operating Room Case Durations. *Anesthesia & Analgesia*, 106(4).

Dexter, F. and Macario, A. (1996). Applications of information systems to operating room scheduling. *Anesthesiology*, 85(6):1232–1234.

Dexter, F., Macario, A., Qian, F., and Traub, R. D. (1999a). Forecasting surgical groups' total hours of elective cases for allocation of block time: application of time series analysis to operating room management. *Anesthesiology*, 91(5):1501–8.

Dexter, F., Macario, A., and Traub, R. (1999b). Which algorithm for scheduling add-on elective cases maximizes operating room utilization? : Use of bin packing algorithms and fuzzy constraints in operating room management. *Anesthesiology*, 91(5):1491–1491.

Dexter, F., Traub, R. D., and Qian, F. (1999c). Comparison of statistical methods to predict the time to complete a series of surgical cases. *Journal of Clinical Monitoring and Computing*, 15(1):45–51.

Embætti landlæknis (2021). Bið eftir völdum skurðaðgerðum. Report, Embætti landlæknis.

Epstein, R. H. and Dexter, F. (2013). Rescheduling of previously cancelled surgical cases does not increase variability in operating room workload when cases are scheduled based on maximizing efficiency of use of operating room time. *Anesthesia & Analgesia*, 117(4):995–1002.

Eshghali, M., Kannan, D., Salmanzadeh-Meydani, N., and Esmaieeli Sikaroudi, A. M. (2024). Machine learning based integrated scheduling and rescheduling for elective and emergency patients in the operating theatre. *Annals of Operations Research*, 332(1):989–1012.

Fei, H., Chu, C., and Meskens, N. (2008). Solving a tactical operating room planning problem by a column-generation-based heuristic procedure with four criteria. *Annals of Operations Research*, 166(1):91.

Fügener, A., Hans, E. W., Kolisch, R., Kortbeek, N., and Vanberkel, P. T. (2014). Master surgery scheduling with consideration of multiple downstream units. *European Journal of Operational Research*, 239(1):227–236.

Guerriero, F. and Guido, R. (2011). Operational research in the management of the operating theatre: a survey. *Health Care Manag Sci*, 14(1):89–114.

Gür, c., Alakaş, H. M., Pınarbaşı, M., and Eren, T. (2024). Stochastic operating room scheduling: a new model for solving problem and an approach for determining the factors that affect operation time variations. *Soft Computing*, 28(5):3987–4007.

Gür, c., Pınarbaşı, M., Alakaş, H. M., and Eren, T. (2023). Operating room scheduling with surgical team: a new approach with constraint programming and goal programming. *Central European Journal of Operations Research*, 31(4):1061–1085.

Gurobi Optimization, LLC (2024). Gurobi Optimizer Reference Manual.

Hans, E., Wullink, G., van Houdenhoven, M., and Kazemier, G. (2008). Robust surgery loading. *European Journal of Operational Research*, 185(3):1038–1050.

Hans, E. W., van Houdenhoven, M., and Hulshof, P. J. H. (2012). A framework for healthcare planning and control. In Hall, R., editor, *Handbook of Healthcare System Scheduling*, pages 303–320. Springer US, Boston, MA.

Harris, S. and Claudio, D. (2022). Current Trends in Operating Room Scheduling 2015 to 2020: a Literature Review. *Operations Research Forum*, 3(1):21.

Ivarsson, B., Larsson, S., and Sjöberg, T. (2004). Postponed or cancelled heart operations from the patient's perspective. *J Nurs Manag*, 12(1):28–36.

Izady, N. and Mohamed, I. (2021). A clustered overflow configuration of inpatient beds in hospitals. *Manufacturing & Service Operations Management*, 23(1):139–154.

Jacobs, R. F., Berry, W. L., Whybark, C., and Vollmann, T. E. (2018). *Manufacturing Planning and Control for Supply Chain Management: The CPIM Reference*. McGraw-Hill Education, New York, second edition. edition.

Jebali, A. and Diabat, A. (2015). A stochastic model for operating room planning under capacity constraints. *International Journal of Production Research*, 53(24):7252–7270.

Jebali, A. and Diabat, A. (2017). A chance-constrained operating room planning with elective and emergency cases under downstream capacity constraints. *Computers & Industrial Engineering*, 114:329–344.

Jebali, A., Hadj Alouane, A. B., and Ladet, P. (2006). Operating rooms scheduling. *International Journal of Production Economics*, 99(1):52–62. Control and Management of Productive Systems.

Jonnalagadda, R., Walrond, E., Hariharan, S., Walrond, M., and Prasad, C. (2005). Evaluation of the reasons for cancellations and delays of surgical procedures in a developing country. *International Journal of Clinical Practice*, 59(6):716–720.

Kamran, M. A., Karimi, B., and Dellaert, N. (2018). Uncertainty in advance scheduling problem in operating room planning. *Computers & Industrial Engineering*, 126:252–268.

Kamran, M. A., Karimi, B., Dellaert, N., and Demeulemeester, E. (2019). Adaptive operating rooms planning and scheduling: A rolling horizon approach. *Operations Research for Health Care*, 22:100200.

Kane, R. L., Shamliyan, T. A., Mueller, C., Duval, S., and Wilt, T. J. (2007). The association of registered nurse staffing levels and patient outcomes: systematic review and meta-analysis. *Medical care*, pages 1195–1204.

Kim, S.-C. and Horowitz, I. (2002). Scheduling hospital services: the efficacy of elective-surgery quotas. *Omega*, 30:335+.

Kroer, L. R., Foverskov, K., Vilhelmsen, C., Hansen, A. S., and Larsen, J. (2018). Planning and scheduling operating rooms for elective and emergency surgeries with uncertain duration. *Operations Research for Health Care*, 19:107–119.

Lamiri, M., Dreo, J., and Xie, X. (2007). Operating room planning with random surgery times. In *2007 IEEE International Conference on Automation Science and Engineering*, pages 521–526.

Lamiri, M., Grimaud, F., and Xie, X. (2009). Optimization methods for a stochastic surgery planning problem. *International Journal of Production Economics*, 120(2):400–410. Special Issue on Introduction to Design and Analysis of Production Systems.

Lamiri, M., Xie, X., Dolgui, A., and Grimaud, F. (2008a). A stochastic model for operating room planning with elective and emergency demand for surgery. *European Journal of Operational Research*, 185(3):1026–1037.

Lamiri, M., Xie, X., and Zhang, S. (2008b). Column generation approach to operating theater planning with elective and emergency patients. *IIE Transactions*, 40(9):838–852.

Landa, P., Aringhieri, R., Soriano, P., Tànfani, E., and Testi, A. (2016). A hybrid optimization algorithm for surgeries scheduling. *Operations Research for Health Care*, 8:103–114.

Lovejoy, W. S. and Li, Y. (2002). Hospital operating room capacity expansion. *Management Science*, 48(11):1369–1387.

Macario, A. (2009). Truth in scheduling: is it possible to accurately predict how long a surgical case will last? *Anesth Analg*, 108(3):681–5.

Makboul, S., Kharraja, S., Abbassi, A., and Alaoui, A. E. H. (2021). A two-stage robust optimization approach for the master surgical schedule problem under uncertainty considering downstream resources. *Health Care Management Science*.

Marques, I. and Captivo, M. E. (2017). Different stakeholders' perspectives for a surgical case assignment problem: Deterministic and robust approaches. *European Journal of Operational Research*, 261(1):260–278.

Marques, I., Captivo, M. E., and Barros, N. (2019). Optimizing the master surgery schedule in a private hospital. *Operations Research for Health Care*, 20:11–24.

Marques, I., Captivo, M. E., and Vaz Pato, M. (2012). An integer programming approach to elective surgery scheduling. *OR Spectrum*, 34(2):407–427.

May, J. H., Spangler, W. E., Strum, D. P., and Luis, V. G. (2011). The surgical scheduling problem: Current research and future opportunities. *Production and Operations Management*, 20(3):392–405.

May, J. H., Strum, D. P., and Vargas, L. G. (2000). Fitting the lognormal distribution to surgical procedure times*. *Decision Sciences*, 31(1):129–148.

McIntosh, C., Dexter, F., and Epstein, R. H. (2006). The impact of service-specific staffing, case scheduling, turnovers, and first-case starts on anesthesia group and operating room productivity: A tutorial using data from an australian hospital. *Anesthesia & Analgesia*, 103(6):1499–1516.

M'Hallah, R. and Visintin, F. (2019). A stochastic model for scheduling elective surgeries in a cyclic master surgical schedule. *Computers & Industrial Engineering*, 129:156–168.

Min, D. and Yih, Y. (2010). Scheduling elective surgery under uncertainty and downstream capacity constraints. *European Journal of Operational Research*, 206(3):642–652.

Ministry of Health (2021). The future development of Landspítali's services. Technical report, Ministry of Health.

Molina-Pariente, J. M., Hans, E. W., and Framinan, J. M. (2018). A stochastic approach for solving the operating room scheduling problem. *Flexible Services and Manufacturing Journal*, 30(1):224–251.

Najjarbashi, A. and Lim, G. J. (2019). A variability reduction method for the operating room scheduling problem under uncertainty using cvar. *Operations Research for Health Care*, 20:25–32.

Neyshabouri, S. and Berg, B. P. (2017). Two-stage robust optimization approach to elective surgery and downstream capacity planning. *European Journal of Operational Research*, 260(1):21–40.

Oliveira, M., Bélanger, V., Ruiz, A., and Santos, D. (2023). A systematic literature review on the utilization of extended operating room hours to reduce surgical backlogs. *Front Public Health*, 11:1118072.

Opit, L. J., Collins, R. E., and Campbell, G. (1991). Use of operating theatres: the effects of case-mix and training in general surgery. *Ann R Coll Surg Engl*, 73(6):389–92; discussion 392–3.

Otten, M., Braaksma, A., and Boucherie, R. J. (2019). Minimizing earliness/tardiness costs on multiple machines with an application to surgery scheduling. *Operations Research for Health Care*, 22:100194.

Pandit, J. J. (2018). *Practical Operating Theatre Management: Measuring and Improving Performance and Patient Experience*. Cambridge University Press, Cambridge.

Pandit, J. J. (2020). Rational planning of operating lists: a prospective comparison of 'booking to the mean' vs. 'probabilistic case scheduling' in urology. *Anaesthesia*, 75(5):642–647.

Pandit, J. J. and Dexter, F. (2009). Lack of sensitivity of staffing for 8-hour sessions to standard deviation in daily actual hours of operating room time used for surgeons with long queues. *Anesthesia & Analgesia*, 108(6):1910–1915.

Pandit, J. J. and Tavare, A. (2011). Using mean duration and variation of procedure times to plan a list of surgical operations to fit into the scheduled list time. *European Journal of Anaesthesiology | EJA*, 28(7):493–501.

Pandit, J. J., Westbury, S., and Pandit, M. (2007). The concept of surgical operating list 'efficiency': a formula to describe the term. *Anaesthesia*, 62(9):895–903.

Parmar, D., Woodman, M., and Pandit, J. J. (2022). A graphical assessment of emergency surgical list efficiency to determine operating theatre capacity needs. *British Journal of Anaesthesia*, 128(3):574–583.

Pattnaik, S., Dixit, S. K., and Bishnoi, V. (2022). The burden of surgical cancellations: A quality improvement study on the importance of preoperative assessment. *Cureus*, 14(1):e21731.

Proudlove, N., Hine, A., Tavare, A., and Pandit, J. J. (2013). Improvements and corrections to estimating probabilities in the formula for planning a list of operations to fit into a scheduled time. *European Journal of Anaesthesiology | EJA*, 30(10):633–635.

Riise, A., Mannino, C., and Burke, E. K. (2016). Modelling and solving generalised operational surgery scheduling problems. *Computers & Operations Research*, 66:1–11.

Runarsson, T. P. and Sigurpalsson, A. O. (2019a). Towards an evolutionary guided exact solution to elective surgery scheduling under uncertainty and ward restrictions. Paper presented at the IEEE Congress on Evolutionary Computation (CEC), Wellington, New Zealand.

Runarsson, T. P. and Sigurpalsson, A. O. (2019b). Towards an evolutionary guided exact solution to elective surgery scheduling under uncertainty and ward restrictions. In *2019 IEEE Congress on Evolutionary Computation (CEC)*, pages 419–425.

Samudra, M., Van Riet, C., Demeulemeester, E., Cardoen, B., Vansteenkiste, N., and Rademakers, F. E. (2016). Scheduling operating rooms: achievements, challenges and pitfalls. *Journal of Scheduling*, 19(5):493–525.

Santibáñez, P., Begen, M., and Atkins, D. (2007). Surgical block scheduling in a system of hospitals: an application to resource and wait list management in a British Columbia health authority. *Health Care Management Science*, 10(3):269–282.

Schneider, A. J. T., Theresia van Essen, J., Carlier, M., and Hans, E. W. (2020). Scheduling surgery groups considering multiple downstream resources. *European Journal of Operational Research*, 282(2):741–752.

Shehadeh, K. S. (2022). Data-driven distributionally robust surgery planning in flexible operating rooms over a wasserstein ambiguity. *Computers & Operations Research*, 146:105927.

Shehadeh, K. S. and Padman, R. (2021). A distributionally robust optimization approach for stochastic elective surgery scheduling with limited intensive care unit capacity. *European Journal of Operational Research*, 290(3):901–913.

Shylo, O. V., Prokopyev, O. A., and Schaefer, A. J. (2013). Stochastic operating room scheduling for high-volume specialties under block booking. *INFORMS Journal on Computing*, 25(4):682–692.

Sigurpalsson, A. O., Runarsson, T. P., and Saemundsson, R. J. (2019a). Elective surgery scheduling at the general surgery speciality. Paper presented at the Biomedical and Health Sciences Conference at the University of Iceland, Reykjavik, Iceland.

Sigurpalsson, A. O., Runarsson, T. P., and Saemundsson, R. J. (2019b). Elective surgery scheduling at the general surgery speciality. Paper presented at the Health Care System Engineering 2019 conference, Montreal, Canada.

Sigurpalsson, A. O., Runarsson, T. P., and Saemundsson, R. J. (2020). Stochastic master surgical scheduling under ward uncertainty. In Bélanger, V., Lahrichi, N., Lanzarone, E., and Yalçındağ, S., editors, *Health Care Systems Engineering*, pages 163–176, Cham. Springer International Publishing.

Sigurpalsson, A. O., Runarsson, T. P., and Saemundsson, R. J. (2021). Rescheduling of elective patients upon arrival of new priority patients. Paper presented at the Biomedical and Health Sciences Conference at the University of Iceland, Reykjavik, Iceland.

Sigurpalsson, A. O., Runarsson, T. P., and Saemundsson, R. J. (2022). Bounding the likelihood of exceeding ward capacity in stochastic surgery scheduling. *Applied Sciences*, 12(17).

Sigurpalsson, A. O., Runarsson, T. P., Saemundsson, R. J., Moller, P. H., and Hallgrimsdottir, V. (2018). Practical general surgery scheduling. Paper presented at the 44th International Conference of the EURO Working Group on Operational Research Applied to Health Services, Oslo, Norway.

Sigurpalsson, A. O., Runarsson, T. P., Saemundsson, R. J., Moller, P. H., and Hallgrimsdottir, V. (2019c). Better scheduling of surgeries. Poster presented at Researchers' Night, Reykjavik, Iceland.

Sigurpalsson, A. O., Saemundsson, R. J., and Runarsson, T. P. (2025). Mind the gap: Strategies for semi-acute patient scheduling in elective surgery. *Submitted*.

Soh, K. W., Walker, C., and O'Sullivan, M. (2017). A literature review on validated simulations of the surgical services. *Journal of Medical Systems*, 41(4):61.

Soh, K. W., Walker, C., O'Sullivan, M., and Wallace, J. (2024). Innovative operating room scheduling metric for creating surgical lists with desirable room utilisation rates. *Operations Management Research*, 17(2):544–567.

Spratt, B. and Kozan, E. (2021). An integrated rolling horizon approach to increase operating theatre efficiency. *Journal of Scheduling*, 24(1):3–25.

Stowell, A., Claret, P. G., Sebbane, M., Bobbia, X., Boyard, C., Genre Grandpierre, R., Moreau, A., and de La Coussaye, J. E. (2013). Hospital out-lying through lack of beds and its impact on care and patient outcome. *Scand J Trauma Resusc Emerg Med*, 21:17.

Strum, D. P., Sampson, A. R., May, J. H., and Vargas, L. G. (2000). Surgeon and type of anesthesia predict variability in surgical procedure times. *Anesthesiology*, 92(5):1454–66.

Stuart, K. and Kozan, E. (2012). Reactive scheduling model for the operating theatre. *Flexible Services and Manufacturing Journal*, 24(4):400–421.

Tamata, A. T. and Mohammadnezhad, M. (2023). A systematic review study on the factors affecting shortage of nursing workforce in the hospitals. *Nurs Open*, 10(3):1247–1257.

van den Broek d'Obrenan, A., Ridder, A., Roubos, D., and Stougie, L. (2020). Minimizing bed occupancy variance by scheduling patients under uncertainty. *European Journal of Operational Research*, 286(1):336–349.

van Essen, J. T., Bosch, J. M., Hans, E. W., van Houdenhoven, M., and Hurink, J. L. (2014). Reducing the number of required beds by rearranging the or-schedule. *OR Spectrum*, 36(3):585–605.

van Oostrum, J. M., Van Houdenhoven, M., Hurink, J. L., Hans, E. W., Wullink, G., and Kazemier, G. (2008). A master surgical scheduling approach for cyclic scheduling in operating room departments. *OR Spectrum*, 30(2):355–374.

Van Riet, C. and Demeulemeester, E. (2015). Trade-offs in operating room planning for electives and emergencies: A review. *Operations Research for Health Care*, 7:52–69.

Wachtel, R. E. and Dexter, F. (2008). Tactical increases in operating room block time for capacity planning should not be based on utilization. *Anesth Analg*, 106(1):215–26.

Wang, J.-J., Dai, Z., Chang, A.-C., and Shi, J. J. (2022). Surgical scheduling by fuzzy model considering inpatient beds shortage under uncertain surgery durations. *Annals of Operations Research*, 315(1):463–505.

Wang, L., Demeulemeester, E., Vansteenkiste, N., and Rademakers, F. E. (2021). Operating room planning and scheduling for outpatients and inpatients: A review and future research. *Operations Research for Health Care*, 31:100323.

Wang, Y., Tang, J., and Fung, R. Y. K. (2014). A column-generation-based heuristic algorithm for solving operating theater planning problem under stochastic demand and surgery cancellation risk. *International Journal of Production Economics*, 158:28–36.

Wang, Y., Zhang, Y., and Tang, J. (2019). A distributionally robust optimization approach for surgery block allocation. *European Journal of Operational Research*, 273(2):740–753.

Wang, Z., Glynn, P. W., and Ye, Y. (2016). Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2):241–261.

Westbury, S., Pandit, M., and Pandit, J. J. (2009). Matching surgical operating capacity to demand using estimates of operating times. *J Health Organ Manag*, 23(5):554–67.

Wullink, G., Van Houdenhoven, M., Hans, E. W., Van Oostrum, J. M., Van Der Lans, M., and Kazemier, G. (2007). Closing emergency operating rooms improves efficiency. *Journal of Medical Systems*, 31(6):543–546.

Yahia, Z., Eltawil, A. B., and Harraz, N. A. (2016). The operating room case-mix problem under uncertainty and nurses capacity constraints. *Health Care Management Science*, 19(4):383–394.

Zhang, J., Dridi, M., and El Moudni, A. (2019). A two-level optimization model for elective surgery scheduling with downstream capacity constraints. *European Journal of Operational Research*, 276(2):602–613.

Zhang, J., Dridi, M., and El Moudni, A. (2020). Column-generation-based heuristic approaches to stochastic surgery scheduling with downstream capacity constraints. *International Journal of Production Economics*, 229:107764.

Zhou, J. and Dexter, F. (1998). Method to assist in the scheduling of add-on surgical cases–upper prediction bounds for surgical case durations based on the log-normal distribution. *Anesthesiology*, 89(5):1228–32.

Zhou, J., Dexter, F., Macario, A., and Lubarsky, D. A. (1999). Relying solely on historical surgical times to estimate accurately future surgical times is unlikely to reduce the average length of time cases finish late. *Journal of Clinical Anesthesia*, 11(7):601–605.

Zhu, S., Fan, W., Yang, S., Pei, J., and Pardalos, P. M. (2019). Operating room planning and surgical case scheduling: a review of literature. *Journal of Combinatorial Optimization*, 37(3):757–805.

Zonderland, M. E., Boucherie, R. J., Litvak, N., and Vleggeert-Lankamp, C. L. A. M. (2010). Planning and scheduling of semi-urgent surgeries. *Health Care Management Science*, 13(3):256–267.

# Appendices

# A Supplementary Data: Statistical Analysis of Surgery Times

*Table A.1. A pairwise comparison between the median surgery time of five frequent types of operations using Wilcoxon's rank sum test.*

| | I | II | III | IV |
|---|---|---|---|---|
| II | $0.000^{***}$ | - | - | - |
| III | $0.000^{***}$ | $0.000^{***}$ | - | - |
| IV | 1.000 | $0.000^{***}$ | $0.000^{***}$ | - |
| V | $0.000^{***}$ | $0.000^{***}$ | $0.000^{***}$ | $0.000^{***}$ |

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

*Table A.2. A pairwise comparison between the median surgery time of the operators for the same type of operation using Wilcoxon's rank sum test.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 2 | 1.000 | - | - | - | - | - | - | - |
| 3 | $0.000^{***}$ | $0.000^{***}$ | - | - | - | - | - | - |
| 4 | $0.000^{***}$ | $0.000^{***}$ | $0.000^{***}$ | - | - | - | - | - |
| 5 | 0.917 | 1.000 | $0.000^{***}$ | 0.020 | - | - | - | - |
| 6 | $0.000^{***}$ | $0.000^{***}$ | 1.000 | $0.001^{**}$ | $0.000^{***}$ | - | - | - |
| 7 | $0.000^{***}$ | $0.000^{***}$ | $0.000^{***}$ | $0.000^{***}$ | $0.000^{***}$ | $0.000^{***}$ | - | - |
| 8 | $0.000^{***}$ | $0.000^{***}$ | 1.000 | $0.000^{***}$ | $0.000^{***}$ | 1.000 | $0.000^{***}$ | - |
| 9 | 1.000 | 1.000 | $0.000^{***}$ | $0.000^{***}$ | 1.000 | $0.000^{***}$ | $0.000^{***}$ | $0.000^{***}$ |

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

# B Practical Heuristics

## B.1 Pattern Scheduling

The Pattern Scheduling consists of two steps. In the first step, all feasible patterns of patients are generated for each surgeon. The feasibility is determined on hand by practical rules, e.g., imposing bounds the number of patients admitted to ICU in a pattern and probabilistic restrictions on overtime, e.g., no pattern generated has more than e.g. 30% chance of exceeding its capacity. In the second step, the feasible patterns are assigned to corresponding blocks by a scheduling model to minimise the number of blocks with regular or extended overtime. Heuristics for practitioners interested in implementing this approach are outlined in the next to sections.

### B.1.1 Pattern Generation

1. **Patients**: Select elective patients from the waiting list for each surgeon that can or should be scheduled in the planning horizon.
2. **Practical Rules:** Define the practical rules that the hospital or the surgical speciality uses when constructing patterns of elective patients. A Pattern is a combination of one or more patients that can be operated in a single block but in no particular order. These practical rules can be, for example, imposing bounds on the number of:
   - Patients per patterns
   - Ward admissions per pattern
   - ICU admissions per pattern

   Additionally, data on possible block capacities should be collected.
3. **Pattern Generation:** Generate all possible patterns for each surgeon (or group of surgeons) using a computer program so that the generated patterns are according to the practical rules defined in Step 2.
4. **Collect historical data:** Gather historical data on surgery times from the hospital's database for all patients selected in Step 1 based on their type of operation

and surgeon. Use the last $n$ data points. If fewer than $n$ data points exist for the combination, combine data from other surgeons. If no data exists, estimate the surgery time. At the same time collect historical data on the time between two surgeries.

5. **Overtime verification:** Simulate each pattern generated in Step 2 with the historical data collected in Step 4 with Monte-Carlo sampling to estimate the probability of exceeding the regular block capacity. If the simulated probability exceeds a specified threshold, eliminate the pattern, as it may cause last-minute cancellations on the day of surgery due to overtime restrictions. An exception must be given to patterns of single surgeries. During the verification collect statistics on the simulated probability of exceeding the block capacity (regular overtime). In addition, simulate each pattern towards the risk of exceeding extended block capacity (extended overtime) and collect those statistics. Extended block capacity refers to the time added to the regular block capacity, e.g. 30 minutes, and carries a high risk of last-minute cancellations.

**Output:** The output of this process is a set of feasible patterns consisting of one or more patients for each surgeon based on the surgeon's list of patients. In addition, statistics for regular and extended overtime are provided for each pattern.

### B.1.2 Scheduling

Assign the patterns from Step 5 to corresponding blocks so that the number of patterns selected with regular or extended overtime is minimised using a mixed integer programming model and solve it with an off-the-shelf solver. Add a higher penalty cost to extended overtime as it carries a higher risk of last-minute cancellations. The model is presented in Section 6.2.1 with Objective (44) and Constraints (45)-(49).

## B.2 Ward Combinations

The ward combinations (WC) are designed to minimise the risk of last-minute cancellations due to uncertainty in the length of stay in the ward following surgery by bounding the risk of exceeding a given capacity to a specified threshold each day in a scheduling model. Several steps must be taken to utilise the WC. First, all possible WCs must be generated and validated against the risk of exceeding all bed availabilities. Next, empirical distributions of each type of operation must be discretised. Finally, the actual WC each day must be linked to the pre-generated WCs through indexing and bed availability in a scheduling model to determine the validity of the combination. Heuristics for practitioners interested in implementing this approach are outlined in the next to sections.

### B.2.1 Generation of Ward Combinations

1. **Ward Scenarios:** Specify the number of possible ward scenarios ($K$). Each ward patient follows a single scenario per day, determining the probability that the patient stays that day following surgery. The probabilities ($\rho_k$) for each scenario can be calculated using the following formula:

$$\rho_k = \frac{k-1}{|K|-1} \quad \forall k \in K$$

2. **Number of Staffed Beds:** Specify the upper limit on the number of staffed ward beds ($M^A$)

3. **Confidence Level:** Specify the threshold $\Omega$. The threshold determines the maximum allowed probability of exceeding a given ward bed capacity. For example, if $\Omega = 0.10$, the risk of exceeding is 10% at most.

4. **Ward Combinations:** With the help of a computer program, pre-generate all possible ward combinations (WC) of patients that can occupy the ward beds in a day with probability $\rho_k$ as defined Step 1 but exclude the first and the last scenario ($K'$) as patients in the first scenario have a 0% chance of occupying a bed. The patients in the last scenario have a 100% chance of occupying a bed. They are later used in the scheduling model, to be presented in Appendix B.2.3, to determine the available bed capacity for patients in other scenarios.

5. **Indexing:** Give each WC a unique index using the following formula:

$$l = \sum_{k \in K'} n_k |A|^{|K'|-k+1}$$

   where $|A|$ is the total number of available beds. That is, if the number of beds is six ($M^A$), then the total number of possible bed availabilities is seven (the size of

the set $\{0,1,\ldots,M^A = 6\}$). The indexing will later be used as a look-up in the scheduling model to be presented in Appendix B.2.3.

6. **Validate:** Verify each WC, indexed with $l$ and generated in Step 4, with Monte-Carlo sampling towards the risk of exceeding a given ward capacity $a \in A$ where $A = \{0,\ldots,M^A\}$. Use the assumption that the number of patients in the ward for each WC is defined as the sum of independent Binomial distributions. Collect statistics for the results of the simulation for each WC. Utilise the results in the binary parameter $F_{a,l}$ so that it is set to 1 if the simulated probability for a given WC is below the threshold $\Omega$, specified in Step 3, otherwise set it to 0.

In Figure B.1, an example of the process described above is given. In the example, it is assumed that the number of scenarios is set to five ($|K| = 5$), that the number of staffed ward beds is six ($M^A$), and that the maximum probability of exceeding a ward bed capacity is 10% ($\Omega = 0.90$). As the first and the last scenarios are excluded, all possible combinations of patients having a 25%, 50% and 75% chance of staying in a ward on a given day are generated and with each WC given a unique index which is used as a lookup in the scheduling model. Each WC is validated towards the risk of exceeding all possible bed availabilities, as shown in the right-hand side of the figure for a single WC indexed with 206. Thus, for this particular WC, the risk of exceeding the six beds is higher than 10% and, thus, carries a high risk of last-minute cancellations.

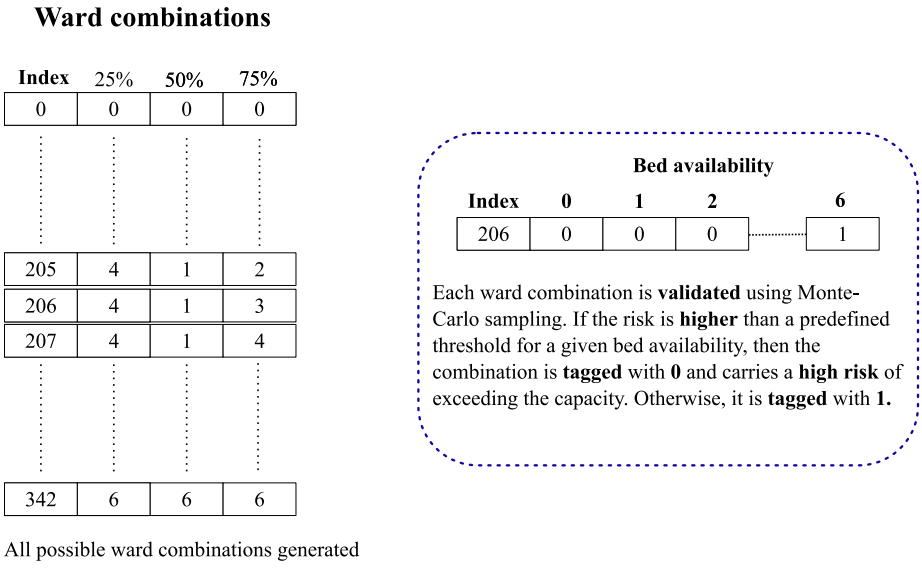



*Figure B.1. Example of how ward combinations are generated and validated towards the risk of exceeding a given bed capacity.*

**Output:** The output of this step is a matrix with the indexes of the WC and if WC is valid (1) or not (0) towards the risk of exceeding a given availability of ward beds as can be seen for a single WC, indexed with 206, in Figure B.1.

### B.2.2 Discretisation

To make it possible to use the WCs bounding the risk of exceeding a given bed capacity in a scheduling model, each ward patient's empirical daily ward probabilities must be discretised to the probabilities associated with each of the scenarios selected in Step 1 in Appendix B.2.1 using the following steps:

1. **Historical Data:** Extract historical data from the hospital's database that show how long each in-patient stayed following surgery. For each patient, collect the length of stay and the type of operation performed.
2. **Ward Distributions:** For each type of operation, generate a probability distribution from the historical data collected in Step 1, with the likelihood of staying in the ward on any given day following surgery. This can be achieved by calculating the ratio of patients staying at least one day, two days, up to $M^w$ days from the total number of patients who underwent this surgery and were admitted to the ward following surgery. Note that a patient who stays three days also stays the days before.
3. **Discretisation:** For each of the ward distribution generated in Step 2, discretise the daily ward probabilities based on the number of scenarios ($|K|$) selected in Step 1 in Appendix B.2.1 using the following procedure:
   - Multiply the ward probability of each day with $|K|-1$
   - Round the outcome to the nearest integer
   - Divide by $|K|-1$

   For example, if the number of scenarios is set to five ($|K|=5$) and the empirical ward probability is 0.65 on a given day, the value is discretised to $\rho_4 = 0.75$.

**Output:** The output of this process is a set of discretised ward distributions for each type of operation made from historical data. An example of an output of this step is given in table B.1 for two hypothetical types of operations that have been discretised using the steps mentioned above. In this example, it is assumed that the number of scenarios is five ($|K|=5$). As a result, the possible values of $\rho_k$ are 0.00, 0.25, 0.50, 0.75 and 1.00 as one can seen in the example.

### B.2.3 Ward Combination Optimisation

This step aims to assign the patterns to corresponding blocks using a scheduling model to minimise the risk of last-minute cancellations due to uncertainty in surgery time and length of stay in the ward. This can be achieved using the following steps:

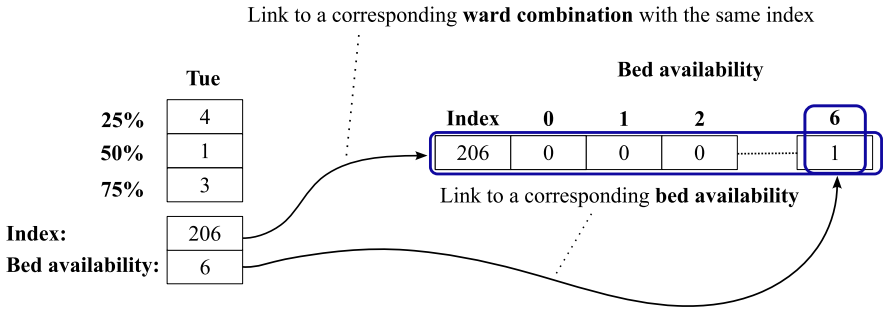
*Table B.1. Example of two specific types of operation where their empirical daily empirical values have been discretised into five possible scenarios.*

| | Days after surgery | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Operation** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | $\cdots$ | $M^W$ |
| Type I | 1.00 | 1.00 | 0.75 | 0.50 | 0.25 | 0.25 | 0.25 | 0.00 | $\cdots$ | 0.00 |
| Type II | 1.00 | 0.75 | 0.25 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | $\cdots$ | 0.00 |

$M^W$ is the maximum number of days that the distribution should be calculated for.

1. **Pattern Generation:** Use the Pattern Scheduling approach described in Appendix B.1 to generate feasible patterns of patients for each surgeon.
2. **Discretised Ward Distributions:** Assign the corresponding discretised ward distribution, generated for each type of operation from historical data as presented in Appendix B.2.2, to each ward patient in each pattern based on their type of operation.
3. **Ward Occupancy:** Generate separate distributions for each ward scenario, selected in Step 1 in Appendix B.2.1, for each pattern but exclude the first as these patients have a 0% chance of being in the ward on a given day. This is achieved by counting the number of ward patients with probability $\rho_k$ of staying in the ward each day following surgery within each pattern. Assign that number to the corresponding ward occupancy distribution for each scenario.
4. **Scheduling:** Use the scheduling model presented in Appendix B.1.2 but add Constraints (30)–(37), as described in section 5.1.2, to the model. These constraints match the pre-generated ward combinations, as presented in Appendix B.2.1, to actual ward combinations of patients each day for a given bed availability and index. The availability each day is determined by subtracting the number of patients with a 100% chance of staying that day from the total number of staffed ward beds. Thus, the task of the pre-generated WC is to inform the scheduling model of whether an actual WC of patients on any given day is valid, as each WC is validated against the risk of exceeding a given capacity. Consequently, the optimisation only provides valid solutions as this is implemented as a hard constraint in the model.

In Figure B.2, an example is given for a single day of how an actual WC consisting of 4 patients having 25% of staying that day, 1 having 50% and 3 having 75% chance is linked to a pre-generated WC. The linking between the actual WC and WC is made possible through indexing, calculated using the equation given in Step 5 in Appendix B.2.1 by the scheduling model and the bed availability. The bed availability specifies if the actual WC is valid. In this example, the WC is valid for the six available beds. Note that the actual WC consists of all ward patients on a given day, not just those admitted that day.




**Index** generated based on the actual combination of patients on a given day by the scheduling model
**Bed availability** is calculated by subtracting the number of patients with a 100% chance of staying from the total number of staffed beds.

*Figure B.2. Linking an actual ward combination of patients to a pre-generated ward combination using indexing and bed availability to determine if the actual combination is valid.*

# C Scheduling and Rescheduling Models

The following appendices provide complete formulations for each of the scheduling and rescheduling models provided in Chapter 6.

## C.1 Front Load

$$\begin{aligned}
\min \quad & \lambda^{ov}\Big( \sum_{d\in D, r\in R} \big(u_{d,r} + \lambda^{ov} v_{d,r}\big) \\
& + \sum_{(d,p,r)\in DPR:} \big(\big[\delta_p\big]^2 + \lambda^{ov}\big[\delta_p^{\Delta}\big]^2\big) x_{d,p,r}\Big) && (44)\\
& + \lambda^{w} w^{\max} && (52)\\
& + \lambda^{RS} \sum_{i\in I} \Delta_i && (65)\\
& + \lambda^{RS} \sum_{\substack{i\in I:\\ i\in S^p}} \Delta_i && (66)\\
& + \lambda^{FL} \sum_{(d,p,r)\in DPR: P_s\subseteq P} dx_{d,p,r} && (53)
\end{aligned}$$

**Subject to:**

$$\sum_{(d,p,r)\in DPR: i\in I_p} x_{d,p,r} = 1, \quad \forall i \in I \tag{45}$$

$$\sum_{r\in R,\ p\in P_o:(d,p,r)\in DPR} x_{d,p,r}\Upsilon_{o,p} \le 1, \quad \forall d \in D, \quad \forall o \in O \tag{60}$$

$$\sum_{p\in P:(d,p,r)\in DPR} x_{d,p,r} \le 1, \quad \forall d \in D, \quad r \in R \tag{46}$$

$$\sum_{(d,p,r)\in DPR:\delta_p\ge\delta'} x_{d,p,r} \le u_{d,r} \tag{48}$$

$$\sum_{(d,p,r)\in DPR:\delta_p^{\Delta}\geq\delta^{\Delta'}} x_{d,p,r} \leq v_{d,r} \tag{49}$$

$$\bar{n}_d^{ICU} + \sum_{\substack{r\in R,p\in P,j\in\{0,1\ldots,M^{ICU}-1\}:\\(d-j,p,r)\in DPR}} Q_{j,p}^{ICU} x_{d-j,p,r} \leq M_{ICU}^{A},\ \forall\ d\in D \tag{50}$$

$$\bar{n}_d^{w} + \sum_{\substack{r\in R,p\in P,j\in\{0,1,\ldots,M^{W}-1\}:\\(d-j,p,r)\in DPR}} Q_{j,p}^{w} x_{d-j,p,r} \leq w^{\max} \qquad \forall d\in D \tag{51}$$

$$\sum_{(d,p,r)\in DPR:i\in I_p} dx_{d,p,r} = \Psi_i, \quad \forall i\in I \tag{61}$$

$$\gamma_i - \Psi_i \leq 0, \quad \forall i\in I \tag{62}$$

$$-\gamma_i + \Psi_i \leq \Delta_i^{+}, \quad \forall i\in I \tag{63}$$

$$M\Delta_i \geq \Delta_i^{+}, \quad \forall i\in I \tag{64}$$

## C.2 Daily Limit

$$\begin{aligned}
\min \quad & \lambda^{ov}\Big( \sum_{d\in D,r\in R} \big(u_{d,r} + \lambda^{ov} v_{d,r}\big) \\
& + \sum_{(d,p,r)\in DPR:} \big(\big[\delta_p\big]^2 + \lambda^{ov}\big[\delta_p^{\Delta}\big]^2\big) x_{d,p,r}\Big) && (44)\\
& + \lambda^{w} w^{\max} && (52)\\
& + \lambda^{RS} \sum_{i\in I} \Delta_i && (65)\\
& + \lambda^{RS} \sum_{\substack{i\in I:\\ i\in S^p}} \Delta_i && (66)\\
& + \lambda^{DL} \sum_{d\in D,r\in R} \bar{R}_{d,r} && (55)
\end{aligned}$$

**Subject to:**

$$\sum_{(d,p,r)\in DPR:i\in I_p} x_{d,p,r} = 1, \quad \forall i\in I \tag{45}$$

$$\sum_{r\in R,\ p\in P_o:(d,p,r)\in DPR} x_{d,p,r}\Upsilon_{o,p} \leq 1, \quad \forall d\in D, \quad \forall o\in O \tag{60}$$

$$\sum_{p\in P:(d,p,r)\in DPR} x_{d,p,r} \leq 1, \quad \forall d\in D, \quad r\in R \tag{46}$$

$$\sum_{(d,p,r)\in DPR:\delta_p\geq\delta'} x_{d,p,r} \leq u_{d,r} \tag{48}$$

$$\sum_{(d,p,r)\in DPR:\delta_p^{\Delta}\geq\delta^{\Delta'}} x_{d,p,r} \leq v_{d,r} \tag{49}$$

$$\bar{n}_d^{ICU} + \sum_{\substack{r\in R,p\in P,j\in\{0,1\ldots,M^{ICU}-1\}:\\(d-j,p,r)\in DPR}} Q_{j,p}^{ICU} x_{d-j,p,r} \leq M_{ICU}^{A},\ \forall\, d\in D \tag{50}$$

$$\bar{n}_d^{w} + \sum_{\substack{r\in R,p\in P,j\in\{0,1,\ldots,M^{W}-1\}:\\(d-j,p,r)\in DPR}} Q_{j,p}^{w} x_{d-j,p,r} \leq w^{\max} \qquad \forall d\in D \tag{51}$$

$$\sum_{(d,p,r)\in DPR:i\in I_p} dx_{d,p,r} = \Psi_i, \quad \forall i\in I \tag{61}$$

$$\gamma_i - \Psi_i \leq 0, \quad \forall i\in I \tag{62}$$

$$-\gamma_i + \Psi_i \leq \Delta_i^{+}, \quad \forall i\in I \tag{63}$$

$$M\Delta_i \geq \Delta_i^{+}, \quad \forall i\in I \tag{64}$$

$$\sum_{p\in P:\ (d,p,r)\in DPR,\ C_p>N^s} x_{d,p,r} \leq \bar{R}_{d,r}, \quad d\in D, \quad r\in R \tag{54}$$

## C.3 Balance ORs

$$\begin{aligned}
\min \quad & \lambda^{ov}\Big( \sum_{d\in D,r\in R} \big(u_{d,r} + \lambda^{ov} v_{d,r}\big) \\
& + \sum_{(d,p,r)\in DPR:} \big(\big[\delta_p\big]^2 + \lambda^{ov}\big[\delta_p^{\Delta}\big]^2\big) x_{d,p,r}\Big) && (44)\\
& + \lambda^{w} w^{\max} && (52)\\
& + \lambda^{RS} \sum_{i\in I} \Delta_i && (65)\\
& + \lambda^{RS} \sum_{\substack{i\in I:\\ i\in S^p}} \Delta_i && (66)\\
& + \lambda^{BO} V^{B} && ((56))
\end{aligned}$$

**Subject to:**

$$\sum_{(d,p,r)\in DPR:i\in I_p} x_{d,p,r} = 1, \quad \forall i\in I \tag{45}$$

$$\sum_{r\in R,\ p\in P_o:(d,p,r)\in DPR} x_{d,p,r}\Upsilon_{o,p} \leq 1, \quad \forall d \in D, \quad \forall o \in O \tag{60}$$

$$\sum_{p\in P:(d,p,r)\in DPR} x_{d,p,r} \leq 1, \quad \forall d \in D, \quad r \in R \tag{46}$$

$$\sum_{(d,p,r)\in DPR:\delta_p\geq\delta'} x_{d,p,r} \leq u_{d,r} \tag{48}$$

$$\sum_{(d,p,r)\in DPR:\delta_p^{\Delta}\geq\delta^{\Delta'}} x_{d,p,r} \leq v_{d,r} \tag{49}$$

$$\bar{n}_d^{ICU} + \sum_{\substack{r\in R,p\in P,j\in\{0,1\ldots,M^{ICU}-1\}:\\(d-j,p,r)\in DPR}} Q_{j,p}^{ICU} x_{d-j,p,r} \leq M_{ICU}^{A},\ \forall\ d \in D \tag{50}$$

$$\bar{n}_d^{w} + \sum_{\substack{r\in R,p\in P,j\in\{0,1,\ldots,M^{W}-1\}:\\(d-j,p,r)\in DPR}} Q_{j,p}^{w} x_{d-j,p,r} \leq w^{\max} \qquad \forall d \in D \tag{51}$$

$$\sum_{(d,p,r)\in DPR:i\in I_p} dx_{d,p,r} = \Psi_i, \quad \forall i \in I \tag{61}$$

$$\gamma_i - \Psi_i \leq 0, \quad \forall i \in I \tag{62}$$

$$-\gamma_i + \Psi_i \leq \Delta_i^{+}, \quad \forall i \in I \tag{63}$$

$$M\Delta_i \geq \Delta_i^{+}, \quad \forall i \in I \tag{64}$$

$$\sum_{(d,p,r)\in DPR:\ D^v=v} x_{d,p,r} \leq V^B, \quad \forall v \in V \tag{56}$$

# D Supplementary Results: Uncertainty in Semi-Acute Elective Arrivals

*Table D.1. Summary statistics for the distributions of surgery times and length of stay (LOS) in the ward for each at General Surgery speciality. Throughput is shown for the year 2019.*

| | All patients | | | | Semi-Acute Arrivals | | |
|---|---|---|---|---|---|---|---|
| | | | Mean | | | Mean | |
| **Month** | **#Pat.** | **#In-Pat** | **Surgery time**$^{\dagger}$ | **LOS**$^{\star}$ | **#Pat.** | **Surgery time**$^{\dagger}$ | **LOS**$^{\star}$ |
| 1 | 116 | 36 | 130.9 | 4.6 | 33 | 112.4 | 5.4 |
| 2 | 135 | 41 | 125.5 | 4.0 | 39 | 108.9 | 5.5 |
| 3 | 112 | 34 | 127.3 | 4.4 | 21 | 111.3 | 6.3 |
| 4 | 99 | 30 | 120.2 | 3.7 | 33 | 107.5 | 4.8 |
| 5 | 114 | 30 | 130.2 | 3.8 | 25 | 116.4 | 4.7 |
| 6 | 75 | 27 | 124.1 | 3.8 | 21 | 122.3 | 4.9 |
| 7 | 51 | 23 | 140.0 | 4.6 | 24 | 144.0 | 5.7 |
| 8 | 39 | 12 | 130.1 | 4.6 | 8 | 111.9 | 8.6 |
| 9 | 100 | 28 | 143.3 | 5.0 | 31 | 121.5 | 6.1 |
| 10 | 91 | 29 | 151.1 | 3.7 | 16 | 132.6 | 4.2 |
| 11 | 104 | 38 | 133.9 | 4.2 | 27 | 108.0 | 6.0 |
| 12 | 105 | 34 | 135.3 | 4.5 | 34 | 137.3 | 6.0 |

Surgery times are shown in minutes. LOS is shown in days. #Pat. is the total number of patients for each category, and #In-Pat. the number of patients admitted to the ward following surgery.

*Table D.2. Results for the actual scheduling data for each month. Results are shown for the base model (Base Schedule) and after rescheduling (Final Schedule).*

| | Base Schedule | | | | | | | | Final Schedule | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Blocks | | | | Overtime | | Ward | | Blocks | | | | Overtime | | Ward | | | SR time | |
| Month | #Open | Utilisation | Reserved Cap. | Empty | #Regular | #Extended | Mean | [Min, Max] | #Open | Utilisation | Reserved Cap. | Empty | #Regular | #Extended | Mean | [min, max] | #Resch. | Median | Mean |
| 1 | 40 | 0.73 | 0.39 | 0.17 | 13 | 6 | 4.1 | [1.6, 7.8] | 46 | 0.85 | 0.81 | 0.04 | 23 | 10 | 5.1 | [2.0, 8.7] | - | - | - |
| 2 | 43 | 0.78 | 0.30 | 0.10 | 12 | 8 | 4.2 | [1.8, 7.0] | 48 | 0.94 | 0.94 | 0.00 | 30 | 13 | 5.5 | [2.1, 9.3] | - | - | - |
| 3 | 41 | 0.78 | 0.33 | 0.15 | 11 | 7 | 4.2 | [1.0, 6.8] | 44 | 0.88 | 0.81 | 0.08 | 17 | 12 | 5.1 | [2.1, 7.5] | - | - | - |
| 4 | 33 | 0.67 | 0.54 | 0.31 | 6 | 4 | 2.6 | [0.5, 6.2] | 38 | 0.84 | 0.67 | 0.21 | 19 | 8 | 3.8 | [0.9, 6.6] | - | - | - |
| 5 | 38 | 0.83 | 0.34 | 0.21 | 18 | 7 | 3.6 | [1.3, 6.8] | 44 | 0.89 | 0.82 | 0.08 | 26 | 12 | 4.4 | [1.4, 8.0] | - | - | - |
| 6 | 27 | 0.62 | 0.65 | 0.44 | 4 | 0 | 2.8 | [0.6, 6.7] | 32 | 0.74 | 0.49 | 0.33 | 11 | 4 | 4.2 | [1.2, 7.6] | - | - | - |
| 7 | 17 | 0.52 | 0.82 | 0.65 | 3 | 1 | 1.5 | [0.2, 4.2] | 26 | 0.68 | 0.37 | 0.46 | 7 | 4 | 3.7 | [1.6, 6.5] | - | - | - |
| 8 | 15 | 0.71 | 0.78 | 0.69 | 5 | 3 | 1.3 | [0.2, 3.2] | 19 | 0.67 | 0.27 | 0.60 | 6 | 3 | 1.8 | [0.5, 3.7] | - | - | - |
| 9 | 38 | 0.69 | 0.45 | 0.21 | 15 | 9 | 3.7 | [1.8, 5.9] | 43 | 0.86 | 0.77 | 0.10 | 25 | 15 | 4.9 | [2.0, 9.6] | - | - | - |
| 10 | 40 | 0.72 | 0.40 | 0.17 | 13 | 8 | 3.7 | [1.6, 6.6] | 42 | 0.82 | 0.72 | 0.13 | 17 | 11 | 4.5 | [2.7, 7.7] | - | - | - |
| 11 | 37 | 0.78 | 0.40 | 0.23 | 11 | 4 | 4.4 | [1.7, 7.9] | 41 | 0.90 | 0.77 | 0.15 | 21 | 7 | 5.6 | [2.5, 9.1] | - | - | - |
| 12 | 37 | 0.70 | 0.46 | 0.23 | 10 | 8 | 3.4 | [1.5, 5.6] | 43 | 0.89 | 0.80 | 0.10 | 21 | 13 | 5.6 | [3.2, 8.6] | - | - | - |

Shown are the number of blocks in use (#Open), the average block utilisation, the reserved capacity from the allocated capacity (Reserved Cap.), and the percentage of empty blocks from the total number of blocks available. Additionally, the number of times surgeons require regular (#Regular) or extended (#Extended) overtime, the average ward bed occupancy (Mean) along with its minimum and maximum [min, max], the number of rescheduling events (#Resch), and the time between scheduled and rescheduled appointments (SR time) are shown.

*Table D.3. Computational results for each month for the Front load strategy.*

| | Base Schedule | | | | | | | | Final Schedule | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Blocks | | | | Overtime | | Ward | | Blocks | | | | Overtime | | Ward | | | SR time | |
| Month | #Open | Utilisation | Reserved Cap. | Empty | #Regular | #Extended | Mean | [Min, Max] | #Open | Utilisation | Reserved Cap. | Empty | #Regular | #Extended | Mean | [min, max] | #Resch. | Median | Mean |
| 1 | 37 | 0.79 | 0.39 | 0.23 | 4 | 2 | 4.0 | [1.0, 5.9] | 41 | 0.96 | 0.82 | 0.15 | 22 | 18 | 5.0 | [1.0, 8.5] | 6 | 6.5 | 5.2 |
| 2 | 39 | 0.88 | 0.29 | 0.19 | 18 | 6 | 4.1 | [1.2, 6.6] | 43 | 1.07 | 0.96 | 0.10 | 34 | 23 | 5.6 | [1.8, 11.0] | 17 | 2.0 | 2.6 |
| 3 | 36 | 0.88 | 0.34 | 0.25 | 17 | 6 | 4.2 | [2.4, 6.4] | 37 | 1.03 | 0.79 | 0.23 | 26 | 19 | 5.1 | [2.3, 7.4] | 7 | 4.0 | 5.5 |
| 4 | 29 | 0.76 | 0.54 | 0.40 | 6 | 3 | 2.9 | [0.7, 4.6] | 34 | 0.93 | 0.66 | 0.29 | 22 | 18 | 4.0 | [0.8, 7.7] | 16 | 2.0 | 2.8 |
| 5 | 36 | 0.87 | 0.35 | 0.25 | 7 | 5 | 3.5 | [1.6, 5.0] | 38 | 1.03 | 0.82 | 0.21 | 25 | 17 | 4.5 | [1.6, 7.3] | 6 | 1.5 | 2.0 |
| 6 | 27 | 0.63 | 0.65 | 0.44 | 0 | 0 | 2.7 | [1.1, 4.5] | 32 | 0.75 | 0.50 | 0.33 | 13 | 7 | 4.3 | [1.2, 7.3] | 3 | 2.0 | 3.7 |
| 7 | 17 | 0.50 | 0.82 | 0.65 | 0 | 0 | 1.6 | [0.3, 2.8] | 26 | 0.67 | 0.36 | 0.46 | 9 | 6 | 3.8 | [0.9, 7.0] | 3 | 1.0 | 1.0 |
| 8 | 17 | 0.60 | 0.79 | 0.65 | 0 | 0 | 1.4 | [0.0, 3.0] | 19 | 0.65 | 0.26 | 0.60 | 2 | 0 | 1.9 | [0.6, 4.4] | 2 | 2.0 | 2.0 |
| 9 | 32 | 0.82 | 0.45 | 0.33 | 7 | 5 | 3.3 | [1.0, 4.9] | 40 | 0.92 | 0.77 | 0.17 | 22 | 19 | 4.6 | [1.6, 7.2] | 9 | 1.0 | 3.6 |
| 10 | 37 | 0.79 | 0.39 | 0.23 | 7 | 3 | 3.7 | [1.3, 6.0] | 41 | 0.84 | 0.72 | 0.15 | 16 | 9 | 4.8 | [2.3, 8.5] | 1 | 4.0 | 4.0 |
| 11 | 37 | 0.78 | 0.40 | 0.23 | 3 | 1 | 4.0 | [1.8, 5.9] | 40 | 0.92 | 0.77 | 0.17 | 16 | 13 | 5.8 | [2.5, 9.4] | 3 | 1.0 | 1.3 |
| 12 | 32 | 0.81 | 0.46 | 0.33 | 6 | 2 | 3.7 | [1.5, 5.3] | 39 | 0.99 | 0.80 | 0.19 | 24 | 19 | 5.7 | [2.3, 10.0] | 16 | 2.0 | 3.7 |

See Table D.2 for full description.

*Table D.4. Computational results for each month for the Daily limit strategy.*

| | Base Schedule | | | | | | | | Final Schedule | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Blocks | | | | Overtime | | | Ward | Blocks | | | | Overtime | | Ward | | | SR time | |
| Month | #Open | Utilisation | Reserved Cap. | Empty | #Regular | #Extended | Mean | [Min, Max] | #Open | Utilisation | Reserved Cap. | Empty | #Regular | #Extended | Mean | [min, max] | #Resch. | Median | Mean |
| 1 | 44 | 0.66 | 0.40 | 0.08 | 3 | 3 | 4.0 | [1.0, 6.2] | 44 | 0.89 | 0.82 | 0.08 | 20 | 16 | 5.2 | [1.0, 9.7] | 8 | 1.5 | 2.1 |
| 2 | 42 | 0.81 | 0.29 | 0.13 | 18 | 6 | 4.1 | [1.8, 6.7] | 45 | 1.01 | 0.95 | 0.06 | 30 | 21 | 5.4 | [2.6, 9.2] | 12 | 1.0 | 2.4 |
| 3 | 39 | 0.81 | 0.34 | 0.19 | 15 | 8 | 4.1 | [1.9, 6.1] | 40 | 0.95 | 0.79 | 0.17 | 24 | 16 | 5.0 | [2.8, 7.1] | 9 | 1.0 | 3.2 |
| 4 | 40 | 0.54 | 0.55 | 0.17 | 6 | 3 | 2.9 | [1.1, 4.8] | 42 | 0.76 | 0.67 | 0.13 | 15 | 8 | 4.0 | [1.4, 7.2] | 3 | 1.0 | 1.3 |
| 5 | 44 | 0.70 | 0.36 | 0.08 | 6 | 5 | 3.7 | [1.7, 5.4] | 44 | 0.89 | 0.82 | 0.08 | 16 | 12 | 4.6 | [2.2, 7.0] | 2 | 1.0 | 1.0 |
| 6 | 34 | 0.49 | 0.65 | 0.29 | 0 | 0 | 2.7 | [1.1, 4.4] | 37 | 0.64 | 0.49 | 0.23 | 4 | 3 | 4.1 | [1.5, 6.9] | 2 | 4.0 | 4.0 |
| 7 | 23 | 0.37 | 0.82 | 0.52 | 0 | 0 | 1.6 | [0.4, 2.7] | 31 | 0.56 | 0.36 | 0.35 | 4 | 2 | 3.7 | [1.4, 8.1] | 1 | 1.0 | 1.0 |
| 8 | 26 | 0.38 | 0.79 | 0.46 | 0 | 0 | 1.4 | [0.1, 2.6] | 27 | 0.45 | 0.25 | 0.44 | 0 | 0 | 1.9 | [0.3, 4.0] | 0 | - | - |
| 9 | 43 | 0.61 | 0.45 | 0.10 | 7 | 5 | 3.4 | [1.8, 5.1] | 45 | 0.83 | 0.78 | 0.06 | 16 | 7 | 4.7 | [2.0, 7.7] | 7 | 1.0 | 1.9 |
| 10 | 42 | 0.69 | 0.40 | 0.13 | 8 | 5 | 4.0 | [1.5, 6.6] | 45 | 0.77 | 0.72 | 0.06 | 11 | 4 | 4.8 | [2.4, 8.8] | 5 | 1.0 | 2.2 |
| 11 | 42 | 0.68 | 0.41 | 0.13 | 5 | 3 | 4.0 | [2.0, 6.2] | 43 | 0.85 | 0.76 | 0.10 | 15 | 4 | 5.5 | [2.8, 8.5] | 4 | 1.0 | 3.0 |
| 12 | 40 | 0.64 | 0.47 | 0.17 | 4 | 3 | 3.69* | [2.2, 5.3] | 43 | 0.90 | 0.81 | 0.10 | 19 | 5 | 5.6 | [2.2, 8.5] | 12 | 2.0 | 3.1 |

See Table D.2 for full description.

*Table D.5. Computational results for each month for the Balance ORs strategy.*

| | Base Schedule | | | | | | | | Final Schedule | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Blocks | | | | Overtime | | Ward | | Blocks | | | | Overtime | | Ward | | SR time | | |
| Month | #Open | Utilisation | Reserved Cap. | Empty | #Regular | #Extended | Mean | [Min, Max] | #Open | Utilisation | Reserved Cap. | Empty | #Regular | #Extended | Mean | [min, max] | #Resch. | Median | Mean |
| 1 | 36 | 0.80 | 0.40 | 0.25 | 4 | 3 | 3.9 | [1.0, 6.1] | 42 | 0.93 | 0.81 | 0.13 | 16 | 14 | 5.1 | [1.00, 10.0] | 5 | 1.0 | 1.8 |
| 2 | 36 | 0.94 | 0.30 | 0.25 | 22 | 7 | 4.1 | [2.4, 6.5] | 44 | 1.04 | 0.95 | 0.08 | 32 | 22 | 5.5 | [3.0, 8.9] | 12 | 1.0 | 2.0 |
| 3 | 36 | 0.87 | 0.35 | 0.25 | 15 | 6 | 4.1 | [2.4, 5.7] | 41 | 0.92 | 0.79 | 0.15 | 19 | 14 | 5.0 | [2.7, 8.4] | 8 | 1.0 | 2.6 |
| 4 | 28 | 0.77 | 0.55 | 0.42 | 6 | 3 | 2.8 | [0.7, 5.0] | 37 | 0.84 | 0.65 | 0.23 | 16 | 10 | 3.9 | [0.9, 6.9] | 3 | 1.0 | 2.7 |
| 5 | 36 | 0.85 | 0.36 | 0.25 | 7 | 5 | 3.6 | [1.5, 5.4] | 41 | 0.94 | 0.80 | 0.15 | 22 | 15 | 4.5 | [1.8, 7.2] | 3 | 1.0 | 1.0 |
| 6 | 24 | 0.70 | 0.65 | 0.50 | 0 | 0 | 2.8 | [0.8, 4.9] | 32 | 0.73 | 0.49 | 0.33 | 8 | 4 | 4.1 | [1.3, 8.3] | 2 | 1.0 | 1.0 |
| 7 | 12 | 0.73 | 0.82 | 0.75 | 2 | 0 | 1.7 | [0.5, 3.5] | 27 | 0.64 | 0.36 | 0.44 | 7 | 5 | 3.8 | [0.9, 8.8] | 2 | 1.5 | 1.5 |
| 8 | 16 | 0.63 | 0.79 | 0.67 | 0 | 0 | 1.4 | [0.1, 3.0] | 18 | 0.68 | 0.26 | 0.63 | 2 | 0 | 1.8 | [0.4, 3.6] | 0 | - | - |
| 9 | 32 | 0.82 | 0.45 | 0.33 | 5 | 5 | 3.4 | [0.1, 5.0] | 40 | 0.92 | 0.77 | 0.17 | 23 | 13 | 4.7 | [0.2, 7.8] | 7 | 1.0 | 1.9 |
| 10 | 36 | 0.80 | 0.40 | 0.25 | 8 | 4 | 3.8 | [1.4, 6.6] | 41 | 0.84 | 0.72 | 0.15 | 13 | 8 | 4.6 | [1.4, 7.5] | 0 | - | - |
| 11 | 36 | 0.79 | 0.41 | 0.25 | 4 | 1 | 4.2 | [1.4, 6.6] | 40 | 0.91 | 0.76 | 0.17 | 17 | 10 | 5.7 | [2.5, 10.0] | 2 | 1.0 | 1.0 |
| 12 | 32 | 0.79 | 0.47 | 0.33 | 5 | 3 | 3.6 | [2.1, 5.5] | 41 | 0.94 | 0.80 | 0.15 | 20 | 14 | 5.6 | [2.5, 9.9] | 8 | 2.0 | 2.1 |

See Table D.2 for full description.