STOCHASTIC MODELING AND DECISION MAKING IN TWO HEALTHCARE APPLICATIONS: INPATIENT FLOW MANAGEMENT AND INFLUENZA PANDEMICS

A Thesis
Presented to
The Academic Faculty

by

Pengyi Shi

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
December 2013

STOCHASTIC MODELING AND DECISION MAKING IN TWO HEALTHCARE APPLICATIONS: INPATIENT FLOW MANAGEMENT AND INFLUENZA PANDEMICS

Approved by:

Professor Jim Dai, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Pinar Keskinocak, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Julie Swann
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Turgay Ayer
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Sundaresan Jayaraman
College of Business and School of Materials Science and Engineering
Georgia Institute of Technology

Date Approved: November 2013

ACKNOWLEDGEMENTS

Pursuing the Ph.D. degree is a long, and most of the time, painful journey. I am deeply grateful to all the people who have helped me during this journey.

First of all, I would like to thank my two advisors, Professors Jim Dai and Pinar Keskinocak. They have dedicated an enormous amount of time, numerous efforts, and great patience to guide me on my thesis research. Without their support and continuous encouragement, I would not be able to finish my thesis and continue my career in academia. They are great mentors not only in academics but also in life; every meeting with them is enjoyable and inspiring. I am truly lucky to have the opportunity to work with and learn from them.

I would like to thank Professor Julie Swann for co-advising me during my first few years, and I appreciate her guidance when I was working on the second part of this thesis. I also thank Professors Turgay Ayer and Sundaresan Jayaraman for serving on my dissertation committee, providing insightful comments to my thesis, and offering generous help during my job searching process.

I would like to thank Professor Hayriye Ayhan for teaching me three important stochastic courses which have benefitted my research. I would like to thank Professors Gary Parker and Paul Kvam and Ms Pam Morrison for their dedicated service to the ISyE graduate program. They are always helpful when we students meet various problems.

I also want to thank my medical collaborators: Mr. Jin Xin and Mr. Joe Sim from the National University Hospital in Singapore, Dr. Franklin Dexter from the University of Iowa, and Dr. Bruce Lee from the Johns Hopkins University. Without their professional knowledge, inspiring discussion and kind support, many of my

research results would not be possible to finish.

I would like to thank my fellow Ph.D. students and friends. Your company and support have brought me a lot of joy during this long journey.

I would like to thank my parents for all the love and unconditional support they give me. I appreciate all the sacrifice they have to make since I cannot accompany them in China and I have not been able to visit them frequently.

Finally, I want to pass my deepest thanks to my husband Linji Yang. He always believes in me and gives me vigorous support during my career pursuit. His encouragement was particularly important when I experienced setbacks during my research. Without him, I may not be able to survive this long journey. He deserves as much credit as I do for completing this thesis.

Support for my Ph.D. studies was provided in part by the National Science Foundation under Grants CMMI 1030589, and the following Georgia Tech benefactors: Andrea Laliberte, Joseph C. Mello, the Nash Family, Claudia L. and J. Paul Raines, Richard "Rick" E. and Charlene Zalesky, and the University Health Care System (Augusta, GA).

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . . . **iii**

LIST OF TABLES . . . . . **xi**

LIST OF FIGURES . . . . . **xiii**

SUMMARY . . . . . **xix**

PART I INPATIENT FLOW MANAGEMENT

I OVERVIEW **1**

1.1 Motivation and research questions 2

1.2 Summary of contributions 6

1.2.1 Outline of Chapters 2 to 4 9

1.3 Literature review 9

II EMPIRICAL STUDY AT NUH **12**

2.1 NUH inpatient department 12

2.1.1 Admission sources 13

2.1.2 Medical specialties 17

2.1.3 Rationales for excluding certain wards 19

2.1.4 Data set 20

2.2 Early discharge campaign 22

2.2.1 Discharge distributions in Periods 1 and 2 22

2.2.2 Implementation of the early discharge policy 23

2.2.3 The changing operating environment 26

2.3 Ward capacity and overflow proportion 28

2.3.1 Basic ward setting in NUH 28

2.3.2 Capacity and BOR 29

2.3.3 Overflow proportion 31

2.3.4 Shared wards 34

2.4 Bed-request process 35
2.4.1 Bed-request rate from ED-GW patients and Arrival rate to ED 35
2.4.2 Testing the non-homogeneous Poisson assumption for ED-GW patients 37
2.4.3 Other admission sources 43
2.5 Length of Stay 47
2.5.1 LOS Distribution 49
2.5.2 AM- and PM-patients 49
2.5.3 LOS distributions according to patient admission source and specialty 53
2.5.4 LOS between right-siting and overflow patients 57
2.5.5 Test iid assumption for LOS 57
2.6 Service times 61
2.6.1 Service time distribution 61
2.6.2 Residual distribution 64
2.7 Pre- and post-allocation delays 67
2.7.1 Transfer process from ED to general wards 68
2.7.2 Pre- and post-allocation delays 71
2.7.3 Distribution of pre- and post-allocation delays 74
2.8 Internal transfers 75
2.8.1 Overall statistics on internal transfers 77
2.8.2 Transfer between GWs and ICU-type wards 80
2.8.3 Transfer between two GWs 82

III HIGH-FIDELITY MODEL FOR HOSPITAL INPATIENT FLOW MANAGEMENT 84

3.1 A stochastic network model for the inpatient operations 84
3.1.1 A stochastic processing network with multi-server pools . . . 85
3.1.2 Critical feature 1: a two-time-scale service time model 87
3.1.3 Critical feature 2: bed assignment with overflow 90
3.1.4 Critical feature 3: allocation delays 91

3.1.5 Service policies 94
3.1.6 Modeling patient transfers between ICU and GW 95
3.2 Populated stochastic model using NUH data 97
3.2.1 Arrivals 97
3.2.2 Server pools and service policy 99
3.2.3 Length of stay and discharge distributions 101
3.2.4 Patient class 102
3.2.5 A dynamic overflow policy 103
3.2.6 Pre- and post-allocation delays 104
3.3 Verification of the populated NUH model 107
3.3.1 The baseline scenario 108
3.3.2 Models missing any of the critical features 110
3.4 Factors that impact ED-GW patients' waiting time 113
3.4.1 Period 2 discharge has a limited impact on reducing waiting time statistics 114
3.4.2 A hypothetical Period 3 policy can have a significant impact on flattening waiting time statistics 115
3.4.3 Policies impact on the daily waiting time statistics and overflow proportion 118
3.4.4 Sensitivity analysis 120
3.4.5 Intuition about the gained insights 121
3.5 Concluding remarks and future research 122
3.5.1 Future work 124
IV A TWO-TIME-SCALE ANALYTICAL FRAMEWORK 126
4.1 A single-pool model without allocation delays 126
4.2 A two-time-scale approach for the single-pool model 128
4.2.1 A two-time-scale approach: step 1, midnight dynamics 128
4.2.2 A two-time-scale approach: step 2, time-of-day queue length dynamics 129
4.2.3 Predict time-dependent queue length and waiting time dynamics 131

4.3 Single-pool model with allocation delays 135
4.3.1 System dynamics 135
4.3.2 Predict the performance measures 136
4.4 Numerical results 139
4.5 Diffusion approximations 140
4.5.1 Approximation for the midnight customer count 141
4.5.2 Approximate the hourly customer count 143
4.5.3 Approximate the time-dependent performance 146
4.5.4 Numerical results on the diffusion approximation 150
4.6 Diffusion limits for the single-pool model 157
4.7 Conclusion and future work 163

PART II INFLUENZA PANDEMIC MODELING AND RESPONSE

V OVERVIEW 167
5.1 Background 167
5.2 Contributions and literature review 169
5.2.1 Contributions 169
5.2.2 Literature review on disease spread models 171
5.3 Basic disease spread model 172
5.3.1 Baseline model settings 172
5.3.2 Simulation logic 173
5.3.3 Data and model calibration 175

VI THE IMPACT OF SEASONALITY AND VIRAL MUTATION ON THE COURSE OF AN INFLUENZA PANDEMIC 181
6.1 Introduction 181
6.2 Modeling seasonality and viral mutation 182
6.2.1 Modeling seasonality 182
6.2.2 Modeling viral mutation 182
6.2.3 Combination of seasonality and viral mutation 184

6.2.4 Simulation runs and sensitivity analysis 184
6.3 Results 185
6.3.1 Seasonality scenarios 186
6.3.2 Mutation scenarios 186
6.3.3 Seasonality and viral mutation scenarios 187
6.4 Discussion 189
6.4.1 Public health implications 191
6.4.2 Conclusions and limitations 192

VII THE IMPACT OF MASS GATHERINGS AND HOLIDAY TRAVELING ON THE COURSE OF AN INFLUENZA PANDEMIC 193

7.1 Introduction 193
7.2 Models 194
7.2.1 Modeling mass social mixing: public gatherings and Holiday travel 194
7.2.2 Force of infection and model calibration 196
7.2.3 Simulation runs and sensitivity analyses 201
7.3 Results 202
7.3.1 The timing of mass travel/public gatherings t^* 202
7.3.2 Impact of Holiday traveling on multiple peaks 206
7.3.3 The duration of the mass traveling period (l) and the proportion of the population traveling (p) under the non-Holiday setting 209
7.3.4 Risk for travelers' families under the non-Holiday setting . . 210
7.3.5 Regional impact of traveling and mass gatherings 211
7.4 Discussion 212
7.4.1 Public health implications 214
7.4.2 Conclusion and future direction 214

APPENDIX A — APPENDIX FOR CHAPTER 2 216

APPENDIX B — APPENDIX FOR CHAPTER 3 230

APPENDIX C — APPENDIX FOR CHAPTER 4 . . . . . **271**

APPENDIX D — APPENDIX FOR CHAPTER 7 . . . . . **273**

REFERENCES **277**

LIST OF TABLES

1 The average waiting time and x-hour service levels for ED-GW patients. 15

2 Discharge time distributions in Periods 1 and 2. 23

3 Primary specialties and BOR for the 19 general wards. 30

4 Bed allocation in shared wards. 35

5 Results for Kolmogorov-Smirnov tests on testing the non-homogeneous Poisson assumption for bed-request processes. 40

6 Average LOS for each specialty and each admission source. 54

7 Average LOS for right-siting and overflow patients. 58

8 Results of the χ^2-tests for testing the identically distributed assumption for LOS. 60

9 Results of the nonparametric tests for testing the serial dependence among patient LOS. 62

10 Proportion of transfer patients for each admission source and medical specialty. 78

11 Decomposition of ED-GW and EL transfer patients by number of transfers and pathways 79

12 Number of patients transferred in and out for each ward. 83

13 Server pool setting in the simulation model. 100

14 Priority of primary and overflow pools in the simulation model. . . . 100

15 Estimated value for the parameter p of the Bernoulli distribution to determine patient classes. 103

16 Simulation and empirical estimates of waiting time statistics for ED-GW patients from each specialty. 109

17 Values of the key parameters in the baseline disease spread simulation model. 179

18 Adjusted parameter values to achieve the attack rates in the 1957 Pandemic. 180

19 Age-specific attack rates from the 1957 pandemic. 180

20 The peak prevalence value in the second wave varies as the mutant strain emerges later. 187

21 Results from Different Mass Gathering Scenarios (Initial $R_0 = 1.5$). . 203
22 Results from Different Mass Gathering Scenarios (Initial $R_0 = 1.3$). . 204
23 Results from Different Mass Gathering Scenarios (Initial $R_0 = 1.8$). . 205
24 The average waiting times and service levels by bed-request hour. . . 219
25 Waiting time statistics for ED-GW patients from each specialty. . . 221
26 Overflow proportion and BOR share for each ward. 224
27 LOS distribution in Periods 1 and 2. 224
28 LOS tail frequencies (start from 31 days, cut-off at 90 days). 225
29 LOS distributions for ED-GW patients admitted in AM and PM. . . 225
30 Simulation and empirical estimates of the waiting time statistics for ED-GW patients requesting beds in each hour of the day. 267

LIST OF FIGURES

1 Hourly waiting time statistics for ED-GW patients. 3

2 Discharge time distributions in Periods 1 and 2. 4

3 Four admission sources to general wards and nine patient specialties. 13

4 Waiting times statistics for each medical specialty. 18

5 Discharge time distributions during and after the implementation of early discharge policy. 25

6 Monthly admission rate, number of beds, and BOR. 27

7 Overflow proportion for each specialty in Periods 1 and 2. 32

8 Overflow proportion for each ward in Periods 1 and 2. 34

9 Arrival rate to ED and bed-request rate. 37

10 Hourly bed request rate from 4 major specialties in Period 1. 38

11 QQ plot and CDF plot of $\{R_j^i\}$ from all intervals in Period 1 for the bed-request process of ED-GW patients. 39

12 Comparison between sample means and sample variances of bed-requests. 42

13 Hourly bed-request rate for each admission source. 43

14 QQ plots and CDF plots of $\{R_j^i\}$ from all intervals in Period 1 for other admission sources. 45

15 Histograms of the inter-bed-request time for ICU-GW and SDA patients using the combined data. The bin size is 10 minutes. 46

16 Histograms of the daily number of admissions for EL patients and daily number of bed-requests for ICU-GW and SDA patients. 48

17 Histograms of the first bed-request time. 48

18 LOS distributions in Periods 1 and 2. 50

19 Average LOS with respect to admission time. 51

20 LOS distribution for ED-AM and ED-PM patients. 52

21 LOS distributions of each specialty. 55

22 Distribution of service times in two time resolutions. 63

23 LOS and day-resolution service time distributions for General Medicine patients. 64

24 Empirical distribution of the residual of service time. 66

25 Admission time and discharge time distributions for same-day discharge patients. 66

26 Empirical distributions of the residual of service times for AM- and PM-admitted ED-GW patients. 66

27 Process flow of the transfer from ED to GW. 68

28 Mean and CV of estimated pre- and post-allocation delays with respect to the delay initiation hour. 74

29 Empirical distributions of allocation delays. 75

30 Fitting the log-transformed data for pre-allocation delay. 76

31 Fitting the log-transformed data for post-allocation delay. 76

32 Transfer-out time distributions for transfer patients from ED-GW and EL sources. 80

33 Estimated LOS distributions for transfer patients. 81

34 Transfer-out time distribution. 83

35 Arrival and server pool configuration in the stochastic model of NUH inpatient department. 85

36 Service time distributions, at hourly resolution, for General Medicine patients that are admitted in afternoons. 90

37 Pre- and post-allocation delays under different scenarios. 92

38 Mean and CV of pre- and post-allocation delays used in the simulation model. 106

39 Baseline simulation output compares with empirical estimates 1. . . . 109

40 Baseline simulation output compares with empirical estimates 2. . . . 110

41 Simulation output from using an iid service time model. 111

42 Simulation output from a model without allocation delays. 112

43 Simulation output from using a static overflow policy. 113

44 Comparing hourly waiting time statistics under the baseline scenario and scenario with Period 2 discharge distribution. 115

45 Period 2 discharge distribution and a hypothetical discharge distribution. 116

46 Comparing hourly waiting time statistics under the baseline scenario and scenario with Period 3 policy. 117

47 Hourly waiting time statistics under three scenarios. 120

48 Numerical results for the steady-state time-dependent mean waiting time and mean queue length. 140

49 Stationary distribution of the midnight customer count from exact Markov chain analysis and diffusion approximation (large systems). . 151

50 Stationary distribution of the midnight customer count from exact Markov chain analysis and diffusion approximation (small systems). . 152

51 Stationary distribution of $X(t)$ from exact Markov chain analysis and diffusion approximation for $N = 525$. 153

52 Stationary distribution of $X(t)$ from exact Markov chain analysis and diffusion approximation for $N = 66$. 153

53 Time-dependent mean queue length from exact Markov chain analysis and diffusion approximation (large systems). 154

54 Time-dependent mean queue length from exact Markov chain analysis and diffusion approximation (small systems). 155

55 Time-dependent mean waiting time and 6-hour service level from exact Markov chain analysis and diffusion approximation ($N = 525$). 156

56 Time-dependent mean waiting time and 6-hour service level from exact Markov chain analysis and diffusion approximation ($N = 132$). 156

57 Time-dependent mean waiting time and 6-hour service level from exact Markov chain analysis and diffusion approximation ($N = 66$). 157

58 Time-dependent mean waiting time from diffusion approximation when feeding in exact stationary distribution of the midnight customer count ($N = 66$ and $N = 132$). 158

59 Time-dependent performance measures under different LOS distributions ($N = 505$). 165

60 Time-dependent performance measures under different LOS distributions ($N = 66$). 166

61 An example of the contact network. 173

62 Plot of R_0 value as function of time. 183

63 Natural disease history with viral mutation. 183

64 Daily prevalence curves for seasonality scenarios. 188

65 Daily prevalence curves for mutation scenarios. 188

66 Reproduced prevalence curve for the 1918 pandemics. 189

67 An example of the contact network during the traveling period. . . . 196

68 Epidemic curves in the Holiday scenarios. 207

69 Epidemic curves in the Holiday and social distancing scenarios. 208

70 Peak prevalence value and peak day in Bibb County. 212

71 Waiting time distributions calculated in the conventional way. 217

72 Hourly waiting times statistics calculated in the conventional way. . . 218

73 Waiting times statistics for each specialty calculated in the conventional way. 220

74 BOR from primary and non-primary specialties for each ward in Period 1 and 2. 223

75 Average duration between bed-request time and bed-allocation time for right-siting and overflow patients. 228

76 Estimated average post-allocation delay with respect to the delay initiation time. 229

77 Independence between admission and discharge hours and between LOS and discharge hours using Period 1 data. 231

78 Estimate average duration between bed-request and bed-allocation for different groups of patients. 234

79 Estimated values of $p(t)$ from empirical data and values used in the baseline simulation. 236

80 Three groups of hypothetical discharge distributions 239

81 Hourly waiting time statistics under two scenarios. 240

82 Hourly waiting time statistics under scenarios with hypothetical discharge distributions of group (a). 241

83 Hourly waiting time statistics under scenarios with hypothetical discharge distributions of group (b). 241

84 Hourly waiting time statistics under scenarios with hypothetical discharge distributions of group (c). 242

85 Hourly waiting time statistics under the midnight discharge scenario. 242

86 Simulation output compares with empirical estimates (Period 2 data). 244

87 Hourly waiting time statistics under the scenarios with the Period 2 discharge distribution and a hypothetical discharge distribution with the peak time at 4-5pm. 246

88 Hourly waiting time statistics under the baseline scenario and scenarios with different choices of arrival models. 248

89 Hourly waiting time statistics under the revised-baseline-arrival1 scenario and other scenarios. 249

90 Hourly waiting time statistics under the revised-baseline-arrival2 scenario and other scenarios. 250

91 Hourly waiting time statistics under the baseline scenario and scenarios with different patient priority settings. 251

92 Hourly waiting time statistics under the revised-baseline-priority1 scenario and other scenarios. 252

93 Hourly waiting time statistics under the revised-baseline-priority2 scenario and other scenarios. 252

94 Hourly waiting time statistics under the baseline scenario and scenarios with different allocation delay distributions. 253

95 Hourly waiting time statistics under the revised-baseline-exponential scenario and other scenarios. 254

96 Hourly waiting time statistics under the revised-baseline-normal scenario and other scenarios. 254

97 Hourly waiting time statistics under the baseline scenario and scenarios with different choices of $p(t)$. 256

98 Hourly waiting time statistics under the revised-baseline-$p(t)$-0 scenario and other scenarios. 257

99 Hourly waiting time statistics under the revised-baseline-$p(t)$-0.5 scenario and other scenarios. 257

100 Hourly waiting time statistics under the revised-baseline-$p(t)$-1 scenario and other scenarios. 258

101 Hourly waiting time statistics under the baseline scenario and the scenario with increased arrival rate. 259

102 Hourly waiting time statistics under the revised-baseline-increase-arrival scenario and other scenarios. 260

103 Hourly waiting time statistics under the revised-baseline-noAMPM scenario and other scenarios. 262

104 Hourly waiting time statistics under the revised-baseline-noAMPM-normal-load scenario and other scenarios. 263

105 Moving average plots from 5 replications. 265

106 Hourly waiting time statistics from each batch. 266

SUMMARY

Delivering health care services in an efficient and effective way has become a great challenge for many countries due to the aging population worldwide, rising health expenses, and increasingly complex healthcare delivery systems. It is widely recognized that models and analytical tools can aid decision-making at various levels of the healthcare delivery process, especially when decisions have to be made under uncertainty. This thesis employs stochastic models to improve decision-making under uncertainty in two specific healthcare settings: inpatient flow management and infectious disease modeling.

In Part I of this thesis, we study patient flow from the emergency department (ED) to hospital inpatient wards. This line of research aims to develop insights into effective inpatient flow management to reduce the waiting time for admission to inpatient wards from the ED. Delayed admission to inpatient wards, also known as ED boarding, has been identified as a key contributor to ED overcrowding and is a big challenge for many hospitals. Part I consists of three main chapters. In Chapter 2 we present an extensive empirical study of the inpatient department at our collaborating hospital. Motivated by this empirical study, in Chapter 3 we develop a high fidelity stochastic processing network model to capture inpatient flow with a focus on the transfer process from the ED to the wards. In Chapter 4 we devise a new analytical framework, *two-time-scale analysis*, to predict time-dependent performance measures for some simplified versions of our proposed model. We explore both exact Markov chain analysis and diffusion approximations.

Part I of the thesis makes contributions in three dimensions. First, we identify

several novel features that need to be built into our proposed stochastic network model. With these features, our model is able to capture inpatient flow dynamics at hourly resolution and reproduce the empirical time-dependent performance measures, whereas traditional time-varying queueing models fail to do so. These features include unconventional non-i.i.d. (independently and identically distributed) service times, an overflow mechanism, and allocation delays. Second, our two-time-scale framework overcomes a number of challenges faced by existing analytical methods in analyzing models with these novel features. These challenges include time-varying arrivals and extremely long service times. Third, analyzing the developed stochastic network model generates a set of useful managerial insights, which allow hospital managers to (i) identify strategies to reduce the waiting time and (ii) evaluate the trade-off between the benefit of reducing ED congestion and the cost from implementing certain policies. In particular, we identify early discharge policies that can eliminate the excessively long waiting times for patients requesting beds in the morning.

In Part II of the thesis, we model the spread of influenza pandemics with a focus on identifying factors that may lead to multiple waves of outbreak. This line of research aims to provide insights and guidelines to public health officials in pandemic preparedness and response. In Chapter 6 we evaluate the impact of seasonality and viral mutation on the course of an influenza pandemic. In Chapter 7 we evaluate the impact of changes in social mixing patterns, particularly mass gatherings and holiday traveling, on the disease spread.

In Chapters 6 and 7 we develop agent-based simulation models to capture disease spread across both time and space, where each agent represents an individual with certain socio-demographic characteristics and mixing patterns. The important contribution of our models is that the viral transmission characteristics and social contact patterns, which determine the scale and velocity of the disease spread, are no longer static. Simulating the developed models, we study the effect of the starting season

of a pandemic, timing and degree of viral mutation, and duration and scale of mass gatherings and holiday traveling on the disease spread. We identify possible scenarios under which multiple outbreaks can occur during an influenza pandemic. Our study can help public health officials and other decision-makers predict the entire course of an influenza pandemic based on emerging viral characteristics at the initial stage, determine what data to collect, foresee potential multiple waves of attack, and better prepare response plans and intervention strategies, such as postponing or cancelling public gathering events.

STOCHASTIC MODELING AND DECISION MAKING IN TWO HEALTHCARE APPLICATIONS: INPATIENT FLOW MANAGEMENT AND INFLUENZA PANDEMICS

PART I

Inpatient Flow Management

by

Pengyi Shi

CHAPTER I

OVERVIEW

Hospital inpatient beds accommodate patients who need to stay in a hospital (usually for one or more nights) for treatment, and these beds are one of the most critical resources in hospitals. Inpatient flow and bed management has crucial impact on hospital operations [69], especially on emergency department (ED) crowdedness [97, 81, 7, 150, 130]. Prolonged *waiting time for admission to inpatient beds*, also known as ED boarding, has been identified as a key contributor to ED overcrowding worldwide [146, 79, 117]. The waiting time for admission to inpatient beds, or simply the *waiting time* in this thesis, is defined as the duration from when ED doctors decide to admit a patient (i.e., the bed-request time of the patient) to when the patient is admitted to an inpatient bed. This waiting time is closely monitored by government agencies. For example, the ministry of health (MOH) of Singapore publishes the daily median of this waiting time from each Singaporean public hospital on its website (see [139]); also see reports and surveys from the department of health in UK [40] and the US general accounting office [146]. According to [146], more than half of the surveyed US hospitals have an average waiting time longer than 4 hours, and 20% of the surveyed hospitals have boarded patients longer than 8 hours on average.

While no patient likes waiting, excessively long waiting time (e.g., 8 hours or more) is extremely undesirable, not only because patients can get very frustrated during the long wait [118], but also because of the adverse outcome associated with it. Liu et al. [98] and Singer et al. [141] have discovered that patients who waited longer than 6 hours after their admission decisions are more likely to experience longer inpatient

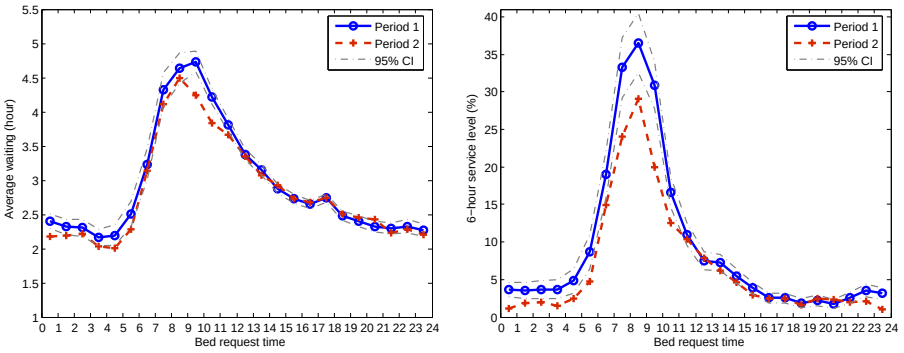
stay, higher mortality rates, and other undesirable events in ED such as suboptimal blood pressure control. In addition, patients continue to occupy ED resources while waiting to be admitted and can block new patients from being treated in the ED, which lead to ED overcrowding and sometimes ambulance diversion [1]. Moreover, recent studies have estimated that as high as 15% of the overall time spent in EDs was by these admitted patients (boarding patients) [22], while just a 1-hour reduction in the mean waiting time for admission to inpatient beds would result in about $10,000 additional daily revenue for hospitals [116]. Thus, it is important for hospitals to eliminate the excessive amount of waiting, especially for morning bed-requests.

Part I of this thesis is dedicated to (i) build a high fidelity model to capture inpatient flow dynamics with a particular focus on the transfer process from ED to inpatient beds, (ii) predict the time-of-day waiting time performance during this transfer process and other important performance measures, and (iii) generate insights into efficient inpatient flow management and identify strategies (from the inpatient side) to reduce the waiting time and eventually alleviate ED overcrowding. This part constitutes three main chapters (Chapters 2 to 4). Before starting the next three chapters, we provides an overview in the rest of this chapter. In Section 1.1, we first introduce the motivation for the research questions we aim to answer in the next three chapters . Then we summarize our major contributions in Section 1.2. Finally, we provide a brief literature review on patient flow models for hospital operations in Section 1.3.

1.1 Motivation and research questions

Our study is motivated by an empirical study at our collaborating hospital in Singapore, National University Hospital (NUH). NUH is one of the major public hospitals in Singapore. It operates a busy ED and a large inpatient department that has about 1000 inpatient beds to serve patients admitted from ED and other sources. These

inpatient beds locate in different wards, and we focus on beds in 19 general wards (GWs) in our research (GW beds are sometimes also referred as *floor beds* in other hospitals, and we give out the precise definition of GWs in Section 2.1.3). At NUH, around 20% of patients visiting ED are admitted into a general ward after finishing the treatment in ED, thereby becoming *ED-GW patients*. From January 1, 2008 to June 30, 2009, called *Period 1* in this thesis, the average waiting time for ED-GW patients at NUH is 2.82 hours (169 minutes), which does not seem to be very long. However, this level of complacency immediately evaporates if we examine the waiting times of patients requesting beds in mornings. The solid curve in Figure 1a shows that the average waiting time is more than 4 hours long for patients who request a bed between 7 and 10am. Moreover, among these patients, Figure 1b shows that more than 30% of them have to wait 6 hours or longer. In this paper, we define the *6-hour service level* as the fraction of patients who have to wait 6 hours or longer.



(a) Average waiting times (b) 6-hour service level

Figure 1: **Hourly waiting time statistics for ED-GW patients**; Period 1: January 1, 2008 to June 30, 2009; Period 2: January 1, 2010 to December 31, 2010. Each dot represents the average waiting time or 6-hour service level for patients requesting beds in that hour. For example, the dot between 7 and 8 represents the value of the hourly statistics between 7am and 8am. The 95% confidence intervals are plotted for Period 1 curves.

The inpatient discharge policy is believed by NUH to have contributed to the prolonged waiting times for ED-GW patients requesting beds in the morning. The

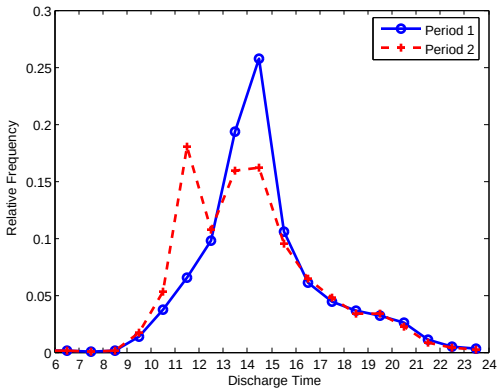


Figure 2: **Discharge time distributions in Periods 1 and 2.** The values in the first 6 hours are nearly zero and are not displayed.

solid curve in Figure 2 plots the discharge distribution of patients from general wards at NUH in Period 1. Clearly, the peak discharge hour is between 2pm and 3pm. Therefore, many admissions must wait until after 3pm, while bed-requests of ED-GW patients can occur during the entire day (e.g., see the solid curve in Figure 13 in Section 3.2.1). In other words, if there is no bed immediately available for a morning bed-request, the incoming patient is likely to wait until afternoon to be admitted.

In fact, the time-dependency of waiting times is not unique at NUH. Similar waiting time curves have been observed in other hospitals (see Figure 30 of [5]), and so have the number of patients waiting at different time of a day [120, 69]. Meanwhile, the bed-request and discharge patterns in many other hospitals are also similar to what we observed at NUH; see, e.g., Table 1 in [120] and Figure 6 in [5]. Studies in literature [9, 154] and government agencies [39] have recommended discharging patients at earlier hours of the day to eliminate the temporary mismatch between bed demand and supply in the morning.

In July 2009, NUH itself launched an early discharge campaign. After six months' implementation, a new discharge pattern emerged in Period 2: January 1, 2010 to December 31, 2010. The dashed curve in Figure 2 displays the new discharge distribution. A morning discharge peak arises, occurring between 11am and noon; 26%

of the patients are discharged before noon in Period 2, doubling the proportion in Period 1 (13%). The daily average waiting time is reduced from 2.82 hours (169 minutes) in Period 1 to 2.77 hours (166 minutes) in Period 2, and the daily 6-hour service level is reduced from 6.52% in Period 1 to 5.13% in Period 2. The dashed curves in Figures 1a and 1b plot the time-dependent hourly average waiting time and 6-hour service level in Period 2, respectively. From these empirical results, we observe that (a) some improvement in reducing the peak hourly 6-hour service level has been achieved in Period 2, and (b) little progress has been made in eliminating the long waiting times for morning bed-requests (*flattening* the hourly waiting time statistics) or reducing the daily waiting time statistics.

These empirical observations raise two issues. First, it is unclear whether the improvements in Period 2 result from the NUH's early discharge campaign. As in many hospitals, the operating environment is continuously changing at NUH. Bed capacity is being increased in response to the rising number of patients seeking treatment. In Period 2, the bed occupancy rate (BOR) has reduced by 2.7% [136]. Therefore, it is difficult to evaluate the impact of the early discharge policy through empirical analysis alone. Second, one wonders if there is any discharge policy, perhaps combined with other operational policies, that can achieve a more significant improvement in flattening or reducing the waiting time statistics. Unfortunately, it is prohibitively expensive for hospitals to experiment with various options in a real operational environment to identify such policies. Therefore, we need a high-fidelity model to (i) capture the inpatient flow dynamics and predict the time-dependent waiting time performance, and (ii) quantify the impact of operational policies such as early discharge and identify strategies to eliminate the long waiting times.

1.2 Summary of contributions

Part I makes three major contributions to the modeling, theory, and practice of inpatient flow management.

Modeling. Based on the comprehensive empirical study we conduct at NUH (see Chapter 2), we develop a new stochastic network model to capture inpatient flow dynamics in Chapter 3. This model can reproduce, at high fidelity, many empirical performance measures at both the hospital and the medical specialty levels. In particular, the model can approximately replicate the time-dependent hourly waiting time performance as shown in Figure 1.

In order for the model to be able to capture the inpatient operations at hourly resolution, we find several key features must be built in. They include a two-time-scale service time model, an overflow mechanism among multiple server pools, and pre- and post-allocation delays which capture the extra amount of delay caused by resource constraints other than bed unavailability during the ED to wards transfer process. Under our two-time-scale service time model, service times of inpatients are not independent and identically distributed (iid). We will elaborate this service time model and other key features in Section 3.1. Time-varying $M_t/GI/n$ queues or their network versions, where the arrival process is Poisson with time-varying arrival rate and the service times are iid, have been used in literature to model hospital operations; see, for example, [62, 89, 1]. Despite our best efforts, we are not able to reproduce the time-dependent performance curves using these models. See Section 3.3.2 for simulation results for models that miss each one of the three key features.

We want to emphasize that studying inpatient flow dynamics at hourly resolution and capturing time-of-day performance are important, especially when one evaluates policies that impact the interface between ED and wards, where hours of waiting matter. For example, our model predicts that certain types of discharge policies can significantly reduce waiting times for morning bed-requests, but have limited impact

on the daily waiting time statistics (see also the second contribution below). By studying the time-of-day performance, we are able to gain insights into the impact of such policies on certain *sub-groups* of patients, in addition to the aggregated impact on all patients. Moreover, as pointed out by Armony et al. [5], understanding the system's behavior at hourly resolution is of particular importance for operational planning when nurses and physicians are modeled as servers, e.g., for planning nurse staffing. Thus, our model can potentially be used to aid other operational decisions that require a understanding of the time-varying dynamics of inpatient flow.

Moreover, our model strikes a proper balance between analytical tractability and fidelity. We are able to analyze some simplified versions of the proposed model while still keeping certain key features, including the two-time-scale service time model and allocation delays. This leads to our second contribution on analytical methods.

Theory. In Chapter 4, we develop an analytical framework, known as the *two-time-scale analysis*, to predict time-dependent performance measures for some simplified versions of the high fidelity stochastic network model we proposed in Chapter 3. Due to the unique features such as the two-time-scale service times and allocation delays, no existing analytical method applies to analyzing our proposed models. Even in the simplest setting with a single server pool, it is challenging to use existing approximation methods to predict the time-dependent hourly performances because the service times are *extremely* long (the average is around 5 days) compared with the time-variations of the arrival rate (arrival period is one day). Our proposed framework can overcome this challenge as well as other difficulties. We focus on analyzing a single-server-pool model with this two-time-scale framework, and we demonstrate both exact analysis and diffusion approximations. The analysis help us generate insights into the impact of different operational policies on both the daily and time-of-day performance measures. This leads to our third contribution in the practice of inpatient flow management.

Practice. Through the two-time-scale analytical methods and simulation analysis of the proposed model, we obtain managerial insights into the impact of early discharge and other operational policies on both the daily and time-of-day waiting time performance. First, consistent with the empirical observations, the Period 2 early discharge alone has little impact on reducing or flattening the waiting time of ED-GW patients. Second, if the hospital is able to (i) move the first discharge peak in Period 2 three hours earlier, to occur between 8am and 9am, and still keep 26% discharge before noon (see the dash-dotted curve in Figure 45) and (ii) meanwhile stabilize the time-varying allocation delays, then the hourly waiting time curves can be approximately flattened (see Figure 46). However, the daily waiting time statistics still show limited reductions. Third, we identify policies that can significantly impact the daily waiting time performance such as increasing bed capacity and reducing the mean allocation delays; these policies do not necessarily flatten the hourly waiting time curves though. Finally, we use the developed two-time-scale analytical framework to provide some intuition to explain the different impacts on the hourly and daily waiting time performance of these policies.

To the best of our knowledge, the model we have developed is the first stochastic model to comprehensively analyze the effect of discharge policy in combination with other strategies such as stabilizing allocation delays. The most relevant work is a recent paper by Powell et al. [120], where the authors propose a deterministic fluid model to analyze the effect of discharge timing on the waiting time for admission to wards. Their model provides a simple method to calculate the hourly mean patient count (number of patients in service and waiting), and this method can actually be supported by a more rigorous argument using our two-time-scale analytical framework. However, the fluid method is not enough to calculate the mean queue length or other performance measures which depend on the *entire distribution* of the hourly patient count. Therefore, some of the managerial insights generated in [120] can be

misleading. For example, the authors find that by shifting the peak inpatient discharge time four hours earlier, the waiting time can be reduced to zero; but zero waiting can hardly be achieved in any hospital with as much as 90% bed utilization and random arrivals and service times. We believe our model is more comprehensive and sophisticated so that it captures inpatient flow operations at hourly resolution and generates insights on many operational policies including discharge timing. Some other relevant works on discharge policies are mostly empirical studies. For example, [86] classifies admission data from 23 Australian hospitals into five categories based on the relative timing of daily admission and discharge curves, and uses statistical analysis to show that days with late discharge peaks contribute significantly to ED overcrowding.

1.2.1 Outline of Chapters 2 to 4

The next three chapters is organized as follows. First, in Chapter 2, we present the empirical study we conduct at the NUH inpatient department. We document statistics for many performance measures which motivate the stochastic network model we develop. Then, in Chapter 3, we introduce the general framework of our proposed stochastic network model, and populate the model with empirical data documented in Chapter 2. We simulate the populated model to generate a number of managerial insights for reducing and flattening waiting times for admission to wards. Finally, in Chapter 4, we introduce the two-time-scale analytical framework to analyze several versions of our proposed model. The analysis will generate further insights for us to understand and improve inpatient flow management.

1.3 Literature review

Hospital patient flow. Hospital patient flow has been studied extensively in the operations research literature. For example, [5] and [70] conduct detailed studies of patient flow in various departments at an Israeli and a US hospital, respectively.

Readers are also referred to the many articles cited in these two papers for further references. Armony et al. [5] do not focus on discharge policies, but they empirically study the transfer process flow from ED to GW (which they call internal wards). Discrete-event simulation and queueing theory are two commonly used approaches for modeling and improving patient flow [59, 82, 157]. Compared to the rich literature on patient flow models of ED, inpatient flow management and the interface between ED and inpatient wards have received less attention; see the same discussion in Section 4 of [5]. Related works on inpatient operations include capacity allocation and flow improvement in specialized hospitals or wards [63, 33, 19, 62], ward nurse staffing [148, 155], bed assignment and overflow [145, 104], and elective admission control and design [74, 75]. Note that Yankovic and Green [155] demonstrate that the admission or discharge *blocking* caused by nurse shortages can have a significant impact on system performance. This insight is consistent with our findings on the allocation delays.

Stochastic network models. Stochastic network models have been a common tool to study manufacturing, communication and service systems [55, 8, 156]. In particular, research motivated by call center operations has extensively studied stochastic systems with time-varying arrivals and time-dependent performance. For example, Feldman et al. [44] and recent work by Liu and Whitt [99] propose staffing algorithms to achieve time-stable performance. Unlike call center models, our hospital model has extremely long service times with an average of about five days. Within the service time of a typical patient, the arrival pattern has gone through five cycles. Therefore, existing approximation methods developed for call center models (such as PSA [60, 149], lagged PSA [58], modified offered-load approximation [105], infinite-server approximation [83], and iteration algorithms [32, 45]) are not applicable to our hospital model. Moreover, the servers in our model are inpatient beds. It is not realistic to adjust the number of beds within a short time window.

Time scales in hospital operations. Previous studies have noticed different time scales in hospital operations [124, 104]. Our two-time-scale analysis is inspired by, but significantly different from, these works. Mandelbaum et al. [104] point out that different time scales arise naturally when hospitals operate in the quality- and efficiency-driven (QED) regime, i.e., the number of servers is large, service times are in days, whereas waiting times are in hours. Ramakrishnan et al. [124] construct a two-time-scale model for ED and wards, where the wards operate on a time scale of days and are modeled by a discrete-time queue, and the ED operates on a much faster time scale and is modeled by a continuous time Markov chain (CTMC). While their discrete-time queue is similar to our discrete-time queue in the single-server-pool setting to be introduced in Section 4.1, their focus is on improving ED operations; our focus is the inpatient department operations and we aim to predict the time-dependent performance during the ED to wards transfer process. We do not explicitly model operations within the ED in this research.

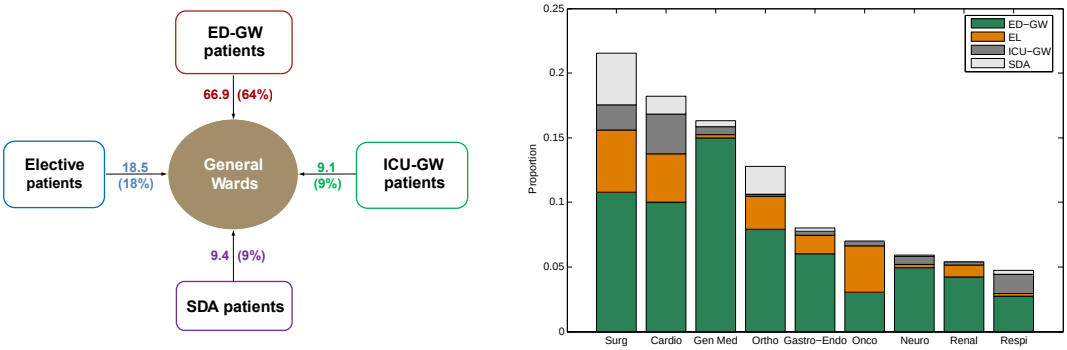
CHAPTER II

EMPIRICAL STUDY AT NUH

This chapter is organized as follows. Section 2.1 gives an overview of NUH's inpatient department. Section 2.2 describes an early discharge campaign implemented in 2009 at NUH, and explains the reason of using two periods (Periods 1 and 2) in the empirical analysis. Sections 2.3 introduces another important performance measure, the overflow proportion. This section also describes the basic organizational unit at NUH, *ward*, and reports ward-level statistics. Sections 2.4 to 2.7 relate to the modeling elements of the proposed stochastic network model in Chapter 3. Section 2.4 discusses the bed-request process (which serves as the arrival process to the stochastic model). Sections 2.5 and 2.6 are for the service time model. Section 2.7 summarizes the motivation of modeling allocation delays and relevant empirical studies. Finally, Section 2.8 presents a supplement study for patients who have been internally transferred.

2.1 NUH inpatient department

This section introduces some basic information of the NUH inpatient department. We introduce different admission sources (Section 2.1.1) and medical specialties (Section 2.1.2), and show ED-GW patient's waiting time performance in Periods 1 and 2. We also describe the data set used in our empirical study in Section 2.1.4. We focus on 19 *general wards*, which have a total number of beds ranging from 555 to 638 between January 1, 2008 and December 31, 2010. They exclude a certain number of wards including intensive-care-unit (ICU) wards, isolation wards, high-dependence wards, pediatric wards, and obstetrics and gynaecology (OG) wards. All exclusions are explained in Section 2.1.3.



(a) Admission sources and the daily admission rates

(b) Patient distribution based on medical diagnosis

Figure 3: **Four admission sources to general wards and nine patient specialties.** Daily admission rates and patient distributions are estimated from data between January 2008 and December 2010.

2.1.1 Admission sources

We classify inpatient admissions to general wards (GWs) into four sources. They are ED-GW, ICU-GW, Elective (EL), and SDA patients. ED-GW patients are those who have completed treatments in the ED and need to be admitted into a general ward. ICU-GW patients are those patients who are initially admitted to ICU-type wards (from either ED or other external resources) and are later transferred to general wards. Most of the Elective (EL) and same-day-admission (SDA) patients come to the hospital to receive surgeries. They are admitted via referrals from clinical physicians, and usually have less urgent medical conditions than ED-GW or ICU-GW patients. Figure 3a shows the four admission sources and their average daily admission rates. Patients admitted to general wards from any of the four sources are called *general patients*.

ED-GW patients and their waiting time performance

The ED of NUH provides treatment to patients in need of urgent medical care, and determine the timely transition to the next stage of definitive care, if necessary. Of the 310519 patients who visit ED between January 2008 and December 2010 (from

either ambulance or walk-in arrivals), 61018 (19.7%) patients are admitted to the GWs and become ED-GW patients. 213078 (68.6%) patients are treated and directly discharged from ED because of death, absconded, admission no show, transferred to other hospital, followed up at Specialist Outpatient Clinic (SOC), Primary Health Care (PHC), General Practitioner (GP), and discharges to home. The remainder are admitted to an ICU-type (ICU, isolation, or high-dependency) ward (12163 patients, 3.9%) for further medical care, or to the EDTU (10180 patients, 3.3%) for further observation, or to other wards such as the Endoscopy ward.

Recall that we define the *waiting time* of an ED-GW patient as the duration between her bed-request time and actual admission time. The average waiting time for all ED-GW patients is 2.82 hours (169 minutes) for Period 1, and 2.77 hours (166 minutes) for Period 2. In addition to the average waiting times, we consider the *x-hour service level*, denoted by $f(W \geq x)$, that is defined as the fraction of ED-GW patients who wait x hour or longer. Here, W denotes the waiting time of a typical ED-GW patient. The overall 6-hour service level is 6.52% in Period 1 and 5.13% in Period 2. Table 1 also reports the 4-, 8-, and 10-hour service levels in the two periods. Note that the 8- and 10-hour service levels show more significant improvement in Period 2 than the average waiting time.

Our definition of waiting time is consistent with the convention in the medical literature [146, 139], except that we use the admission time to wards as the end point of the waiting period while literature usually use the time when the patient exits ED. Thus, our reported waiting time in this thesis is a slight overestimation of the value computed in the conventional way. (The gap between patient exiting ED and admission to ward is about 18 minutes on average at NUH.) Table 1 shows the waiting time statistics calculated in both ways. In Appendix A, we will report hourly waiting time statistics calculated in the conventional way as well as more distributional statistics for the waiting time.

Table 1: **The average waiting time ($\bar{W}$) and x-hour service levels ($f(W \geq x)$) for ED-GW patients.** We demonstrate the waiting time statistics calculated in two ways. One is using the duration between bed-request time and time of admission to wards, and the other method is using the duration between bed-request time and time of exiting ED. We use the former way to report waiting time statistics in this thesis, while the latter way is often used in medical literature [146, 139]. Note that the sample size differs in the two periods. This is because Period 1 contains 18 months whereas Period 2 contains 12 months. The average monthly number of bed requests is 1970 and 2107 for Period 1 and 2, respectively.

	Period 1		Period 2	
sample size	35452		25285	
	use admit. time	use ED-exit time	use admit. time	use ED-exit time
$\bar{W}$ (hour)	2.82	2.52	2.77	2.46
$f(W \geq 4)$	18.91%	15.73%	18.56%	15.15%
$f(W \geq 6)$	6.52%	5.34%	5.13%	3.97%
$f(W \geq 8)$	2.30%	1.90%	1.26%	0.86%
$f(W \geq 10)$	0.98%	0.79%	0.22%	0.09%

Elective patients

Most of the Elective (EL) patients come to NUH to receive surgeries, and they are admitted at least one day prior to surgery. The daily number of admissions from EL patients are pre-scheduled (with an average of 18.5 patients per day). The beds for these scheduled patients are usually reserved so that patients need not wait for their beds when they arrive at the hospital. Moreover, the arrival times of EL patients (the time when presenting at wards) are also scheduled as the patients are typically advised to come in the afternoon. As a result, there is no meaningful time stamp for EL patient's bed-request time.

ICU-GW and SDA patients

ICU-GW and SDA patients sometimes are also referred as internal transfer patients since they are initially admitted to a non-general ward and then transferred to a general ward. Of the 13988 patients initially admitted to ICU-type wards (from either ED or other admission sources) between 2008 and 2010, 8282 (59.2%) of them

transfer to GWs later. The remaining patients are discharged directly from an ICU-type ward.

Same-day-admission (SDA) patients first go to the operating rooms for surgical procedures, usually in the morning, occupy a temporary bed until recovery, and are finally admitted to a GW. An SDA patient is similar to an EL patient except that the EL patient is admitted into a GW *before* the day of surgery, whereas the SDA patient is admitted to a GW *after* the surgery. Therefore, it is expected that an EL patient typically stays in a general ward bed at least one day longer than an SDA patient.

For an ICU-GW or a SDA patient, although there is a delay between the bed-request time and the departure time from the ward she currently stays, this waiting time is taken less seriously than that of ED-GW patients. This claim is supported by our empirical observations that the average waiting time is more than 7 hours for ICU-GW patients and about 3.5 hours for SDA patients, both longer than that of ED-GW patients. The major reason could be (a) the ICU-GW and SDA patients have been receiving care at the current ward, thus this waiting time is not an issue unless there is a bed shortage in ICU-type wards or the SDA ward; (b) the Ministry of Health (MOH) of Singapore does not monitor this performance measure, so the NUH has less incentive to improve it than the waiting time statistics for ED-GW patients.

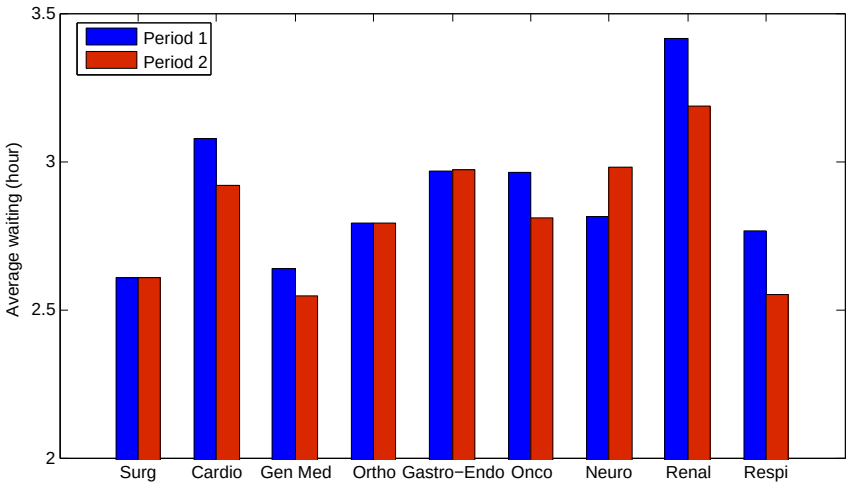
Besides the four admission sources we described above, there are a few patients (around 2.5% of the total admissions to GWs) who are admitted to general wards from other sources. For example, some patients are transferred from EDTU or Endoscopy ward to a GW. In our empirical study, we lump these patients into the SDA admission source due to their similar admission patterns and length of stay (LOS) distributions. In Figure 3a, the daily admission rate for "SDA patients" already includes these patients.

2.1.2 Medical specialties

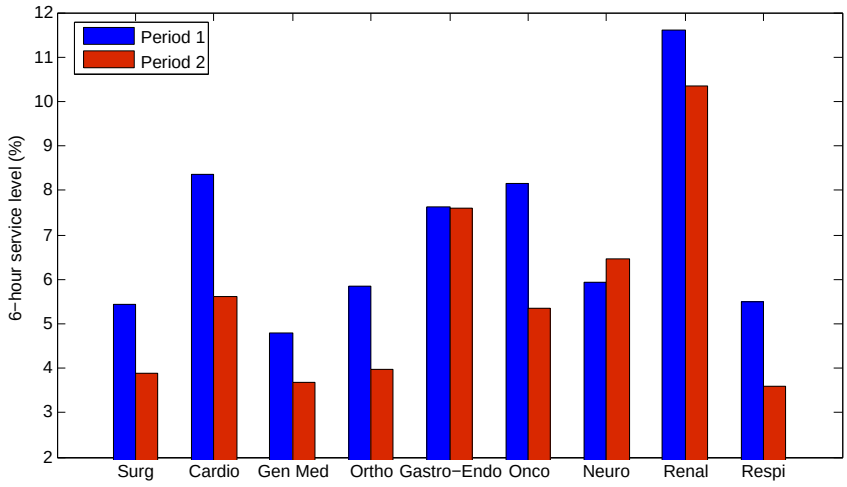
General patients are classified by one of nine medical specialities based on diagnosis at time of admission as an inpatient: Surgery, Cardiology, Orthopedic, Oncology, General Medicine, Neurology, Renal Disease, Respiratory, and Gastroenterology-Endocrine. Although Gastroenterology and Endocrine are two different medical specialties, we group them together and denote as Gastroenterology-Endocrine (Gastro-Endo or Gastro for short). The grouping is based on the fact that patients from these two specialties share the same ward and have similar LOS distributions. See [144] for the same classification. We group Dental, Eye, and ENT patients into Surgery for similar reasons. As explained in Section 2.4 of [136], two other specialties, Obstetrics and Gynaecology (OG) and Paediatrics are excluded from our study.

Figure 3b plots the distribution of general patients among different specialties and admission sources. Different specialties show very different admission-source distributions. For example, the majority of General Medicine patients are admitted from ED, while a significant proportion of Surgery patients are EL and SDA patients.

Figures 4a and 4b plot the average waiting time and 6-hour service level for ED-GW patients from each specialty in the two periods of study. Renal patients show the longest average waiting time, and their 6-hour service level is more than 10% in both periods. Surgery, General Medicine and Respiratory patients have better performances on the waiting time statistics than other specialties. Comparing the two periods, the average waiting time remains similar for each specialty, but the 6-hour service levels show a more significant reduction in Period 2 for most specialties, especially for Cardiology and Oncology. These observations suggest that the small fraction of patients with long waiting times benefit more in Period 2 than other patients.



(a) Average waiting times



(b) 6-hour service level

Figure 4: **Waiting times statistics for each medical specialty.**

2.1.3 Rationales for excluding certain wards

The entire inpatient department of NUH has 38 wards in total. We exclude 13 special care units from our defined general wards, i.e., 5 ICU wards, 5 high-dependency units (HD), 2 isolation units (ISO), and a delivery ward. It is because these wards are dedicated to patients with special needs and therefore have different performance expectations from GWs. We call ICU, HD, and ISO wards *ICU-type wards*. We consider the interface between GW and ICU-type wards through ICU-GW patients.

We exclude four Pediatric wards because they act independently from the rest of the hospital. The hospital rarely assigns an adult inpatient to a Pediatric ward (1% incidence), and Pediatric inpatients rarely stays in adult wards (0.8% incidence). Moreover, the hospital has a dedicated children's emergency department with its own admission process and a Pediatric intensive care unit (PICU) for critically ill newborns and children. Thus, Pediatric patients have few interactions with adult patients, and their performances are not the focus of our study.

Finally, we exclude two OG wards for a similar reason. Less than 1% OG patients stay in non-OG wards, and less than 0.5% non-OG adult patients are admitted to OG wards. Moreover, OG patients have very different admission patterns from other adult patients. Most of them come to deliver babies, so they go to the delivery ward or SDA ward first, and then transfer to OG wards; a few of them are directly admitted from ED. Their length of stay (LOS) in the hospital is also significantly shorter than other patients.

In summary, we focus on the remaining 19 general wards in the empirical study. We refer inpatient beds in these 19 GWs as *general beds*. The 19 GWs are designated to serve patients from different medical specialties, and we will give out more details in Section 2.3.2.

2.1.4 Data set

We obtain four raw data sets from NUH, i.e., admission data, discharge data, emergency attendance data and internal transfer data. Each of the data sets contains data entries from January 1, 2008 to December 31, 2010. We combine the four data sets into one *merged data set* using patient ID and case number as identifiers. Each record in the merged data contains a patient's entire inpatient care history and the following information:

1. The admission related information includes patient gender and age, admission date and time, allocated ward and bed number, and medical specialties.

2. The discharge related information includes patient discharge date and time, discharge ward number, and diagnostic code.

3. Based on whether there is a matched case ID in the ED attendance data, we classify each patient record as "visited ED" or "No visit to ED". For a patient who have visited ED, the ED attendance related information includes "Trauma Start" time (time of inpatient bed request) and "Trauma End" time (time of leaving ED).

4. Based on whether there is a matched case ID in the internal transfer data, we classify each patient record as "having been transferred" or "no transfer". For a patient who has gone through at least one transfer, the transfer related information includes his/her transfer frequency, transfer in and out time for each transfer, and target ward and bed in each transfer.

The merged data covers from January 1, 2008 to December 31, 2010. During our empirical study, we exclude 6-month data, from July 1, 2009 to December 31, 2009, which is when the early discharge campaign was implemented at NUH. Thus, the data set is separated into two periods. Period 1 is from January 1, 2008 to June 30,

2009, and Period 2 is from January 1, 2010 to December 31, 2010. Period 1 is one and half year long (547 days) and Period 2 is one year long (365 days). In the rest of this chapter, we will compare a number of performance measures between Periods 1 and 2. When there is no need to separate the data, the *combined data set*, which combines the data from these two periods, is used. In Section 2.2, we will further explain the reason of excluding the 6-month data from our empirical study.

An extra data set on bed request information

To better understand the delay during the ED to wards transfer process, we obtain an extra data set which contains detailed bed-request information. In this data set, each entry represents a bed-request that is processed by the bed management unit (BMU) at NUH, and the patient associated with the bed request can be from various sources, e.g., an ED-GW or an ICU-GW patient requesting a GW bed, or a patient requesting to be transferred from one GW to another GW. Each entry contains the following time stamps:

(a) Bed-request time: the date and time when the bed-request is submitted to the BMU;

(b) Bed-allocation time: the date and time when a bed is allocated for the requesting patient;

(c) Bed-confirmation time: the date and time when the allocated bed is confirmed by nurses in the current unit (e.g., confirmed by ED nurses if the requesting patient is an ED-GW patient);

(d) Request completion time: the date and time when the requesting patient is admitted to the allocated bed and the bed-request is completed.

This bed-request data set was extracted from an external IT system that is different from the NUH's system where we obtain the other four raw data sets. Due to

resource constraints, we only obtained bed-request data from June 1, 2008 to December 31, 2008, and June 1, 2010 to December 31, 2010 (14 months in total). Through patient ID and case number, we are able to link this 14-month data set with the merged data set.

2.2 Early discharge campaign

From July 2009 to December 2009, NUH started a campaign to discharge more patients before noon. This early discharge campaign gathered momentum and by December 2009, a new and stable discharge distribution emerged. In Section 2.2.1, we show more empirical statistics for the discharge distributions in Periods 1 and 2. In Section 2.2.2, we describe the measures that NUH introduced in the second half of 2009 to achieve the new discharge distribution. We also explain the reason for choosing Period 1 and Period 2 data in our empirical study. In Section 2.2.3, we discuss the changes in the operating environment between 2008 and 2010 and why they limit us from using the empirical comparison of performance measures between Periods 1 and 2 to directly evaluate the impact of early discharge policy.

2.2.1 Discharge distributions in Periods 1 and 2

Figure 2 plots the hourly discharge distributions in the two periods. Table 2 lists the corresponding numbers for the two discharge distributions. In Period 1, 12.7% of the patients are discharged before noon, and there is a single discharge peak between 2pm and 3pm. In Period 2, 26.1% of the patients are discharged before noon, more than double the percentage in Period 1. It is evident from Figure 2 that there is a new discharge peak between 11am to 12pm in Period 2. In terms of the number of patients, in Period 1, as many as 26.3 patients are discharged per hour during the peak time (2-3pm). In Period 2, the peak number of discharges is reduced to 21 patients between 11am and 12pm, and the average number of patients discharged in the original peak hour (2-3pm) is reduced to 18.7 patients. The average discharge

Table 2: **Discharge time distributions from general wards:** Period 1: January 1, 2008 to June 30, 2009; Period 2: January 1, 2010 to December 31, 2010.

Dis. time	Period 1	Period 2
0-1	0.15%	0.12%
1-2	0.15%	0.15%
2-3	0.11%	0.11%
3-4	0.09%	0.11%
4-5	0.10%	0.08%
5-6	0.11%	0.11%
6-7	0.15%	0.12%
7-8	0.07%	0.08%
8-9	0.16%	0.16%
9-10	1.32%	1.68%
10-11	3.69%	5.35%
11-12	6.55%	17.99%
12-13	9.77%	10.75%
13-14	19.39%	15.91%
14-15	25.74%	16.17%
15-16	10.56%	9.49%
16-17	6.08%	6.49%
17-18	4.46%	4.74%
18-19	3.68%	3.36%
19-20	3.24%	3.34%
20-21	2.55%	2.22%
21-22	1.06%	0.85%
22-23	0.47%	0.37%
23-24	0.32%	0.22%

hour is moved from 14.6 to 14.1, a half-hour earlier. These statistics indicate that NUH has obtained a satisfactory compliance rate in discharging more patients before noon in Period 2.

2.2.2 Implementation of the early discharge policy

The discharge process at NUH is similar to many other hospitals [63, 4, 138]. Discharge planning usually begins a day or two prior to the anticipated discharge date. On the day of discharge, the attending physician makes the morning round, confirms the patient's condition, and writes the discharge order. The nurses document the order and prepare the patient for discharge. Finally, pharmacy delivers discharge

medication if needed. Obviously, a variety of factors can affect the actual discharge time, such as when the doctor performs the rounds, when pharmacy delivers the medication, and transportation arrangements to send the patient home or to step-down facilities.

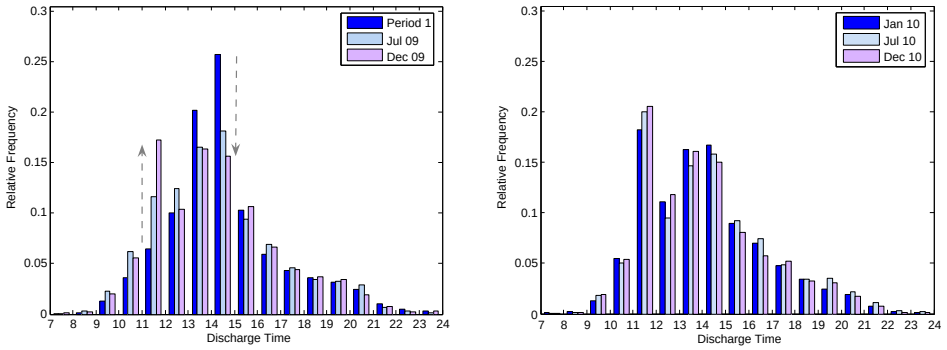
To expedite the discharge process and have more patients discharge before noon, NUH began an early discharge campaign from July 2009. The campaign initially started with a small number of wards, and was later expanded to the entire inpatient department. By December 2009, the early discharge was completely in effect. Hospital managers have worked closely with physicians, nurses, and patients to promote the campaign. Some of the initiatives include:

(i) Two discharge rounds: physicians in some specialties do two discharge rounds per day (instead one morning round). They try to finish the first round before 10 am, so that some patients can leave before 12 noon. The second round begins at about 2-3pm, and more patients can be discharged in late afternoon.

(ii) Discharge lounges: a few discharge lounges are added to several wards. Patients waiting for medicines or transportation can wait in the lounge instead of occupying hospital beds.

(iii) Day-minus-1-discharge plans: physician and nurses identify discharge needs as early as possible and prioritize tests (or other clearance) accordingly. Nurses begin to prepare discharge documents and medicine before the day of discharge.

The early discharge policy was not only costly to implement, but also required time to attain a high rate of compliance. Indeed, we observe a "stabilizing" process in the discharge patterns when the new policy was being implemented in NUH. Figure 5a compares the Period 1 discharge distribution with the distributions for July and December 2009. As early as May 2009, the peak discharge value decreases from 25.7% to 20.0% comparing to other months in Period 1, while more patients are discharged



(a) Distributions in Period 1, and July and December of 2009

(b) Distributions in Jan, July, and Dec of 2010

Figure 5: **Discharge time distributions during and after the implementation of early discharge policy at NUH.**

between 11am and noon. However, at that time there is no explicit second peak in the discharge distribution. From May through September, the value of the original peak (between 2 and 3pm) keeps decreasing, and the proportion of patients discharged between 11am and noon keeps increasing. Till September 2009, a new peak between 11am and noon with a peak value of 14.4% emerge. In December 2009, the new peak is even higher (peak value 17.3%) than the 2-3pm peak value (16.4%). Figure 5b compares the discharge distributions in some selected months of 2010. We can see the distribution stabilizes in 2010. The above observations explain why we choose Periods 1 and 2, since they correspond to before and after implementation of the early discharge policy. We exclude July through December 2009 to avoid potential bias resulting from discharge distribution instability.

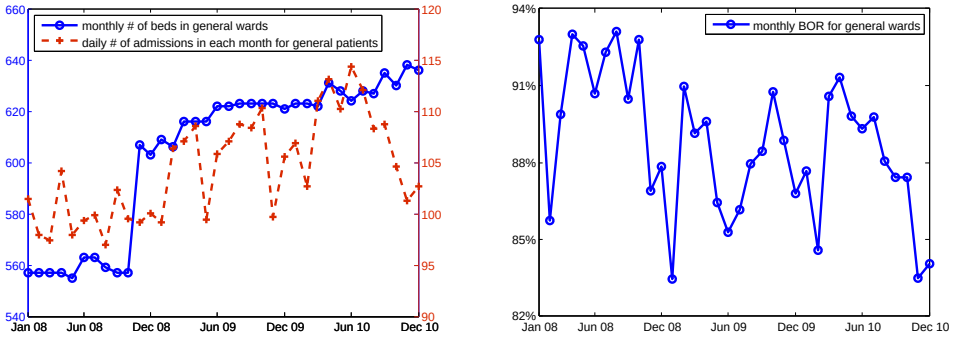
As mentioned in the previous chapter, early discharge policy has been recommended by many previous studies [9, 154] and government agencies [39]. However, few hospitals have reported to implement the policy with any success. For example, studies mention "limited success in achieving discharges by noon" in certain hospitals [142, 154], or that the policy was only experimented in a few wards [39]. Several hospitals claim that they have implemented or tried to implement the early discharge policy [154, 39, 128, 80, 147, 132], but its impact on hospital performances has not

been well documented. To our best knowledge, NUH is one of the first few hospitals that have successfully implemented the early discharge policy in the *entire* hospital and achieved satisfactory compliance rate as of December 2009.

2.2.3 The changing operating environment

Although NUH has successfully implemented early discharge in 2010 and high-fidelity data is available for us to empirically compare the performance measures before and after the implementation of early discharge, we note that pure empirical comparisons cannot fully quantify the effectiveness of the early discharge policy due to changes in the operating environment.

As in many hospitals, the operating environment is continuously changing at NUH. The number of admitted general patients has been increasing from 2008 to 2010 (the total numbers of admissions to GWs are 36473 in 2008, 38509 in 2009, and 39429 in 2010). To meet the increasing demand, NUH has increased general bed capacity over the three years. Figure 6a plots the daily admission rate of each month (the red curve) from January 2008 to December 2010. The blue curve in Figure 6a plots the monthly average number of beds in GWs. As a result, we observe a change in the bed occupancy rate (BOR). BOR is a key performance measure which reflects the utilization of beds in a specified period (see the end of this section for a rigorous definition). Figure 6b plots the monthly BORs of the GWs from 2008 to 2010. The average BOR is 90.3% in Period 1, and 87.6% in Period 2. Period 2 has a 2.7% reduction of BOR. From queueing theory, we know that reduced bed utilization can lead to a reduction in waiting time. Thus, even the waiting time is reduced in Period 2, we cannot conclude that the reduction is purely from implementing the early discharge policy. Therefore, we need a high fidelity data to capture inpatient operations and evaluate the impact of early discharge and other operational policies on system's performance. This will be the main focus of Chapter 3.



(a) Monthly admission rate and number of general beds

(b) Monthly BOR

Figure 6: **Monthly admission rate, number of beds, and BOR.** The two figures show the monthly admission rate from general patients, the monthly average number of general beds, and monthly BOR from January 2008 to December 2010.

Definition for BOR: BOR is always defined for a specific group of beds. The group can be all beds in a ward or all beds in all general wards. In this thesis, our default group is all beds in all general wards if no group is specified. For a given group of beds and a given period, BOR is defined as (see Page 10-11 of [114]):

$$\text{BOR} = \frac{\text{Total Inpatient Days of Care}}{\text{Total Bed Days Available}} \times 100, \tag{1}$$

where the *total inpatient days of care* equals the sum of patient days among all patients who have used a bed in the group in the specified period, and patient days of a patient equals the number of days within the period that a patient occupies any bed in the group. Patient days of a patient is almost equal to patient length of stay (LOS; see Section 2.5), except that the patient day of a same-day discharge patient is 1, while the LOS equals 0; also LOS may include days outside the given time period. *Total bed days available* is equal to the sum of *bed days available* among all beds in the group, where bed days available of a bed is the number of days within the time period that bed is available to be used for patients. Note that BOR is a slightly different concept from bed utilization, since we use service time (not the integer-valued patient days) to calculate utilization. From our NUH data, BOR is slightly higher than the

corresponding utilization (for all beds and for most wards), but the two values are very close, typically differing by only 1% to 2%. Thus, we focus on reporting BOR in the empirical study.

2.3 Ward capacity and overflow proportion

In this section, we report ward-level statistics for the 19 GWs. In particular, we introduce an important performance measure, the *overflow proportion*. We first given an overview of NUH's ward setting in Section 2.3.1. Then in Section 2.3.2, we report the ward-level BOR. In Section 2.3.3, we report the overflow proportion from both ward and specialty levels. Finally in Section 2.3.4, we present some supplementary statistics for *shared wards* which serve patients from multiple specialties.

2.3.1 Basic ward setting in NUH

In NUH, each GW contains a number of beds in close proximity. The wards are relatively independent of each other, with each having its dedicated nurses, cleaning team and other staff members. There are usually multiple rooms in each ward. A room is equipped with 1 to 8 beds, depending on the ward "class", and is shared by patients of the same gender. In general, class C wards have 8 beds per room, class B wards 4 or 6 beds per room, and class A wards 1 or 2 beds per room (see details in [111]). Stays in class B2 or C wards are eligible for heavy subsidy from the government, thus the daily expenses in these subsidized wards are much less than the expenses in class A or B1 wards. As a result, there is a much greater demand for the subsidized wards. Table 3 lists the number of beds in each of the 19 wards.

Physicians always prefer to have their patients stay in the same wards to save rounding time. Hospital can also achieve a better match between patient needs and nurse competencies by doing so. Therefore, NUH designates each general ward to serve patients from only one or two (rarely three) specialties. We call the ward's designated specialty its *primary* specialty. Table 3 lists the primary specialties for

the 19 wards.

Note that around September to December 2008, NUH changed the primary specialties for several wards to better match the demand and supply of bed for each specialty, a reaction to the big capacity increase in late 2008 (see Figure 6a). Since most of our reported statistics in this section relate to the ward primary specialties, we exclude the period before the re-designated specialties became operational for consistency. The term "reduced Period 1", therefore, refers to the remaining time in Period 1 after the re-designated specialties took effect. The start time for the reduced Period 1 for each affected ward depends on the time of specialty re-designation; the end time is fixed at June 30, 2009. Thus, the duration of the reduced Period 1 may differ for each ward, since the re-designated specialty could take effect at different times. Table 3 lists the start month of (reduced) Period 1 for each ward. For example, Ward 52 was re-designated as an Orthopedic ward from November 2008, and Ward 54 a Surgery/Orthopedic ward from March 2009.

For wards with no changes in their primary specialties, we use data points from the entire Period 1 to calculate ward-level statistics; otherwise, we use the reduced Period 1. We calculate ward-level statistics for Period 2 using data points from the entire Period 2, because no speciality re-designation occurred.

2.3.2 Capacity and BOR

Figure 6b plots the monthly BOR for all general wards from January 2008 to December 2010, from which we can see the monthly BOR fluctuates between 80% and 95%. The average BOR for all GWs is 90.3% for Period 1 and 87.6% for Period 2. In fact, if we exclude January to October 2008, the average BOR for the remaining Period 1 is about 87.4%, which is similar to Period 2. This suggests that NUH has successfully increased its bed capacity, resulting in BOR stabilization despite significant increases in patient admissions from January 2008 to December 2010. The total

Table 3: **Primary specialties and BOR for the 19 general wards.** The start time of Period 1, if not January 2008, corresponds to when the re-designated specialties took effect for wards having changed the primary specialties.

Ward	Prim. specialty	# of beds		Per 1 start	BOR (%)	
		Jan 08	Dec 10		Per 1	Per 2
41	Surg, Card	44	44	Feb 09	90.9	92.0
42	GM, Respi	33	44	Nov 08	86.4	92.2
43	Surg	44	44	Jan 08	93.4	88.9
44	Respi, Surg	14	44	Mar 09	79.0	80.3
51	Ortho	39	39	Jan 08	76.7	67.5
52	Ortho	22	26	Nov 08	74.4	75.3
53	GM, Neuro	46	46	Jan 08	96.8	97.1
54	Surg, Ortho	50	50	Mar 09	80.7	77.6
55	Renal	44	33	Jan 08	91.7	86.5
56	Card	17	17	Nov 08	90.1	95.1
57	Neuro	14	14	Jan 08	97.3	96.5
57O	Onco	24	24	Jan 08	93.9	93.2
58	Onco	24	24	Jan 08	90.2	91.7
63	Card	43	44	Jan 08	95.5	96.1
64	Gastro	46	50	Jan 08	94.2	92.8
66	Med, Surg	31	34	Feb 09	86.9	86.8
76	Med, Card	18	18	Jan 08	90.0	94.6
78	Onco, Surg, Ortho	25	25	Mar 09	83.0	82.8
86	Onco	8	14	Jan 08	89.6	87.5
Total General Beds		555	638		90.3	87.6

number of general beds increased from 555 beds as of January 1, 2008 to 638 beds as of December 31, 2010.

Not surprisingly, BOR is ward dependent. Table 3 the BOR in Periods 1 and 2 for each ward. The BOR for all 19 wards are also plotted in Figure 74. We make the following observations: (i) BORs of dedicated wards (Wards 43, 56, 57, 58, 63, 64) are generally high, most exceeding 90%, with the exceptions of Orthopedic wards 51 and 52 which have much lower BORs for both periods; (ii) class A/B1 wards (Wards 66, 76, 78, 86) have lower BORs than other wards because they are not government-subsidized; (iii) Ward 44 has a much lower BOR than other Medicine wards, mainly because half of its capacity serves infectious respiratory patients who cannot share rooms with other patients; and (iv) comparing the BORs for the two periods shows no consistent pattern of increase or decrease.

2.3.3 Overflow proportion

Usually patients are assigned to their designated wards. However, when an ED-GW patient has waited for several hours in the ED, but no bed from the primary wards is available or expected to be available in the next few hours, NUH may overflow the patient to a non-primary ward as a temporary expedient. Overflow events may also occur among patients admitted from other sources, such as when ICU-type wards need to free up capacity, ICU-GW patients may be overflowed. We define the *overflow proportion* as the number of patients admitted to non-primary wards divided by the total number of admissions. The admissions here include both the initial admission and transfer to general wards, e.g., a transfer from ICU to GW is counted as a different admission in addition to the initial admission.

Obviously, there is a trade-off between patient waiting time and overflow proportion. On the one hand, the waiting time can always be reduced by overflowing patients more aggressively since overflow acts as resource pooling. On the other hand,

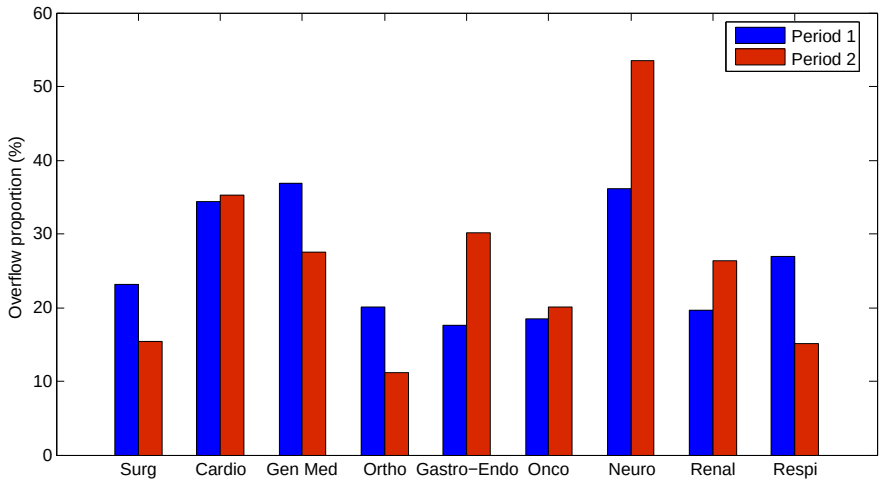


Figure 7: **Overflow proportion for each specialty in Periods 1 and 2.**

overflow decreases the quality of care delivered to patients and increases hospital operational costs [144]. In NUH, the average overflow proportion among all patients is 26.95% and 24.99% for Periods 1 and 2, respectively. The overflow proportion for all ED-GW patients is 29.91% in Period 1 and 28.54% in Period 2, slightly higher than the values for all patients. The reduction of overflow proportion in Period 2 indicates that the reduced waiting time for ED-GW patients in Period 2 does not result from a more aggressive overflow policy.

Next, we show overflow proportions on both the specialty level and ward level.

Overflow proportion for each specialty

The overflow proportion for a specialty is defined as the number of overflow admissions from this specialty divided by the total number of admissions from this specialty. Figure 7 compares the overflow proportion for each specialty in Periods 1 and 2. Note that (i) Cardiology, General Medicine, and Neurology patients have significant higher overflow proportions than other specialties, which suggests that these specialties may not have enough beds allocated to them; (ii) the overflow proportions of Surgery, General Medicine, Respiratory, and Orthopedic show significant reductions in Period 2, whereas Gastro-Endo and Neurology show a big increase in Period 2.

Overflow proportion

The overflow proportion for a ward is defined as the number of overflow admissions to this ward divided by the total number of admissions to this ward. Figure 8 compares the overflow proportions for GWs in Periods 1 and 2. Table 26 in Appendix A.3 lists the corresponding numerical values.

We observe that dedicated wards (serving only one specialty) generally have a lower overflow proportion than the shared wards (serving multiple specialties). Comparing the two periods, most of the wards show reduced overflow proportions in Period 2, with some showing significant reductions (mostly dedicated wards); some wards show a small increase. The only exceptions are Wards 44 and 52, which show significant increases in the overflow proportions.

Moreover, comparing the BOR (Table 3) and the overflow proportion for each ward, we can see it is generally true that if the ward has a lower BOR, its overflow proportion will be higher; examples are Wards 51, 52, and 54. The only exception is Ward 44, which has a low BOR and a low overflow proportion at the same time. In practice, the BMU prefers to overflow class A/B1 patients to a non-primary class A/B1 ward instead of downgrading them to a lower-class primary ward. This also explains why class A/B1 wards have higher overflow proportions than most class B2/C wards, since class A/B1 wards are "pooled" together more often.

Note that overflow proportion only takes patient count into consideration. It does not differentiate between an overflow patient with a long LOS and an overflow patient with a short LOS, where the latter is always preferred for the right-siting of care. In Section A.3 of the appendix, we introduce another statistics: the BOR share, which is the proportion of BOR contributed from primary patients and overflow patients. This statistic takes patients LOS into consideration since the BOR calculation involves LOS.

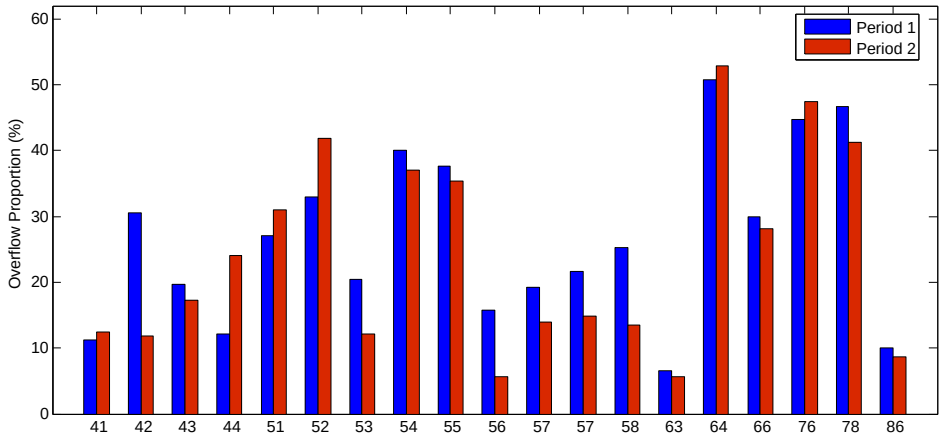


Figure 8: **Overflow proportion for each ward in Periods 1 and 2.**

2.3.4 Shared wards

Excluding class A/B1 wards, NUH has five shared wards (Ward 41, 42, 44, 53, and 54) serve two primary specialties; see Table 3. Each bed in the shared wards is nominally allocated to a certain specialty, but nurses in these wards have the flexibility to care for patients from either speciality.

For each of the shared wards and for each period, we calculate (i) the ratio between the BORs of the two primary specialties and (ii) the ratio between their admission numbers. We compare these two ratios with the nominal capacity allocation. Table 4 lists these three sets of statistics (in Columns 4-5, 6-7, 8, respectively). First, we can see that the ratios of the BORs and the ratios of admission numbers are close for each ward, except for ward 44 in Period 2 and ward 54 in Period 1. The closeness indicates that the average LOS of the two primary specialties are close. Second, we can see that the ratios in Columns 4-7 are mostly above 80%, and generally exceed the ratios of the nominal bed allocation (last column). This indicates that each ward is predominantly used by patients from one certain specialty, regardless of the nominal allocation.

Table 4: **Bed allocation in shared wards.** The ratio of BOR is defined as the BOR from Prim.1 specialty divided by the sum of BORs from its primary specialties. The ratio of admissions or the ratio of allocated beds is defined similarly by just changing BOR to the number of admissions or the number of allocated beds, respectively. The ratios of allocated beds are estimated from the average number of beds in both periods; the nominal bed allocation is unknown for Ward 53.

Ward	Specialty		Ratio of BOR		Ratio of admissions		Ratio of alloc beds
	Prim. 1	Prim. 2	per 1	per 2	per 1	per 2	
41	Surg	Card	81.45	81.10	81.97	80.27	72.09
42	Gen Med	Respi	94.93	95.66	92.34	93.94	77.27
44	Respi	Surg	72.62	69.26	67.48	59.39	53.33
53	Gen Med	Neuro	86.49	93.50	82.09	89.50	unknown
54	Ortho	Surg	86.53	84.90	73.83	80.31	66.67

2.4 Bed-request process

In this section, we study the bed-request processes from the four admission sources with a focus on the bed-request process from ED-GW patients. In Section 2.4.1, we show the hourly bed-request pattern of ED-GW patients and its connection with the arrival process to the emergency department. In Section 2.4.2, we test whether the bed-request process from ED-GW patients follows a non-homogeneous Poisson process. Finally, in Section 2.4.3 we study the bed-request processes from the other three admission sources.

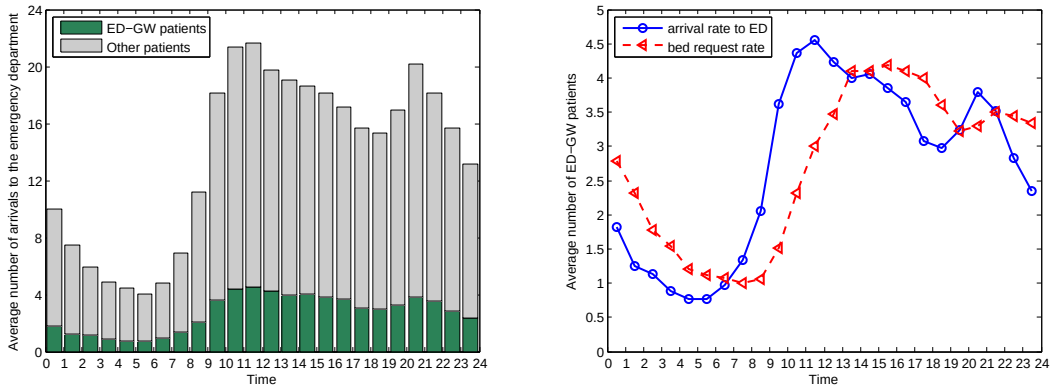
2.4.1 Bed-request rate from ED-GW patients and Arrival rate to ED

Recall that *bed-request time* for an ED-GW patient is when ED physicians decide to admit and request an inpatient bed for this patient (the patient has finished treatment in ED); it corresponds to the Trauma Start time in our data set. Only about 20% of the arrivals to ED at NUH are admitted to the GWs and become ED-GW patients. Figure 9a plots the hourly arrival rate to ED in Period 1. The green bars represent the hourly arrival rates from patients who will eventually be admitted to a general ward (i.e., ED-GW patients). The grey bars represent the arrival rates from all other patients, who will be directly discharged from the ED or admitted to other wards.

From the figure, we can see the total hourly arrival rates from all patients (sum of green and grey bars) begins to increase from 7 am, followed by two peaks: a peak between 11am and noon (21.7 per hour) and a peak between 8pm and 9pm (20.2 per hour). This pattern is similar to those observed in many hospitals of other countries (e.g., see Figure 1 of [61] and Figure 2 of [157]), indicating that the arrival rate pattern to NUH's emergency department is not unique. Moreover, Figure 9a shows that the proportion of the green and grey bars does not change much throughout the day. About 17% to 22% of patients arriving at the ED become ED-GW patients in each hour, which suggests that the patient mix (ED-GW patients versus other patients) is quite stable.

Figure 9b demonstrates the connection between ED arrival rate and bed-request rate of ED-GW patients. The solid curve shows the arrival rate to ED from ED-GW patients, which is identical to the green bars in Figure 9a. The dashed curve shows the average number of bed requests from ED-GW patients during each hour. We use the term *hourly bed-request rate* to denote the number of beds requested by ED-GW patients in each hour. The bed-request rate starts to increase from 7am, and reaches three or more per hour between noon and midnight. The peak is between 1 pm and 5pm (4.2 per hour). If we compare the two curves in Figure 9b, we can see their shapes are similar and the dashed curve seems to be a horizontal shift of the solid curve. This depicts the relationship between the arrival process to ED and the bed-request process of ED-GW patients: when an ED-GW patient arrives at the emergency department, it takes about two hours to receive treatment (plus the possible waiting time) before a physician decides to admit him/her and makes a bed-request.

Figure 10 compares the hourly bed request rate from ED-GW patients among four specialties in Period 1: Medicine, Surgery, Cardiology, and Orthopedic. In this figure, we aggregate the five medical specialties belonging to the Medicine cluster (General Medicine, Neurology, Renal Disease, Respiratory, and Gastroenterology-Endocrine)



(a) Arrival rate to emergency department

(b) Arrival rate and bed request rate of ED-GW patients

Figure 9: **Hourly arrival rate to the emergency department and bed-request rate of ED-GW patients.** In subfigure (b), the arrival rate to ED is from patients who will eventually be admitted into general wards (ED-GW patients). Period 1 data is used.

into one and omit Oncology due to its small volume. This figure shows that the proportion of the specialties changes little over time, suggesting that patient-mix is stable in each hour. It is also consistent with our observation that the bed-request rate curves from each specialty have similar shapes (figures not shown here).

We use Period 1 data to plot Figures 9 and 10. Using Period 2 data show similar patterns/phenomena, while the average arrival rate and bed-request rates both increase in Period 2, since more patients visit the hospital in Period 2 (also see Section 2.2.3).

2.4.2 Testing the non-homogeneous Poisson assumption for ED-GW patients

Brown et al. [16] proposed a method to test non-homogeneous Poisson arrival processes. We apply this method to NUH data to test the bed-request process from ED-GW patients. The null hypotheses of our test is that the bed-requests of ED-GW patients form an inhomogeneous Poisson process with piecewise-constant arrival rates.

To perform the test, we follow the procedures described in [16]. First, we divide

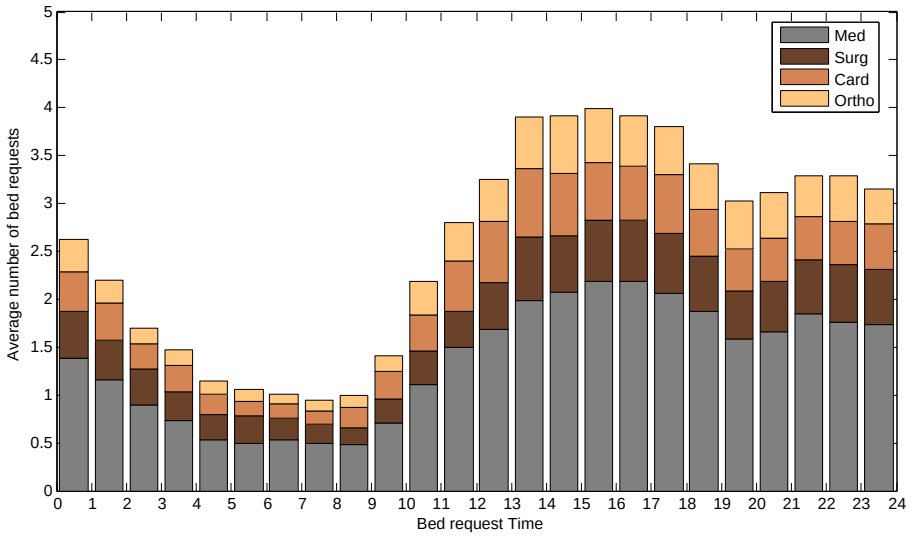
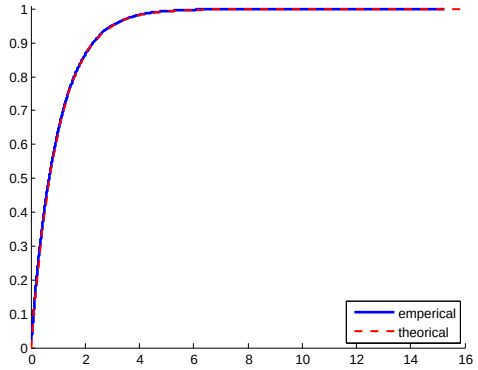


Figure 10: **Hourly bed request rate from 4 major specialties in Period 1.** The plot aggregates the five specialties belong to the Medicine cluster and omits Oncology.

each day into 7 time blocks: 0am-2am, 2am-4am, 4am-9am, 9am-11am, 11am-13pm, 13pm-18pm, and 18pm-0am. Note that we do not use blocks of equal length. We choose these blocks so that within each of them, the hourly arrival rates are close for the included hours. We call a block on a certain day a *time interval*, e.g., 2am-4am on May 1, 2008 is a time interval. The blocks we choose also ensure that we have enough data points in each time interval. Second, for each time interval i, we collect the bed-request time stamps belonging to that interval and transform the bed-request time in the same way as introduced in [16]. That is, let T_j^i denote the j^{th} ordered bed-request time in the i^{th} interval $[T_{\text{start}}^i, T_{\text{end}}^i)$, $i = 1, \ldots, I$, where I denotes the total number of intervals. Let $J(i)$ denote the total number of bed-requests in the i^{th} interval, and define $T_0^i = T_{\text{start}}^i$ and $T_{J(i)+1}^i = T_{\text{end}}^i$. We have $T_{\text{start}}^i = T_0^i \leq T_1^i \leq \cdots \leq T_{J(i)}^i < T_{J(i)+1}^i = T_{\text{end}}^i$. The transformed variable R_j^i is defined as

$$R_j^i = -\Big(J(i) + 1 - j\Big) \cdot \log\left(\frac{T_{J(i)+1}^i - T_j^i}{T_{J(i)+1}^i - T_{j-1}^i}\right), \qquad j = 1, \ldots, J(i).$$

(a) QQ plot against the exponential assumption



(b) CDF plot against the exponential assumption

Figure 11: **QQ plot and CDF plot of $\{R_j^i\}$ from all intervals in Period 1 for the bed-request process of ED-GW patients.**

Under the null hypothesis that the bed-request rate is constant within each time interval, the $\{R_j^i\}$s are independent standard (with rate 1) exponential random variables (see the derivation in [16]). Third, we aggregate the transformed values of $\{R_j^i\}$ from intervals in a certain set of days and perform the Kolmogorov-Smirnov (K-S) test on the assumption of standard exponential distribution.

The second column in Table 5 shows the K-S test results on testing the bed-request process for each month of Periods 1 and 2. That is, we aggregate $\{R_j^i\}$ from all intervals belonging to each month (there are about $7 \times 30 = 210$ time intervals in a month), and perform 30 sets of K-S test for the 30 months. We can see that at significant level of 5%, 24 null hypotheses (out of 30) are not rejected.

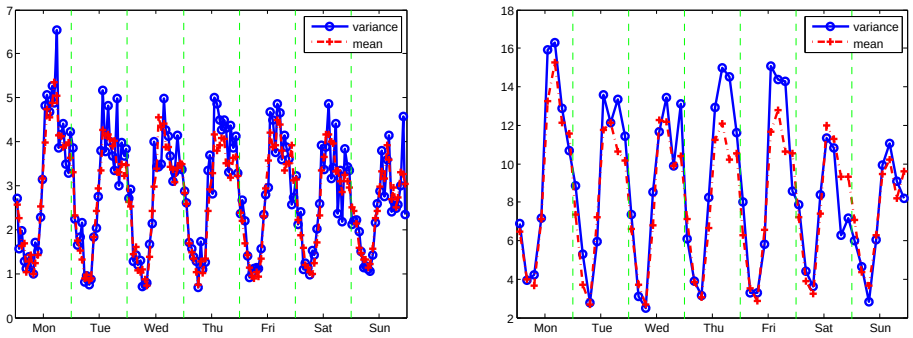
We also perform K-S tests for longer time windows, e.g., aggregating all intervals from the 18 months in Period 1. Due to the large sample sizes (more than 35000 samples in Period 1), the p-value of K-S test at significance level 5% is very close to zero, so it is difficult to pass the test. However, the Q-Q plot and CDF plot in Figure 11 show that the distribution of the transformed values $\{R_j^i\}$ from all intervals in Period 1 is still visually close to the standard exponential distribution.

Table 5: **Results for Kolmogorov-Smirnov tests on testing the non-homogeneous Poisson assumption for bed-request processes of ED-GW, SDA, and ICU-GW patients and admission process of EL patients.** A test is passed at the significance level of 5% if the reported value is larger than 0.05.

month	ED-GW	EL	SDA	ICU
2008-01	0.0134	0.0071		
2008-02	0.0638	0.0849		
2008-03	0.0479	0.1802		
2008-04	0.1842	0.0178		
2008-05	0.0062	0.0228		
2008-06	0.215	0.0002	0.0053	0.0000
2008-07	0.1028	0.1148	0	0.0001
2008-08	0.1949	0.0388	0	0.0000
2008-09	0.1064	0.0253	0.0055	0.0000
2008-10	0.1253	0.0256	0.0279	0.0003
2008-11	0.2442	0.0026	0.0001	0.0005
2008-12	0.3092	0.091	0.0098	0.0000
2009-01	0.1218	0.421		
2009-02	0.0694	0.0061		
2009-03	0.186	0.0925		
2009-04	0.1112	0.0091		
2009-05	0.0565	0.0729		
2009-06	0.018	0.1876		
2010-01	0.3259	0.0018		
2010-02	0.9596	0.5737		
2010-03	0.0851	0.0007		
2010-04	0.6379	0.0004		
2010-05	0.2684	0.0338		
2010-06	0.003	0.1048	0.2959	0.0028
2010-07	0.0065	0.4903	0	0.0000
2010-08	0.0546	0.0064	0.0103	0.0000
2010-09	0.4329	0.4402	0.0004	0.0000
2010-10	0.795	0.0472	0.0485	0.0000
2010-11	0.0563	0.0005	0.0064	0.0000
2010-12	0.1996	0.3198	0.0127	0.0000

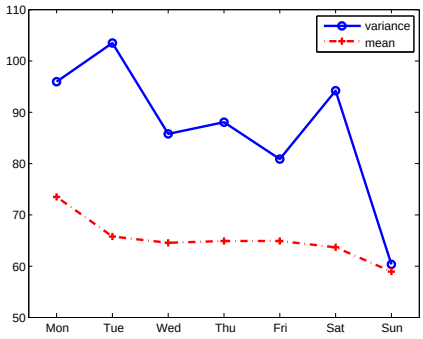
The above test results suggest that it is reasonable for us to assume the bed-request process from ED-GW patients is a non-homogeneous Poisson process with piecewise-constant arrival rates. But note that the null hypothesis in the test does not contain any assumption on the bed-request rates of different intervals being equal or having a certain relationship. In particular, the test results do not suggest that the bed-request rate function is periodic. On the contrary, we find that the bed-request process is *not* a periodic Poisson process if using one *day* or even one *week* as a period. Figures 12a and 12b clearly show that the bed-request rates depend on the day of week, so the bed-request process cannot be periodic Poisson with one day as a period. We then examine whether the bed-request process is periodic Poisson with one week as a period. If this assumption were valid, then for each day of the week, the daily bed-request on that day in all weeks would have formed an iid sequence following a Poisson distribution. As a consequence, the mean and variance of the daily bed-request on that day of the week would be equal or close. However, Figure 12c shows that the sample variances are significantly larger than the sample means for each day of the week except for Sunday, which indicates that the bed-request process is not a periodic Poisson process with one week as a period. We conjecture that the high variability comes from the seasonality of bed-requests (e.g., February has a lower bed-requests rate than other months; see the red curve in Figure 6a)) and the overall increasing trend in the bed demand (see discussions in Section 2.2.3).

Furthermore, Figure 12c demonstrates that, under the 1-day resolution, the bed-request process shows over-dispersion, a term that was coined in Maman [103] and means that the arrival process has "significantly larger values of the sampled CV's compared to the CV's one would expect for data generated by a Poisson distribution." Unlike the 1-day resolution case, we observe from Figures 12a and 12b that, under the 1-hour and 3-hour resolutions, the sample means and sample variances are close for most intervals. This observation is consistent with the findings in Section 3.3



(a) 1-hour resolution

(b) 3-hour resolution



(c) 24-hour resolution

Figure 12: **Comparison between sample means and sample variances of bed-requests.** Three different resolutions are used: 1 hour, 3 hours, and 24 hours. Period 1 data is used.

of [103] and suggests that variability of bed-request rates at these two resolutions is close to (or somewhat larger than) the variability of iid Poisson random variables. Note that we have differentiated among seven days in a week in Figures 12a and 12b to account for the day-of-week variations; Maman [103] did the same when testing the arrival process to ED (see Section 3.3 in her paper). If we do not differentiate, the over-dispersion phenomenon would be more prominent. Maman [103] also gave a possible explanation for the phenomenon that the difference between the empirical and Poisson CV's increases when one decreases the time resolution (see Remark 3.3 there).

2.4.3 Other admission sources

We now study the bed-request processes from SDA and ICU-GW patients and admission process from EL patients (i.e., using EL patient's admission time stamp). We study the admission process of EL patients because there is no meaningful time stamps for EL patient's bed-request time in the NUH data. Figure 13 plots the hourly bed-request rates for SDA and ICU-GW patients and the hourly admission rate for EL patients.

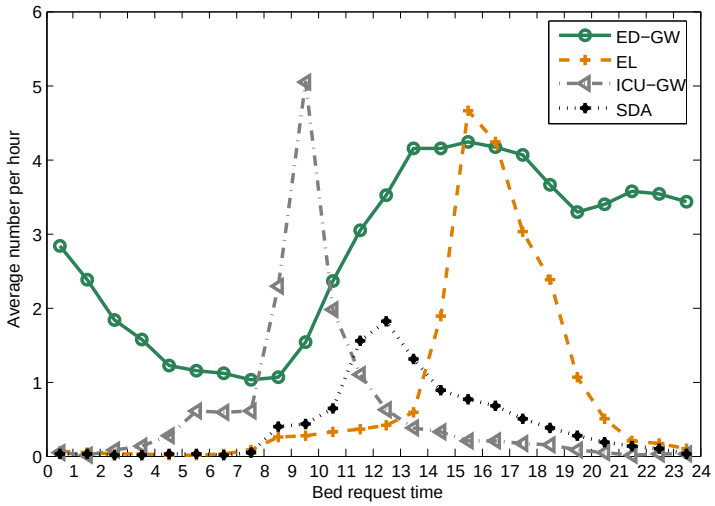


Figure 13: **Hourly bed-request rate for each admission source.** The curve for EL patients is plotted from using the admission time, so it is the hourly admission rate for EL patients. Period 1 data is used.

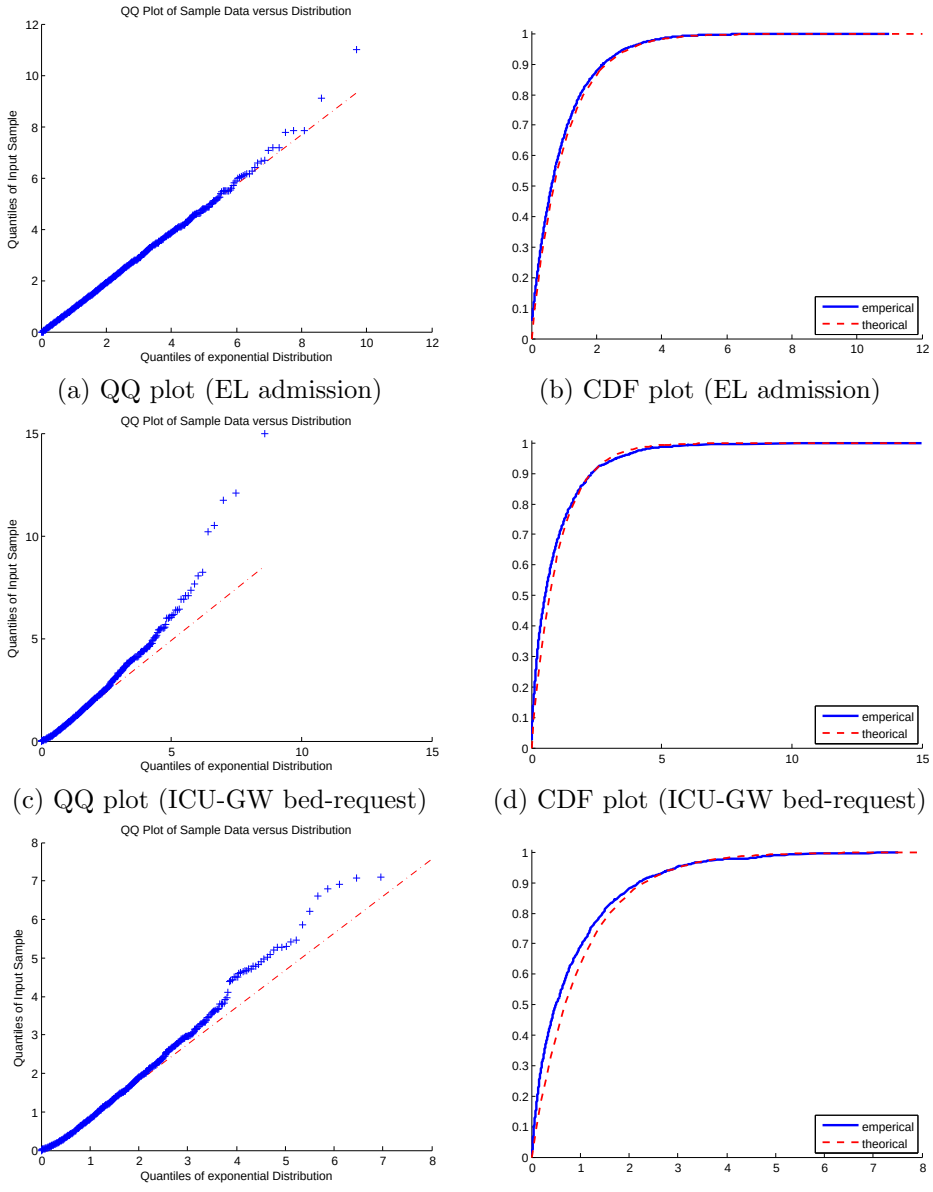
We first test the non-homogeneous Poisson assumption for the bed-request processes from SDA and ICU-GW patients and admission process from EL patients. The fourth to sixth columns of Table 5 show the K-S test results using the monthly data in Periods 1 and 2. Note that we only have 14-month data for the bed-request times of SDA and ICU-GW patients (see explanation in Section 2.1.4). Thus, the last two columns of Table 5 only display the K-S test results for these 14 months. From the table, we see at the significance level of 5%, 17 null hypotheses out of 30 are rejected for the EL admission process, and nearly all the null hypotheses are rejected for SDA and ICU-GW bed-request processes (13 and 14, out of 14, are rejected for SDA and

ICU-GW, respectively). Similar to Figure 11, Figure 14 shows the Q-Q plots and CDF plots for the transformed values $\{R_j^i\}$ for the EL admission process and the SDA and ICU-GW bed-request processes. In the figure, $\{R_j^i\}$ from all intervals in Period 1 are aggregated. We observe that the distribution of the transformed values is still visually close to the standard exponential distribution for EL admission process, but not for the other two tested processes.

Two levels of random fluctuations

A closer look at the bed-request times of ICU-GW and SDA patients reveals a *batching* phenomenon. Figure 15 plots the histogram of the inter-bed-request time between two consecutive bed-requests within the same day for ICU-GW and SDA patients. From the figures we can see that most bed-requests are less than 30 minutes away from the previous bed-request. In particular, about half of the ICU-GW inter-bed-request times are less than 10 minutes .

We talked to the NUH staff to understand the batching phenomenon and the bed-request processes of ICU-GW and SDA patients. In practice, the ICU physicians decides which patients should be transferred to general wards after the morning rounds each day, and these patients to be transferred become ICU-GW patients according to our definition. Thus, the number of bed-requests from ICU patients on a day is determined first, and then ICU nurses submit these bed-requests to BMU, usually in a batch. Similarly, the SDA surgeries each day are scheduled in advance, and the number of bed-requests from SDA patients on a day is also pre-determined. The SDA nurses submit bed-requests for SDA patients after they finish receiving surgeries on each day. In addition, we understand that the EL admission process can also be viewed as a two-step process in a similar way, although we do not observe a batching phenomenon there. The elective admissions are pre-scheduled on a daily basis, while within a day, when the elective patients arrive at the hospital and are



(a) QQ plot (EL admission)

(b) CDF plot (EL admission)

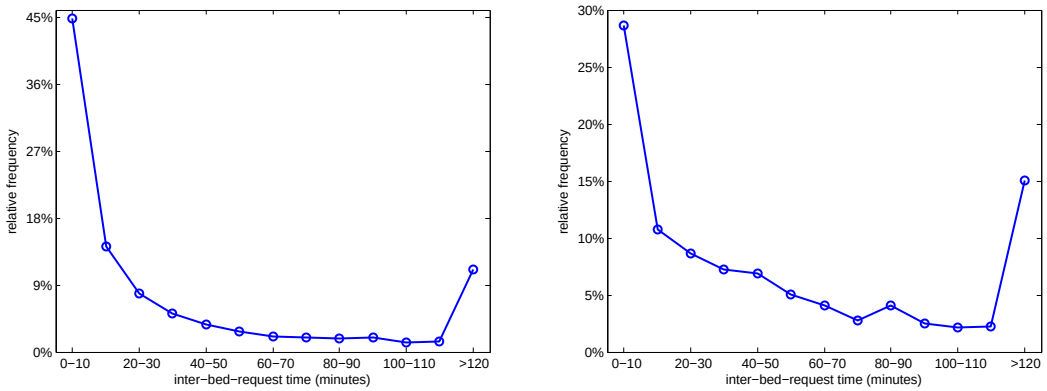
(c) QQ plot (ICU-GW bed-request)

(d) CDF plot (ICU-GW bed-request)

(e) QQ plot (SDA bed-request)

(f) CDF plot (SDA bed-request)

Figure 14: **QQ plots and CDF plots of $\{R_j^i\}$ from all intervals in Period 1 for the admission process of EL patients and the bed-request processes of ICU-GW and SDA patients.**



(a) ICU-GW bed-requests

(b) SDA bed-requests

Figure 15: Histograms of the inter-bed-request time for ICU-GW and SDA patients using the combined data. The bin size is 10 minutes.

admitted depends on the patient and staff schedules.

Thus, there are two levels of randomness in the bed-request processes from ICU-GW and SDA patients and in the EL admission process: (i) the number of bed-requests or admissions each day, and (ii) when nurses submit bed-requests or when (EL) patients are admitted within a day. Figure 16 plots the empirical distributions of the daily number of bed-requests from ICU-GW and SDA patients and the daily number of admissions from EL patients. From the figure, we see a two-peak shape in the distributions of EL and SDA patients. The reason is that elective and SDA surgeries are usually performed on weekdays, and few EL and SDA patients are admitted on weekends. After we plot the daily number of admissions or bed-requests for EL and SDA patients on weekdays and weekends separately, the two-peak shape no longer appears.

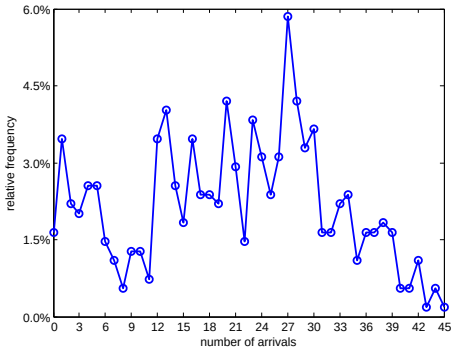
For the second level of randomness, the empirical distributions of the admission times for EL patients and bed-request times for ICU-GW and SDA patients can be calculated from the corresponding hourly admission rate or bed-request rate in Figure 13. In addition, we plot in Figure 17 the histogram of the first admission time each day for EL patients and the first bed-request time each day for ICU-GW

and SDA patients. We can see that nurses usually submit the bed-requests for ICU-GW and SDA patients in the morning, while most EL patients are admitted in the afternoon.

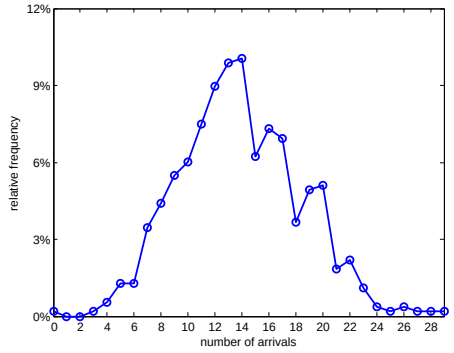
2.5 Length of Stay

In this thesis, the length of stay (LOS) of an inpatient is defined as the number of *nights* the patient stays in the hospital, or equivalently, day of discharge minus day of admission. In this section, we present empirical statistics of LOS in the two periods. We first show the LOS distributions for all patients in Section 2.5.1. In Sections 2.5.2 and 2.5.3, we demonstrate that the LOS distribution depends on patient admission source, speciality, and admission time. In Section 2.5.4, we compare the average LOS between overflow patients and right-siting patients. Finally, in Section 2.5.5, we test the iid assumptions among patient LOS.

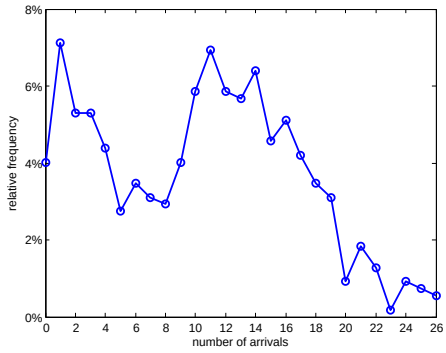
We want to emphasize three points before starting the subsections. First, LOS is a different concept from *service time*, which refers to the duration between patient admission time and discharge time. A patient's LOS takes only *integer* values, while service time can take any real values. But LOS constitutes the majority of service time, and the difference between the two is usually a few hours only. In Chapter 3, we will show that a critical feature for our proposed stochastic networks is the new service time model, in which a patient's service time is no longer modeled as an exogenous iid random variable, but an endogenous variable depending on LOS and other factors. Second, our definition of LOS is consistent with the definition adopted by most hospitals and the medical literature, except for same-day discharge patients. We assume the LOS of same-day discharge patients is zero, while most hospitals adjust their LOS to be 1 for billing purposes (see, for example, the National Hospital Discharge Survey [68, 28]). Third, for all the reported statistics in this section, we include in the samples only patients who did *not* transfer to ICU-type wards after



(a) EL admissions

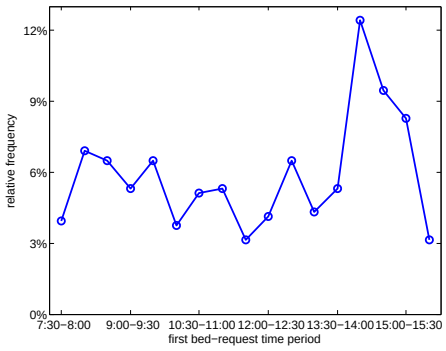


(b) ICU-GW bed-requests

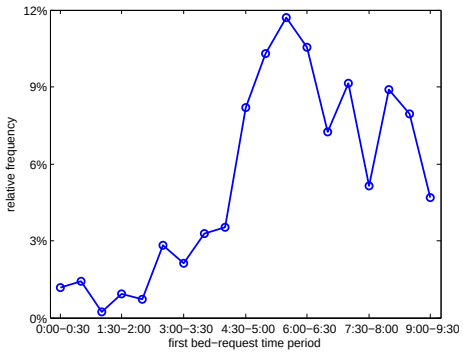


(c) SDA bed-requests

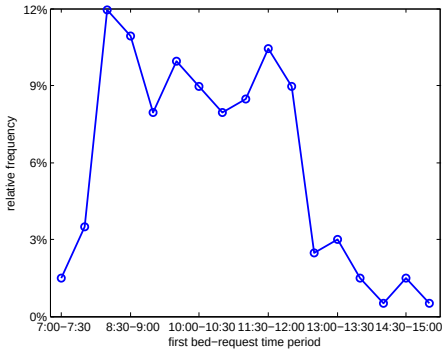
Figure 16: **Histograms of the daily number of admissions for EL patients and daily number of bed-requests for ICU-GW and SDA patients using Period 1 data.**



(a) EL admissions



(b) ICU-GW bed-requests



(c) SDA bed-requests

Figure 17: **Histograms of the first admission time each day for EL patients and the first bed-request time each day for ICU-GW and SDA patients.** Period 1 data is used. The bin size is 30 minutes.

their initial admission to GWs. Transfer patients have different LOS distributions, and we will discuss them in Section 2.8.2.

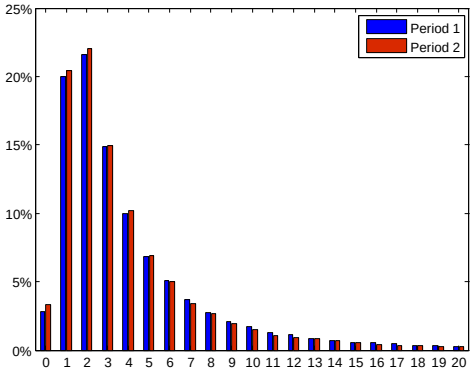
2.5.1 LOS Distribution

Figure 18a plots the LOS distributions in two periods with the cut-off value at 30 days. The means (without truncation) for Periods 1 and 2 are 4.55 and 4.37 days, respectively. The coefficients of variations (CVs), which is defined as the standard deviation divided by the mean, are 1.28 and 1.29, respectively. More than 95% of the patients have LOS between 0 and 15 days in both periods. The two distributions are both right-skewed. About 0.78% and 0.73% of the patients stay in NUH for more than 30 days in Periods 1 and 2, respectively, although the average LOS is only about 4.5 days for both periods. The maximum LOS is 206 days for Period 1 and 197 days for Period 2. Tables 27 and 28 in Appendix A.4 show the numerical values of the empirical LOS distributions and the tail frequencies of LOS after 30 days for the two periods.

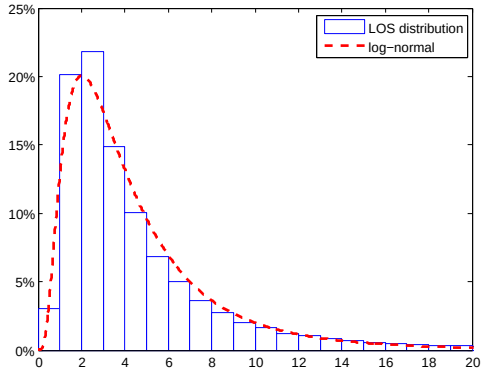
From the figures and tables, we can see there is little difference in the LOS distributions between Periods 1 and 2. We now use the combined data of the two periods to report statistics in the next few subsections. Figure 18b plots the empirical LOS distribution curve from the combined data, which visually resembles a log-normal distribution (with mean 4.65 and standard deviation 4).

2.5.2 AM- and PM-patients

Empirical evidence suggests that ED-GW patients' LOS depends on admission times. Figure 19a plots the average LOS for ED-GW patients admitted during each hour (using combined data). We observe that patients admitted before 10am have similar average LOS, and so are patients admitted after 12 noon. There is also a spike from 10am to noon. Given these interesting features, we categorize ED-GW patients into two groups: those admitted before noon, and those admitted after noon. For



(a) LOS distributions in Periods 1 and 2



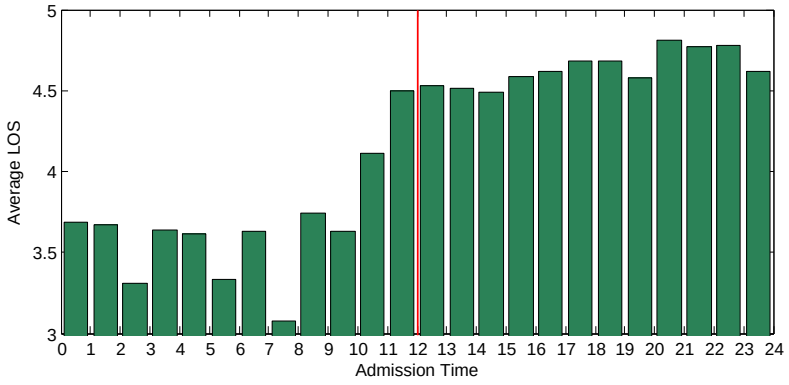
(b) Fitting the LOS distribution with a log-normal distribution (mean 4.65 and std 4)

Figure 18: **LOS distributions in Periods 1 and 2.**

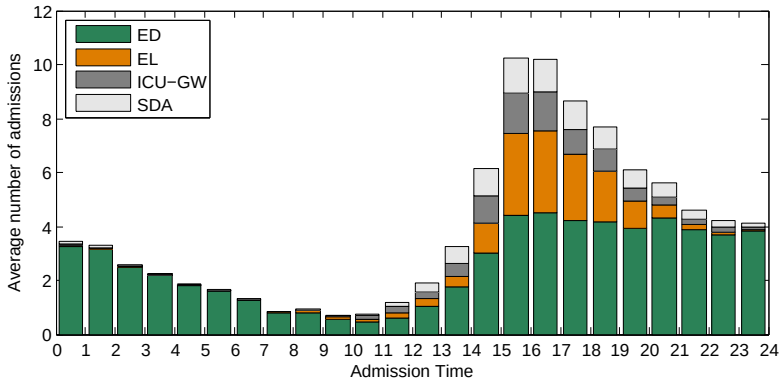
convenience, from now on we refer to them as ED-AM patients and ED-PM patients, respectively.

Figure 19b also provides the admission time distributions of the four admission sources. Around 69% of the ED-GW patients, 95% of the EL patients, 94% of the ICU-GW patients, and 92% of the SDA patients are admitted after noon. This suggests that for the purpose of comparing the differences of LOS between AM and PM admissions, we should focus on ED-GW patients, since patients from other sources comprise a very small portion of those admitted before noon.

The LOS distributions for ED-AM patients and ED-PM patients are substantially different. Figure 20a plots their LOS distributions with the cut-off value of 20 days. The sample size of ED-PM patients is 2.2 times that of ED-AM patients. Note that around 11% to 13% of the ED-AM patients are same-day discharge patients (i.e., those with LOS=0), whereas nearly 0% of the ED-PM patients are discharged same day in the two periods. Tables 29 in Appendix A.4 lists the total sample sizes and numerical values of the LOS distributions for ED-AM and ED-PM patients.

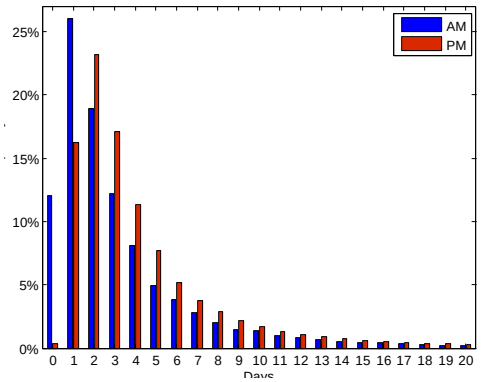


(a) Average LOS for ED-GW patients admitted in each hour

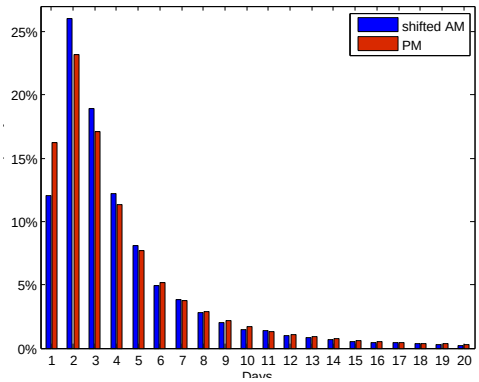


(b) Average number of admissions for ED-GW, EL, ICU-GW, and SDA patients in each hour

Figure 19: **Average LOS with respect to admission time.** The combined data is used.



(a) Original LOS distribution for ED-AM and ED-PM patients



(b) LOS distribution for ED-AM patients is shifted to the righthand side of x-axis by 1

Figure 20: **LOS distribution for ED-AM and ED-PM patients.**

One-day difference

Close examination reveals a difference of about one day between the average LOS for ED-AM and ED-PM patients. Using combined data, the average LOS is 3.60 days for all ED-AM patients and 4.66 days for all ED-PM patients. In fact, the two LOS distributions in Figure 20a are similar in shape when we do a shift. Figure 20b shows the comparison between the LOS distribution for ED-PM patients and the *shifted* LOS distribution for ED-AM patients, where the shifted distribution means that we shift the LOS distribution to the right-hand side of x-axis by 1 (e.g., value 1 in the shifted distribution corresponds to value 0 in the original distribution for ED-AM patients). We omit ED-PM patients with LOS=0 in Figure 20b due to the negligible proportion, so the plots start from value 1. After the shift, the two distribution curves are indeed close. Furthermore, the one-day difference in average LOS between ED-AM and ED-PM patients persists when we look into each specialty; see Table 6 in Section 2.5.3 below.

We speculate a potential reason for the one-day difference between ED-AM and ED-PM patients is staff schedules, i.e. most tests, consulting, and treatment occur

between 7am and 5pm (the regular working hours). ED-AM patients can be subjected to these tests and treatment since most of them are admitted in early morning (before 6am), whereas ED-PM patients must wait until the following day since most admissions are after 4pm. In Appendix A.4, we use two hypothetical scenarios to further illustrate this speculation.

2.5.3 LOS distributions according to patient admission source and specialty

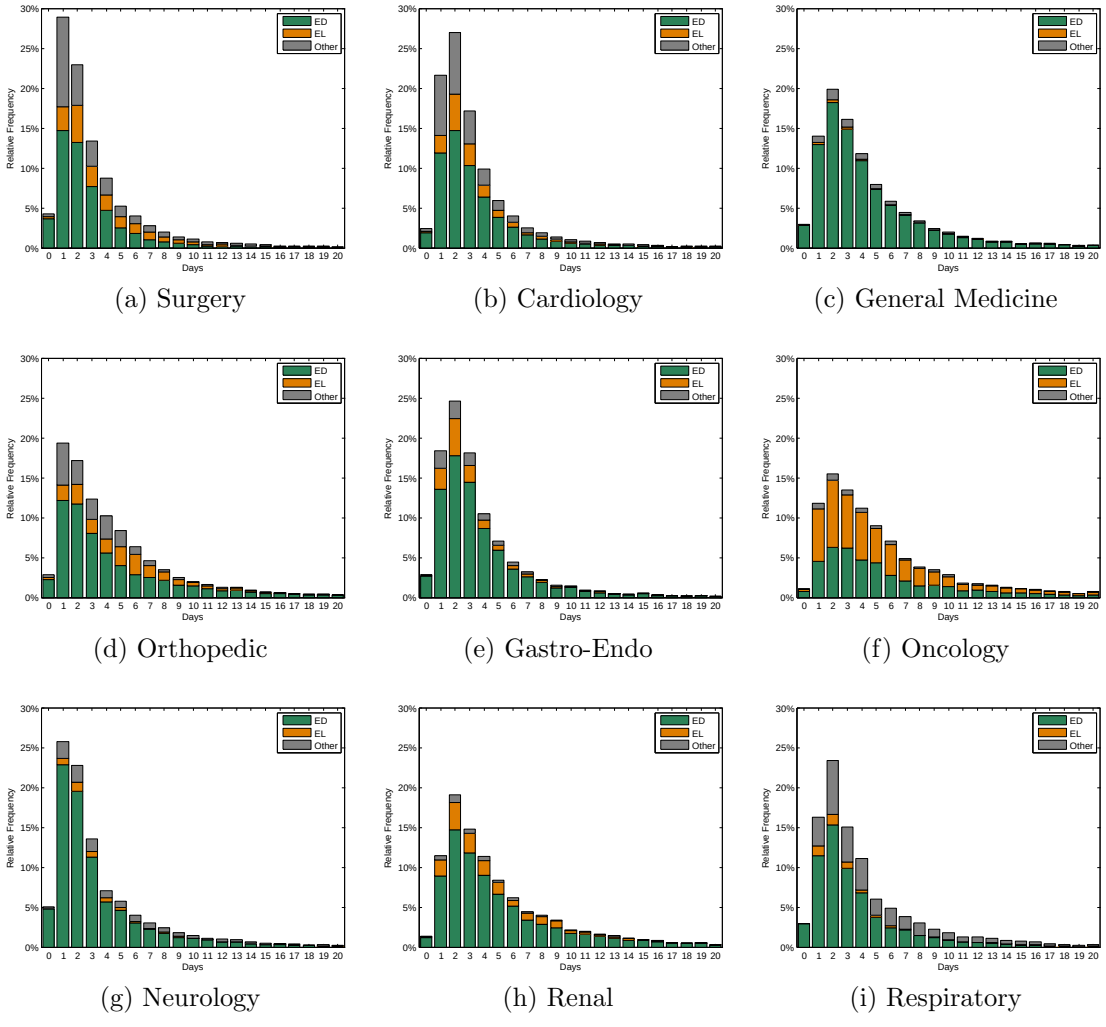
Table 6 reports the average and standard deviation of the LOS for each specialty and for each admission source in Periods 1 and 2. From the table, we can clearly see that the average LOS is both admission-source and specialty dependent. Moreover, consistent with Section 2.5.2, the one-day difference in average LOS between ED-AM and ED-PM patients exists across all specialties. Using the combined data, we plot the LOS distributions for each specialty and for each admission source in Figure 21. From Table 6 and the figures, we observe the following:

1. Comparing across specialties, Oncology, Orthopedic and Renal patients record a longer average LOS. Surgery and Cardiology patients demonstrate a shorter average LOS. The LOS distributions of each specialty exhibit a similar shape, which resembles a log-normal distribution. Oncology and Renal patients tend to have a longer tail. Both have a high proportion of patients staying longer than 14 days (9.93% for Oncology, and 7.59% for Renal, compared with 4.95% for all patients). The Coefficients of Variation (CV) for most combinations of specialty and admission source are between 1 and 2 in both periods. ICU-GW patients from specialties belonging to the Medicine cluster show a large CV (e.g., General Medicine, Respiratory), due to their small sample sizes.

2. Comparing across all admission sources, SDA patients in general have a shorter average LOS (about 2-3 days); ICU-GW patients, however, have a much longer

Table 6: **Average LOS for each specialty and each admission source.** The LOS is measured in days. The number in each parentheses is the standard deviation for the corresponding average.

Cluster	Period	ED-GW(AM)	ED-GW(PM)	EL	ICU-GW	SDA
Surg	1	2.36 (2.93)	3.27 (3.43)	4.55 (6.55)	9.58 (12.60)	2.59 (4.72)
	2	2.37 (3.04)	3.25 (3.40)	4.71 (6.11)	10.12 (13.32)	3.63 (8.09)
Card	1	2.95 (3.75)	3.83 (3.93)	4.15 (5.08)	5.22 (6.78)	2.55 (3.38)
	2	3.02 (3.93)	4.01 (4.68)	4.15 (5.64)	5.15 (7.47)	2.75 (4.26)
Gen Med	1	3.94 (4.76)	5.25 (5.87)	5.32 (5.79)	10.43 (18.43)	3.17 (2.62)
	2	4.09 (5.41)	5.24 (5.35)	5.47 (6.20)	8.82 (13.69)	3.15 (2.26)
Ortho	1	5.45 (8.22)	6.04 (7.04)	6.27 (6.19)	10.82 (13.32)	3.41 (4.32)
	2	3.27 (4.52)	4.65 (5.64)	6.15 (7.04)	13.49 (13.82)	4.62 (6.49)
Gastro	1	3.32 (3.91)	4.48 (4.47)	3.70 (4.39)	8.33 (12.25)	3.24 (3.99)
	2	3.51 (6.14)	4.18 (5.10)	3.55 (3.32)	6.97 (8.76)	3.27 (5.24)
Onco	1	5.93 (7.58)	7.03 (7.14)	6.45 (7.95)	8.62 (9.02)	4.10 (4.18)
	2	5.56 (6.15)	6.62 (6.69)	6.32 (8.22)	7.65 (9.06)	4.38 (5.40)
Neuro	1	3.23 (5.22)	4.07 (4.69)	4.06 (4.69)	7.56 (7.67)	2.59 (2.40)
	2	2.98 (6.69)	3.51 (4.52)	4.50 (4.77)	9.16 (11.85)	2.45 (1.85)
Renal	1	5.75 (6.55)	6.51 (6.90)	5.70 (6.20)	10.22 (12.91)	2.08 (1.16)
	2	4.63 (6.56)	5.40 (6.01)	5.06 (5.80)	8.65 (12.20)	3.30 (3.27)
Respi	1	3.21 (5.10)	4.29 (4.26)	4.45 (6.27)	7.86 (10.71)	2.33 (3.33)
	2	2.89 (3.65)	4.28 (4.27)	3.68 (3.81)	7.36 (9.70)	3.43 (2.07)
All	1	3.70 (5.25)	4.78 (5.45)	5.17 (6.47)	7.59 (10.82)	2.84 (4.29)
	2	3.46 (5.10)	4.48 (5.11)	5.11 (6.57)	7.62 (10.77)	3.66 (6.63)



(a) Surgery (b) Cardiology (c) General Medicine

(d) Orthopedic (e) Gastro-Endo (f) Oncology

(g) Neurology (h) Renal (i) Respiratory

Figure 21: **LOS distributions of each specialty.** In each plot, ED-AM and ED-PM patients are aggregated under the group "ED"; ICU-GW and SDA patients are aggregated under the group "Other". The combined data is used.

average LOS than patients from other sources for most specialties. Comparing EL and ED-GW patients, we find that EL patients tend to have a longer average LOS than ED-GW for most specialties.

3. Comparing between the two periods, most specialties show similar average LOS for the two periods with two exceptions. Renal patients show a significant decrease in average LOS (a reduction of about 1 day) in Period 2 for all admission sources except SDA patients. Orthopedic also shows a significant reduction in the average LOS for ED-GW patients. In particular, there were fewer long-stay patients in Period 2 from these two specialties indicated by our tail distribution plots (figures omitted here).

The heterogeneity of LOS among specialties is expected, since the underlying medical conditions for patients from different specialties are markedly different. Moreover, the patient admission source also influences the average LOS. In particular, we note that the average LOS of EL patients is longer than that of ED-GW patients from Surgery, Cardiology, and Orthopedic. This is somewhat counter-intuitive, since ED-GW patients generally have more urgent and complicated conditions than EL patients and need longer treatment time. One possible explanation is that most EL patients (from these specialties) undergo surgical procedures during their stay, but their priority in surgery scheduling is lower than that of ED-GW patients. EL patients usually are admitted at least one day earlier before the day of surgery, while ED-GW patients may have their surgeries done on the same day of admission due to the urgency. We note that hospitals in other countries report similar dependency of average LOS on admission sources (ED-admitted patients have shorter LOS than elective patients), e.g., UK [112]. However, some studies also report shorter average LOS for elective patients, e.g., Canada (see Page 14 of [21]) and US [17, 76]. The difference could probably be the result of financial incentives and related factors in place.

2.5.4 LOS between right-siting and overflow patients

As introduced in Section 2.3, NUH sometimes overflows patients to non-primary wards. We call a patient who is assigned to a non-primary ward an *overflow* patient, otherwise a *right-siting* patient. In this section, we compare the LOS between right-siting and overflow patients.

Considering the dependence of LOS on admission source and specialty, Table 7 compare the average LOS for right-siting and overflow patients for each specialty and admission source (and admission period for ED-GW patients). Specialties belonging to the Medicine cluster are aggregated to get a more reliable estimation (with a larger sample size). From the table, we observe that the average LOS are close between right-siting and overflow patients for Medicine, Surgery, and Cardiology. Overflow patients from Orthopedic show a longer average LOS than that of right-siting patients for each admission source with the exception of SDA. In contrast, Oncology overflow patients show a shorter average LOS than that of right-siting patients. However, given the sample sizes of Orthopedic and Oncology overflow patients are small (as well as the high standard deviation), we cannot definitively conclude that overflow patients have a significant longer or shorter LOS than right-siting patients.

2.5.5 Test iid assumption for LOS

In this section, we test whether it is reasonable to assume the patients' LOS are iid random variables. Similar to the previous section, we test the iid assumption for each admission source and medical specialty (and admission period for ED-GW patients). We use the Period 1 data for the tests.

Test the identically distributed assumption

To test whether the LOS within a patient class are identically distributed, we further separate the Period 1 data into 6 groups. Each group containing the LOS of patients admitted within one of the six quarters in Period 1 (one and half years in total). We

Table 7: **Average LOS for right-siting and overflow patients.** Period 1 data is used. Numbers in parentheses are standard deviations. Specialties belonging to the Medicine cluster are aggregated for a larger sample size.

Cluster	Source	right-siting		overflow	
		#	ALOS	#	ALOS
Med	ED-AM	2010	3.45 (4.09)	2325	3.05 (4.15)
	ED-PM	6360	4.61 (4.95)	4296	4.56 (4.87)
	EL	952	3.86 (4.52)	605	4.62 (5.50)
	ICU	756	7.08 (10.26)	779	6.64 (9.11)
	SDA	274	3.01 (2.25)	229	2.30 (1.53)
Surg	ED-AM	1364	2.23 (2.55)	537	1.85 (2.21)
	ED-PM	3040	3.04 (2.87)	590	3.04 (2.83)
	EL	1642	4.21 (6.23)	296	4.37 (5.37)
	ICU	869	7.65 (9.02)	53	8.87 (6.83)
	SDA	1894	2.26 (3.00)	281	2.12 (1.96)
Card	ED-AM	590	2.77 (3.08)	693	2.67 (3.54)
	ED-PM	1653	3.70 (3.65)	1550	3.57 (3.62)
	EL	710	3.70 (3.65)	509	4.16 (5.25)
	ICU	1332	3.95 (4.73)	237	4.27 (3.98)
	SDA	459	1.92 (1.97)	249	2.08 (2.38)
Ortho	ED-AM	971	4.62 (6.18)	155	6.55 (10.57)
	ED-PM	2363	5.53 (6.41)	488	6.46 (7.93)
	EL	1041	5.57 (4.90)	195	7.59 (7.55)
	ICU	62	8.53 (9.59)	19	11.63 (18.79)
	SDA	906	3.17 (3.03)	139	2.96 (2.42)
Onco	ED-AM	249	5.69 (7.35)	73	2.99 (2.83)
	ED-PM	645	6.81 (7.15)	171	4.09 (4.15)
	EL	1348	6.10 (7.64)	214	4.35 (7.10)
	ICU	148	7.43 (6.53)	19	6.89 (8.90)
	SDA	7	3.43 (2.64)	2	1.50 (0.71)

denote the 6 groups as 08Q1, 08Q2, 08Q3, 08Q4, 09Q1, and 09Q2, respectively, We use the quarter setting since it allows us to conduct a modest number of tests for each combination of admission source and specialty and meanwhile ensures enough sample points within each group.

Our null hypothesis is that the samples (LOS) from two consecutive quarter-groups follow the same distribution, and we adopt the χ^2-test to test the null hypothesis (see Test 43 in [84]). Table 8 lists the values of the test statistics and the critical value at the significance level 5% for the five groups of tests in each patient class. Note that the sample points for Oncology SDA patients are too few to conduct reliable tests. Thus, we do not perform the tests for them and we leave the entries belonging to the Oncology SDA group blank in the table. We can see that among the 120 performed tests, the majority of them cannot reject the null hypothesis (with the test statistics less than the critical values). The only two exceptions are highlighted in red. These test results indicate that it is reasonable for us to assume the LOS are identically distributed within a patient class.

Test the independence assumption

We adopt a nonparametric test proposed in [30] to investigate the serial dependence of the LOS. We focus on testing the dependence between the LOS of two patients admitted consecutively. The main idea of this test is to examine whether the L_1 distance between the estimates of the samples' joint density and the estimates of the product of individual marginals is small enough. Because under the null hypothesis of independence, the joint density of the samples should be equal to the product of the individual marginals.

Similar to what we did for the identically distributed assumption, we test the serial dependence for LOS within each quarter-group of each combination of admission source and specialty (and admission period for ED-GW patients). Table 9 lists the

Table 8: **Results of the χ^2-tests for testing the identically distributed assumption for LOS.** In the table, ts denotes for test statistics, and cv denotes for critical values at the significance level 5%. The samples for Oncology SDA patients are too few to conduct reliable tests, and the corresponding entires are left blank. Specialties belonging to the Medicine cluster are aggregated for a larger sample size.

Cluster	Data	ED-AM		ED-PM		EL		SDA		ICU	
		ts	cv	ts	cv	ts	cv	ts	cv	ts	cv
Med	08Q1 vs. 08Q2	29.64	46.19	42.74	59.30	32.71	36.42	13.35	19.68	42.84	48.60
	08Q2 vs. 08Q3	27.69	43.77	47.02	60.48	29.70	44.99	16.36	21.03	38.79	51.00
	08Q3 vs. 08Q4	26.51	43.77	63.69	62.83	23.96	43.77	10.24	22.36	47.71	53.38
	08Q4 vs. 09Q1	24.04	40.11	39.91	61.66	14.17	35.17	9.02	19.68	52.34	53.38
	09Q1 vs. 09Q2	33.01	43.77	47.85	58.12	17.61	36.42	10.62	21.03	30.46	54.57
Surg	08Q1 vs. 08Q2	15.69	28.87	14.24	31.41	40.93	40.11	23.31	28.87	46.23	49.80
	08Q2 vs. 08Q3	18.89	28.87	33.56	37.65	28.73	38.89	21.50	28.87	30.96	47.40
	08Q3 vs. 08Q4	17.09	26.30	33.49	40.11	30.73	41.34	20.73	31.41	43.34	51.00
	08Q4 vs. 09Q1	20.52	31.41	23.68	37.65	32.65	43.77	24.44	37.65	42.85	52.19
	09Q1 vs. 09Q2	22.83	31.41	23.25	35.17	28.47	40.11	23.48	35.17	31.26	52.19
Card	08Q1 vs. 08Q2	27.61	31.41	39.46	41.34	20.26	38.89	11.12	21.03	23.18	41.34
	08Q2 vs. 08Q3	20.97	31.41	30.67	41.34	31.04	41.34	10.00	21.03	16.05	37.65
	08Q3 vs. 08Q4	17.31	31.41	18.69	40.11	29.35	37.65	7.69	23.68	18.70	37.65
	08Q4 vs. 09Q1	23.32	31.41	18.78	40.11	23.42	31.41	10.16	23.68	23.99	38.89
	09Q1 vs. 09Q2	22.47	30.14	24.97	41.34	20.89	33.92	10.87	19.68	29.64	41.34
Ortho	08Q1 vs. 08Q2	35.24	47.40	38.26	58.12	20.60	44.99	11.04	27.59	18.00	25.00
	08Q2 vs. 08Q3	32.67	43.77	48.98	58.12	23.21	43.77	20.74	27.59	9.25	23.68
	08Q3 vs. 08Q4	25.74	41.34	32.36	54.57	24.60	42.56	20.03	28.87	13.85	26.30
	08Q4 vs. 09Q1	21.20	42.56	48.95	55.76	22.99	40.11	19.17	26.30	9.64	22.36
	09Q1 vs. 09Q2	32.88	38.89	42.71	53.38	23.68	40.11	11.31	23.68	16.62	27.59
Onco	08Q1 vs. 08Q2	13.69	26.30	24.35	44.99	38.87	52.19			28.64	30.14
	08Q2 vs. 08Q3	15.33	32.67	25.61	42.56	36.85	48.60			13.50	30.14
	08Q3 vs. 08Q4	16.27	32.67	26.96	43.77	48.19	52.19			25.05	31.41
	08Q4 vs. 09Q1	15.34	32.67	27.87	41.34	37.92	48.60			17.27	28.87
	09Q1 vs. 09Q2	21.51	36.42	27.31	42.56	27.89	47.40			11.43	27.59

values of the test statistics and the critical value at the significance level 5% for all the 144 tests we have done. Again we do not perform the tests for the Oncology SDA patients because of the small sample size. From the table, we can see that the majority of the tests cannot reject the null hypothesis of independence (with the test statistics less than the critical values). The seven exceptions are highlighted in red. These test results indicate that it is reasonable for us to assume the LOS are independent within a patient class.

2.6 Service times

In this section, we present empirical findings on patient service times, which motivate our new service time model to be introduced in Chapter 3. As mentioned in the previous section, LOS constitutes the majority of a patient's service time, and it is natural that service time is also specialty- and admission-source-dependent. Thus, in this section we focus on service time distributions for all patients (from all admission sources and specialties). We first show service time distributions at both hourly and daily resolutions in Section 2.6.1 and observe a clustering phenomenon under the hourly resolution. Then in Section 2.6.2, we take a closer look at the residual distribution of service time to explain the clustering phenomenon. The samples for statistics reported in this sections are the same as those used in reporting LOS distributions, i.e., we exclude transfer patients who transfer to ICU-type wards after their initial admissions to GWs.

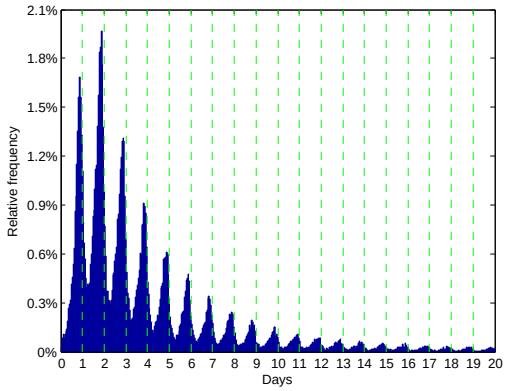
2.6.1 Service time distribution

Hourly resolution

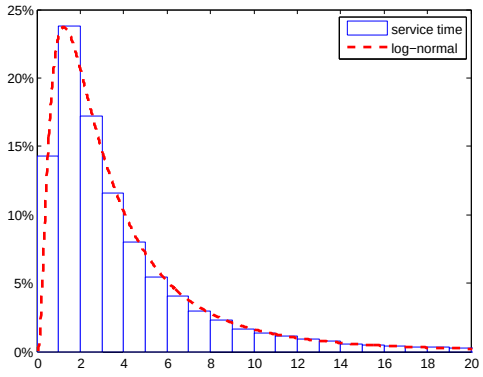
Like LOS distributions, the service time distributions for the two periods are not significantly different. Therefore, we plot them using the combined data. Figure 22a shows the histogram of the service time for all patients. The bin size is 1 hour, and each green line on the horizon axis represents a 24-hour (1 day) increment.

Table 9: **Results of the nonparametric tests for testing the serial dependence among patient LOS.** In the table, ts denotes for test statistics, and cv denotes for critical values at the significance level 5%. The samples for Oncology SDA patients are too few to conduct reliable tests, and the corresponding entires are left blank. Specialties belonging to the Medicine cluster are aggregated for a larger sample size.

Cluster	Data	ED-AM		ED-PM		EL		SDA		ICU	
		ts	cv	ts	cv	ts	cv	ts	cv	ts	cv
Med	08Q1	0.367	0.357	0.285	0.288	0.539	0.573	0.519	0.615	0.819	0.848
	08Q2	0.375	0.382	0.250	0.269	0.609	0.631	0.346	0.440	0.749	0.735
	08Q3	0.325	0.379	0.237	0.270	0.614	0.641	0.484	0.602	0.739	0.799
	08Q4	0.262	0.332	0.248	0.269	0.539	0.586	0.429	0.547	0.807	0.844
	09Q1	0.348	0.327	0.229	0.257	0.507	0.538	0.437	0.581	0.871	0.894
	09Q2	0.320	0.360	0.250	0.262	0.479	0.541	0.515	0.603	0.726	0.775
Surg	08Q1	0.271	0.332	0.231	0.276	0.362	0.434	0.195	0.234	0.954	0.978
	08Q2	0.404	0.397	0.238	0.292	0.467	0.505	0.175	0.220	0.878	0.990
	08Q3	0.308	0.399	0.283	0.316	0.459	0.500	0.159	0.238	0.927	1.008
	08Q4	0.327	0.353	0.284	0.324	0.557	0.589	0.253	0.315	0.951	1.040
	09Q1	0.357	0.374	0.294	0.317	0.454	0.518	0.419	0.368	1.150	1.216
	09Q2	0.337	0.366	0.225	0.286	0.535	0.565	0.279	0.314	1.035	1.144
Card	08Q1	0.359	0.446	0.352	0.355	0.557	0.668	0.259	0.290	0.547	0.551
	08Q2	0.524	0.548	0.391	0.382	0.585	0.602	0.221	0.300	0.416	0.465
	08Q3	0.507	0.552	0.361	0.362	0.526	0.591	0.296	0.366	0.420	0.532
	08Q4	0.401	0.493	0.302	0.342	0.397	0.515	0.293	0.442	0.453	0.493
	09Q1	0.359	0.438	0.296	0.348	0.503	0.586	0.228	0.416	0.518	0.594
	09Q2	0.407	0.466	0.320	0.338	0.501	0.533	0.400	0.467	0.396	0.465
Ortho	08Q1	0.770	0.826	0.547	0.612	0.668	0.745	0.458	0.532	2.000	2.000
	08Q2	0.865	0.953	0.599	0.618	0.752	0.803	0.395	0.461	1.736	1.750
	08Q3	0.691	0.726	0.572	0.592	0.607	0.711	0.422	0.444	1.560	1.590
	08Q4	0.747	0.829	0.550	0.550	0.739	0.807	0.469	0.588	1.391	1.529
	09Q1	0.756	0.795	0.514	0.546	0.806	0.796	0.447	0.494	1.636	1.686
	09Q2	0.646	0.651	0.461	0.519	0.848	0.870	0.464	0.543	1.728	1.778
Onco	08Q1	0.877	1.078	0.901	0.963	0.647	0.708			1.237	1.332
	08Q2	1.272	1.373	1.080	1.097	0.713	0.765			1.264	1.420
	08Q3	1.092	1.099	0.898	1.025	0.701	0.709			1.667	1.682
	08Q4	0.982	1.047	0.804	0.906	0.653	0.687			1.617	1.672
	09Q1	1.272	1.300	0.882	0.934	0.678	0.746			1.723	1.756
	09Q2	1.085	1.213	0.927	0.945	0.700	0.727			1.293	1.410



(a) Distribution, in hourly resolution, of the service times from both periods; each green dashed line corresponds to a 24-hour increment



(b) Distribution, in daily resolution, of the service times from both periods; a log-normal distribution fits the histogram (mean 5.02, std 6.32)

Figure 22: **Distribution of service times in two time resolutions.**

This histogram demonstrates some unique features. First, most of the data points "cluster" around the integer values (the green lines), with multiple peaks appearing at integer values which represent Day 1, Day 2, ... In fact, such clustering phenomenon has been observed in other hospitals using the same 1-hour time resolution; see, for example, [5]. Second, we note that connecting the peak points gives a curve with a shape similar to the LOS distribution in Figure 18a. This indicates that there is a strong dependence between service time and LOS, although they are two different concepts.

Daily resolution

Figure 22b plots the histogram of the service times using the combined data, but in daily resolution, i.e., the bin size is 1 day. Like the LOS distribution, this plot resembles a log-normal distribution, which is consistent with the observations from [5]. However, Figure 23 shows that the LOS distribution and the day-resolution service time distribution can be significantly different.

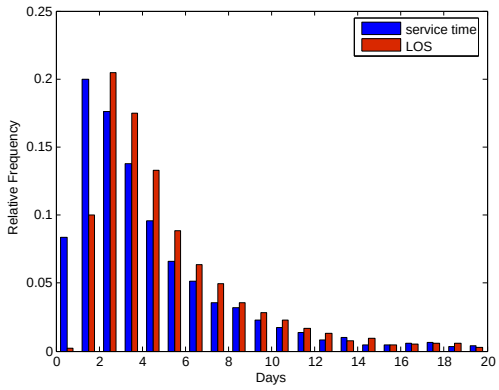


Figure 23: **LOS and day-resolution service time distributions for General Medicine patients.** The combined data is used.

2.6.2 Residual distribution

To better understand the clustering phenomenon in service time distribution under hourly resolution (see Figure 22), we focus on the distribution pattern around the integer values. We use $\lfloor x \rfloor$ to denote the floor of a real number x, i.e., the largest integer value r that is smaller than or equal to x. Using the time unit of day, we define the residual of service time S as

$$\text{res}(S) = S - \lfloor S \rfloor. \tag{2}$$

In the rest of this thesis, we always use the time unit of day for service time and residual, unless otherwise specified.

Figure 24a shows the empirical distributions of residuals in Periods 1 and 2. Clearly, the distributions are both U-shaped, with most residuals beging close to 0 (or 1 from periodicity). In fact, in both periods, more than 65% of the residuals are located between 0.58 and 1 day, and another 9% are located between 0 and 0.1 day. Since $\lfloor S \rfloor$ takes integer values, this U-shape residual distribution results in the clustering phenomenon we observe in Figure 22.

We now show the relationship between res(S) and admission/discharge time and explain why the residual distribution has such U-shape. Let T_{adm} and T_{dis} be the

admission time and discharge time of a patient, respectively (all in the unit of days). We then have

$$\begin{aligned} \text{res}(S) &= S - \lfloor S \rfloor \\ &= T_{\text{dis}} - T_{\text{adm}} - \lfloor (T_{\text{dis}} - T_{\text{adm}}) \rfloor \\ &= (T_{\text{dis}} - \lfloor T_{\text{dis}} \rfloor - (T_{\text{adm}} - \lfloor T_{\text{adm}} \rfloor)) \bmod 1, \end{aligned} \tag{3}$$

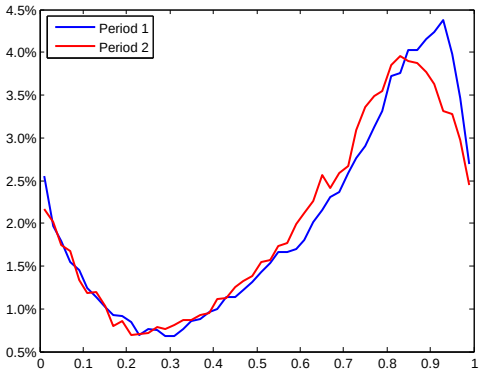
where for two real numbers x and $y \neq 0$, $x \bmod y = x - \lfloor x/y \rfloor \cdot y$.

The *time-of-day* distributions of admission and discharge (i.e., distributions of $T_{\text{adm}} - \lfloor T_{\text{adm}} \rfloor$ and $T_{\text{dis}} - \lfloor T_{\text{dis}} \rfloor$) jointly determine the residual distribution. We know that the majority of patients (more than 60%) are admitted between 2pm and 10pm (see Figure 19b), and discharged between noon and 4pm (see Figure 2) each day. Thus, the admission hour ($T_{\text{adm}} - \lfloor T_{\text{adm}} \rfloor$) is mostly distributed between 0.58 and 0.92 day, and the discharge hour ($T_{\text{dis}} - \lfloor T_{\text{dis}} \rfloor$) is mostly distributed between 0.5 and 0.67 day. According to (3), the residual should mostly be distributed between 0.58 and 1 day, with some distributed between 0 and 0.09 day. This matches our observation from Figure 24a. In summary, since most admissions occur after previous discharges, the residual is close to 0 (or 1 from periodicity) and thus leads to the clustering phenomenon in the service time distribution.

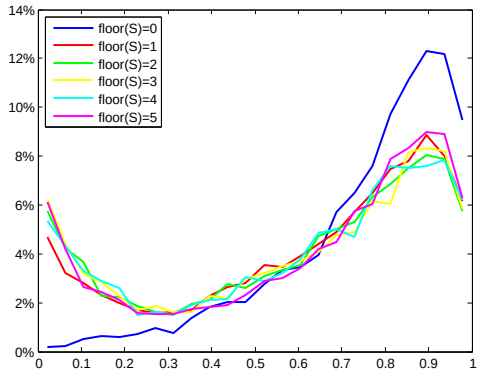
Next, we present additional empirical findings on the residual distribution.

Independence on the value of $\lfloor S \rfloor$

We examine whether the residual distribution depends on the value of $\lfloor S \rfloor$. Figure 24b shows the histogram of the residuals conditioning on the values of $\lfloor S \rfloor$ with Period 1 data. The bin size is 1 hour. Except for the case conditioning $\lfloor S \rfloor = 0$, the conditional residual distributions look similar and they resemble the aggregated one (the blue one) in Figure 24a. We observe the same phenomenon when we plot the conditional residual histogram using Period 2 data.



(a) Empirical distribution in Periods 1 and 2



(b) Empirical distribution conditioning on floor of service times using Period 1 data

Figure 24: **Empirical distribution of the residual of service time.** The bin size is 0.02 day (30 minutes).

When $\lfloor S \rfloor = 0$, the conditional distribution curve is significantly different from other conditional distributions. This difference, which can also be explained using (3), is mainly due to the admission and discharge distributions of same-day discharge patients (see Figure 25), which are very different from those of other patients.

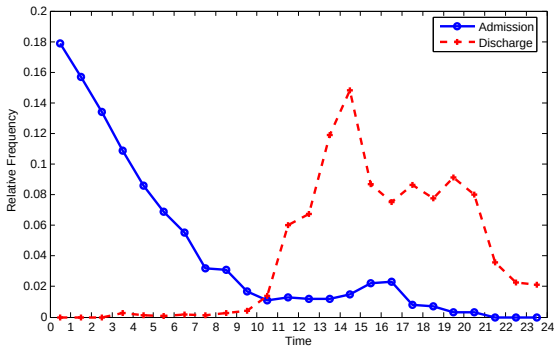


Figure 25: **Admission time and discharge time distributions for same-day discharge patients**. The combined data is used.

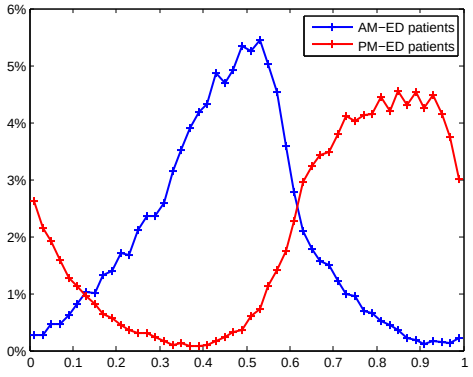


Figure 26: **Empirical distributions of the residual of service times for AM- and PM-admitted ED-GW patients.** Period 1 data is used.

Residual distribution for AM and PM admissions

Recall that in Section 2.5.2, we find that the average LOS of ED-AM and ED-PM patients almost differ by 1 day. The service time, however, shows less difference between ED-AM and ED-PM patients. The average service times are 4.15 and 3.89 days for ED-AM patients, and 4.61 and 4.30 days for ED-PM patients in Periods 1 and 2, respectively. Thus, the difference in the average service times is about 0.25 to 0.31 day (around 6-7 hours) between ED-AM and ED-PM patients, less than the one-day difference in the average LOS.

Moreover, we find that the difference in the average service time mainly comes from the difference in the residual distribution between ED-AM and ED-PM patients. Figure 26 shows the residual distributions between ED-AM and ED-PM patients, which are significantly different. The reason can still be explained by (3). The majority of ED-AM patients (around 60%) are admitted between midnight and 4am (see Figure 19b) and discharged between noon and 4pm (see Figure 2), thus, their residuals are mostly distributed between 0.33-0.5 day, matching the blue curve in Figure 26; while the majority of ED-PM patients are admitted between 2pm and 10pm and discharged between noon and 4pm, so the residual distribution is close to the aggregated one in Figure 24a. The empirical distributions of $\lfloor S \rfloor$ for ED-AM and ED-PM patients, on the other hand, are close to each other.

2.7 Pre- and post-allocation delays

In this section, we take a closer look at the ED to wards transfer process and understand the bottlenecks within this transfer process. Often, the inpatient bed unavailability is regarded as a major bottleneck within the transfer process; while in this section, we show that secondary bottlenecks, such as the unavailability of physicians, ward nurses and ED porters, also have a significant effect on the waiting time of ED-GW patients. We first provide a comprehensive description of the process flow of

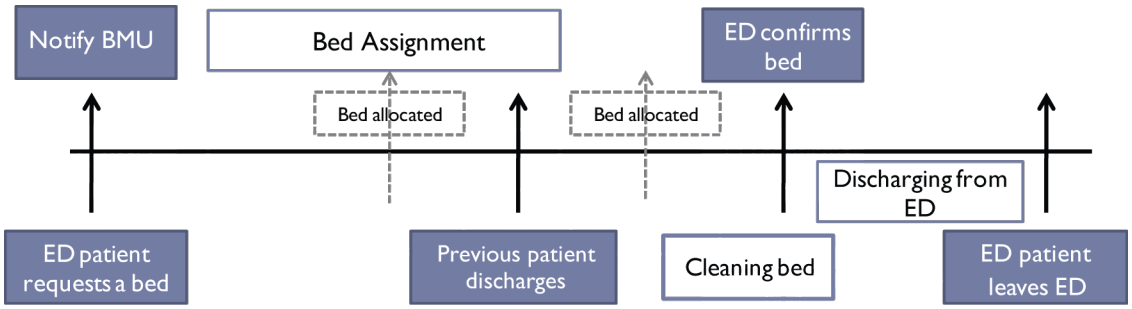


Figure 27: **Process flow of the transfer from ED to GW.**

a typical patient transfer from ED to a GW in Section 2.7.1. This process flow motivates us to separate an ED-GW patient waiting time into two parts: pre-allocation delay and post-allocation delay (see Section 2.7.2). We use the two allocation delays to capture delays caused by secondary bottlenecks. Finally in Section 2.7.3, we show empirical distributions for the two allocation delays.

2.7.1 Transfer process from ED to general wards

When a patient finished receiving treatment in ED and physicians decide to admit him/her, ED nurses send a bed-request to the bed management unit (BMU) for this patient. Then BMU staff initiate bed search and allocate an appropriate bed for the patient. After a bed is allocated, ED confirms the bed allocation and then transfers the patient to the allocated bed. Figure 27 illustrates an example of the process flow for transferring an ED-GW patient to a GW. In the next two subsections, we give detailed explanation of this figure and describe (i) the bed allocation process and (ii) the discharge process from ED after bed allocation.

Bed allocation process

At NUH, the BMU controls all types of inpatient bed allocations (including bed allocation for ED-GW patients) during the day time, from 7am to 7pm. During the night, a nurse manager is in charge of all bed allocations. The allocation process for a bed-request from an ED-GW patient usually has four steps:

1. After BMU receives the bed request, one of the BMU staff makes a tentative bed allocation, trying to match all the criteria for the patient, such as gender, medical specialty, class of bed.

2. The staff member then checks/negotiates with the ward nurses who are in charge of the allocated bed in order to secure acceptance. If the ward nurses reject the request, then the staff member has to make another tentative allocation and repeats the negotiation process until one ward agrees to accept the patient.

3. Once a ward accepts the patient, BMU notifies the ED nurses about the bed allocation and waits for ED's confirmation. Usually, ED confirms the allocation. But sometimes the bed requirements might change because the patient's medical condition changes. Under such circumstances, ED cannot confirm the bed allocation and has to submit a new bed-request to BMU. BMU then repeats steps 1 and 2 to effect a new allocation.

4. After ED's confirmation, the bed is officially allocated to the patient.

The bed allocation process is similar for elective and internal transfer patients, except that when the receiving ward agrees to accept the incoming patient, the bed allocation is confirmed via other ways (no longer through ED nurses).

In the last step, when a bed is allocated to the patient, the allocated bed may be in different status: still occupied by the patient who is going to discharge soon, or in cleaning, or ready to be used. The bed can be allocated to a patient even if it is still being occupied because BMU allocates beds based on two types of bed information:

(i) Real-time information: BMU has the status of all inpatient beds in real-time, e.g., whether a bed is currently vacant, being cleaned, or being occupied;

(ii) "Planned" discharged information: BMU also has access to the planned discharge information, allowing its staff to know which patients are going to be

discharged today. The planned discharge information also includes the ward nurses' estimate of the expected discharge time for each planned discharge.

Thus, when the bed allocation is made from planned discharge information, the bed can be allocated even before the bed becomes available (i.e., before the current occupying patient discharges).

We note that the majority of time in the bed allocation process is spent on BMU staff searching for appropriate beds and negotiating with ward nurses. In addition to bed unavailability, insufficient number of BMU agents, especially during the peak hours when a large number of bed requests are presented (usually from 1pm to 7pm), can cause delay in the bed allocation process and thus become a bottleneck. Another bottleneck comes from ward nurse unavailability, i.e., when nurses are busy with other activities (e.g., doing morning rounds with physicians), they cannot communicate with BMU staff to confirm bed-requests.

Discharging from ED and transfer to wards

When to start the ED discharge process for an ED-GW patient depends on the status of his/her allocated bed. ED nurses monitor the real-time bed status via several big screens in ED. If a ED-GW patient's allocated bed is still being occupied, ED nurses usually wait until the bed becomes available (and status changes to in clean) to start the discharge process; otherwise, ED nurse starts to prepare the patient's discharge process immediately. It usually takes the following steps to discharge and transfer a patient to a GW:

a) ED nurses ensure that all test results are complete and the patient needs no further treatment in ED;

b) ED nurses check for vital symptoms to ensure patient's medical stability;

c) ED physicians give written instructions for discharge from ED;

d) ED arranges a porter (patient's escort) and an ED nurse to transport the patient;

e) Ward nurses admit the patient to the allocated bed and complete the admission.

Delays can occur in each of the above steps even if the bed is ready to use. For example, the patient cannot exit ED if his/her test results are not ready for release or the patient is not medically stable. ED physicians or nurses may be busy attending to other patients, and do not have time to prepare for the ED discharge. Similarly, if ward nurses are busy, the patient cannot be admitted to bed. Moreover, porters may not be available to transport patients, especially during peak hours. We see that ED physicians, nurses, and porters may all become bottlenecks during the ED exit and transfer process.

2.7.2 Pre- and post-allocation delays

From the previous section, we see the bed allocation process and the ED discharge and transfer process are the two major processes prior to a ED-GW patient's admission to a GW. Delays can not only be caused by bed unavailability, but also secondary bottlenecks such as ED nurses and physicians. Thus, to understand the proportion of delays caused by *secondary bottlenecks* in the entire waiting time of an ED-GW patient, we use the bed-allocation time to divide the waiting time of an ED-GW patient into two parts: (i) *pre-allocation delay*: the duration from bed-request time to bed-allocation time; and (ii) *post-allocation delay*: the duration from bed-allocation time to admission time. We intend to use the pre-allocation delay to capture secondary bottlenecks in the bed-allocation process, and reflect the minimum amount of time that BMU needs to search and negotiate a bed for an incoming patient. We use the post-allocation delay to capture secondary bottlenecks in the ED discharge and transfer process. Next, we show how we empirically estimate these two delays from certain subgroups of patients and the empirical results.

Estimate pre-allocation delay

Our intention is to use the pre-allocation delay to reflect the minimum amount of time that BMU needs to search and negotiate a bed for an incoming patient. Thus, we need to eliminate the effect of bed unavailability and other factors when estimating the pre-allocation delay. For example, if there is no bed available upon a patient's bed-request time and no planned discharge information is available, BMU has to wait until an appropriate bed becomes available to start the negotiation process. When we estimate the pre-allocation delay, the ideal situation is to use the duration from when BMU starts the actual bed searching process till the bed-allocation time. Unfortunately, NUH data does not register any time stamp to reflect the start time of the bed-allocation process. Thus, we impose the first condition to select patients in our samples for estimating pre-allocation delay: *the allocated bed is available before the patient's bed-request time.*

Moreover, when BMU wants to overflow an incoming patient to a non-primary ward, the staff may wait a few hours before starting the negotiation process in order to control the overflow proportion. our empirical evidence also shows that the negotiation process could take longer than allocating a primary bed (see details in Section A.5 in the appendix). Thus, we impose the second condition to select patients in the samples: *the allocated bed comes from the primary ward for the patient.* We use these two conditions to eliminate the impact of bed unavailability and specialty mismatch on the pre-allocation delay, so that our estimation could better reflect the time that BMU needs to search and negotiate a bed. For the included ED-GW patients, their pre-allocation delay starts from the bed-request time and ends at the bed-allocation time.

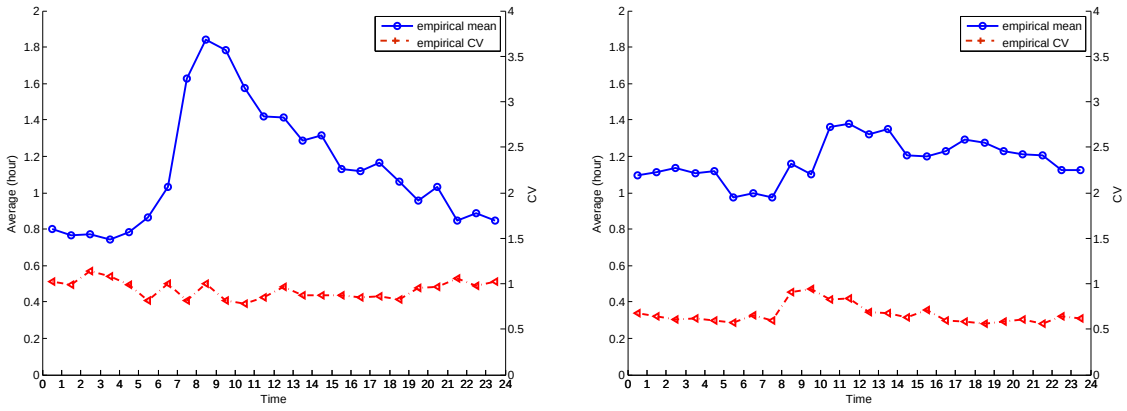
Estimate post-allocation delay

For the post-allocation delay, since a bed may not be available upon the bed-allocation time, we consider two scenarios to eliminate the impact of bed unavailability: (a) if bed-available time (i.e., the discharge time of the previous patient occupying the bed) is earlier than the bed-allocation time, the post-allocation delay starts from bed-allocation time and ends at admission time; (b) otherwise, the post-allocation delay starts from bed-available time and ends at admission time. All ED-GW patients are included in the samples.

Note that we use the previous patient's discharge time as the bed-available time without taking the bed cleaning time into account. It is because bed cleaning generally takes less than 20 minutes at NUH, and nurses usually start the ED discharge and transfer process when bed is being cleaned. Section A.5 in the appendix provides more empirical support.

Time-dependence

Figures 28a and 28b plot the empirical estimates of the mean and CV for the pre-allocation and post-allocation delays, respectively. The empirical curves in Figure 28 clearly demonstrate a time-dependent feature of both allocation delays. The average delays are longer if the delay initiation time is in the morning, especially for the pre-allocation delay. The longer pre-allocation delay in the morning may stem mainly from the ward side. At NUH the ward physicians and nurses are busy with morning rounds, and therefore it may take the BMU longer time to search and negotiate for beds. The longer post-allocation delay in the morning may stem mainly from the ED side. The ED at NUH is usually congested in late mornings, so it is likely that ED physicians and nurses are busy with newly arrived patients and have less time to discharge and transfer admitted patients to wards.



(a) Pre-allocation delay

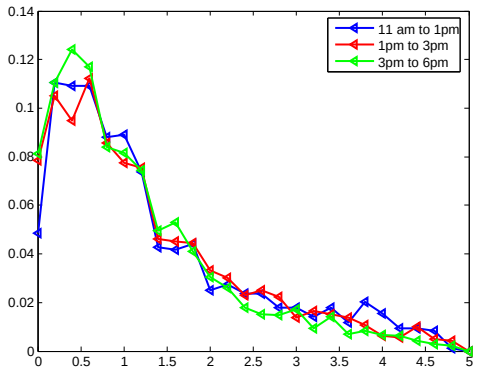
(b) Post-allocation delay

Figure 28: **Mean and CV of estimated pre- and post-allocation delays with respect to the delay initiation hour.** Left vertical axis is for the average; right vertical axis is for the CV. The scale of the right vertical axis is deliberately chosen to be large, so that the four curves are not crossed over.

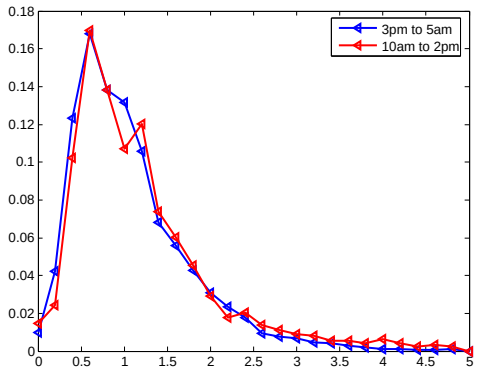
2.7.3 Distribution of pre- and post-allocation delays

Figure 28 shows that the pre- and post-allocation delays depend on the delay initiation time. Thus, to estimate the distributions for the two allocation delays, we group patients into several sub-groups according to the delay initiation hour, so that within each sub-group, the averages of pre- or post-allocation delay for each of the aggregated hours are close. For pre-allocation delay, we create 7 sub-groups: 1-3am, 3-5am, 11am-1pm, 1pm-3pm, 3pm-6pm, 6pm-9pm, and 9pm-1am (the next day). For post-allocation delay, we create two sub-groups: 10am-2pm, and 2pm-5am (the next day). These aggregations allow a moderate sample size for each sub-group. We exclude patients whose pre-allocation delay initiates between 5am and 11am, and patients with post-allocation initiation times between 5am and 10am due to the small sample sizes in these time intervals. Patients selected in the samples satisfy the same conditions as we mentioned above.

Figure 29a shows empirical distributions for selected pre-allocation sub-groups, and Figure 29b shows empirical distributions for two post-allocation sub-groups. We observe that all the plotted distributions resemble log-normal distributions. Plots for



(a) Distribution of pre-allocation delay for delay-initiation time between 11 am and 1 pm, 1 pm and 3 pm, and between 3 pm and 6 pm



(b) Distribution of post-allocation delay for delay-initiation time between 2 pm and 5 am (next day), and between 10 am and 2 pm

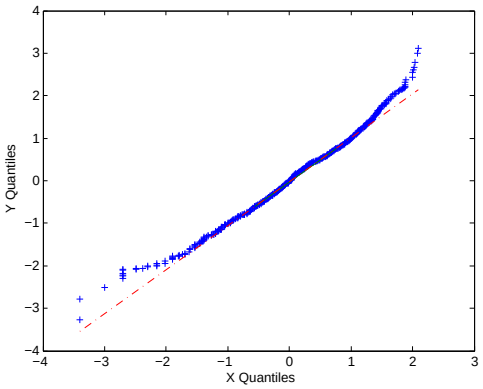
Figure 29: **Empirical distributions of allocation delays.** Bin size is 0.2 hour (12 minutes).

some other time intervals have a similar shape.

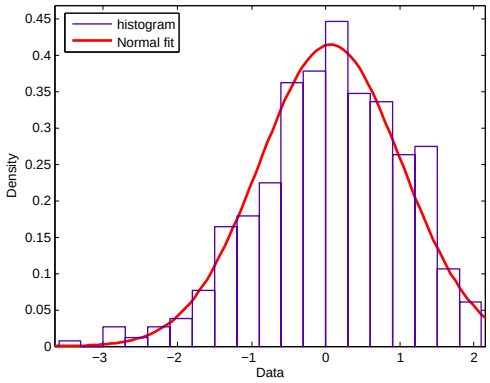
To test the reasonableness of the log-normal assumption, we perform log-transformation on the data points in each sub-group (for both allocation delays). Figures 30a and 31a show the Q-Q plot of the log-transformed data against normal distribution for the selected sub-groups. Figures 30b and 31b show histograms of the log-transformed data and fitted normal distributions. The figures suggest that the normal distribution curves are visually close to the empirical distribution curves. We observe similar features when analyzing the log-transformed data for other sub-groups. Although the fitting results cannot pass rigorous statistical tests (e.g., K-S test), these figures indicate that the log-normal assumption for pre- and post-allocation delays is still reasonable and is a good starting point for building models.

2.8 Internal transfers

Patients may go through one or more internal transfers after their initial admissions to GWs. Since we focus on GWs, we mainly consider two types of transfers: transfers between GWs and non-GWs (mostly ICU-type wards), and transfers between two GWs. In this section, we conduct empirical analyses on patients who have gone through at

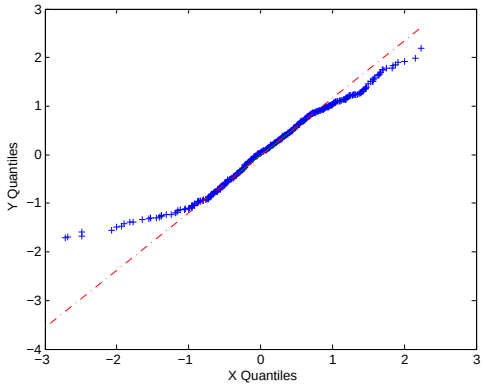


(a) Q-Q plot

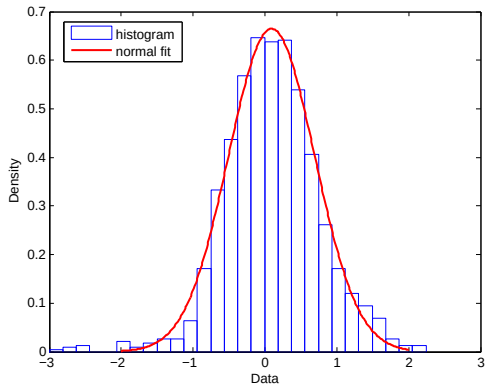


(b) Histogram and fitting with a normal distribution with mean 0.066 and std 0.96

Figure 30: **Fitting the log-transformed data for pre-allocation delay.** The delay-initiation time between 11 am and 1 pm.



(a) Q-Q plot



(b) Histogram and fitting with a normal distribution with mean 0.093 and std 0.60

Figure 31: **Fitting the log-transformed data for post-allocation delay.** The delay-initiation time between 10 am and 2 pm.

least one such transfer after the initial admissions. We first given an overview of all transfer patients in Section 2.8.1 with a focus on ED-GW and EL transfer patients. Then in Section 2.8.2, we show empirical results for patients transferred between GWs and ICU-type wards. Finally in Section 2.8.3, we focus on patients transferred between two GWs, including those who are initially admitted to a non-primary ward, and then transferred to a primary ward.

2.8.1 Overall statistics on internal transfers

Out of the total 94786 general patients admitted in Periods 1 and 2, 79687 (84%) patients have not been transferred after their initial admissions. The remaining 14840 patients (16%) have gone through at least one transfer. We call patients who has been transferred at least once the *transfer patients*.

Table 10 shows the proportion of transfer patients for each admission source and for each specialty. Clearly, the proportion of transfer patients depends on both the admission source and specialty. Comparing across admission sources, SDA patients have the smallest proportion of transfer patients (except Oncology). Comparing across specialties, Cardiology and Surgery have the highest proportion of EL transfer patients, whereas Oncology and Renal have a relative higher proportion of ED-GW transfer patients than other specialties. Also note that the proportions of transfer patients are close for ED-AM and ED-PM patients for most specialties except four belonging to the Medicine cluster: General Medicine, Renal, Neurology, and Gastro-Endo show a higher proportion of transfer patients for AM admissions than that of PM admissions.

Overall, transfer patients admitted from ICU-GW and SDA sources make up only a small proportion of all transfer patients (1500 out of 14840, or about 10%), and account for only 1.6% of all general patients. Thus, in the following analysis, we focus on transfer patients from ED-GW and EL sources.

Table 10: **Proportion of patients who have gone through at least one internal transfer for each admission source and for each speciality**. The combined data is used.

Specialty	ED-AM	ED-PM	EL	ICU	SDA
Surg	10.96%	9.35%	27.28%	15.18%	3.93%
Cardio	19.17%	15.94%	42.15%	11.59%	9.28%
General Med	17.80%	10.97%	15.71%	12.56%	2.94%
Ortho	14.84%	14.83%	21.03%	23.29%	5.94%
Gastro-Endo	22.19%	13.07%	8.20%	9.96%	6.64%
Onco	28.34%	23.96%	19.20%	11.21%	15.79%
Neurology	16.21%	12.23%	14.06%	12.87%	2.31%
Renal disease	30.46%	19.61%	14.78%	18.14%	6.67%
Respiratory	14.74%	14.00%	16.99%	10.09%	6.50%
Total	17.76%	13.63%	25.17%	12.62%	5.42%

ED-GW and EL transfer patients

Out of the total 77904 ED-GW and EL patients admitted in Periods 1 and 2, 13340 (17%) of them have been transferred at least once after initial admission. Out of these 13340 patients, 7285 patients (54.61%) have gone through one transfer; 4428 patients (33.19%) two transfers; and 905 patients (6.78%) three transfers. The remaining 722 patients (5.41%) have been transferred more than four times, and constitute less than 0.8% of the total general patients. Therefore, we also exclude them from the analysis below.

Now we study the "transfer paths" of these ED-GW and EL patients with one, two, and three transfers. Table 11 summarizes the patient count for each transfer path. We use "1" to denote a general ward, and "0" to denote a non-general ward. For example, path "1-0-1" means the patient is initially admitted to a GW, then transferred to a non-GW, transferred back to a GW, and finally discharged from the GW. We can see

(i) *One-time transfer:* Of the 7285 patients who transferred once, 1667 patients are transferred to a non-general ward (more than 60% to an ICU-type ward). The other 5618 are transferred to another general ward.

Table 11: **Decomposition of ED-GW and EL transfer patients by number of transfers and pathways**. The combined data is used. We use 1 to denote a general ward, and 0 to denote a non-general ward. In the last row, group I contains paths 1-0-0-0, 1-1-0-0, and 1-1-1-0; group II contains paths 1-0-0-1, 1-0-1-1, and 1-1-0-1.

# of transfers	total count	path and count			
1	7285	1-0		1-1	
		1667		5618	
2	4428	1-0-0	1-0-1	1-1-0	1-1-1
		114	4036	102	176
3	905	group I	group II	1-0-1-0	1-1-1-1
		44	707	130	24

(ii) *Two-time transfer:* Of the 4428 patients who transferred twice, the majority (4036 patients, 91%) follow the path of "1-0-1." In fact, more than 95% of the non-general wards (i.e., "0" in the path) belong to one of the ICU-type wards. Thus, we sometimes refer these 4036 patients as *GW-ICU-GW* patients. The remaining patients with paths "1-0-0" and "1-1-0" are those who initially stayed in GWs, and finally are discharged from a non-GW. Very few patients make two transfers between three GWs (following path "1-1-1").

(iii) *Three-time transfer:* Eight possible paths exist for the 905 three-time transfer patients. We aggregate some paths when displaying the patient count in Table 11. Group I (paths "1-0-0-0", "1-1-0-0", and "1-1-1-0") represents those patients who are initially admitted to a GW but discharged from a non-GW. There is no back and forth between GWs and non-GWs. Group II (paths "1-0-0-1", "1-0-1-1", and "1-1-0-1") represents those patients who are initially admitted to a GW, transferred to a non-GW during the stay, and finally discharged from a GW. Group II constitutes the majority of the 905 patients. Finally, the remaining two paths, 1-0-1-0 and 1-1-1-1, form their own group. Again, we can see that patients rarely make three transfers between four general wards.

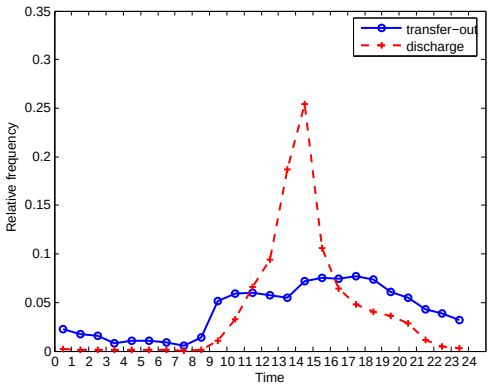


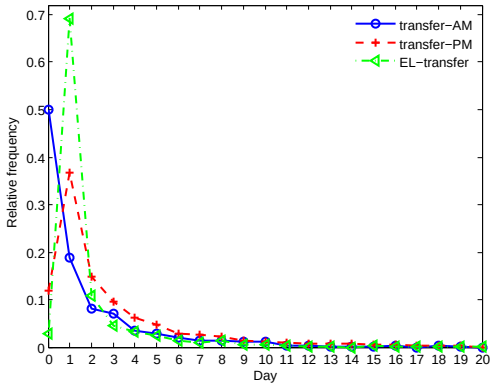
Figure 32: **Transfer-out time distribution for transfer patients from ED-GW and EL sources.** The transfer-out time distribution is estimated from the combined data. For reference, we also plot the Period 1 discharge distribution.

2.8.2 Transfer between GWs and ICU-type wards

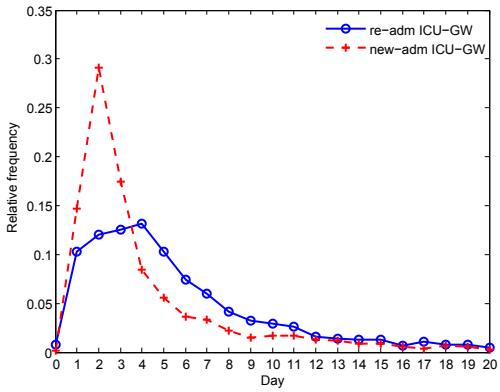
In this subsection, we focus on transfer between GWs and ICU-type wards. Specifically, we consider ED-GW or EL sourced patients who transfer once or twice between GWs and ICU-type wards after their initial admissions, i.e., patients following transfer paths "1-0" and "1-0-1." For each of these patient, his/her first visit to a GW starts from the initial admission time and ends at the first transfer-out time to a ICU-type ward. If he/she transfers twice, the second visit to a general ward starts from the transfer-in time (from ICU to GW) and ends at the final discharge time.

Figure 32 plots the empirical distribution of the transfer-out time (from GW to ICU) for these patients. We also plot the discharge time distribution in Period 1 for comparison. Unlike the discharge distribution which has a 2-3pm peak, the transfer-out time distribution spreads more evenly after 9am, indicating that operational factors such as staff schedule have less influence on the transfer-out time. This is reasonable since transfer from GW to ICU usually occurs when patient medical condition gets worse and is often very urgent.

Figure 33 plots the empirical LOS distributions for these transfer patients. We separately estimate the LOS of the first- and second-visit to GWs, i.e., number of nights in the corresponding visit. Specifically, Figure 33a shows the three first-visit



(a) First-visit LOS distributions for transfer ED-AM, ED-PM patients (ED-GW source), and EL-transfer patients



(b) Second-visit LOS distribution

Figure 33: **Estimated LOS distributions for transfer patients.** The combined data is used. For references, LOS distribution for non-transfer ICU-GW Cardiology patients is plotted in (b).

LOS distribution curves for transfer patients from ED-AM, ED-PM, and EL sources. Figure 33b shows the second-visit LOS distribution for all transfer patients who transferred twice. Clearly, the first-visit LOS is mostly 0-2 days, much shorter than the LOS of non-transfer ED-GW or EL patients. For the second-visit LOS, we also plot the LOS distribution curve from non-transfer ICU-GW patients for comparison. Although the path for the second visit and non-transfer ICU-GW patients is the same (from ICU to a GW), their LOS distributions are significantly different.

When empirically estimating the first-visit and second-visit LOS distributions, we exclude data entries from Orthopedic and Oncology specialties and aggregate entries from all other specialties together. We do the aggregation because (i) the empirical LOS distributions are close for patients from all specialties with the exception of Orthopedic and Oncology, and (ii) we do not have enough data points to get reliable estimation separately (for each specialty).

2.8.3 Transfer between two GWs

In this subsection, we consider the 5618 ED-GW and EL patients who transferred once between two GWs. We further separate these patients into two groups. The first group consists of those patients who are initially admitted to non-primary wards and are later transferred to a primary ward. The second group consists of the remaining patients, who are likely patients transferred from a wrong class ward to a right class ward (e.g., a subsidized patient transfers from a class A bed to a class B2 bed). The first group comprises 3133 patients; and the second group 2485 patients. Each group constitutes about 3% of the total volume of general patients.

Table 12 shows the number of patients transferred in ("flow-in") and transferred out ("flow-out") for each ward among these 5618 patients. Most wards show a balanced flow-in and flow-out volume. Certain wards, such as Ward 53, 55, and 63, receive more patients transferred in than the patients transferred out. For Ward 53 and 63, it is because these two often receive Medicine and Cardiology patients who are medically complicated from other Medicine and Cardiology wards, respectively. Orthopedic wards (Ward 51, 52, and 54) transfer out more patients than they receive, possibly because they tend to place the overflow patients back to the primary wards (recall the Orthopedic wards have high overflow proportions; see Section 2.3.3). The observation here suggests that the occupancy level in each ward would not be affected significantly by these transfers due to the balance between transfer-in and transfer-out volume, especially considering the total volume (5618 patients) is small comparing to the total volume of general patients.

Figure 34 plots the transfer-out time distribution for these 5618 patients. We can see more than 85% of the transfers occur between 2pm and 10pm, the same period when most discharges occur. This observation is consistent with NUH's policy to avoid non-urgent and unnecessary transfers unless there is a surfeit of beds.

Table 12: **Number of patients transferred in and out for each ward.** Here, we only consider the 5618 patients who transferred once between two GWs. The combined data is used. Two non-GWs, Ward 48 and 96, are included because some Surgery patients overflow to them.

Ward	41	42	43	44	48	51	52	53	54	55	56
Flow out	287	599	353	275	46	338	352	377	566	256	124
Flow in	87	42	288	147	70	57	118	800	197	785	151
Ward	57	57O	58	63	64	66	76	78	86	96	**total**
Flow out	126	165	267	141	669	277	101	199	80	20	**5618**
Flow in	220	458	434	398	632	298	190	156	81	9	**5618**

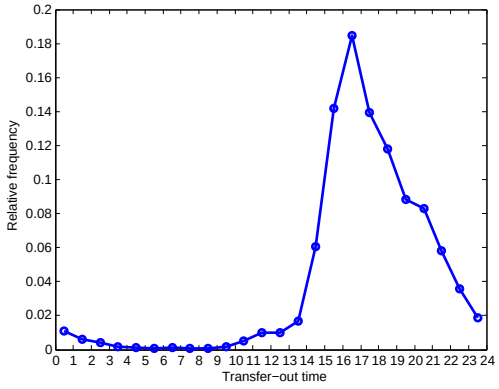


Figure 34: **Transfer-out time distribution for the 5618 patients who transferred once between two GWs.**

CHAPTER III

HIGH-FIDELITY MODEL FOR HOSPITAL INPATIENT FLOW MANAGEMENT

This chapter is organized as follows. In Section 3.1, we introduce the general framework of our proposed stochastic network model that captures the inpatient flow operations. In Section 3.2, we populate the proposed stochastic network model with NUH data. In Section 3.3, we verify the populated model by comparing the model output with empirical performance. In Section 3.4, we use the populated model to generate a number of managerial insights for reducing and flattening waiting times for admission to wards. This chapter concludes in Section 3.5.

3.1 A stochastic network model for the inpatient operations

In this section, we describe a general framework of our proposed stochastic model. Although the model is built upon the extensive empirical study of NUH inpatient operations we presented in Chapter 2, the framework could be adapted to other hospitals. We first give an overview of the basic ingredients of the stochastic processing network and the basic patient flow in Section 3.1.1. Then in Sections 3.1.2 to 3.1.4, we specify the details of three modeling features that are critical to capture inpatient operations. These features are a non-iid, two-time-scale service time model, an overflow mechanism, and pre- and post-allocation delays that create additional delay during patient's admission. Finally, we discuss service policies and an adjustment to incorporate patient transfer in Sections 3.1.5 and 3.1.6, respectively.

Under a specified service policy and a specification of input parameters estimated

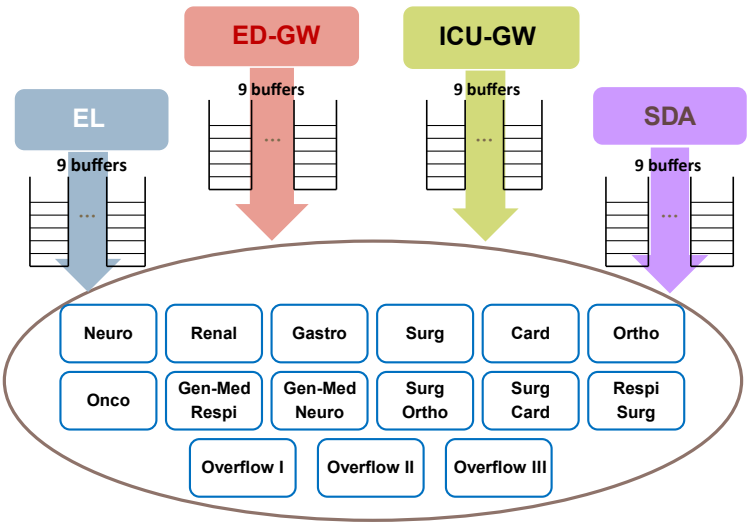


Figure 35: **Arrival and server pool configuration in the stochastic model of NUH inpatient department.**

from a hospital data set, the proposed stochastic model can be populated and simulated on a computer. Section 3.2 details how we populate the model using NUH data. Section 3.3 verifies the populated model by comparing the simulation output against the empirical estimates. We will see that our proposed stochastic model can approximately replicate waiting time performance, even at hourly resolution, from the empirical data.

3.1.1 A stochastic processing network with multi-server pools

Our proposed stochastic model is a variant of a stochastic processing network that was proposed in Harrison [72] and precisely specified in Dai and Lin [38]. A stochastic processing network processes incoming customers (patients) of various classes. The basic ingredients of a stochastic processing network are *servers*, *buffers*, *activities*, and *service policies*. Figure 35 depicts a stochastic processing network representation of the NUH inpatient department.

Servers. In this paper, general ward beds play the role of servers, and these servers are grouped into J *parallel* server pools. Each server pool models a general ward or a group of similar wards. We use n_j to denote the number servers in pool j,

$j = 1, \ldots, J$. These n_j servers are assumed to be identical. The J server pools serve customers from K different classes. Here, the customers are patients who need to receive hospital care in a general ward, and a customer *class* can be a combination of an admission source and a medical specialty, sometimes with other criteria such as admission time. Customers in the same class are homogeneous, following the same arrival process, service time specification, and service priority.

Buffers. In our model, each admission source is associated with an arrival process, which is used to model the patient bed-request process. In the rest of this chapter, we use patient and customer, bed-request and arrival, and bed and server interchangeably. Each arriving patient (from any of the admission sources) is assigned to a specialty with a certain probability that depends both on the source and the arrival hour. Each arriving patient is held in a buffer, waiting to be assigned a bed and later to be admitted into the bed. The patients waiting in these buffers are processed following certain priorities which are specified by a *service policy*.

Activities and service policies. Each server pool is designated to serve patients from one or more medical specialties, and we call the pool a *primary pool* for patients from the designated specialties. We assume each class of patients can potentially be assigned to any of the J server pools in the model. If a patient is assigned to a primary server pool, we say she is *right-sited*, otherwise, *overflowed*. Adapting the stochastic processing network terminology to the hospital setting, an *activity* is the binding of a server pool serving a particular class of patients. When the server pool is a primary pool for the class, the corresponding activity is said to be a *primary* activity. Clearly, primary activities are more desirable because they avoid patient overflow. However, to reduce waiting time, it is sometimes necessary to activate non-primary activities. A *service policy* dictates which activities should be initiated at a decision time point. In the hospital setting, a service policy is also known as a *bed assignment policy* that dictates which beds should be assigned to

which waiting patients at a decision time point. The decision time points have three categories: the arrival time of a patient, the departure time of a patient, and the overflow trigger time of a patient. A patient can be overflowed only when her waiting time exceeds her pre-assigned overflow trigger time. The service policy also dictates the choice of the overflow trigger time for each patient.

Basic patient flow. After a bed is assigned to a patient, she has to experience extra delays (pre- and post-allocation delays) before she can be admitted to the bed. Thus, a patient's admission time is different from her bed assignment time in our model. Once a patient is admitted, she occupies the bed until departure. The duration of occupation is called the patient's *service time*. The service time of each patient is random and follows the two-time-scale model (4) below. At the end of the service time, the patient departs from the system. Thus, our proposed stochastic network model has a *single-pass* structure. The departure times for most patients in our model corresponds to their discharge times from the hospital, and we use departure and discharge interchangeably in the rest of the paper.

3.1.2 Critical feature 1: a two-time-scale service time model

The service time, S, of a patient is the duration between the admission time (T_{adm}) and the discharge time (T_{dis}). We use day as the time unit for service times unless specified otherwise. Clearly, the service times of patients are random. Both the patient's medical condition and hospital operational policies can affect the service time. We adopt model (4) to separate different sources of influence on service times:

$$\begin{aligned} S &= T_{\text{dis}} - T_{\text{adm}} \\ &= (\lfloor T_{\text{dis}} \rfloor - \lfloor T_{\text{adm}} \rfloor) + (T_{\text{dis}} - \lfloor T_{\text{dis}} \rfloor) - (T_{\text{adm}} - \lfloor T_{\text{adm}} \rfloor) \\ &= \text{LOS} + h_{\text{dis}} - h_{\text{adm}}. \end{aligned} \tag{4}$$

We will discuss the rationale for using service time model (4) in Section 3.1.2 below. Here, LOS stands for length of stay and is equal to the number of midnights that the

patient spends in a ward, or equivalently, day of discharge minus day of admission, and h_{dis} and h_{adm} stand for the time of day when the patient is admitted and discharged, respectively. The time of day is between 0 and 1, with midnight being 0 day and 12pm (noon) being .5 day. For a patient who is discharged on the same day of admission, recall that our definition of her LOS is equal to 0, whereas when hospitals report occupancy level or some other statistics [68, 28], the LOS of such same-day discharge patients is adjusted to 1 for accounting and cost recovery purposes.

Non-iid service times

Based on the extensive empirical study on LOS and service time (see Sections 2.5 and 2.6 in Chapter 2), we make the following assumptions for the service time model in (4):

(a) The discharge hour h_{dis} is independent of LOS and of h_{adm}; Section B.2 in the appendix provides some empirical evidence for this assumption.

(b) LOS distributions are class dependent. Patients from different medical specialties or admission sources follow different LOS distributions.

(c) For each class of patients, their LOS forms a sequence of iid random variables following a discrete distribution. One can use an empirical LOS distribution directly estimated from data, or a discrete version of the log-normal distribution based on our empirical fitting results and similar findings in [5].

(d) The discharge hours h_{dis} for each class of patients forms another sequence of iid random variables following a certain discharge distribution. See Figure 2 in Section 1.1 for an example of NUH's discharge distribution.

(e) We assume all iid sequences of LOS and h_{dis} are independent of each other, i.e., there is no dependency among classes.

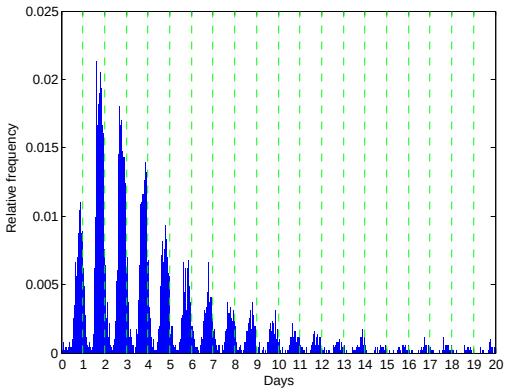
Note that for a class of patients, their admission hours h_{adm} are ordered and thus cannot be iid. Though the LOS and h_{dis} of these patients are two independent iid sequences, it follows from (4) that their service times are no longer exogenous variables and are *not* iid.

Separation of time scales

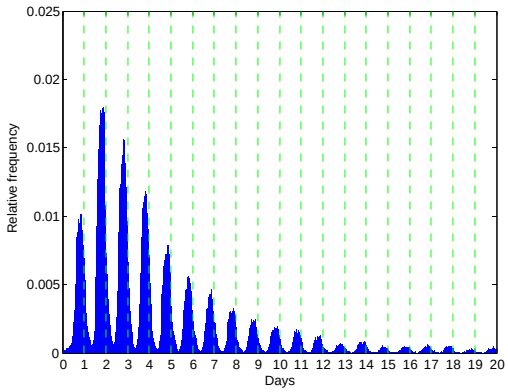
In the service time model (4), we use LOS to capture the number of nights that a patient *needs* to spend in the hospital, as a consequence of her medical conditions. We use the other two terms to capture the extra amount of time that is caused by operational factors. In particular, the discharge hour h_{dis} depends on discharge patterns that are mainly the results of schedules and behaviors of medical staff (also see Section 2.2.2). The way we model the service time allows us to evaluate a variety of policies that may affect the two parts of the service time (LOS versus $(h_{\text{dis}} - h_{\text{adm}})$) jointly or separately. For example, the early discharge policy implemented at NUH aims to reduce the operational bottlenecks and move the discharge hour h_{dis} to an earlier time of the day without affecting the patient's medical conditions (LOS), whereas expanding the capacity at a nursing home or a step-down care facility to ensure timely discharge of patients in need of long-term care will mainly affect the LOS term [14]. In Section 3.4, we use simulation to gain managerial insights into the impact of early discharge and other policies on the waiting time performance.

Moreover, this service time model captures an interesting phenomenon, the separation of time scales: the LOS is in the order of days, while $(h_{\text{dis}} - h_{\text{adm}})$ is in the order of hours. Indeed, we can observe these two time scales from Figures 36a below, which plots the empirical service time distribution at hourly resolution (also see the clustering phenomenon we discussed in Section 2.6). On the one hand, the distribution peaks at integer values representing 1, 2, 3, ... days, which is captured by the LOS. On the other hand, the sample points distribute around the integers mostly within

the range of a few hours, which is captured by the term $(h_{\text{dis}} - h_{\text{adm}})$. Figure 36b illustrates that our proposed service time model (4) can produce the distributions that resemble empirical distributions. The two time scales (hour versus day) have been discovered in other studies of hospital operations [5, 104, 125] and appointment scheduling [6].



(a) Empirical service time distribution for Period 1; each green dashed line corresponds to a 24-hour increment



(b) Service time distribution from simulation output; LOS and discharge distributions in the simulation are empirically estimated from Period 1 data

Figure 36: **Service time distributions, at hourly resolution, for General Medicine patients that are admitted in afternoons.**

3.1.3 Critical feature 2: bed assignment with overflow

In this section, we spell out the details for bed assignment under a specified service policy. In particular, we described the overflow mechanism in our model.

When a patient makes a bed-request, if a primary bed is available, that bed is assigned to the patient. When more than one primary pool has such a bed, a priority policy included in the service policy is used to decide which primary pool to select from.

If no primary bed is available at the bed-request time, the patient waits in a buffer and is assigned with an overflow trigger time T. The trigger time T may depend on the bed-request time, the admission source, and the specialty of the patient. An

overflow policy dictates the choice of T. The patient waits for a primary bed before her waiting time reaches T. After that, the patient can be assigned to either a primary bed or an overflow bed, whichever becomes available first.

Queueing implication and QED regime

Patients can be overflowed to a non-primary server pool only if her waiting time exceeds the trigger time T. When T is not 0, a bed can be idle even if a patient from a non-primary specialty has been waiting. Therefore, in our model the overflow policies are in general *idling*, which is different from the non-idling policies employed in many existing queueing models [90].

Overflow is an important measure for hospitals to balance the random demand and supply of different beds and to admit patients in a reasonably short time, given that it is difficult to adjust bed capacity among various specialties and wards in a short time window (this is in contrast to call center operations where the agents can be added or removed in a matter of hours). NUH data shows that the *partial* resource sharing from such overflow provides enough flexibility for hospitals to run in the Quality-and-Efficiency Driven (QED) regime, in which the average patient waiting time (in the order of few hours) is a small fraction of the average service time (in the order of days) and the bed utilization is high, say, $> 90\%$. A QED regime is usually gained by pooling a large number of servers (e.g., hundreds of beds) working in parallel and is difficult to be achieved by a small number of servers (e.g., 30 beds in a ward).

3.1.4 Critical feature 3: allocation delays

Motivated by the empirical observations in Section 2.7, we explicitly model operational delays that are caused by resource constraints (e.g., ED and ward nurses) other than bed unavailability during the ED to wards transfer process. Each patient in the model, even if a primary bed is available for her upon arrival, has to experience a

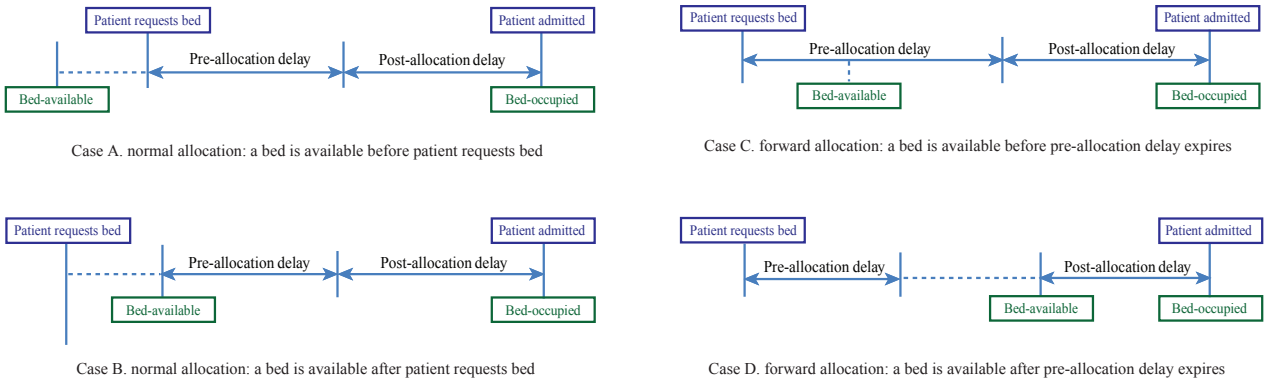


Figure 37: **Pre- and post-allocation delays under different scenarios.**

pre-allocation delay first, and then a *post-allocation delay* before being admitted to the bed. We first describe the process flow from a patient's bed-request to her admission to a bed in our model, and then explain the rationale of modeling the two allocation delays. Figures 37 illustrates the process with two allocation delays under various scenarios.

Patient flow from bed-request to admission

In our model, when a patient makes a bed-request, we assume two bed-allocation modes: *normal allocation* and *forward allocation*. The two modes differ from each other with respect to when the patient starts to experience a pre-allocation delay. In a normal allocation, the patient starts to experience a pre-allocation delay immediately at the bed-request time if a primary bed is available at that time (Case A in Figure 37). If no primary bed is available, the patient waits in a buffer for a bed. When a bed becomes available and is assigned to her, following the bed assignment policy described in Section 3.1.3, she starts to experience a pre-allocation delay (Case B in Figure 37). In a normal allocation, this pre-allocation delay always begins at or after the bed-available time.

A forward allocation is used only when there is no primary bed available at the patient's bed-request time (Cases C and D in Figure 37). The patient starts to experience a pre-allocation delay immediately at her bed-request time. In other words,

a pre-allocation delay always begins before a bed becomes available in the model. Therefore, sometimes a bed may still be unavailable when the patient finishes her pre-allocation delay stage.

In general, a patient starts to experience a post-allocation delay when the pre-allocation delay expires. The only exception is when the forward allocation mode is used and a patient finishes experiencing a pre-allocation delay but a bed is still unavailable (Case D in Figure 37). In this case, the patient waits until a bed becomes available for her, and a post-allocation delay starts at the bed-available time. When the post-allocation expires, the patient is admitted into the bed, completing the bed-request process.

We assume that a bed-request at time t, if there is no primary bed available, has probability $p(t)$ to be a normal allocation and probability $1-p(t)$ to be a forward allocation. We assume that the pre- and post-allocation delays are independent random variables following certain continuous distributions. The means of the distributions can be time-dependent, depending on when the patient requests a bed and starts to experience the allocation delays.

Rationale for modeling and other remarks

As we can see from Section 2.7, allocating a bed to an incoming patient is a process. We use the pre-allocation delay to model the time needed for the BMU to search and negotiate a bed for a patient from an appropriate ward, and use the post-allocation delay to model the delay from ED side after a bed is allocated and available to use for an incoming patient. The start and end points of the pre-allocation delay correspond to when a BMU agent starts and finishes the bed-allocation process, respectively. At the end of the bed-allocation process, a bed is allocated to the patient and NUH registers this time as the *bed-allocation* time. However, this bed-allocation time does not necessarily correspond to the time when a bed is assigned to a patient in our

model; the bed assignment in our model is specified in Section 3.1.3 and always happens at a patient's bed-request time, overflow trigger time, or discharge time. For example, if a primary bed is available upon a patient's bed-request, the bed assignment is instantaneously done in our model before the patient starts to experience the pre-allocation delay. The start point of the post-allocation delay corresponds to the allocation-completion time or the bed-available time, whichever is later, while the end point corresponds to the patient's admission time in practice.

Among the time stamps mentioned in the previous paragraph, NUH does not record when the bed-allocation process starts. According to our interviews, BMU agents normally wait until a bed becomes available before starting the bed-allocation process (which is close to the normal-allocation mode), or sometimes they can forward-allocate a bed based on the planned discharge information (which is close to the forward-allocation mode). We use the normal- and forward-allocation modes to approximate the reality. Additional empirical analysis in Section B.3 also supports this model setting. An alternative setting is to randomly assign this bed-allocation start time to occur between the bed-request and bed-available times following a certain distribution. We leave this extension to a future study.

3.1.5 Service policies

A service policy governs all of the decisions regarding bed assignments at various decision time points. It has four components: (i) how to pick a bed from a primary pool upon an arrival, (ii) how to pick a bed from a non-primary pool when a patient's overflow trigger time is reached; (iii) how to set an overflow trigger time; and (iv) how to pick a waiting patient from a group of eligible patients upon the departure of another patient. We elaborate each component below.

Component (i) specifies the priority of primary pools for each of the specialties having more than one primary pool. In general, dedicated pools (pools serving one

specialty) have higher priorities than shared pools (pools serving multiple specialties). Therefore, when seeking a primary bed for a patient, we start from the dedicated pools. If there is no dedicated bed free, we then search in shared pools.

Component (ii) specifies the priority of non-primary pools in overflowing patients. The priority depends on the specialty of the patient to be overflowed. In general, pools that serve similar specialties have high priority. Shared pools have higher priority than dedicated pools. Both components (i) and (ii) need to be estimated based on the actual configuration in the particular hospital being modeled.

Section 3.1.3 has introduced an overflow mechanism in our model. Component (iii) sets the overflow trigger time T for patients who have to wait because of the unavailability of primary beds upon their arrivals. When a patient's waiting time reaches the trigger time T, component (ii) is used to search for a non-primary bed for her. Different hospitals may adopt different overflow policies, and we will specify the time-dependent dynamic overflow policy adopted at NUH in Section 3.2.5.

Component (iv) is a patient priority list, which is used when a bed becomes available and needs to be assigned to one of the *eligible* patients. The eligible patients consist of both the primary patients and the overflow patients whose waiting times are greater than their overflow trigger times. Again, this component needs to be estimated according to each hospital's own situation. Generally speaking, patients who have waited longer than their overflow trigger times have a higher priority than those who have not.

3.1.6 Modeling patient transfers between ICU and GW

In a hospital, a *real* patient can be transferred between a GW and an ICU-type ward multiple times after her initial admission to the GW (see empirical results in Section 2.8). Since our proposed network has a single-pass structure, we do the following adjustments to incorporate such patient flows between GWs and ICU-type

wards.

We determine an arriving patient to be a *non-transfer* or a *transfer* patient upon her arrival according to certain Bernoulli distributions. A non-transfer patient corresponds to a real patient in the hospital who does not transfer between a GW and an ICU-type ward. The transfer patient construct is used to model the first stay in a GW of a real patient who transfers to an ICU-type ward after the initial admission. Thus, the discharge (departure) time of a transfer patient in the model corresponds to the real patient's transfer-out time, and her LOS and service time are adjusted accordingly.

A real patient who transfers back to a GW after her first transfer will have a second stay in the GW. To model that second stay, we create a pseudo-patient in the model. The admission time of this pseudo-patient corresponds to the transfer-in time (from an ICU-type ward to a GW) of the modeled real patient, and the discharge time of this pseudo-patient corresponds to the final discharge time of the real patient or the next transfer-out time if the real patient transfers out of the GW again. Thus, the service time of the pseudo-patient corresponds to the duration of the second stay of the real patient. Additional pseudo-patients can be created to accommodate triple or more transfers in a similar way.

In the model, we treat the pseudo-patients as ICU-GW patients regardless of the initial admission source of the corresponding real patients. That is because the admission process and admission time distribution of these pseudo-patients are close to those of the other ICU-GW patients according to our empirical analysis. To differentiate the two streams of ICU-GW patients, we call the pseudo-patients the *re-admitted* ICU-GW patients, and the others the *newly-admitted*. When an arrival from the ICU-GW source occurs in the model, we determine the arriving patient being newly-admitted or re-admitted according to certain Bernoulli distributions.

Note that a real patient can also transfer between two GWs, while our proposed

stochastic model does not incorporate such transfers. We leave this extension to future studies; also see discussion in Section 3.5.

3.2 Populated stochastic model using NUH data

Based on the empirical study presented in Chapter 2, we populate the proposed stochastic network model with NUH data, which we refer to as the *NUH model* in the rest of this chapter. In this section, we discuss how we empirically estimate all the necessary input for the NUH model. Unless stated otherwise, we always use Period 1 data to estimate the input, and the resulting NUH model is called the baseline scenario. Section 3.2.1 introduces the arrival processes for the four admission sources. Section 3.2.2 describes the server pool setting and the service policy. Section 3.2.3 discusses the empirical LOS and discharge distributions, while Section 3.2.4 introduces classification of patients based on the observations of LOS distributions. Sections 3.2.5 and 3.2.6 illustrates a dynamic overflow policy and time-varying allocation delays for the NUH model, respectively.

3.2.1 Arrivals

Time-varying arrival rates

As shown in Figure 35, patient arrivals to our model derive from four sources. For each source, the arrival rate depends on the time of day. For ED-GW, ICU-GW, and SDA patients, we use their empirical, hourly bed-request rates as their arrival rates in the NUH model. For EL patients, their arrivals are pre-scheduled. NUH has their admission times but lacks meaningful records of bed-request times. Thus, we use their empirical, hourly *admission* rates as their arrival rates in the NUH model. We assign EL patients the highest priority and set their allocation delays to be zero. In this way, the waiting times of EL patients in the NUH model are negligible, and hence their admission times are close to their bed-request times. Figure 13 shows the estimated hourly arrival rates for the four admission sources in the course of a

day. Note that in this figure, the daily arrival rate of each source is close to its daily admission rate shown in Figure 3a, except for the ICU-GW source since re-admitted patients are included here.

Arrival processes

A time-nonhomogeneous Poisson process for ED-GW patients. For the empirical bed-request process of ED-GW patents, we conducted a detailed study to test the assumption that it is a time-nonhomogeneous Poisson process (see Section 2.4.2). The test results suggest it is reasonable to assume that the bed-request process for ED-GW patients is nonhomogeneous Poisson. However, we find that the bed-request process is not a *periodic* Poisson process with either one day or one week as a period. In particular, the empirical coefficient of variation (CV) of the daily arrival rate for each day of week is much higher than 1, the theoretical CV under the Poisson assumption. We conjecture that the high variability comes from the seasonality of bed-requests and the overall increasing trend in the bed demand. In the NUH model, we assume that the ED-GW patient's arrival process is time-nonhomogeneous Poisson. We further assume that it is periodic with one day as a period. The arrival rate function of the periodic Poisson process is constant in each hour and is plotted as the solid curve in Figure 13. Note that setting one week as a period is another reasonable choice, and we discuss this extension as a future study to capture the day-of-week phenomenon in Section 3.5.1.

A non-Poisson arrival process model for other sources. The number of EL admissions each day is pre-scheduled at NUH. The bed-requests of ICU-GW or SDA patients are departures from the ICU-type or SDA wards, and their volumes are in a way also pre-scheduled on a daily basis: ICU physicians determine the number of patients to be transferred to general wards after the morning rounds each day, and then ICU nurses submit the bed-requests for these ICU-GW patients; similar

to EL patients, the SDA surgeries each day are scheduled in advance, and the SDA nurses submit bed-requests after SDA patients finish receiving surgeries on that day. (Also see the discussions in Section 2.4.3.) Based on this observation, we propose a non-Poisson arrival model for EL, ICU-GW, and SDA patients. We first generate a total number of A_k^j arrivals (to arrive in day k) from admission source j ($j = 1, 2, 3$, denoting EL, ICU-GW, and SDA, respectively) at the beginning of day k ($k = 0, 1, \dots$), where the value of A_k^j is randomly generated from the empirical distribution of the daily number of bed-requests for $j = 2, 3$ or daily admissions for $j = 1$. We then randomly assign the arrival times of A_k^j arrivals according to order statistics that draw from the empirical distribution of bed-request (or admission) times of source j. These distributions can be estimated from the arrival rate curves in Figure 13. Note that if the daily number of arrivals follows a Poisson distribution, the generated process is in fact a time-nonhomogeneous Poisson process with one day as the period [95].

3.2.2 Server pools and service policy

In the NUH model, there are 15 server pools. Table 13 lists the number of servers and the primary specialties for each server pool. This table is based on our empirical study at NUH (see Section 2.3). We slightly adjust the number of servers in certain server pools. The details of the adjustment are explained in Section B.4 of the appendix.

The service policy is built based on NUH's internal guideline [110], our empirical observations, and discussions with BMU staff. Specifically, Table 14 gives the priority table for components (i) and (ii) of the service policy discussed in Section 3.1.5. Component (iii), the overflow policy, will be elaborated in Section 3.2.5.

The priority list of component (iv) is given below. First, patients who have waited longer than their overflow trigger times have a higher priority than those who have not. This is aligned with NUH's goal of improving the 6-hour service level. Second,

Table 13: **Server pool setting in the simulation model.** Each row lists the server pool index, primary specialty, and number of servers. The three overflow wards are explained in Section B.4 of the appendix.

pool ID	primary specialty	no. of servers
0	Gen Med, Respi	41
1	Gen Med, Neuro	40
2	Renal	33
3	Neuro	12
4	Gastro-Endo	39
5	Surg	42
6	Card	40
7	Ortho	50
8	Onco	43
9	Respi, Surg	25
10	Surg, Ortho	38
11	Surg, Card	30
12	Overflow ward I	39
13	Overflow ward II	43
14	Overflow ward III	48
Total		563

Table 14: **Priority of primary and overflow pools in the simulation model.** In each row, pool numbers are ordered in decreasing priority.

Specialty	Primary	Overflow
Surg	5, 10, 11, 9	14, 12, 13, 7, 4, 1, 0, 2, 3
Card	6, 11	13, 14, 12, 4, 10
Gen Med	0, 1	14, 13, 4, 2, 3, 9, 10, 12, 8, 7, 11, 5, 6
Ortho	7, 10	12, 5, 14, 13, 4, 1, 2
Gastro-Endo	4	14, 13, 1, 0
Onco	8	13, 14, 1
Neuro	3, 1	14, 13, 4, 2, 0, 9, 10, 8, 7, 11
Renal	2	1, 4
Respi	9, 0	14, 13, 1, 4, 2, 3, 10, 8, 7, 11, 5

among the patients waiting longer than their overflow trigger times, those from the primary specialties have a higher priority than the ones from overflow specialties. Third, among patients from the same specialty, the ED-GW patients have a higher priority than ICU-GW and SDA patients, while ICU-GW and SDA have the same priority. This is based on the empirical observation that at NUH, ICU-GW and SDA patients have a much longer average waiting time than ED-GW patients (see Section 2.1.1). Also see [120] for a similar priority setting. Moreover, our model assumes that EL patients have the highest priority among all admission sources to account for using admission times as a proxy for bed-request times; see reasons in Section 3.2.1. Fourth, when patients are waiting in multiple buffers with the same priority or in a single buffer, we choose the patient with the longest waiting time.

3.2.3 Length of stay and discharge distributions

Non-transfer patients

Table 6 lists the empirically estimated mean and standard deviation of LOS for non-transfer patients from different admission sources and specialties. We use the empirical LOS distributions estimated from Period 1 data in the NUH model. As discussed in Section 2.5.3, admission source and specialty affect patient's LOS. Moreover, LOS distributions are also admission-period dependent for ED-GW patients. In the NUH model, we use empirical discharge distributions estimated from the data. The discharge distributions in the two periods are plotted in Figure 2.

Transfer patients

Section 3.1.6 explained how to incorporate the patient flows between GW's and ICU-type wards into the model. The transfer patients we include in the NUH model are real ED-GW or EL patients at NUH who transfer once or twice between GW's and ICU-type wards after the initial admission. We do not model (i) the real patients who are initially admitted from ICU-GW or SDA source and have been transferred;

and (ii) the real ED-GW or EL patients who have transferred more than two times. We exclude them because the volume of these patients is small. Therefore, only an ED-GW or EL patient in the NUH model will be assigned to be a transfer or non-transfer type upon her arrival. An ICU-GW patient, however, will be assigned to be newly-admitted or re-admitted upon her arrival.

We use the first-visit LOS and transfer-out times of the modeled real patients to estimate the LOS distributions and discharge distributions for the transferred ED-GW or EL patients, respectively. We use the second-visit LOS of the real patients who transferred twice to estimate the LOS distributions for the re-admitted ICU-GW patients. These LOS distributions are plotted in Figure 33 of Section 2.8.2. The discharge distribution of the re-admitted ICU-GW patients is the same as the one for the non-transfer patients (as in Figure 2). Figure 32 plots the discharge (transfer-out) distribution for all the transfer ED-GW and EL patients. We do not observe a significant difference between the two periods.

3.2.4 Patient class

Patients belonging to the same class are homogeneous, having the same LOS and discharge distributions. Our empirical evidence has shown that the LOS distributions depend on admission source, medical specialty, admission period (for ED-GW patients), and whether patients are transferred or not. We proceed in the following steps to determine a patient's class in the NUH model:

1. When an arrival from one of the four admission sources occurs, we assign this patient to one of the nine medical specialties, following an empirical distribution that depends on both the bed-request hour and admission source. Figure 3b plots the daily distributions of specialties and admission sources. After assigning the specialty, the service priority of the patient is determined. The following two steps make sure the LOS and discharge distributions are the same within

Table 15: **Estimated value for the parameter p of the Bernoulli distribution to determine patient classes.** For ED-GW and EL patient types, p represents the probability of being a transfer patient; for ICU-GW, p represents the probability of being a re-admitted patient. Parameters for specialties belonging to the Medicine cluster (Gen Med, Gastro-Endo, Neuro, Renal, Respi) are estimated together due to the limited number of data points, and we use Med to represent this group.

p	Surg	Card	Med	Ortho	Onco
ED-GW	4.58%	11.52%	4.78%	9.42%	5.69%
EL	23.46%	39.95%	4.53%	17.04%	6.01%
ICU-GW	45.10%	43.86%	16.98%	79.69%	39.86%

a class.

2. Next, we determine whether (i) an ED-GW or EL patient is a non-transfer or a transfer patient, (ii) an ICU-GW patient is newly-admitted or re-admitted, following a Bernoulli distribution which depends on the specialty. The parameters for these Bernoulli distributions are empirically estimated based on the relations between the patients in the model and real patients who have transferred (see Sections 3.1.6 and 3.2.3), and are listed in Table 15.

3. Finally, at an ED-GW patient's admission time, we determine her admission period (AM or PM). By now, the patient's class is fully determined.

3.2.5 A dynamic overflow policy

At NUH, there is a general guideline [110] on when and how to overflow a patient. Consistent with this guideline, an empirical study [144] suggests that the hospital overflows patients more aggressively during late night and early morning (before 7am). That is, NUH will overflow a patient almost immediately upon finding that no primary bed is available. The reason is that few discharges happen in this time period, so there is little chance that a primary bed will become available in the next few hours. Thus, there is no need to let the patient wait for another hour. In contrast, during other times, the hospital tends to be more conservative, and allows a patient to

wait some time prior to overflow in anticipation that a primary bed may become available soon. In this way, NUH has better control on the overflow proportion, another important performance metric being monitored (see Section 2.3.3). The preceding discussion suggests that the trigger time T should depend on the bed-request time. It is reasonable to assume that T is low when a bed-request occurs during late night or early morning, and high during other times.

Based on these observations, we use a simple dynamic overflow policy in the NUH model: when a patient requests a bed from 7am to 7pm, the overflow trigger time T is set to be $t_2 = 5.0$ hours, and for bed-requests in all other time periods, T is set to be $t_1 = 0.2$ hour. We choose 7am and 7pm as the starting and ending point to adopt the long overflow trigger time, respectively. This choice is based on observations from [144] and the practice at NUH. 7pm to 7am the next day is the night-shift period at NUH. A nurse manager is in charge of dealing with all bed-requests in this period. She has the authority to overflow patients without negotiation. The values of t_1 and t_2 are obtained through trial-and-error so that the simulation output curves in Figure 39 are as close to the empirical curves in the figure as possible. It is important to note that overflow decisions are very complicated [144], sometimes subjective, in practice. There is no data available for us to get an accurate estimation of the overflow trigger time. Thus, our proposed dynamic policy is an approximation of the real situation. Other variants of the overflow policy are possible, e.g., triggering an overflow event when the number of waiting patients exceeds a specified threshold, selecting the value of T based on the remaining service times of patients who are in service. We leave these extensions for future study.

3.2.6 Pre- and post-allocation delays

In this section, we focus on estimating allocation delays for ED-GW patients. We first explain how to model allocation delays for other patients. We assume the allocation

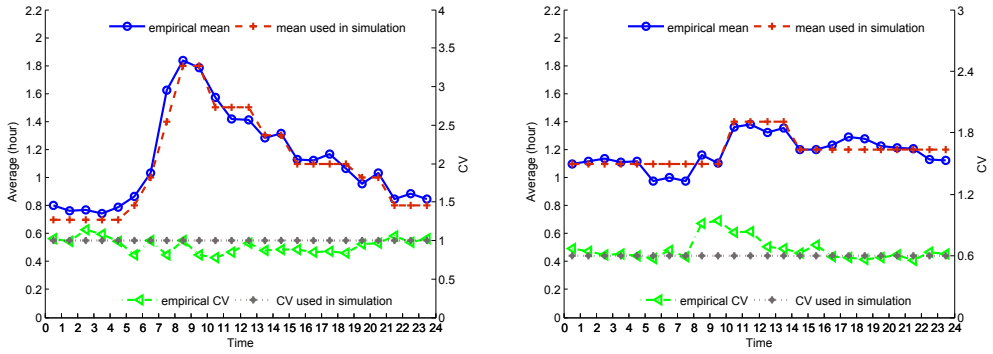
delays of the EL patients to be zero in the model, having explained the rationale of doing so in Section 3.2.1. For ICU-GW and SDA patients, we do not have good time stamps to estimate of their pre- and post-allocation delays reliably. We simply assume their allocation delays follow the same distributions as the ones used to generate the allocation delays for ED-GW patients. Sensitivity analysis shows that a moderate amount of change to the allocation delay distributions of ICU-GW and SDA patients will not affect the overall performance of ED-GW patients.

Distributions of the time-dependent allocation delays

In the NUH data set, at the bed-request time of an ED-GW patient either (i) the allocated bed is already available for the patient or (ii) the bed is not available and is still occupied by another patient. Case (i) corresponds to Case A in Figure 37, and we select a subset of case (i) patients in the data set to estimate the pre-allocation delay distribution. The subset consists of case (i) patients whose allocated beds are from their primary wards. By selecting this group of patients, we try to minimize the influence of bed shortage and specialty mismatch on pre-allocation delay so that our estimation can reflect the minimum time needed for BMU agents to allocate a bed. For the post-allocation delay, there is no such influence and we include all ED-GW patients to estimate its distribution. Also see Section 2.7.3 in Chapter 2 for The selected samples and estimating details.

The empirical histograms and distributional fitting results suggest that using a log-normal distribution is a good starting point for modeling each of the allocation delays. Thus, our model assumes the pre- or post-allocation delay initiated within each hour of a day to be a iid random variable that follows a log-normal distribution. The mean and CV of the log-normal distribution depends on the initiation hour (i.e., the hour when the allocation-delay starts).

Figures 38a and 38b plot the empirical estimates of the mean and CV for the



(a) Pre-allocation delay

(b) Post-allocation delay

Figure 38: **Mean and CV of pre- and post-allocation delays used in the simulation model.** Left vertical axis is for the average; right vertical axis is for the CV. The scale of the right vertical axis is deliberately chosen to be large, so that the four curves are not crossed over.

pre-allocation and post-allocation delays, respectively. In our baseline simulation scenario, we use the two dashed curves denoted with a plus sign as the inputs for the time-dependent mean and CV for each allocation delay, respectively. These two curves are slightly smoother than (but still within the 95% confidence intervals of) the corresponding empirical curves, which have random noise since the sample sizes in certain time intervals are small, particularly between 8am and 10am.

Estimating the normal allocation probabilities $p(t)$

Recall from Section 3.1.4 that in the model, when a patient makes a bed-request at time t and there is no primary bed available at the time, we assume with probability $p(t)$ the allocation for the patient is a normal allocation, meaning this patient will wait until a bed is available before starting to experience the pre-allocation delay. Unfortunately, the NUH data set do not have accurate time stamps to allow us

estimate $p(t)$ reliably. In our baseline scenario, we choose

$$p(t) = \begin{cases} 0 & h(t) \in [0, 6), \\ .25 & h(t) \in [6, 8), \\ 1 & h(t) \in [8, 12), \\ .75 & h(t) \in [12, 14), \\ .5 & h(t) \in [14, 20), \\ 0 & h(t) \in [20, 24), \end{cases} \tag{5}$$

where $h(t)$ stands for the hour of the day of the bed-request time t. The choice of $p(t)$ is based on the current practice at NUH and empirical estimation of the proportion of patients whose bed-allocation process approximately corresponds to the normal-allocation mode in the model. Section B.3 in the appendix discusses the details of estimating $p(t)$ in different time intervals. We realize that, despite our best efforts, our choice of $p(t)$ using (5) is still ad hoc. We report a sensitivity analysis of the choice of $p(t)$ in Section 3.4.4.

3.3 Verification of the populated NUH model

Recall that the populated NUH model, using the input described in Sections 3.2.1 to 3.2.6, is referred to as the *baseline scenario*. In Section 3.3.1, we first show the simulation output from the baseline scenario matches several key empirical performance measures. Then in Section 3.3.2, we show that the simulation output from each model which misses one of the three critical features introduced in Section 3.1 cannot replicate the empirical performance measures.

To implement these models, we wrote simulation code in `C++` language. For each simulation run, we start from an empty system and simulate for a total of 10^6 days. We then divide the simulation output into 10 batches. The performance measures are calculated by averaging the last 9 batches, with the first batch discarded to eliminate

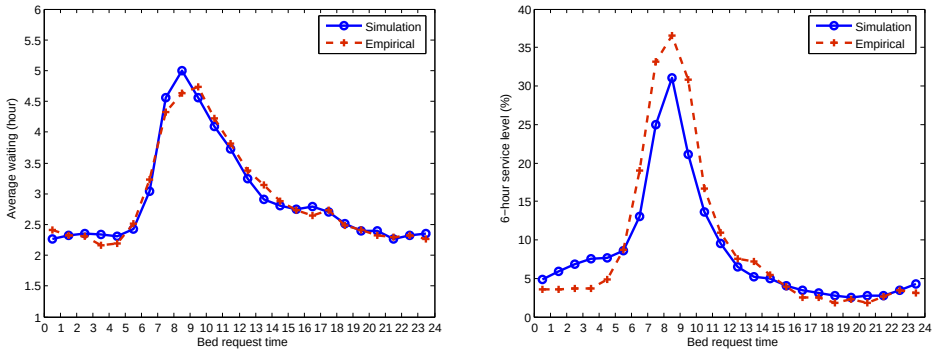
transient effects. Unless otherwise specified, all simulation estimates in this paper are from simulation runs under this setting. The choice of the simulation setting is justified following standard techniques in the literature [91]. Note that in this and the next section, we rely on simulation to obtain the desired performance measures, because there is no existing analytical tool to analyze the proposed stochastic model either exactly or approximately. As mentioned in the introduction, this chapter focuses on establishing a high fidelity model that can capture the inpatient flow dynamics at hourly resolution. We develop a two-time-scale analytical framework to analyze some simplified version of the proposed model in Chapter 4.

3.3.1 The baseline scenario

Recall that the inputs for the baseline scenario are estimated from NUH Period 1 data. Thus, we compare the outputs from this scenario against the empirical performance in Period 1 to verify the NUH model. From simulating the baseline scenario, the daily average waiting time for all ED-GW patients is 2.82 hours and the daily 6-hour service level is 6.29%, close to what we observed empirically in Period 1. Furthermore, Figure 39 shows that the simulation estimates approximately replicate the empirical estimates of the time-of-day (hourly) waiting time performance for all ED-GW patients. Table 16 compares the simulation estimates with the empirical estimates of the average waiting time and the 6-hour service level for each specialty. We can see that the waiting time statistics, even at the specialty level, can be approximately replicated by our simulation.

Besides the waiting time, we can also approximately replicate other key performance measures. The utilization rate is 89.2% from simulation, a little bit higher than the 88.0% empirical utilization in Period 1. Figure 40a plots the hourly average queue length for all ED-GW patients for both simulation and empirical estimates.

We point out that our model cannot perfectly replicate the overflow proportion.

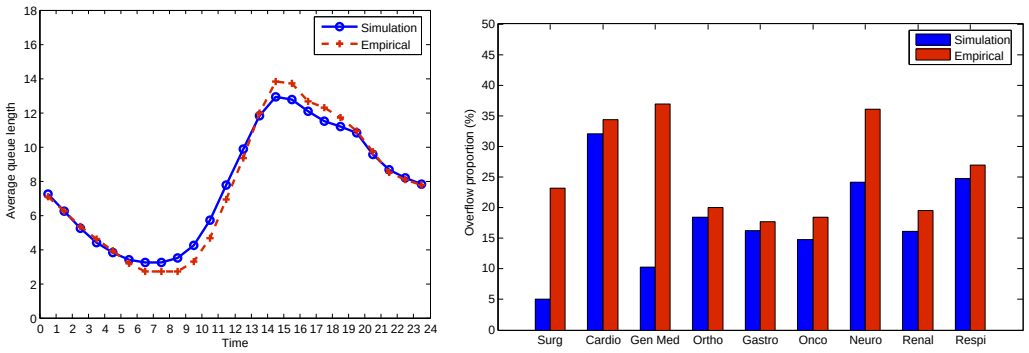


(a) Hourly average waiting time (b) Hourly 6-hour service level

Figure 39: **Baseline simulation output compares with empirical estimates.** This figure compares hourly average waiting time and 6-hour service level between empirical estimates and simulation estimates from the baseline scenario. Empirical estimates are from Period 1 data.

Table 16: **Simulation and empirical estimates of waiting time statistics for ED-GW patients from each specialty.** The simulation estimates are from simulating the baseline scenario, and the empirical estimates are from Period 1 data. The numbers in the parentheses are for the 95% confidence interval of the corresponding value. The confidence intervals for the simulation output are calculated following the batch mean method [91]; the confidence intervals for the empirical statistics are calculated with the standard deviations and sample sizes from the actual data.

	average waiting time (hour)		6-h service level (%)	
Specialty	simulation	empirical	simulation	empirical
Surg	2.64 (2.63, 2.64)	2.61 (2.56, 2.65)	4.85 (4.80, 4.90)	5.45 (4.87, 6.02)
Card	2.97 (2.97, 2.98)	3.08 (3.03, 3.13)	6.81 (6.75, 6.87)	8.36 (7.63, 9.10)
Gen Med	2.73 (2.72, 2.74)	2.64 (2.60, 2.68)	5.39 (5.34, 5.44)	4.79 (4.32, 5.26)
Ortho	2.73 (2.72, 2.73)	2.79 (2.74, 2.85)	5.22 (5.17, 5.28)	5.84 (5.16, 6.53)
Gastro	2.88 (2.88, 2.89)	2.97 (2.90, 3.04)	8.07 (8.00, 8.14)	7.64 (6.73, 8.56)
Onco	2.88 (2.87, 2.88)	2.96 (2.86, 3.07)	7.58 (7.53, 7.64)	8.15 (6.81, 9.50)
Neuro	2.84 (2.83, 2.85)	2.81 (2.75, 2.88)	6.49 (6.43, 6.55)	5.93 (5.04, 6.83)
Renal	3.23 (3.22, 3.24)	3.41 (3.32, 3.51)	10.5 (10.4, 10.5)	11.6 (10.3, 12.9)
Respi	2.82 (2.81, 2.82)	2.77 (2.68, 2.85)	6.25 (6.18, 6.31)	5.50 (4.36, 6.63)
All	**2.82** (2.81, 2.82)	**2.82** (2.80, 2.84)	**6.29** (6.24, 6.34)	**6.52** (6.26, 6.78)



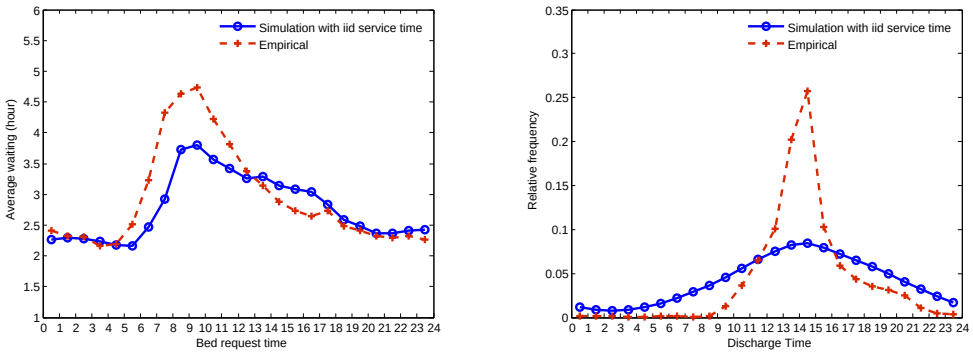
(a) Average queue length (b) Overflow proportion

Figure 40: **Baseline simulation output compares with empirical estimates.** This figure compares hourly average queue length and overflow proportion between empirical estimates and simulation estimates from the baseline scenario. Empirical estimates are from Period 1 data.

Although the simulated overflow proportions for most specialties are close to their empirical counterparts (see Figure 40b), the baseline simulation underestimates the overflow proportions for Surgery, General Medicine, and Neurology specialties. The underestimation in these three specialties leads to an overall underestimation of overflow proportion across all specialties (16.35% in the baseline versus 26.95% from Period 1 data). Moreover, there are certain performance measures that we choose not to calibrate in the model, including the waiting time statistics for ICU-GW and SDA patients. As mentioned, the waiting time statistics for these patients are not our primary focus. Moreover, sensitivity analysis shows that whether or not we can accurately replicate their waiting times has little impact on the waiting time statistics of ED-GW patients. Readers are referred to Sections B.9 and B.6for more discussion on the challenges in calibrating overflow proportions and results of sensitivity analysis, respectively.

3.3.2 Models missing any of the critical features

To show the necessity of modeling the three critical features discussed in Section 3.1 (i.e., the two-time-scale service times, overflow mechanism, and allocation delays), we



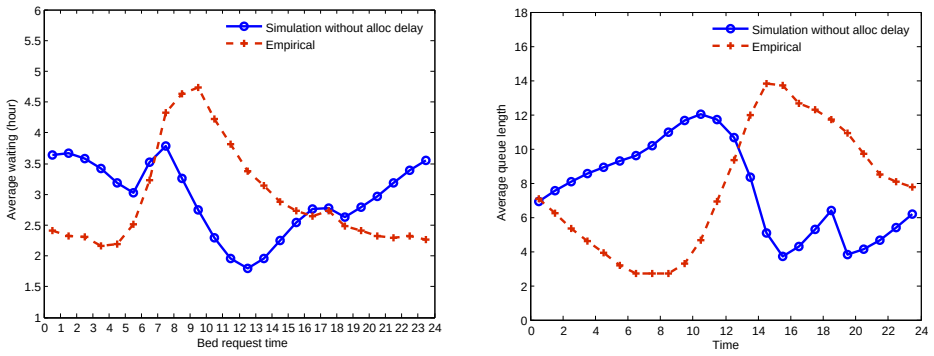
(a) Average waiting time (b) Discharge distribution

Figure 41: **Simulation output from using an iid service time model.**

simulate three versions of the model, each missing one of the critical features. All other input settings for the three versions remain the same as we simulate the baseline scenario unless otherwise specified. Again, we compare the simulation estimates against the empirical performance in Period 1.

Model with conventional iid service times

The two-time-scale service time model proposed in Section 3.1.2 is contrary to the exogenous, iid service time model often used in the queueing literature. We compare an iid service time model with our proposed non-iid model. The iid model assumes the service time S to be the sum of two independent random variables: an integer variable corresponding to the floor of service time $\lfloor S \rfloor$, and a residual variable corresponding to $(S - \lfloor S \rfloor)$. For patients from the same class, we assume their integer parts and residual parts each form an iid sequence based on the empirical evidence. Since the two sequences are independent, the service times are iid. Even though this iid exogenous service time model can reproduce service time distributions such as the one in Figure 36a, it is not able to reproduce the discharge distribution and hourly waiting time statistics; see the simulation output in Figure 41 for an illustration. Therefore, we believe that our new two-time-scale service time model is an important feature to capture inpatient flow operations. Sections 2.6.2 and B.1 contain detailed



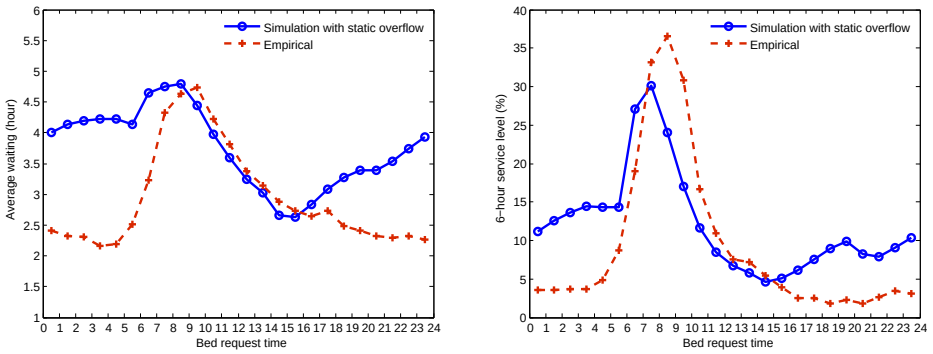
(a) Average waiting time (b) Average queue length

Figure 42: **Simulation output from a model without allocation delays.**

empirical observations and discussion of the iid service time model.

Model without allocation delays

Figure 42 compares the simulation and empirical estimates of the hourly average waiting time and hourly average queue length for ED-GW patients. In the simulation setting, *no* allocation delays are modeled. We can see that the hourly performance curves from simulation are completely different from the empirical estimates. In particular, note that the solid curve in Figure 42b, which shows a rapid drop in the simulated average queue length between 11am and 3pm, contrasts sharply with the empirical (dashed) curve, which drops slowly after 2pm. The main reason for the rapid drop in the solid curve is that in Period 1, between 11am and 3pm, the discharge rate increases in each hour until reaching the peak at 2-3pm (see Figure 2), and a waiting patient in the simulation is admitted into service immediately once a discharge occurs. Thus, Figure 42 suggests the existence of extra delays after bed discharges. In this simulation scenario, to make the daily average waiting time comparable to the estimate from the baseline scenario (2.82 hours) we decrease the numbers of servers listed in Table 13 while keeping all other settings the same as the baseline scenario.



(a) Average waiting time (b) 6-hour service level

Figure 43: **Simulation output from using a static overflow policy.** In the static overflow policy, we use a fixed overflow trigger time $T = 4.0$ hours.

Model without the dynamic overflow policy

Section 3.1.3 has discussed the important role of overflow in achieving a QED regime for hospital operations. Furthermore, we find that adopting a dynamic overflow policy is also critical to replicate the empirical performance at NUH. Figure 43 compares empirical estimates of the hourly waiting time statistics with simulation estimates from a model with a static overflow trigger time $T = 4.0$ hours. Clearly, the model with a static overflow policy fails to capture the dynamics in NUH inpatient operations. In particular, note that under the static overflow policy, the simulation estimates of the average waiting time for patients arriving in the night (10pm to 5am the next day) are about 4 hours, much higher than the empirical estimates. It is because in the simulation these night arrivals have to wait at least 4 hours for an overflow bed if no primary bed is available upon their arrivals, even though a new primary bed is unlikely to become available within 4 hours due to the discharge pattern.

3.4 Factors that impact ED-GW patients' waiting time

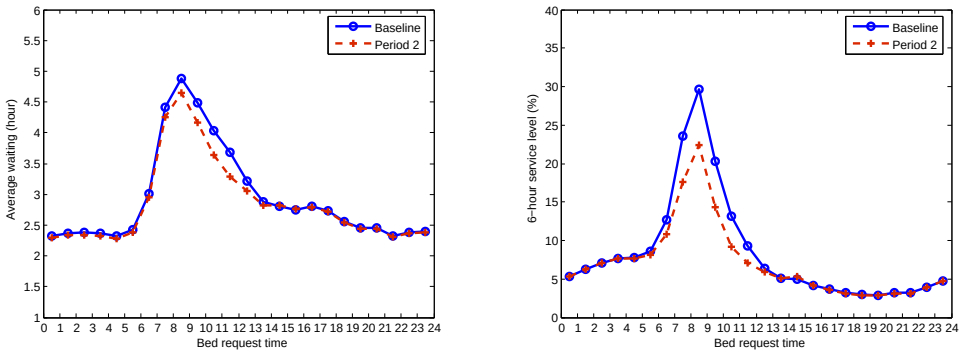
We do "what-if" analyses in this section and address the research questions raised in the introduction (see Section 1.1), i.e., (i) quantify the impact of the NUH Period 2 early discharge policy and (ii) identify operational policies that can reduce the waiting

times of ED-GW patients. We focus on the impact of the tested policies on both the daily and hourly waiting time performance. In Section 3.4.1, we show that early discharge in Period 2 has little impact on the daily and hourly waiting time statistics. In Section 3.4.2, we show that a hypothetical Period 3 policy can flatten the hourly waiting time performance, but has limited impact on reducing the daily waiting time statistics. In Section 3.4.3, we study policies that mainly impact the daily waiting time performance, such as increasing bed capacity and reducing LOS. In Section 3.4.4, we show that most of our gained insights are robust under sensitivity analysis. Finally, we explain in Section 3.4.5 why these policies have different impact on the daily and hourly waiting time performance.

3.4.1 Period 2 discharge has a limited impact on reducing waiting time statistics

To evaluate the impact of NUH's Period 2 early discharge policy, we simulate a scenario with the same inputs as in the baseline scenario, but using the discharge distribution estimated from Period 2 data (i.e., using the dashed curve in Figure 2 instead of the solid curve). Figure 44 compares the simulation estimates of hourly waiting time statistics with those from the baseline scenario. From Figure 44a, the hourly average waiting times show little difference between the two scenarios. From Figure 44b, the hourly 6-hour service level exhibits some reduction for bed-requests between 7am and 11am, e.g, the peak value is now 22% compared to 30% in the baseline scenario, but the values for other hours are almost identical in both scenarios. Not surprisingly, other performance measures from these two scenarios are almost identical. The daily average waiting time under this early discharge scenario is 2.75 hours, a 4-minute reduction, versus 2.82 hours in the baseline scenario. The 6-hour service level is 5.64% versus 6.29% in the baseline scenario. The overflow proportion is 16.26%, not significantly different from the baseline value of 16.35%.

To summarize, our model predicts that the Period 2 early discharge policy has



(a) Average waiting time (b) 6-hour service level

Figure 44: **Comparing hourly waiting time statistics under the baseline scenario and scenario with Period 2 discharge distribution.**

limited impact on reducing daily waiting time statistics and overflow proportions at NUH, and that this policy alone cannot flatten the waiting time performance throughout the day even though it helps to reduce the peak hourly 6-hour service level. This prediction is consistent with our empirical observations of performance in Periods 1 and 2; e.g., see Figure 1.

3.4.2 A hypothetical Period 3 policy can have a significant impact on flattening waiting time statistics

We consider a hypothetical discharge distribution, which still discharges 26% of patients before noon as in Period 2, but shifts the first discharge peak time to 8-9am, i.e., three hours earlier than the first discharge peak time in Period 2. Figure 45 plots this hypothetical discharge distribution. In addition, we assume a hypothetical allocation delay model: each allocation delay (pre- or post-allocation delay) follows a log-normal distribution with a *constant* mean, which is estimated from the empirical daily average. The estimated means of the pre- and post-allocation delays are 1.07 and 1.20 hours, respectively. We keep the same values of CV as in the baseline scenario, i.e., CV = 1 and 0.6 for the pre- and post-allocation delays, respectively. We call the combination of the hypothetical discharge distribution and the hypothetical allocation delay model a *Period 3 policy*. The Period 3 policy has not been

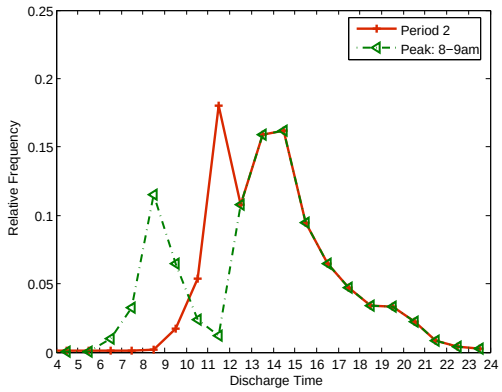
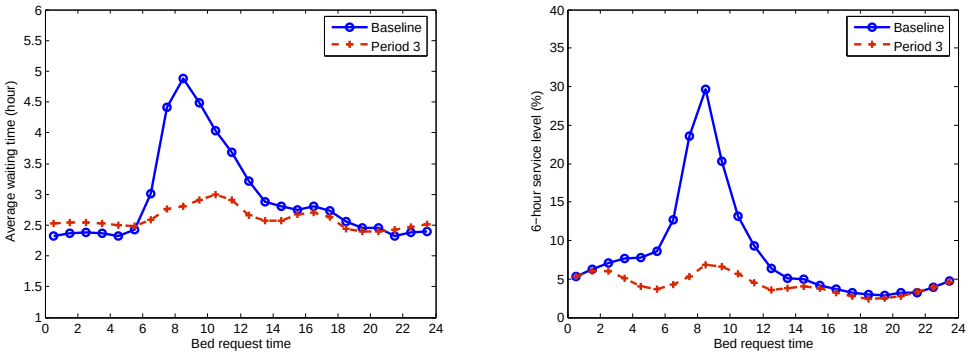


Figure 45: **Period 2 discharge distribution and a hypothetical discharge distribution.** In the hypothetical discharge distribution, the first peak is at 8-9am and 26% patients discharge before noon. Each dot represents the fraction of patients who are discharged during that hour. The values in the first 4 hours are nearly zero and are not displayed.

implemented yet at NUH and may not be fully practical. We call it Period 3 policy because it has the potential to be implemented in the future. For consistency, we call the combination of the Period 2 discharge distribution (which has been implemented) and the time-varying allocation delay model (see Section 3.2.6) the *Period 2* policy.

Figure 46 compares the hourly waiting time statistics between the baseline scenario and the hypothetical Period 3 scenario. Under the Period 3 policy, patients requesting beds in the morning (7am to noon) experience similar average waiting times (2.76 to 2.99 hours) as the daily average (2.59 hours), but the daily average is only 13 minutes lower than the daily average in the baseline scenario. The peak value of the hourly 6-hour service level drops from 30% under the baseline scenario to 6.9% under the Period 3 policy, with the daily 6-hour service level dropping from 6.29% to 4.02%. The overflow proportion drops slightly, from 16.35% under the baseline scenario to 15.69% under the Period 3 policy.

Compared to the Period 2 policy, the Period 3 policy requires achieving both a more aggressive early discharge distribution and allocation delays that are time-stable with constant means throughout the day. Simulation results show that when either component is missing (only implementing the aggressive early discharge policy or



(a) Average waiting time (b) 6-hour service level

Figure 46: **Comparing hourly waiting time statistics under the baseline scenario and scenario with Period 3 policy.** Period 3 policy: hypothetical discharge distribution with first peak at 8-9am and constant mean allocation delays.

only stabilizing the allocation delays), the average waiting times for morning bed-requests are still about 1-2 hours longer than the daily average and the waiting time performance is not approximately flattened.

In summary, our model predicts that this hypothetical Period 3 policy can eliminate the excessively long waiting times for ED-GW patients requesting beds in the morning. Simultaneous implementation of both the aggressive early discharge policy and allocation delay stabilization is necessary for the Period 3 policy to achieve an approximately time-stable performance in waiting times. However, the Period 3 policy has limited impact on reducing the daily waiting time statistics and overflow proportion in the NUH setting.

Findings from other early discharge scenarios

To obtain more insights into the impact of discharge timing, we test other hypothetical discharge distributions combined with the time-varying or constant-mean allocation delay models. Section B.5 in the appendix details the tested policies and simulation results. We highlight two main observations here that are consistent with what we see from the Period 3 policy.

First, an early discharge policy mainly impacts the time-of-day pattern of the

waiting time performance. Several tested combinations of early discharge distributions and constant-mean allocation delays can flatten the waiting time performance. Moreover, we find that the timing of the first discharge peak has a major impact on flattening the performance. For example, if the hospital retains the first discharge peak time between 11am and noon as in Period 2, even pushing 75% patients to discharge before noon and stabilizing the allocation delays cannot approximately flatten the waiting time performance.

Second, an early discharge policy has limited impact on the daily average waiting time and overflow proportion in the NUH setting. In particular, we test a discharge distribution with every patient discharged at as early as midnight to study the largest improvement that an early discharge policy might bring. When the mean allocation delays are constant, the daily average waiting time under this extreme early discharge scenario is 2.42 hours (a 24-minutes reduction from the baseline scenario) and the overflow proportion is 14.00%.

3.4.3 Policies impact on the daily waiting time statistics and overflow proportion

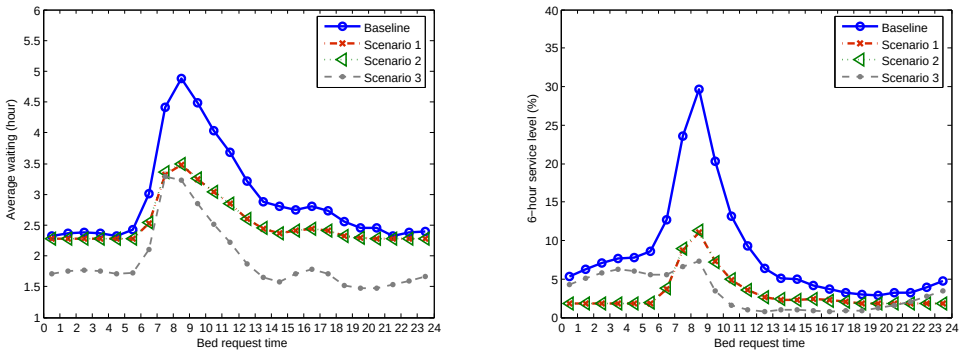
In this section, we show three policies that can significantly reduce the daily waiting time statistics and overflow proportions. They are increasing the bed capacity, reducing the LOS, and reducing the mean allocation delays. Specifically, we consider three scenarios. In the first one, we increase the number of servers from 563 (baseline) to 632 so that the utilization rate is reduced from 89.2% to 79.4% (a 10 percentage point reduction). In the second one, we eliminate excessively long LOS by limiting each patient to stay in the hospital for a maximum of 14 days. The utilization rate is thereby reduced to 78.5%, close to that in the first scenario. In the third scenario, we reduce the mean pre- and post-allocation delays by 30 minutes each. In each scenario, we use the baseline (Period 1) discharge distribution and assume the constant-mean allocation delay model; all other settings not specified here remain the same as in the

baseline scenario.

The daily average waiting times are 2.45, 2.46, and 1.80 hours in the first, second, and third scenario, respectively. The daily 6-hour service levels are 2.60%, 2.60%, and 2.31%, respectively; and the overflow proportions are 8.19%, 8.17%, and 15.94%, respectively. Figure 47 plots the hourly waiting time statistics under the three scenarios. Comparing to the baseline scenario, a 10% capacity increase results in a significant reduction in the overflow proportion (a 8% *absolute* reduction) but only reduces the daily average waiting time by 22 minutes. Reducing the LOS shows a similar impact since it is essentially equivalent to creating more capacity. A total one hour reduction in the mean pre- and post-allocation delays leads to a one hour reduction in the daily average waiting time, while it has limited impact on reducing the overflow proportion. In all three scenarios, the daily 6-hour service levels are significantly reduced.

Furthermore, we see from the figure that in all three scenarios, the hourly average waiting time is not stabilized, i.e., the average waiting time for patients requesting beds between 7am and 11am is still about 1-2 hours longer than the daily average. The hourly 6-hour service level, though, appears to be more time-stable than the average waiting time for each scenario, especially considering that the peak value is 30% in the baseline. Note that until we increase the bed capacity to 707 beds (utilization rate reduces to 71.0%), the waiting time curves can be approximately stabilized; whereas reducing the mean allocation delays down to 0 still cannot achieve a time-stable performance.

In summary, our model predicts that reducing the mean allocation delays can significantly reduce the daily average waiting times, while increasing the bed capacity or reducing the LOS mainly impact the overflow proportion in the NUH setting. Moreover, these policies have less impact on the time-of-day pattern of waiting time performance and they do not necessarily flatten the hourly waiting time performance.



(a) Average waiting time (b) 6-hour service level

Figure 47: **Hourly waiting time statistics under three scenarios.** Scenario 1: increasing bed capacity by 10%; Scenario 2: assuming each patient can stay in the hospital for a maximum of 14 days; Scenario 3: reducing the mean pre- and post-allocation delays by 30-minutes for each. In each scenario, we use the Period 1 discharge distribution and assume the constant-mean allocation delay model.

Our results suggest that at NUH the 2-3 hours average waiting time mainly comes from secondary bottlenecks other than bed unavailability, such as inadequate nurse staffing. This finding is consistent with the observation in a recent paper that the long waiting time of ED-GW patients may not be caused by a lack of inpatient beds but rather by other inefficiencies which slow the transitions of care between different hospital units [116]. In the following section, we will evaluate the impact of increasing capacity and reducing LOS in a more capacity-constrained setting.

3.4.4 Sensitivity analysis

To examine the robustness of the insights we have gained so far, we test five policies – the Period 2 and Period 3 policies and the three alternative policies described in the previous section – under different model settings for sensitivity analysis. These settings include using alternative arrival models, changing the priority among ICU-GW, SDA and ED-GW patients, and choosing different values for the normal allocation probability $p(t)$. The simulation results show that the insights we gained are robust under the tested model variations.

In addition, to evaluate the five policies when the system load is high, we increase

the daily arrival rates of ED-GW patients by 7%, similar to the increase we empirically observed from Period 1 to Period 2. When all other settings are kept the same as in the baseline, utilization under the increased arrival rate assumption becomes 93%, and the daily average waiting time and 6-hour service level become 4.37 hours and 18.60%, respectively. We test the five policies under the increased arrival rate assumption. We find that the insights we have gained are still robust in this capacity-constrained setting, except that (a) increasing capacity by 10% or reducing the LOS now shows a significant reduction in daily average waiting times (reduce from 4.37 to about 2.5 hours); (b) the Period 3 policy can also greatly reduce the daily waiting time statistics because of its side effect of reducing the LOS, which results from using different LOS distributions between AM- and PM-admitted ED-GW patients (see Table 6). Sections B.6 and B.7 in the appendix detail all the experiment setting and simulation results of the sensitivity analysis discussed in this section.

3.4.5 Intuition about the gained insights

Our evaluated policies show different impacts on the daily and hourly waiting time performance. The reason lies in the separation of time scales, which is captured by our two-time-scale service time model in Equation (4). We now provide some intuition to explain the findings we have obtained so far. A more mathematical explanation can be obtained through the two-time-scale analytical framework developed in Chapter 4.

There are two types of waiting in our model. (i) The total number of discharges in one day is less than the total number of arrivals in that day, and therefore some patients have to wait till next day to get a bed. (ii) Within a day, the discharge timing is too late so that morning arrivals have to wait several hours (till the afternoon) to get a bed. The first type of waiting, reflected in the daily waiting time performance, can be affected by the daily arrival rate, LOS distributions and bed capacity. The second type of waiting, reflected in the time-of-day (hourly) waiting times, can be

affected by the time-varying arrival rates and discharge patterns.

Clearly, merely shifting the discharge timing earlier can eliminate or reduce the second type of waiting, not the first. Thus, early discharge policies can flatten hourly waiting time curves, but has limited impact on further reducing daily waiting times (when there is no side-effect in reducing the LOS). Moreover, to achieve the flattening effect, the early discharge policy needs to ensure a modest number of patients be discharged early enough (before 9-10am in the NUH setting as suggested by our simulation results) so that beds become available before the queue starts to build up in the morning. This also explains why Period 2 policy has a limited impact on flattening the hourly waiting time curves, since its first discharge peak starts at 11am, which is not early enough.

In comparison, increasing capacity or reducing LOS helps reduce the first type of waiting and thus can reduce daily waiting time statistics. This effect is particularly significant in a capacity-constrained setting. Even when bed utilization is low, but not excessively low (not lower than 71% in the NUH setting), many morning arrivals can still experience the second type of waiting due to not enough patients being discharged early enough. This is why increasing capacity does not necessarily flatten the hourly waiting time curves.

3.5 Concluding remarks and future research

We have proposed a high-fidelity stochastic network model for inpatient flow management, which can be used as a tool to quantify the impact of various operational policies. In particular, the model captures time-of-day waiting time performance for ED-GW patients and enables us to identify policies that can reduce or flatten waiting times. Our model predicts that a hypothetical Period 3 policy (and similar policies with certain early discharge distributions and constant-mean allocation delays) can achieve time-stable waiting time statistics throughout the day, but has limited impact

on the daily average waiting time and overflow proportion in the NUH setting. Our model also predicts that reducing the mean allocation delay significantly reduces the daily average waiting time, and increasing bed capacity or reducing LOS can greatly reduce the overflow proportion; however, these three policies have less impact on the time-of-day pattern of waiting time performance and they do not necessarily stabilize waiting time statistics. These insights can help hospital managers choose among different policies to implement, depending on the choice of objective, such as to reduce the peak waiting in the morning or to reduce daily waiting time statistics.

Readers should be aware of two issues when interpreting these findings. First, we focus on evaluating the impact of discharge policies and other policies on the waiting time performance of ED-GW patients in this paper. There could be other benefits of these policies that our paper has not modeled. For example, it is believed that early discharge allows more flexibility to transfer patients from ICU to GWs when ICU wards become congested. Second, regarding the impact on the waiting times, the evaluation of these policies is based on predictions from the populated NUH model and comparison to the baseline scenario. Thus, our findings may not be always generalizable to other hospital settings. Section B.5.4 in the appendix shows an example where the Period 2 policy can have more significant benefits in a different setting.

On the implementation level, we recognize the challenges in implementing the Period 3 policy in practice. On the one hand, discharging patients as early as 8-9am is difficult since physicians and nurses are busy with the morning rounds at about this time. On the other hand, stabilizing allocation delays also requires coordination throughout the entire hospital and proper staffing at different units at various hours of the day. Though the Period 3 policy is purely hypothetical and may not be completely practical, we believe it can serve as a goal for hospital managers to aim at if they intend to eliminate excessively long waiting times for morning bed-requests.

More importantly, our model provides an efficient tool to evaluate the impact of a spectrum of policies that lie between Period 2 and Period 3 policies. Based on the outcomes and costs of implementation, hospital managers can choose the desired levels of effort to implement these policies. Besides the discharge policy, our model allows hospital managers to evaluate the trade-off between the benefit of reducing ED overcrowding and the cost of implementing a number of operational and strategic polices.

3.5.1 Future work

Our proposed model on inpatient flow management in this paper can be used for other studies that intend to integrate the ED and inpatient department operations together. Our model can potentially be extended in several directions.

First, we use pre- and post-allocation delays as two black boxes to model all possible secondary bottlenecks including ward nurse and BMU staff shortage at certain times of the day, partly because we want to maintain the tractability of the proposed model. Detailed studies are needed to further understand these secondary bottlenecks so that we can explicitly incorporate them into the model and identify strategies to reduce allocation delays. The two-queue model proposed in [155] appears relevant to this line of research.

Second, our proposed model is a parallel-server system with a single-pass routing structure. In particular, we do not model ICU-type wards and patient flows within ICU-type wards in our system, because the data requirements to model them would be at another level and are beyond the scope of this paper. An extension would be to build upon this paper and recent studies on ICU management [29, 87] to model both ICU-type wards and general wards as a stochastic network that has internal routings between these wards. The extended model could study waiting times for ED-GW and ICU-GW patients as well as waiting times for ED-ICU patients or GW-ICU patients,

and can better evaluate the impact of early discharge and other policies on patients besides ED-GW patients. Besides, we do not model internal patient transfers between two general wards (e.g., see Section 2.8.3). Future studies can extend our proposed model to incorporate transfer activities among different server pools.

Third, considering day-of-week phenomena is another important extension to make the model more realistic. Currently, we assume that ED-GW patients have a stable daily arrival volume without differentiating days of a week. We also assume that elective admissions are stationary by day, while recent work pointed out that elective schedule is actually the main source of daily occupancy variation in many hospitals [74]. Our model can be extended to predict day-of-week performance and help design a better elective schedule.

Finally, to obtain structural insights into the impact of many policies such as discharge timing and overflow trigger time, simulation alone is difficult. There is a need to develop analytical methodology, not purely simulations, to predict performance measures that depend on hour-of-day. In the next chapter, we introduce a new analytical framework to analyze models with some of the critical features we discover in this chapter. We believe the model proposed in this chapter and our preliminary analytical tools will stimulate more analytical research to develop tools to study a new class of stochastic models.

CHAPTER IV

A TWO-TIME-SCALE ANALYTICAL FRAMEWORK

In this chapter, we illustrate a new analytical framework, *two-time-scale analysis*, to analyze a class of time-varying queueing models that are motivated by the stochastic model we proposed in Chapter 3. The two important features for this class of queueing models is that (a) the service time is no longer iid but follows (4), and (b) there are allocation delays during the patient (customer) admission process. We focus on analyzing single-pool models with these two features. Note that in Chapter 3 we also demonstrated that the multi-pool structure with the overflow mechanism is another important feature in capturing hospital inpatient flow; however, in this chapter, we do not incorporate this third feature in our analysis. Developing analytical methods for multi-pool models will be left in a future project.

This chapter is organized as follows. In Section 4.1, we first illustrate a single-pool model without allocation delay and introduce the time-dependent performance measures we focus on. We then demonstrate the basic idea of the two-time-scale approach on this single-pool model in Section 4.2. Then in Section 4.3, we analyze a single-pool model with allocation delays. We show numerical results from the two-time-scale analysis in Section 4.4. To devise efficient numerical algorithm to compute different time-dependent performance measures, we explore diffusion approximations in Section 4.5. These approximations are based on limit theorems, and we prove the limit theorems in Section 4.6. This chapter concludes in Section 4.7.

4.1 A single-pool model without allocation delays

We study a single-pool model denoted as an $M_{\text{peri}}/G_{2\text{timeScale}}/N$ queue. There are N servers in the server pool, modeling N inpatient beds. The "M_{peri}" denotes the

patient arrival process, which is a periodic Poisson process with one day as a period. The "$G_{\text{2timeScale}}$" denotes the service time model, which is unconventional and follows (4):

$$S = \text{LOS} + h_{\text{dis}} - h_{\text{adm}}.$$

The service time of a patient corresponds to the duration that the patient occupies an inpatient bed. It has been shown in Chapter 3 that this two-time-scale service time model is critical to predict steady-state, time-dependent performance measures for hospital inpatient flow management.

In our model, the customers are to model patients. We assume the customers are from a single class with a common arrival rate function and common LOS and discharge distributions. Thus, the model we study here is a *single-class single-pool* model. We use $\lambda(t)$ to denote the common arrival rate function. It satisfies $\lambda(t) = \lambda(t+1)$ for any $t \geq 0$. The daily arrival rate is $\Lambda = \int_0^1 \lambda(t)dt$. The LOS of each patient follows a *geometric* distribution which takes values on $1, 2, \ldots$ and has a success probability of μ (equivalently, the mean LOS is $m = \frac{1}{\mu}$). For notational simplicity, we assume that there is no arrival or discharge at the exact point of midnight each day. In the rest of this chapter, we still use servers and beds, and customers and patients, interchangeably.

This single-pool model has the following dynamics. Upon a patient arrival, we check the server-pool status. If there is a free bed, we admit the patient into service (i.e., the patient starts to occupy the bed); otherwise, we move the patient to the common buffer (waiting area). At the discharge time of a patient, we check the buffer status. If the buffer is empty, the bed becomes idle; otherwise, we assign the bed to the first patient waiting in the buffer following the first-come-first-out (FIFO) rule. After a patient is admitted, she occupies the assigned bed until discharge. The duration between admission and discharge is the patient's service time S.

Performance measures. We focus on predicting the steady-state time-dependent

performance measures for the single-pool model. The performance curves we are interested in are the time-dependent mean queue length $\mathbb{E}_\infty[Q(t)]$ (similar to those in Figure 40a), time-dependent mean virtual waiting time $\mathbb{E}_\infty[W(t)]$ (similar to those in Figure 39a), and time-dependent 6-hour service level $\mathbb{P}_\infty(W(t) \geq 6)$ (similar to those in Figure 39b). Here, $Q(t)$ denotes the queue length at time t, and $W(t)$ is the waiting time for a virtual customer arriving at time t. We use the subscript ∞ to denote the steady-state expectation or probability.

4.2 A two-time-scale approach for the single-pool model

The two-time-scale approach essentially has two steps. First, we consider the queueing system day by day, and obtain the stationary distribution of the so-called *midnight customer count*. Then, based on the midnight customer count, we calculate the distribution of the customer count for different time of day. In Sections 4.2.1-4.2.2, we illustrate these two steps of the two-time-scale approach. Once we get the distribution of the time-dependent customer count, we can calculate different performance measures. In Section 4.2.3, we demonstrate how to predict the time-dependent performance measures in the single-pool model we introduced in Section 4.1.

4.2.1 A two-time-scale approach: step 1, midnight dynamics

Let X_k denote the number of customers in system at the midnight (zero hour) of day k, i.e., the midnight customer count. Let A_k and D_k denote the total number of arrivals and discharges within day k, respectively. We then have the following relationship between X_k and X_{k+1}:

$$X_{k+1} = X_k + A_k - D_k, \quad k = 0, 1, \ldots. \tag{6}$$

Under our arrival process assumption, A_k is a Poisson random variable with rate Λ. Since there is no same-day discharge (LOS is at least one day), the number of discharges on day k only depends on the number of customers admitted before

day k. Recall that LOS follows a geometric distribution. Then D_k is in fact a binomial random variable with parameters $(\min\{X_k, N\}, \mu)$, only depending on X_k. Moreover, we assume that the arrival process is independent of the LOS and discharge distributions. Therefore, $\{X_k : k = 0, 1, \ldots\}$ forms a discrete-time Markov chain (DTMC), and its transition probability from state i to state j is:

$$P_{ij} = \sum_{k=\max\{0,i-j\}}^{\min\{i,N\}} g(i,k) \cdot f(k+j-i). \tag{7}$$

Here $f(k) = \frac{\Lambda^k}{k!} e^{-\Lambda}$ is the probability mass function (pmf) at k for the Poisson distribution with rate Λ, and $g(i,k)$ is the pmf for the binomial distribution.

The Markov chain is irreducible. Moreover, one can prove that this Markov chain is positive recurrent with a unique stationary distribution π if

$$\rho = \frac{\Lambda}{N\mu} < 1. \tag{8}$$

The argument is simple by checking the Foster-Lyapunov criterion [15], i.e., when $x > N$, we have

$$E[X_{k+1} - X_k | X_k = x] = E[A_k - D_k | X_k = x] = \Lambda - N\mu = -N\mu(1-\rho) < 0.$$

Under condition (8), we can compute the stationary distribution π numerically.

4.2.2 A two-time-scale approach: step 2, time-of-day queue length dynamics

For $t \geq 0$, we define $X(t)$ to be the total number of customers in the system at time t, i.e., time-of-day customer count. We assume that $X(0)$ follows the stationary distribution π of $\{X_k : k = 1, 2, \ldots\}$. Similar to (6), for $t \geq 0$, $X(t)$ can be expressed in the following form

$$X(t) = X(0) + A_{(0,t]} - D_{(0,t]}. \tag{9}$$

Here, $A_{(0,t]}$ denotes the cumulative number of arrivals in the period $(0, t]$, and $D_{(0,t]}$ denotes the cumulative number of discharges in the period $(0, t]$. We first argue that

the process $X = \{X(t), t \geq 0\}$ is periodic with one day as a period. We start from showing that $X(t) =^d X(t+1)$ for $t \geq 0$ with $=^d$ denoting equal in distribution. Since $X(0)$ follows the stationary distribution π, we know the customer count at each midnight, $X(k)$, $k = 1, 2, \ldots$, also follows the stationary distribution π. Meanwhile, for any $t \geq 0$, $X(t)$ and $X(t+1)$ can be further represented as

$$\begin{aligned} X(t) &= X(k_t) + A_{(k_t,t]} - D_{(k_t,t]}, \\ X(t+1) &= X(k_t+1) + A_{(k_t+1,t+1]} - D_{(k_t+1,t+1]}, \end{aligned}$$

where $k_t = \lfloor t \rfloor$ is the most recent midnight before time t. Since the arrival process has the period of one day, both $A_{(k_t,t]}$ and $A_{(k_t+1,t+1]}$ follow a Poisson distribution with the same mean $\int_0^{t-k_t} \lambda(s)ds$ and are independent of $X(k_t)$ and $X(k_t+1)$, respectively. The two discharge quantities, $D_{(k_t,t]}$ and $D_{(k_t+1,t+1]}$, follow binomial distributions with parameters $(\min\{X(k_t), N\}, \mu q_{(0,t-k_t]})$ and $(\min\{X(k_t+1), N\}, \mu q_{(0,t-k_t]})$, respectively; here $q_{(0,t-k_t]}$ is the common cumulative discharge distribution from 0 to $t - k_t$. Note that $D_{(k_t,t]}$ only depends on $X(k_t)$ and $D_{(k_t+1,t+1]}$ only on $X(k_t+1)$, while $X(k_t)$ and $X(k_t+1)$ have the same distribution π. Thus, the two discharge quantities also have the same distribution. Eventually, we can see that $X(t)$ and $X(t+1)$ have the same distribution.

The above argument can be generalized to prove $(X(t_1), X(t_2)) =^d (X(t_1+1), X(t_2+1))$ for $t_1, t_2 \in [k_t, k_t+1)$. Then, for any finite K-dimensional joint distribution, we can show $(X(t_1), \ldots, X(t_K)) =^d (X(t_1+1), \ldots, X(t_K+1))$ for $t_1, \ldots, t_K \in [k_t, k_t+1)$. Eventually, we have $\{X(t), k_t \leq t < k_t+1\} =^d \{X(t+1), k_t \leq t < k_t+1\}$, and thus, it is sufficient for us to focus on the dynamics of $X(t)$ for $0 \leq t < 1$. Conditioning on $X(0)$, we can see from the above argument that the distribution of $X(t)$ is a convolution between a Poisson and a binomial distributions. Un-conditioning on $X(0)$ with the stationary distribution π, we can obtain the distribution for $X(t)$.

4.2.3 Predict time-dependent queue length and waiting time dynamics

Mean queue length

For any $t \geq 0$, given the stationary distribution of $X(t)$, the steady-state mean queue length at t is

$$\mathbb{E}_\infty[Q(t)] = \mathbb{E}_\infty[(X(t) - N)^+], \tag{10}$$

where, for a real number a, $a^+ = \max(a, 0)$. Since $X(t)$ is periodic, it is obvious that $\mathbb{E}_\infty[Q(t)]$ is also periodic with one day as a period.

Mean waiting time

We now consider the steady-state mean virtual waiting time $\mathbb{E}_\infty[W(t)]$ for a virtual customer arriving at time t. Again, because $X(t)$ is periodic, we can argue that $\mathbb{E}_\infty[W(t)]$ is also periodic with one day as a period (below also gives more details on how $\mathbb{E}_\infty[W(t)]$ depends on $X(t)$). As a result, we focus on $0 \leq t < 1$.

To illustrate the calculation, we further assume the discharge hour h_{dis} follows a discrete distribution, taking values on a finite number of points $t_1,\ t_2, \ldots,\ t_n$ with probabilities $q_{t_1},\ q_{t_2}, \ldots,\ q_{t_n}$, respectively. Under this discharge distribution, the discharges are in batches. Next, we focus on illustrating the case when $t < t_1$, where t_1 is the first discharge point.

1. Given the stationary distributions of $X(t)$, we know that there is a chance of $p_0(t) = \mathbb{P}_\infty(X(t) < N)$ that the virtual customer does not need to wait, and thus $W(t) = 0$. With probability $(1 - p_0(t))$, this customer cannot enter service immediately and need to wait.

2. Conditioning on the latter case that this customer needs to wait, we know there are $X(t) - N$ customers wait in front of her. Given the value of $X(0) = x_0$, the number of discharges between t and t_1, $D_{(t,t_1]}$, is a binomial random variable with parameters $(x_0, q_{t_1}\mu)$ since there is no discharge from 0 to t. Thus, we can

compute

$$\begin{aligned} p_{t_1}(t) &= \mathbb{P}_\infty(X(t) \geq N, X(t) - N < D_{(t,t_1]}) \\ &= \mathbb{P}_\infty(0 \leq X(0) + A_{(0,t]} - N < D_{(0,t_1]}), \end{aligned} \tag{11}$$

which is the probability that at least $X(t) - N + 1$ customers will be discharged by t_1. The second equation in (11) follows from the fact that $X(t) = X(0) + A_{(0,t]}$ and $D_{(t,t_1]} = D_{(0,t_1]}$ for $t < t_1$. With the chance of $p_{t_1}(t)$, the virtual customer will enter service at t_1 and $W(t) = t_1 - t$.

3. Next, we compute

$$\begin{aligned} p_{t_2}(t) &= \mathbb{P}_\infty(X(t) - N \geq D_{(t,t_1]}, X(t) - N < D_{(t,t_2]}) \\ &= \mathbb{P}_\infty(D_{(0,t_1]} \leq X(0) + A_{(0,t]} - N < D_{(0,t_2]}), \end{aligned} \tag{12}$$

which is the probability that at least $X(t) - N + 1$ customers will be discharged by t_2 but less than $X(t) - N + 1$ customers were discharged before t_1. Here, the number of discharges by t_2, $D_{(t,t_2]} = D_{(0,t_2]}$ follows a binomial distribution with parameters $(x_0, (q_{t_1} + q_{t_2})\mu))$. Then, with the chance of $p_{t_2}(t)$, the virtual customer will enter service at t_2 and $W(t) = t_2 - t$.

4. We repeat this procedure iteratively and calculate until

$$\begin{aligned} p_{t_n}(t) &= \mathbb{P}_\infty(X(t) - N \geq D_{(t,t_{n-1}]}, X(t) - N < D_{(t,t_n]}) \\ &= \mathbb{P}_\infty(D_{(0,t_{n-1}]} \leq X(0) + A_{(0,t]} - N < D_{(0,t_n]}), \end{aligned} \tag{13}$$

which is the chance that at least $X(t) - N + 1$ customers will be discharged today by t_n but less than $X(t) - N + 1$ customers were discharged before t_{n-1}. Similarly, with probability $p_{t_n}(t)$, the customer will enter service at t_n, and the waiting time $W(t) = t_n - t$; otherwise, this virtual customer needs to wait till at least the next day to be admitted.

5. The probability that a customer needs to wait till t_1 the next day to be admitted is:

$$\begin{aligned} p^1_{t_1}(t) &= \mathbb{P}_\infty(X(t) - N \geq D_{(t,1]}, X(t) - N < D_{(t,1+t_1]}) \\ &= \mathbb{P}_\infty(D_{(0,1]} \leq X(0) + A_{(0,t]} - N < D_{(0,1+t_1]}). \end{aligned}$$

The number of customers discharged by the next day at time t_1, $D_{(0,1+t_1]}$ is the sum of $D_{(0,1]}$ and $D_{(1,1+t_1]}$. Note that $D_{(0,1]}$ follows a binomial distribution with parameters $(X(0), \mu)$, while $D_{(1,1+t_1]}$ follows a binomial distribution with parameters $(N, q_{t_1}\mu)$ since the server pool is full at the beginning of the next day. Similarly, we can obtain $p^1_{t_2}(t), \ldots, p^1_{t_n}(t)$, which denote the probabilities that the virtual customer enters service at time t_2, $\ldots$, t_n the next day, respectively.

Sequentially, we can perform the same procedure for discharges k days later. In general, for a given day k ($k = 1, 2, \ldots$) and discharge time $t_i \geq t_2$, the probability $p^k_{t_i}(t)$ can be expressed as

$$\begin{aligned} p^k_{t_i}(t) &= \mathbb{P}_\infty(X(t) - N \geq D_{(t,k+t_{i-1}]}, X(t) - N < D_{(t,k+t_i]}) \\ &= \mathbb{P}_\infty(D_{(0,k+t_{i-1}]} \leq X(0) + A_{(0,t]} - N < D_{(0,k+t_i]}),\ t_i \geq t_2, \end{aligned} \quad (14)$$

while for $t_i = t_1$,

$$\begin{aligned} p^k_{t_1}(t) &= \mathbb{P}_\infty(X(t) - N \geq D_{(t,k]}, X(t) - N < D_{(t,k+t_1]}) \\ &= \mathbb{P}_\infty(D_{(0,k]} \leq X(0) + A_{(0,t]} - N < D_{(0,k+t_1]}). \end{aligned} \quad (15)$$

The waiting time associated with the probability $p^k_{t_i}$ is $W(t) = (k + t_i - t)$. The discharge quantity, $D_{(0,k+t_i]}$ is the sum of $D_{(0,1]}$ and $D_{(1,k+t_i]}$, which follow two binomial distributions with parameters $(X(0), \mu)$ and parameters $(N, \mu(k + \sum_{j=1}^{i} q_{t_j}))$, respectively.

Eventually, we can numerically evaluate the mean waiting time $\mathbb{E}_\infty[W(t)]$ for

$0 \leq t < t_1$ using the following equation:

$$\mathbb{E}_\infty[W(t)] = \sum_{i=1}^{n} p_{t_i}(t) \cdot (t_i - t) + \sum_{k=1}^{\infty} \sum_{i=1}^{n} p_{t_i}^k(t) \cdot (k + t_i - t). \tag{16}$$

The mean waiting time for $t_1 \leq t < t_n$ can be calculated in a similar manner:

$$\mathbb{E}_\infty[W(t)] = \sum_{i=i^*}^{n} p_{t_i}(t) \cdot (t_i - t) + \sum_{k=1}^{\infty} \sum_{i=1}^{n} p_{t_i}^k(t) \cdot (k + t_i - t), \tag{17}$$

where t_{i^*} is the first discharge time point after t; while for $t_n \leq t < 1$, the mean waiting time is simply

$$\mathbb{E}_\infty[W(t)] = \sum_{k=1}^{\infty} \sum_{i=1}^{n} p_{t_i}^k(t) \cdot (k + t_i - t). \tag{18}$$

Note that the set of probabilities $p_{t_i}(t)$ and $p_{t_i}^k(t)$ in (17) and (18) have consistent expressions as those for the case of $0 \leq t < t_1$, i.e., the displays in (11) to (14). In particular, note that for $t_1 \leq t < t_n$, if $t < t_{i-1} < t_i$ (or equivalently, $t_i > t_{i^*}$ is *not* the first discharge point after t), then

$$\begin{aligned} p_{t_i}(t) &= \mathbb{P}_\infty(X(t) - N \geq D_{(t,t_{i-1}]}, X(t) - N < D_{(t,t_i]}) \\ &= \mathbb{P}_\infty(D_{(0,t_{i-1}]} \leq X(0) + A_{(0,t]} - N < D_{(0,t_i]}), \end{aligned} \tag{19}$$

consistent with (12) and (13); the second equation follows from the fact that $X(t) = X(0) + A_{(0,t]} - D_{(0,t]}$. If $t_i = t_{i^*}$ is the first discharge point after t ($t_1 \leq t < t_n$), then

$$\begin{aligned} p_{t_{i^*}}(t) &= \mathbb{P}_\infty(X(t) - N \geq 0, X(t) - N < D_{(t,t_{i^*}]}) \\ &= \mathbb{P}_\infty(D_{(0,t]} \leq X(0) + A_{(0,t]} - N < D_{(0,t_{i^*}]}), \end{aligned} \tag{20}$$

consistent with (11).

6-hour service level

The 6-hour service level can be obtained in a similar way as the mean virtual waiting time. For example, for a virtual customer arriving at time $0 \leq t < t_1$, the 6-hour service level is

$$\mathbb{P}_\infty(W(t) \geq 6) = \sum_{i=1}^{n} p_{t_i} \mathbb{1}_{\{t_i - t \geq 6/24\}} + \sum_{k=1}^{\infty} \sum_{i=1}^{n} p_{t_i}^{(k)}, \tag{21}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function. The 6-hour service level for the case $t_1 \leq t < t_n$ or $t_1 \leq t < t_n$ can be adapted from (17) or (18) in a similar way.

4.3 Single-pool model with allocation delays

In this section, we introduce a revised $M_{\text{peri}}/G_{\text{2timeScale}}/N$ model with the second critical feature: allocation delays. We first describe the queueing dynamics for this revised system in Section 4.3.1. Then in Section 4.3.2 we show how to adapt the two-time-scale analysis to predict time-dependent performance measures in this revised system.

4.3.1 System dynamics

In a single-pool $M_{\text{peri}}/G_{\text{2timeScale}}/N$ model with allocation delays (see the description of allocation delays in Sections 2.7 and 3.1.4), we assume each patient needs to experience an extra amount of delay after a bed becomes available for her. Thus, we assume there are two buffers working in series: the *waiting-bed queue*, and the *allocation-delay queue*. These two buffers holding waiting customers in different status.

Specifically, upon each patient arrival, we first check the server-pool status. If there is a free bed, we assign the bed to the patient and move the patient to the allocation-delay queue; otherwise, we move the patient to the waiting-bed queue. At the discharge time of a previous patient, we check the waiting-bed queue. If it is not empty, the newly-freed bed is assigned to the first patient waiting in the waiting-bed queue (following the FIFO rule), and we then move this patient to the allocation-delay queue. Otherwise, the bed becomes idle.

When a patient is moved to the allocation-delay queue, she stays there for a random amount of time, T_{alloc}. After this T_{alloc} time expires, she is admitted into service. The service time of a patient in this revised single-pool model still follows a form that is similar to (4), except that we replace the admission hour h_{adm} with the bed-assignment hour h_{alloc}, and LOS is slightly modified to represent the number of

midnights between the bed-assignment and discharge times. Recall that a patient's bed-assignment time is either at her arrival time or at the discharge time of a previous patient. We assume that the allocation delays of each patient forms a sequence of iid random variables and are independent of the allocation times (i.e., not time-dependent).

Note that comparing to the high-fidelity model we developed in Chapter 3, here we use a simplified setting for modeling allocation delays. Each patient in this revised single-pool model adopts the *normal-allocation mode* as introduced in Section 3.1.4, i.e., the patient starts to experience allocation delays only after a bed is assigned to her (see Cases A and B in Figure 37). The T_{alloc} amount of delay represents the sum of pre- and post-allocation delays, and corresponds to the total delays that (i) BMU needs to search and negotiate for a bed and (ii) ED needs to discharge the patient from ED and transport her to inpatient wards. Moreover, we focus on the case that T_{alloc} is not time-dependent, whereas in the high-fidelity model, both pre- and post-allocation delays have time-varying means.

4.3.2 Predict the performance measures

Mean queue length

The steady-state, time-dependent mean queue length $\mathbb{E}_\infty[Q(t)]$ equals to the sum of the two buffer sizes at time t, i.e.,

$$\mathbb{E}_\infty[Q(t)] = \mathbb{E}_\infty[Q_{\text{bed}}(t)] + \mathbb{E}_\infty[Q_{\text{alloc}}(t)],$$

where $Q_{\text{bed}}(t)$ and $Q_{\text{alloc}}(t)$ denote the number of customers waiting in the waiting-bed queue and the allocation-delay queue, respectively. We separately calculate each of the mean queue lengths. Below, we mainly focus on calculating the mean allocation-delay queue length.

To obtain $\mathbb{E}_\infty[Q_{\text{alloc}}(t)]$, we further separate the "arrivals" to the allocation-delay

queue into two streams (to differentiate, we refer the patients arriving to the single-pool system the *external arrivals*). These two arrival streams are: (i) the external arrivals who get a free bed immediately upon arriving to the single-pool system; and (ii) the patients in the waiting-bed queue who get a free bed at the discharge times of previous patients. It is easy to argue that the arrivals from the first stream still follow a periodic Poisson process with one day as a period, because this stream of arrivals is essentially a "thinning" of the external arrival process, and the thinning probability $p_0(t) = \mathbb{P}_\infty(X(t) < N)$ (i.e., the non-blocking probability) is independent of the external arrivals coming at or after time t. We use $A^{(1)}_{\text{alloc}}$ to denote the first stream of arrivals to the allocation-delay queue.

To illustrate the second stream of arrival process (denote as $A^{(2)}_{\text{alloc}}$), we again consider the batch discharge policies, where the discharge hour h_{dis} follows a discrete distribution taking values on a finite number of points $t_1,\ t_2, \ldots,\ t_n$ with probabilities $q_{t_1},\ q_{t_2}, \ldots,\ q_{t_n}$, respectively. Then, this arrival stream forms a batch arrival process, and at a discharge point t_i, the number of patients arriving at the allocation-delay queue equals

$$A^{(2)}_{\text{alloc}}(t_i) = (X(t_i^-) - N)^+ - (X(t_i) - N)^+,$$

where $X(t_i^-)$ denotes the left limit of $X(\cdot)$ at time t_i.

Next, we separately calculate the mean allocation-delay queue length from the two arrival streams, i.e.,

$$\mathbb{E}_\infty[Q_{\text{alloc}}(t)] = \mathbb{E}_\infty[Q^{(1)}_{\text{alloc}}(t)] + \mathbb{E}_\infty[Q^{(2)}_{\text{alloc}}(t)]. \tag{22}$$

For the queue length formed by customers from the first arrival stream, we utilize the infinite-server queue theory since we can consider these customers as receiving "service" in an $M_{\text{peri}}/GI/\infty$ system with the service times being T_{alloc} and the arrival process being $A^{(1)}_{\text{alloc}}$. Let $F_{\text{alloc}}(\cdot)$ denotes the CDF of the allocation delay T_{alloc}. From

Theorem 1 of [41], we know that

$$\mathbb{E}_\infty[Q^{(1)}_{\text{alloc}}(t)] = \mathbb{E}[\lambda^{(1)}_{\text{alloc}}(t - T_{\text{alloc},e})]E[T_{\text{alloc}}]. \tag{23}$$

Here, $T_{\text{alloc},e}$ denotes the random variable associated with the equilibrium-residual-lifetime CDF of the allocation delay T_{alloc}

$$F_{\text{alloc},e}(t) = \mathbb{P}(T_{\text{alloc},e} \leq t) = \frac{1}{\mathbb{E}[T_{\text{alloc}}]} \int_0^t (1 - F_{\text{alloc}}(u))du, \quad t \geq 0,$$

and $\lambda^{(1)}_{\text{alloc}}(\cdot) = p_0(\cdot)\lambda(\cdot)$ denotes the arrival rate function of the arrival process $A^{(1)}_{\text{alloc}}$.

For the queue length formed by customers from the second arrival stream, we know that among all the customers who arrived before t, the chance that such a customer is still in the allocation-delay queue is $P(T_{\text{alloc}} > t - t_{\text{arr}}) = 1 - F(t - t_{\text{arr}})$. Here, t_{arr} denotes one of the batch arrival time before t. Therefore, we know that

$$\mathbb{E}_\infty[Q^{(2)}_{\text{alloc}}(t)] = \sum_{t_{\text{arr}} \leq t} \mathbb{E}[A^{(2)}_{\text{alloc}}(t_{\text{arr}})](1 - F(t_{\text{arr}} - t)). \tag{24}$$

Combining (22) to (24), we can calculate the mean allocation-delay queue length $\mathbb{E}_\infty[Q_{\text{alloc}}(t)]$ for any time t.

Finally, for $\mathbb{E}_\infty[Q_{\text{bed}}(t)]$, it can be calculated in the same way as introduced in Section 4.2.3. Because the waiting-bed queue has the same dynamics as the single-pool system without allocation delays.

Mean waiting time and 6-hour service level

It is easy to see from the description in Section 4.3.1 that the total waiting time for a virtual customer arriving at time t in the revised single-pool system is

$$W(t) = W_{\text{bed}}(t) + T_{\text{alloc}}.$$

Here, $W_{\text{bed}}(t)$ is the time that this customer needs to wait before getting a bed assigned. Note that $W_{\text{bed}}(t)$ can be calculated in the same way as illustrated in Section 4.2.3, because the waiting-bed queue has the same dynamics as the single-pool system without allocation delays.

Therefore, we know the mean waiting time

$$\mathbb{E}_\infty[W(t)] = \mathbb{E}_\infty[W_{\text{bed}}(t)] + \mathbb{E}_\infty[T_{\text{alloc}}], \tag{25}$$

and the 6-hour service level

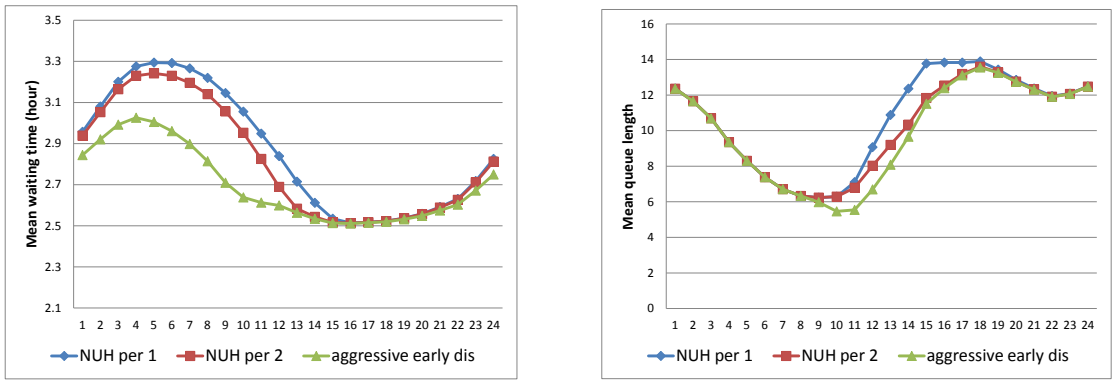
$$\begin{aligned} \mathbb{P}_\infty\left(W(t) \geq 6\right) &= \mathbb{P}_\infty\left(W_{\text{bed}}(t) + T_{\text{alloc}} \geq 6\right) \\ &= \int_0^6 \mathbb{P}_\infty\left(W_{\text{bed}}(t) \geq 6 - x\right) f(x) dx, \end{aligned} \tag{26}$$

where $f(x)$ is the pdf of the allocation delay T_{alloc}.

4.4 Numerical results

In this section, we present some numerical results for the single-pool model. In the numerical experiments, the arrival rate function $\lambda(t)$ is assumed to be piecewise-constant, i.e., the arrival rate is constant in each hour. We proportionally enlarged the hourly arrival rates represented by the solid curve of Figure 13, so that the daily arrival rate $\Lambda = 90.95$ is close to the empirical daily arrival rate from the four admission sources. We do this adjustment because the single-pool model only has one class of customers. The parameter for the LOS distribution is chosen to be $p_{\text{LOS}} = 1/5.3$, i.e., the mean LOS is 5.3 days. The total number of servers is $N = 525$. The allocation delay, T_{alloc}, follows a log-normal distribution with mean = 2.5 hours and CV = 1.

To evaluate the impact of discharge timing, we test three discharge distributions: NUH Period 1 and Period 2 discharge distributions, and the aggressive early discharge distribution in the hypothetical Period 3 policy (see description of these discharge policies in Section 3.4). Figure 48 plots the steady-state time-dependent mean waiting time curves and mean queue length curves under the three discharge scenarios. We can see that (i) the Period 2 early discharge policy has limited impact on stabilizing the waiting time performances, (ii) the hypothetical Period 3 discharge policy shows a more significant impact on stabilizing the waiting time performances, and (iii) early



(a) Mean waiting time

(b) Mean queue length

Figure 48: **Numerical results for the steady-state time-dependent mean waiting time and mean queue length.** Three discharge distributions are tested: NUH Period 1 and Period 2 discharge distributions, and the aggressive early discharge distribution in the hypothetical Period 3 policy.

discharge policies mainly impact the hourly performance measures, not the daily performance. These insights are consistent with our simulation study of the high-fidelity multi-pool model; see summary of the simulation findings in Section 3.4.

4.5 Diffusion approximations

Sections 4.2 and 4.3 have introduced how to predict the time-dependent performance measures under the two-time-scale analysis framework. A key step for the prediction is to numerically solve the stationary distribution π for the midnight customer count process. However, when the system size is large and heavily utilized, it becomes inefficient to compute the stationary distribution this way. Moreover, since the numerical solutions do not have closed-form expressions, it is still difficult to gain structural insights into understanding the system dynamics. Thus, in this section we explore diffusion approximations to devise efficient numerical algorithms to obtain the midnight count stationary distribution and to predict the time-dependent performance measures, which eventually help us provide more insights into improving hospital inpatient operations.

In this section, we focus on studying the single-pool model *without* allocation

delays, since the adjustment to incorporate allocation delays is fairly standard and does not involve intensive computations. We develop diffusion approximations for the midnight customer count process, the hourly customer count process, and the time-dependent performance measures in Sections 4.5.1 to 4.5.3. We then demonstrate in Section 4.5.4 that these approximations are good even when the system size is small (N around 60-70) and the utilization is moderate. The diffusion approximations proposed in this section are for systems with fixed N, and these approximations are rooted in limit theorems, i.e., stochastic processes convergence when $N \to \infty$. We will prove limit theorems in Section 4.6.

4.5.1 Approximation for the midnight customer count

We first explore diffusion approximations for the midnight customer count process $\{X_k, k = 0, 1, 2 \ldots\}$. Let $A_{(0,k]} = \sum_{i=0}^{k-1} A_i$ be the *cumulative* number of arrivals and $D_{(0,k]}$ be the *cumulative* number of departures from 0 until the midnight (zero hour) of day k. Assume that the system starts from an initial state X_0. Under the assumption that the LOS distribution is geometric, it follows from (6) that

$$\begin{aligned} X_k &= X_0 + A_{(0,k]} - D_{(0,k]} \\ &= X_0 + A_{(0,k]} - \sum_{i=1}^{Z_0+\ldots+Z_{k-1}} \xi_i, \end{aligned} \tag{27}$$

where $\{\xi_i : i = 1, 2, \ldots\}$ is a sequence of iid Bernoulli random variables with probability $\mu = 1/m$ taking value 1, and $Z_i = \min(X_i, N)$ is the number of busy servers at the midnight of day i. After proper centering, it follows from (27) that

$$X_k - N = Y_k + \mu \sum_{i=0}^{k-1} (X_i - N)^-, k = 0, 1, 2 \ldots \tag{28}$$

where, for an $x \in \mathbb{R}$, $x^- = \max(-x, 0)$,

$$Y_k = X_0 - N + (A_{(0,k]} - k\Lambda) - \sum_{i=1}^{Z_0+\ldots+Z_{k-1}} (\xi_i - \mu) + k(\Lambda - \mu N), \tag{29}$$

and we have used the fact that $(N-Z_i)$ is equal to the number of idle servers $(X_i-N)^-$ at the midnight of day i.

We define the diffusion-scaled processes $\tilde{X}_k = (X_k - N)/\sqrt{N}$. It follows from (28) that

$$\tilde{X}_k = \tilde{Y}_k + \mu \sum_{i=0}^{k-1} (\tilde{X}_i)^- \quad k = 0, 1, 2 \ldots, \tag{30}$$

where $\tilde{Y}_k = \frac{1}{\sqrt{N}} Y_k$. When the daily arrival rate Λ is high and the number of servers N is large so that $\beta = \sqrt{N}(1-\rho)$ is moderate (where ρ is defined in (8)), we propose to use a discrete-time diffusion process $\{\tilde{X}_k^* : k = 0, 1, 2, \ldots\}$ to replace the diffusion-scaled midnight customer count process $\{\tilde{X}_k : k = 0, 1, 2, \ldots\}$. The discrete-time diffusion process $\{\tilde{X}_k^* : k = 0, 1, 2, \ldots\}$ satisfies the following equation:

$$\tilde{X}_k^* = \tilde{Y}_k^* + \mu \sum_{i=0}^{k-1} (\tilde{X}_i^*)^-, \quad k = 0, 1, 2, \ldots. \tag{31}$$

Here $\tilde{Y}_k^* = \tilde{Y}^*(k)$ for $k = 0, 1, 2, \ldots$, and $\{\tilde{Y}^*(t), t \geq 0\}$ is a Brownian motion with mean $-\mu\beta$ and variance $\sigma^2 = \rho\mu(2-\mu)$. The discrete-time process $\{\tilde{Y}_k^* : k = 0, 1, 2 \ldots\}$ is an embedding of the corresponding Brownian motion. In other words, it is a random walk with the step sizes being iid following a normal distribution with mean $-\mu\beta$ and variance $\sigma^2 = \rho\mu(2-\mu)$. This approximation is based on the limit theorem we will prove in Section 4.6.

Stationary distribution of the the discrete-time diffusion process

Ideally, we want to obtain an explicit formula for the stationary density $\pi^*(\cdot)$ of the discrete-time diffusion process $\{\tilde{X}_k^* : k = 0, 1, 2, \ldots\}$. Using $\pi^*(\cdot)$, we can approximate the stationary distribution of the unscaled midnight customer count by

$$\mathbb{P}(X_\infty = x) \approx \pi^* \left(\frac{x - N}{\sqrt{N}} \right) / \sqrt{N}. \tag{32}$$

At the current stage, it is challenging to obtain the exact formula for $\pi^*(x)$. Alternative, we propose an approximation for π^*. The approximate formula $\tilde{\pi}(x)$ is

given in (33) below:

$$\tilde{\pi}(x) = \begin{cases} \alpha_1 \exp(-\frac{2\mu\beta x}{\sigma^2}), & x \geq 0; \\ \alpha_2 \exp(-\frac{(2\mu-\mu^2)(x+\beta)^2}{2\sigma^2}), & x < 0; \end{cases} \tag{33}$$

where α_1 and α_2 are normalizing constants that make $\tilde{\pi}(x)$ continuous at zero. In Section 4.5.4, we show that this approximate density $\tilde{\pi}(x)$ can already produce remarkably good approximation for the stationary distribution of the midnight count as well as the time-dependent performance measures of our interest.

To explain the rational of proposing the approximate formula (33), note that discrete-time diffusion process $\{\tilde{X}_k^* : k = 0, 1, 2 \ldots\}$ can be seen as a discrete-time analog of the following (continuous-time) piecewise-linear diffusion process $\{\check{X}(t), t \geq 0\}$ which satisfies

$$\check{X}(t) = \check{Y}(t) + \mu \int_0^t (\check{X}(s))^- ds, \quad t \geq 0, \tag{34}$$

where $\{\check{Y}(t), t \geq 0\}$ is a Brownian motion. It is well known that the diffusion limits for $GI/M/n$ queues in the Quality- and Efficiency-Driven (QED) regime have the same form as $\{\check{X}(t), t \geq 0\}$ [66]; also see more discussion on the QED regime in Section 4.6. Based on this analogy, we can see that $\{\tilde{X}_k^* : k = 0, 1, 2 \ldots, \}$ behaves as a discrete version of the Ornstein-Uhlenbeck (OU) process on $(-\infty, 0]$ and as a reflected random walk on $[0, \infty)$. Motivated by the technique introduced in [18] and based on the fact that the stationary density of the discrete-time OU process has a Gaussian form and that the reflected random walk has an exponential tail [88, 137, 11], we thus propose using $\tilde{\pi}(x)$ to approximate the stationary density $\pi^*(x)$. The proof that the stationary density of the discrete-time OU process has a Gaussian form is provided in Appendix C.1.

4.5.2 Approximate the hourly customer count

Recall the hourly customer count at time t can be represented as

$$X(t) = X_k + A_{(k,t]} - D_{(k,t]},$$

where $k = \lfloor t \rfloor$ is the most recent midnight before time t so that $k \leq t < k+1$. Moreover, $A_{(k,t]}$ denotes the total number of arrivals from the midnight of day k to time t. Its mean equals $\Lambda\psi(t)$ with $\psi(t) = \int_0^{t-k} \lambda(s)ds / \int_0^1 \lambda(s)ds$. We use $D_{(k,t]}$ denote the total number of discharges from the midnight of day k to time t, which equals

$$D_{(k,t]} = \sum_{i=0}^{Z_k} \xi_{i,t}$$

where $\xi_{i,t}$ is sequence of iid Bernoulli random variables with success probability $\mu\mathbb{P}(h_{\text{dis}} \leq t-k)$, h_{dis} is the discharge time (time of the day), and Z_k is the number of busy servers at the midnight of day k. We introduce $\phi(t) = \mathbb{P}(h_{\text{dis}} \leq t-k)$, and the mean of $D_{(k,t]}$ equals $Z_k^N \mu\phi(t)$. We can see that the arrival rate pattern determines $\psi(t)$, while the discharge distribution determines $\phi(t)$.

Similar to the midnight count process, we first do proper centering for $X(t)$:

$$X(t) - N - N\mu(\psi(t) - \phi(t)) = (X_k - N) + Y(t) + \mu\phi(t)(X_k - N)^-, \tag{35}$$

where

$$Y(t) = \big(A_{(k,t]} - \Lambda\psi(t)\big) - \big(D_{(k,t]} - Z_k\mu\phi(t)\big) + (\Lambda - N\mu)\psi(t),$$

and we again use the fact that $(N - Z_k) = (X_k - N)^-$ is the number of idle servers at the midnight of day k.

Let

$$\begin{aligned}
\tilde{X}(t) &= \frac{1}{\sqrt{N}}\left(X(t) - N - N\mu(\psi(t) - \phi(t))\right), \\
\tilde{A}_{(k,t]} &= \frac{A_{(k,t]} - \Lambda\psi(t)}{\sqrt{N}}, \\
\tilde{D}_{(k,t]} &= \frac{D_{(k,t]} - Z_k\mu\phi(t)}{\sqrt{N}}.
\end{aligned}$$

Dividing both sides of (35) by $\sqrt{N}$, we know the diffusion-scaled hourly customer count $\tilde{X}(t)$ satisfies the following:

$$\tilde{X}(t) = \tilde{X}_k + \tilde{Y}(t) + \mu\phi(t)\left(\tilde{X}_k\right)^-, \tag{36}$$

where $\tilde{X}_k$ is the diffusion-scaled midnight customer count, and

$$\tilde{Y}(t) = \tilde{A}_{(k,t]} - \tilde{D}_{(k,t]} + \sqrt{N}\mu\psi(t)(\rho - 1).$$

When the daily arrival rate Λ is high and the number of servers N is large so that $\beta = \sqrt{N}(1-\rho)$ is moderate, we can use then density $\tilde{\pi}$ to approximate the stationary distribution of $\tilde{X}_k$ as described in Section 4.5.1. Moreover, the diffusion-scaled arrival $\tilde{A}_{(k,t]}$ and discharge $\tilde{D}_{(k,t]}$ can be approximated by two normal random variables by the central limit theorem (CLT). Thus, the term $\tilde{Y}(t)$ can be approximated by a normal random variable with mean $-\mu\beta\psi(t)$ and variance

$$\gamma^2 = \rho\mu\psi(t) + \rho\mu\phi(t)(1 - \mu\phi(t)).$$

Note that when both $\psi(t)$ and $\phi(t)$ equal 1, $\gamma^2 = \sigma^2 = \rho\mu(2 - \mu)$, where σ^2 is the variance in the diffusion approximations for the midnight count process.

Eventually, the distribution of $\tilde{X}(t)$ can be approximated by a convolution of the diffusion-scaled midnight count (with distribution $\tilde{\pi}$) and a normal random variable. To spell out the details, we know from (36) that conditioning on the value of $\tilde{X}_k = x$, when $x \geq 0$,

$$\tilde{X}(t) = x + \tilde{Y}(t),$$

while for $x < 0$,

$$\tilde{X}(t) = (1 - \mu\phi(t))x + \tilde{Y}(t).$$

Let $h(x)$ denote the pdf of the normal distribution with mean $-\mu\beta\psi(t)$ and variance γ^2. Thus, we propose using the following formula to approximate the stationary density of $\tilde{X}(t) = z$:

$$f(z) = \int_0^\infty h(z - x)\tilde{\pi}(x)dx + \int_{-\infty}^0 h(z - (1 - \mu\phi(t))x)\tilde{\pi}(x)dx. \tag{37}$$

Let $\Upsilon_1 = \int_0^\infty h(z - x)\tilde{\pi}(x)dx$, and $\Upsilon_2 = \int_{-\infty}^0 h(z - (1 - \mu\phi(t))x)\tilde{\pi}(x)dx$. After

doing some algebra, we have

$$\begin{aligned}
\Upsilon_1 &= \int_0^\infty \frac{\alpha_1}{\sqrt{2\pi}\gamma} \exp\left(-\frac{(x-m)^2}{2\gamma^2}\right) \exp\left(-\theta x\right) dx \\
&= \int_0^\infty \frac{\alpha_1}{\sqrt{2\pi}\gamma} \exp\left(-\frac{2m\theta - \gamma^2\theta^2}{2}\right) \exp\left(-\frac{[x-(m-\gamma^2\theta)]^2}{2\gamma^2}\right) dx \\
&= \alpha_1 \exp\left(-\frac{\theta}{2}(2m-\gamma^2\theta)\right)\left(1-\Phi\left(-\frac{m-\gamma^2\theta}{\gamma}\right)\right), \qquad (38)
\end{aligned}$$

where

$$\theta = \frac{2\mu\beta}{\sigma^2}, \quad m = y + \mu\beta\psi(t),$$

$\Phi(x)$ is the CDF of standard normal distribution, and α_1 (as well as α_2 below) are the normalizing constants in $\tilde{\pi}$.

We also have

$$\begin{aligned}
\Upsilon_2 &= \int_{-\infty}^0 \frac{\alpha_2}{\sqrt{2\pi}\gamma} \exp\left(-\frac{(x-l)^2}{2w^2}\right) \exp\left(-\frac{(x+\beta)^2}{2s^2}\right) dx \\
&= \int_{-\infty}^0 \frac{\alpha_2}{\sqrt{2\pi}\gamma} \exp\left(-\frac{\left[x-\frac{s^2l-w^2\beta}{s^2+w^2}\right]^2}{\frac{2w^2s^2}{(s^2+w^2)}}\right) \exp\left(-\frac{(l+\beta)^2}{2(s^2+w^2)}\right) dx \\
&= \frac{\alpha_2}{\gamma}\frac{ws}{\sqrt{w^2+s^2}} \exp\left(-\frac{(l+\beta)^2}{2(s^2+w^2)}\right)\Phi\left(-\frac{s^2l-w^2\beta}{ws\sqrt{w^2+s^2}}\right), \qquad (39)
\end{aligned}$$

where

$$w^2 = \frac{\gamma^2}{(1-\mu\phi(t))^2}, \quad s^2 = \frac{\sigma^2}{2\mu-\mu^2}, \quad l = \frac{z+\mu\beta\psi(t)}{1-\mu\phi(t)}.$$

Plugging (38) and (39) back to (37), we can evaluate $f(z)$ for any given value of $\tilde{X}(t) = z$. Till now, we have obtained closed-form expressions to approximate the stationary distributions for both the midnight customer count and the hourly customer count.

4.5.3 Approximate the time-dependent performance

In Section 4.5.2, we have obtained the approximate stationary distribution for the hourly customer count $X(t)$. Thus, applying (10) again, we can get the approximate mean queue length for any $t \geq 0$.

Next, we focus on introducing how to approximate the hourly mean waiting time $E_\infty[W(t)]$. We briefly introduce how to approximate the 6-hour service level at the end of this sub-section.

Mean waiting time

Consistent with Section 4.2.3, we adopt the batch discharge policy when calculating the mean waiting time, i.e., the discharge hour h_{dis} follows a discrete distribution, taking values on a finite number of points $t_1, t_2, \ldots, t_n$ with probabilities $q_{t_1}, q_{t_2}, \ldots, q_{t_n}$, respectively. As from Section 4.2.3, we can see the key step to get the mean waiting time is to calculate the set of probabilities $p_{t_i}(t)$ and $p_{t_i}^k(t)$. Recall that $p_{t_i}(t)$ denotes the probability that a virtual customer arriving at time t needs to wait till time t_i (the same day) to be admitted, i.e., for $0 \leq t < t_i$, if t_i is the first discharge point after t,

$$p_{t_i}(t) = \mathbb{P}_\infty \left(D_{(0,t]} \leq X(0) + A_{(0,t]} - N < D_{(0,t_i]} \right);$$

otherwise,

$$p_{t_i}(t) = \mathbb{P}_\infty \left(D_{(0,t_{i-1}]} \leq X(0) + A_{(0,t]} - N < D_{(0,t_i]} \right).$$

The probability $p_{t_i}^k(t)$ denotes the probability that a virtual customer arriving at time t needs to wait till time t_i on k days later ($k = 1, 2, \ldots$) to be admitted, i.e.,

$$p_{t_i}^k(t) = \mathbb{P}_\infty \left(D_{(0,k+t_{i-1}]} \leq X(0) + A_{(0,t]} - N < D_{(0,k+t_i]} \right).$$

Thus, we introduce how to approximate these probabilities.

Similar to Section 4.2.3, we focus on discussing the case when $0 \leq t < t_1$. We first show how to approximate $p_{t_1}(t) = \mathbb{P}_\infty \left(0 \leq X(0) + A_{(0,t]} - N < D_{(0,t_1]} \right)$. This probability can be further written as

$$\begin{aligned} p_{t_1}(t) &= \mathbb{P}_\infty \{ X(0) - N + A_{(0,t]} < D_{(0,t_1]} \} \\ &\quad - \mathbb{P}_\infty \{ X(0) - N + A_{(0,t]} < D_{(0,t_1]}, X(0) - N + A_{(0,t]} < 0 \} \\ &= \mathbb{P}_\infty \{ X(0) - N + A_{(0,t]} < D_{(0,t_1]} \} - \mathbb{P}_\infty \{ X(0) - N + A_{(0,t]} < 0 \}. \end{aligned}$$

Let $p^{\mathrm{L}}_{t_1}(t)$ denote $\mathbb{P}_\infty\{X(0) - N + A_{(0,t]} < D_{(0,t_1]}\}$, and $p^{\mathrm{R}}_{t_1}(t)$ denote $\mathbb{P}_\infty\{X(0) - N + A_{(0,t]} < 0\}$. We have that

$$\begin{aligned} p^{\mathrm{L}}_{t_1}(t) &= \mathbb{P}_\infty\big((X(0) - N) + A_{(0,t]} - \Lambda\psi(t) \\ &\quad < (D_{(0,t_1]} - Z(0)\mu\phi(t_1)) + \mu\phi(t_1)(Z(0) - N) + N\mu\phi(t_1) - \Lambda\psi(t)\big) \\ &= \mathbb{P}_\infty\left(\tilde{X}(0) + \tilde{A}_{(0,t]} < \tilde{D}_{(0,t_1]} - \mu\phi(t_1)(\tilde{X}(0))^- + \sqrt{N}\mu\phi(t_1) - \sqrt{N}\mu\rho\psi(t)\right). \end{aligned}$$

and

$$\begin{aligned} p^{\mathrm{R}}_{t_1}(t) &= \mathbb{P}_\infty\big((X(0) - N) + A_{(0,t]} - \Lambda\psi(t) < -\Lambda\psi(t)\big) \\ &= \mathbb{P}_\infty\left(\tilde{X}(0) + \tilde{A}_{(0,t]} < -\sqrt{N}\mu\rho\psi(t)\right). \end{aligned}$$

Conditioning on the value of the diffusion-scaled midnight count $\tilde{X}(0) = x$, we have that

$$\begin{aligned} p_{t_1}(t)|x &= p^{\mathrm{L}}_{t_1}(t)|x - p^{\mathrm{R}}_{t_1}(t)|x \\ &= \mathbb{P}\left(\tilde{A}_{(0,t]} - \tilde{D}_{(0,t_1]} < -x - \mu\phi(t_1)x^- + \sqrt{N}\mu\phi(t_1) - \sqrt{N}\mu\rho\psi(t)\right) \\ &\quad - \mathbb{P}\left(\tilde{A}_{(0,t]} < -x - \sqrt{N}\mu\rho\psi(t)\right). \end{aligned}$$

Recall that $\tilde{A}_{(0,t]}$ and $\tilde{D}_{(0,t_1]}$ are independent and we can use two normal random variables to approximate $\tilde{A}_{(0,t]}$ and $\tilde{A}_{(0,t]} - \tilde{D}_{(0,t_1]}$, respectively. Thus, we can approximate $p_{t_1}(t)|x$ by

$$p_{t_1}(t)|x \approx \Phi\left(\frac{g_1(x)}{\delta_1}\right) - \Phi\left(\frac{g_0(x)}{\delta_0}\right).$$

where Φ is the CDF of the standard normal distribution,

$$g_0(x) = -x - \sqrt{N}\mu\rho\psi(t), \quad g_1(x) = -x - \mu\phi(t_1)x^- + \sqrt{N}\mu\phi(t_1) - \sqrt{N}\mu\rho\psi(t),$$

$\delta_0 = \sqrt{\rho\mu\psi(t)}$, and $\delta_1 = \sqrt{\rho\mu\psi(t) + \rho\mu\phi(t_1)(1 - \mu\phi(t_1))}$. Then, we can approximate the unconditional probability $p_{t_1}(t)$ by using the approximate stationary density $\tilde{\pi}$ for $\tilde{X}(0)$, i.e.,

$$p_{t_1}(t) \approx \int_{-\infty}^{\infty} \Phi\left(\frac{g_1(x)}{\delta_1}\right)\tilde{\pi}(x)dx - \int_{-\infty}^{\infty} \Phi\left(\frac{g_0(x)}{\delta_0}\right)\tilde{\pi}(x)dx. \tag{40}$$

Next, we show how to approximate a general $p_{t_i}(t)$ for $0 \leq t < t_{i-1} < t_i$. Similarly, this probability can be further written as

$$\begin{aligned} p_{t_i}(t) &= \mathbb{P}_\infty\left((X(0)-N)+A_{(0,t]} < D_{(0,t_i]}\right) \\ &\quad - \mathbb{P}_\infty\left((X(0)-N)+A_{(0,t]} < D_{(0,t_{i-1}]}\right) \\ &= \mathbb{P}_\infty\left(\tilde{X}(0)+\tilde{A}_{(0,t]} < \tilde{D}_{(0,t_i]} - \mu\phi(t_i)(\tilde{X}(0))^- + \sqrt{N}\mu\phi(t_i) - \sqrt{N}\mu\rho\psi(t)\right) \\ &\quad - \mathbb{P}_\infty\left(\tilde{X}(0)+\tilde{A}_{(0,t]} < \tilde{D}_{(0,t_{i-1}]} - \mu\phi(t_{i-1})(\tilde{X}(0))^- + \sqrt{N}\mu\phi(t_{i-1}) - \sqrt{N}\mu\rho\psi(t)\right). \end{aligned}$$

Conditioning on the value of the diffusion-scaled midnight count $\tilde{X}(0) = x$ and using two normal random variables to approximate $\tilde{A}_{(0,t]} - \tilde{D}_{(0,t_i]}$ and $\tilde{A}_{(0,t]} - \tilde{D}_{(0,t_{i-1}]}$, we have

$$\begin{aligned} p_{t_i}(t)|x &= \mathbb{P}\left(\tilde{A}_{(0,t]} - \tilde{D}_{(0,t_i]} < -x - \mu\phi(t_i)(x)^- + \sqrt{N}\mu\phi(t_i) - \sqrt{N}\mu\rho\psi(t)\right) \\ &\quad - \mathbb{P}\left(\tilde{A}_{(0,t]} - \tilde{D}_{(0,t_{i-1}]} < -x - \mu\phi(t_{i-1})(x)^- + \sqrt{N}\mu\phi(t_{i-1}) - \sqrt{N}\mu\rho\psi(t)\right) \\ &\approx \Phi\left(\frac{g_i(x)}{\delta_i}\right) - \Phi\left(\frac{g_{i-1}(x)}{\delta_{i-1}}\right). \end{aligned}$$

where

$$g_i(x) = -x - \mu\phi(t_i)x^- + \sqrt{N}\mu\phi(t_i) - \sqrt{N}\mu\rho\psi(t),$$

and $\delta_i = \sqrt{\rho\mu\psi(t) + \rho\mu\phi(t_i)(1-\mu\phi(t_i))}$. Then similarly, we can approximate the unconditional probability $p_{t_i}(t)$ by using the approximate stationary density $\tilde{\pi}$ for $\tilde{X}(0)$, i.e.,

$$p_{t_i}(t) \approx \int_{-\infty}^{\infty} \Phi\left(\frac{g_i(x)}{\delta_i}\right)\tilde{\pi}(x)dx - \int_{-\infty}^{\infty} \Phi\left(\frac{g_{i-1}(x)}{\delta_{i-1}}\right)\tilde{\pi}(x)dx. \tag{41}$$

Finally, we show how to how to approximate the overnight waiting probabilities. We illustrate with $p^1_{t_i}(t)$ for $0 \leq t < t_1$. This probability can be further written as

$$\begin{aligned} p^1_{t_i}(t) &= \mathbb{P}_\infty\left((X(0)-N)+A_{(0,t]} < D_{(0,1+t_i]}\right) \\ &\quad - \mathbb{P}_\infty\left((X(0)-N)+A_{(0,t]} < D_{(0,1+t_{i-1}]}\right) \\ &= \mathbb{P}_\infty\left(\tilde{X}(0)+\tilde{A}_{(0,t]} < \tilde{D}_{(0,1+t_i]} - \mu(\tilde{X}(0))^- + \sqrt{N}\mu(1+\phi(t_i)) - \sqrt{N}\mu\rho\psi(t)\right) \\ &\quad - \mathbb{P}_\infty\left(\tilde{X}(0)+\tilde{A}_{(0,t]} < \tilde{D}_{(0,1+t_{i-1}]} - \mu(\tilde{X}(0))^- + \sqrt{N}\mu(1+\phi(t_{i-1})) - \sqrt{N}\mu\rho\psi(t)\right). \end{aligned}$$

Conditioning on the value of the diffusion-scaled midnight count $\tilde{X}(0) = x$, and using two normal random variables to approximate $\tilde{A}_{(0,t]} - \tilde{D}_{(0,1+t_i]}$ and $\tilde{A}_{(0,t]} - \tilde{D}_{(0,1+t_{i-1}]}$, we have

$$\begin{aligned} p_{t_i}(t)|x &= \mathbb{P}\left(\tilde{A}_{(0,t]} - \tilde{D}_{(0,1+t_i]} < -x - \mu x^- + \sqrt{N}\mu(1+\phi(t_i)) - \sqrt{N}\mu\rho\psi(t)\right) \\ &\quad - \mathbb{P}\left(\tilde{A}_{(0,t]} - \tilde{D}_{(0,1+t_{i-1}]} < -x - \mu x^- + \sqrt{N}\mu(1+\phi(t_{i-1})) - \sqrt{N}\mu\rho\psi(t)\right) \\ &\approx \Phi\left(\frac{g_i^1(x)}{\delta_i^1}\right) - \Phi\left(\frac{g_{i-1}^1(x)}{\delta_{i-1}^1}\right). \end{aligned}$$

where

$$g_i^1(x) = -x - \mu x^- + \sqrt{N}\mu(1+\phi(t_i)) - \sqrt{N}\mu\rho\psi(t),$$

and $\delta_i^1 = \sqrt{\rho\mu\psi(t) + \rho\mu(1-\mu) + \mu\phi(t_i)(1-\mu\phi(t_i))}$. Then similarly, we can approximate the unconditional probability $p_{t_i}^1(t)$ by using the approximate stationary density $\tilde{\pi}$ for $\tilde{X}(0)$, i.e.,

$$p_{t_i}^1(t) \approx \int_{-\infty}^{\infty} \Phi\left(\frac{g_i^1(x)}{\delta_i^1}\right)\tilde{\pi}(x)dx - \int_{-\infty}^{\infty} \Phi\left(\frac{g_{i-1}^1(x)}{\delta_{i-1}^1}\right)\tilde{\pi}(x)dx. \tag{42}$$

Unfortunately, there is no closed-form expression for the probabilities in (40) to (42). Because $\int_{-\infty}^{0} \Phi\left(\frac{g_i(x)}{\delta_i}\right)\tilde{\pi}(x)dx$ or $\int_{-\infty}^{0} \Phi\left(\frac{g_i^1(x)}{\delta_i^1}\right)\tilde{\pi}(x)dx$ involves the convolution between Φ and a normal density, which in general does not have a closed-form expression. Thus, we need to numerically evaluate these probabilities.

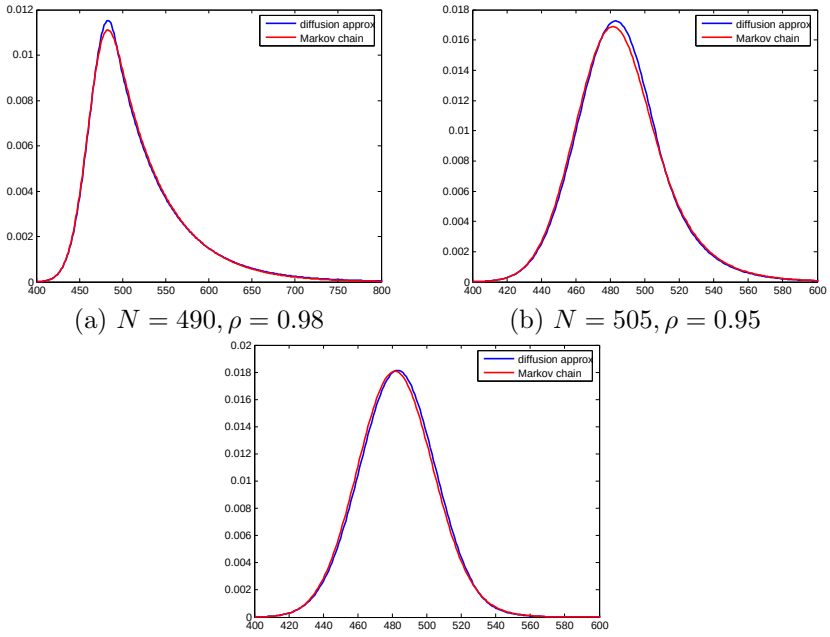
6-hour service level

Once we have the set of the probabilities $p_{t_i}(t)$ and $p_{t_i}^k(t)$, we can apply (21) to obtain the 6-hour service level. When allocation delay is presented, we then just need to apply (26).

4.5.4 Numerical results on the diffusion approximation

Midnight count approximation

Figure 49 compares the stationary distributions of the midnight customer count solved from the exact Markov chain analysis (following the algorithm in Section 4.2.1) and



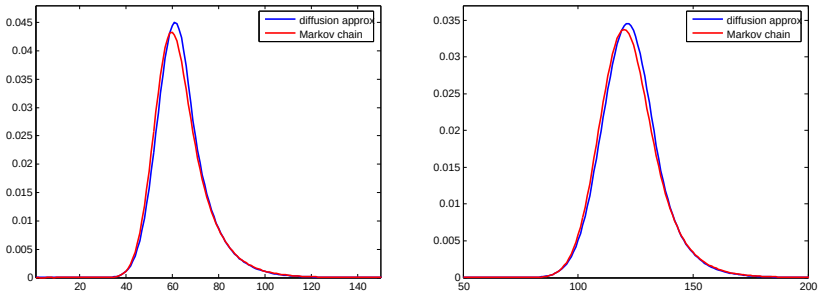
(a) $N = 490, \rho = 0.98$ (b) $N = 505, \rho = 0.95$

(c) $N = 525, \rho = 0.91$

Figure 49: **Stationary distribution of the midnight customer count from exact Markov chain analysis and diffusion approximation (large systems).** Red curves are computed through the Markov chain analysis described in Section 4.2.1; the blue curves are computed through the diffusion approximations in Section 4.5. Parameters for the $M_{\text{peri}}/G_{\text{2timeScale}}/N$ queue: $\Lambda = 90.95$, $\mu = 1/5.3$, and $\lambda(\cdot)$ has the shape as the solid curve in Figure 13. No allocation delay is included.

from using the density $\tilde{\pi}$ and (32). The parameter setting for these experiments remain the same as we introduced in Section 4.4 except that no allocation delay is included. We do not specify the discharge distribution either since the midnight customer count distribution is *not* sensitive to the discharge distribution. We test three sets of N ($N = 490$, 505, 525) with the utilization ρ ranging from 91% to 98%. It is clear from Figure 49 that the approximation based on (32) and (33) is quite accurate.

We have also computed the stationary distributions of the midnight count on smaller systems with $N = 66$ and 132. In these experiments, we fix the mean LOS = 5.3 days and proportionally scaled the hourly arrival rate, so that the daily arrival rate $\Lambda = 11.37$ for $N = 66$ and $\Lambda = 22.74$ for $N = 132$. Figure 50 show the corresponding plots of the midnight customer count distribution. We can see that even when N



(a) $N = 66, \Lambda = 11.37, \rho = 0.91$ (b) $N = 132, \Lambda = 22.74, \rho = 0.91$

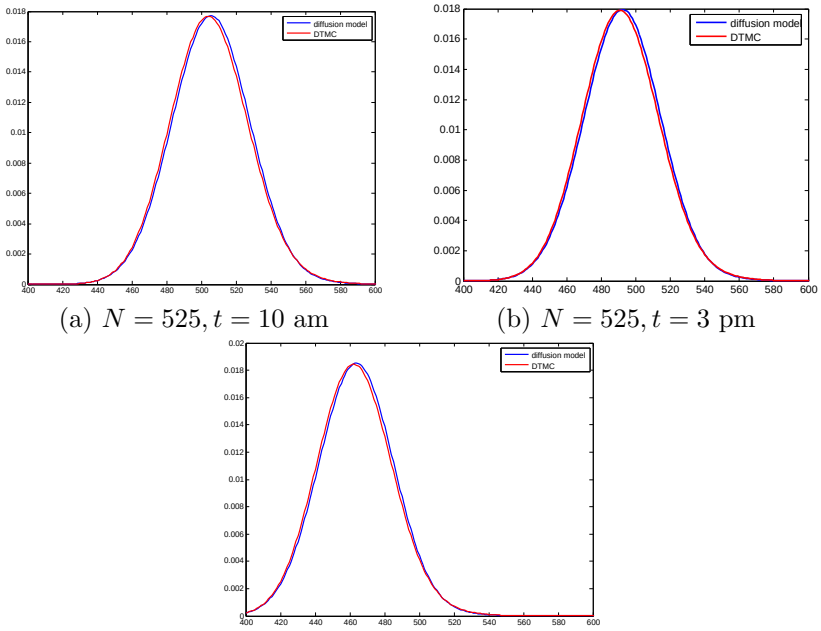
Figure 50: **Stationary distribution of the midnight customer count from exact Markov chain analysis and diffusion approximation (small systems).** Red curves are computed through the Markov chain analysis described in Section 4.2.1; the blue curves are computed through the diffusion approximations in Section 4.5. Parameters for the $M_{\text{peri}}/G_{\text{2timeScale}}/N$ queue: $\mu = 1/5.3$, $\lambda(\cdot)$ has the same shape as the solid curve in Figure 13 but is proportionally adjusted so that $\Lambda = 11.37$ for $N = 66$ and $\Lambda = 22.74$ for $N = 132$. No allocation delay is included.

is only 66 and the utilization is 91%, the stationary distribution computed from the diffusion approximation is still very close to the one solved from the Markov chain analysis.

Hourly count approximation

Figure 51 compares the stationary distribution of the hourly customer count $X(t)$ for certain time t ($t = 10/24, 15/24, 20/24$, corresponding to 10am, 3pm, and 8pm, respectively) when $N = 525$. We adopt the Period 1 discharge distribution, while other parameter settings remain the same as in the experiments for the midnight count. From Figure 51, we can see that the distribution curves from the diffusion approximations are again very close to those obtained from exact Markov chain analysis for the three t we tested.

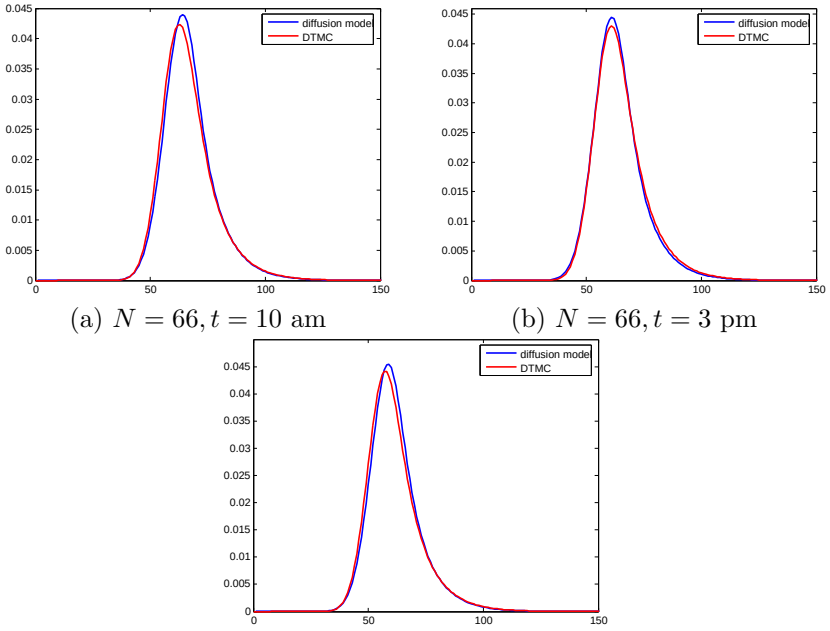
Similarly, we have compared the hourly count distributions on smaller systems. Figure 52 show that the diffusion approximation still performs well when the system size is small ($N = 66$). The distribution curves for $X(t)$ are still very close to each other for the three t values we have tested.



(a) $N = 525, t = 10$ am (b) $N = 525, t = 3$ pm

(c) $N = 525, t = 8$ pm

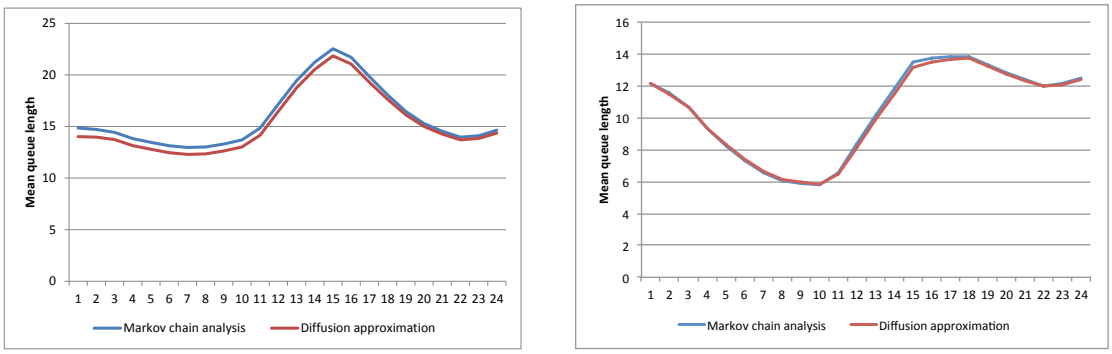
Figure 51: **The stationary distribution of $X(t)$ from exact Markov chain analysis and diffusion approximation for $N = 525$.** Mean LOS is 5.3 days; $\Lambda = 90.95, \rho = 0.92$; no allocation delay is modeled. Period 1 discharge distribution is used.



(a) $N = 66, t = 10$ am (b) $N = 66, t = 3$ pm

(c) $N = 66, t = 8$ pm

Figure 52: **The stationary distribution of $X(t)$ from exact Markov chain analysis and diffusion approximation for $N = 66$.** Mean LOS is 5.3 days; $\Lambda = 11.37, \rho = 0.91$; no allocation delay is modeled. Period 1 discharge distribution is used.



(a) $N = 505, \rho = 0.95$ (b) $N = 525, \rho = 0.92$

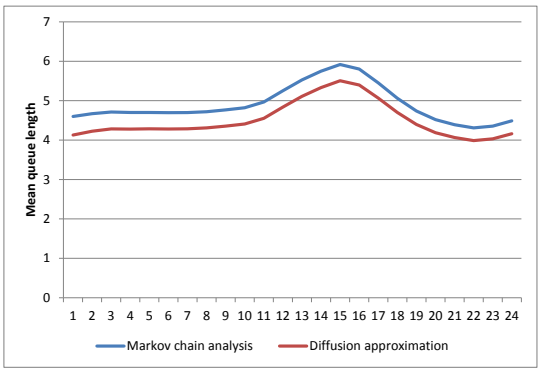
Figure 53: **Time-dependent mean queue length from exact Markov chain analysis and diffusion approximation (large systems).** $N = 505, 525$, $\Lambda = 90.95$, mean LOS is 5.3 days; no allocation delay is modeled. Period 1 discharge distribution is used.

Time-dependent performance measures

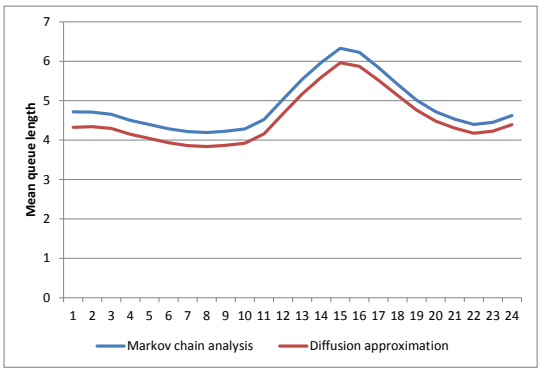
Below we compare the hourly performance measures obtained from the exact analysis and diffusion approximation. The parameter setting for each experiment remains the same as the corresponding setting used in plotting the hourly customer count distribution. In addition, we include the allocation delays so that the figures are comparable to those demonstrated in Section 4.4. The allocation delay T_{alloc} follows a log-normal distribution with mean 2.5 hours and CV 1.

Mean queue length. Figure 53 compares the time-dependent mean queue length curves for $N = 505$ and $N = 525$. We can see the approximate time-dependent mean queue length curve is very close to the one obtained from the exact analysis, especially when $N = 525$. Figure 54 show that the mean-queue length curves for smaller system size ($N = 66$ or 123). In smaller systems, the mean queue length curves demonstrate more differences between the approximation and exact analysis than what we see in the large systems, but the two curves are still fairly close to each other (the maximum difference is less than 0.5).

Mean waiting time and 6-hour service level. Figures 55 to 57 plots the time-dependent mean waiting time and 6-hour service level curves for $N = 525$,



(a) $N = 66, \Lambda = 11.37, \rho = 0.91$

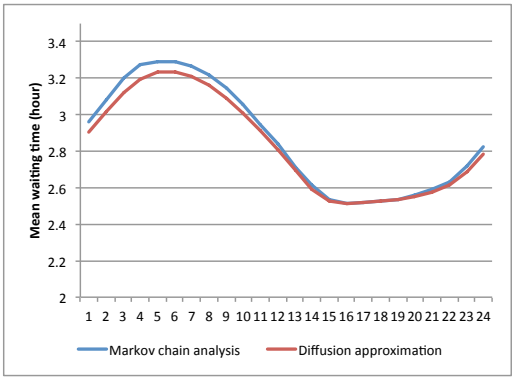


(b) $N = 132, \Lambda = 22.74, \rho = 0.91$

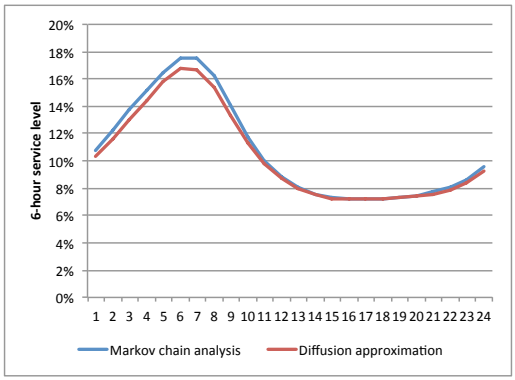
Figure 54: **Time-dependent mean queue length from exact Markov chain analysis and diffusion approximation (small systems).** $N = 66, 132$, mean LOS is 5.3 days; no allocation delay is modeled. Period 1 discharge distribution is used.

$N = 132$, and $N = 66$, respectively. Generally speaking, the diffusion approximation works better on approximating the 6-hour service level than approximating the mean waiting time. Also, it is not surprising that the approximations are better for larger systems ($N = 505$ and $N = 132$) than for small systems ($N = 66$).

Note that the gap between the diffusion approximation and Markov chain analysis can come from two sources: (i) we do not have the exact formula for the stationary density π^* of the diffusion-scaled midnight count and have to use the approximate density $\tilde{\pi}$ in (33); (ii) even if we feed in the exact stationary density obtained from Markov chain analysis, the approximation itself may not be accurate since we use normal random variables to approximate the hourly arrival and discharge. To separate these two sources of inaccuracy, in an additional set of experiments, we feed in the exact stationary distribution for the midnight customer count (which are obtained from Markov chain analysis), and then use the approximation introduced in Section 4.5.3 to get the hourly mean waiting time and 6-hour service level. We find that when we eliminate the first source of inaccuracy, the diffusion approximation produces performance curves that are almost identical to those obtained from Markov chain analysis. For example, as illustrated in Figure 58, for $N = 66$, the gap in the mean waiting time

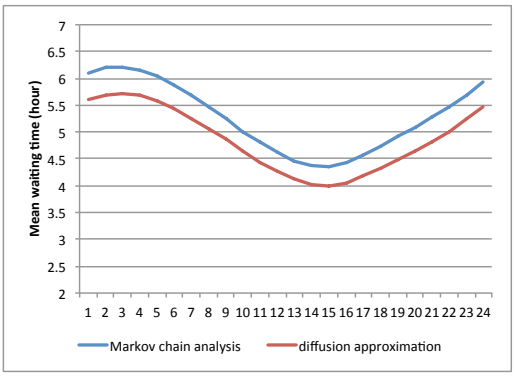


(a) Mean waiting time

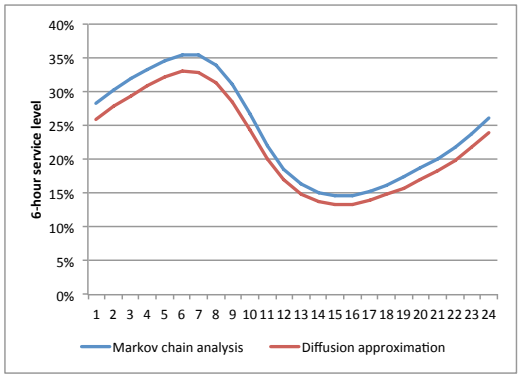


(b) 6-hour service level

Figure 55: **Time-dependent mean waiting time and 6-hour service level from exact Markov chain analysis and diffusion approximation ($N = 525$).** $N = 525$, $\Lambda = 90.95$, mean LOS is 5.3 days; allocation delay follows a log-normal distribution with mean 2.5 hours and CV 1. Period 1 discharge distribution is used.

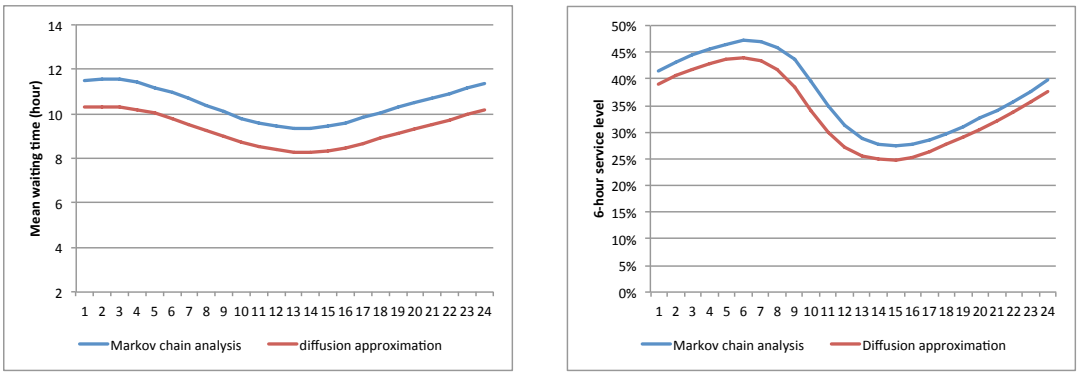


(a) Mean waiting time



(b) 6-hour service level

Figure 56: **Time-dependent mean waiting time and 6-hour service level from exact Markov chain analysis and diffusion approximation ($N = 132$).** $N = 132$, $\Lambda = 22.74$, mean LOS is 5.3 days; allocation delay follows a log-normal distribution with mean 2.5 hours and CV 1. Period 1 discharge distribution is used.



(a) Mean waiting time

(b) 6-hour service level

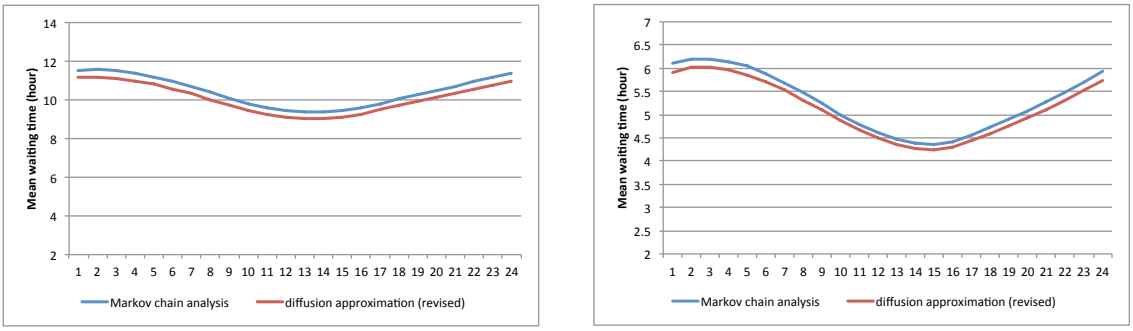
Figure 57: **Time-dependent mean waiting time and 6-hour service level from exact Markov chain analysis and diffusion approximation ($N = 66$).** $N = 66$, $\Lambda = 11.37$, mean LOS is 5.3 days; allocation delay follows a log-normal distribution with mean 2.5 hours and CV 1. Period 1 discharge distribution is used.

between Markov chain analysis and diffusion approximation reduces from 1.2 hours to only around 0.35 hour after using the exact midnight distribution. Therefore, the results indicate that the gap we observed in Figures 55 to 57 are mainly from the first source, i.e., $\tilde{\pi}$ is only an approximation to π^*.

4.6 Diffusion limits for the single-pool model

In Section 4.5, we have demonstrated the diffusion approximations based on Equations 32 and 33 provide an efficient tool to approximate the stationary distribution of the customer count and predict the time-dependent performance measures. These approximations are based on the convergence of stochastic processes. In this section, we prove the limit theorem that supports the diffusion approximation.

Instead of fixing the number of servers N, we now consider a sequence of $M_{\text{peri}}/G_{2\text{timeScale}}/N$ queues indexed by N, i.e., a sequence of single-pool models described in Section 4.1. Again, we do not consider allocation delays in this section. Let Λ^N be the daily arrival rate of the Nth system. Let $m = 1/\mu$, the mean LOS, be fixed and $\rho^N = (\Lambda^N m)/N$ be the traffic intensity of the Nth system. We assume



(a) $N = 66, \Lambda = 11.37, \rho = 0.91$ (b) $N = 132, \Lambda = 22.74, \rho = 0.91$

Figure 58: **Time-dependent mean waiting time from diffusion approximation when feeding in exact stationary distribution of the midnight customer count ($N = 66$ and $N = 132$).** Blue curves are computed through the Markov chain analysis described in Section 4.2.1; red curves are computed through the hourly diffusion approximations in Section 4.5.3 while the midnight count distributions are obtained from Markov chain analysis. $\mu = 1/5.3$, $\Lambda = 11.37$ for $N = 66$ and $\Lambda = 22.74$ for $N = 132$; allocation delay follows a log-normal distribution with mean 2.5 hours and CV 1. Period 1 discharge distribution is used.

that

$$\lim_{N\to\infty} \Lambda^N/N = \Lambda^*, \text{ and } \lim_{N\to\infty} \sqrt{N}(1-\rho^N) = \beta^* \text{ for some } \beta^* > 0. \tag{43}$$

Analogous to the conventional many-server queues that model customer call centers [54], when condition (43) holds, the sequence of $M_{\text{peri}}/G_{\text{2timeScale}}/N$ systems is said to be in the *Quality- and Efficiency-Driven* (QED) regime.

We use X_k^N to denote the midnight customer count at the midnight of day k in the Nth system. We consider the diffusion-scaled midnight customer count processes $\{\tilde{X}_k^N : k = 0, 1, 2, \ldots\}$ for the sequence of singe-pool systems. Recall that $\tilde{X}_k^N = (X_k^N - N)/\sqrt{N}$ satisfies the following relationship:

$$\tilde{X}_k^N = \tilde{Y}_k^N + \mu \sum_{i=0}^{k-1} (\tilde{X}_i^N)^- \quad k = 0, 1, 2 \ldots, \tag{44}$$

where

$$\tilde{Y}_k^N = \tilde{X}_0^N + \frac{1}{\sqrt{N}} \sum_{i=0}^{k-1} \left(A_i^N - \Lambda^N\right) - \frac{1}{\sqrt{N}} \sum_{i=1}^{Z_0^N + \ldots + Z_{k-1}^N} (\xi_i - \mu) + k\sqrt{N}\mu(\rho^N - 1),$$

$\{A_i^N\}$ is a sequence of Poisson random variables with mean Λ^N, and $\{\xi_i\}$ is a sequence of Bernoulli random variables with success probability $1/\mu$. We assume the initial

condition

$$\tilde{X}_0^N \Rightarrow X_0^* \text{ as } N \to \infty, \tag{45}$$

where $\Rightarrow$ denotes convergence in distribution. Under the many-server heavy-traffic framework (e.g., see [37]), we prove the following limit theorem:

Theorem 1 *Consider a sequence of $M_{peri}/G_{2timeScale}/N$ single-pool systems that satisfies (43) and (45). Then for any integer $K > 0$, $\tilde{X}_\cdot^N \Rightarrow X_\cdot^*$ on the compact set $[0, K]$ as $N \to \infty$. The discrete-time limit process X^* satisfies*

$$X_k^* = Y_k^* + \mu \sum_{i=0}^{k-1} (X_i^*)^-, \quad k = 0, 1, 2, \ldots, K \tag{46}$$

where $Y_k^* = Y^*(k)$ for $k = 0, 1, \ldots,$ and $\{Y^*(t), t \geq 0\}$ is a Brownian motion starting from X_0^* and having mean $-\mu\beta$ and variance $\Lambda^* + \mu(1-\mu)$. Here $\tilde{X}_\cdot^N \Rightarrow X_\cdot^*$ on the compact set means the convergence of the joint K-dimensional distributions for any given K, i.e.,

$$\left(\tilde{X}_0^N, \tilde{X}_1^N, \ldots, \tilde{X}_K^N\right) \Rightarrow (X_0^*, X_1^*, \ldots, X_K^*) \text{ as } N \to \infty. \tag{47}$$

The proof of this theorem is as below. The key step is to prove that $\{\tilde{Y}_k^N, k = 0, 1, \ldots\}$ converges to $\{Y_k^*, k = 0, 1, \ldots\}$ on any given compact set $[0, K]$. Then, the convergence of $\tilde{X}^N$ to X^* naturally follows because of the linear form in (46).

Recall that $\{\tilde{Y}_k^N, k = 0, 1, \ldots\}$ and $\{Y_k^*, k = 0, 1, \ldots\}$ are two random walks, we want to show

$$\tilde{Y}^N \Rightarrow Y^* \text{ as } N \to \infty \tag{48}$$

on any given compact set $[0, K]$ ($K \in \mathbb{Z}^+$), or equivalently,

$$\left(\tilde{Y}_0^N, \tilde{Y}_1^N, \ldots, \tilde{Y}_K^N\right) \Rightarrow (Y_0^*, Y_1^*, \ldots, Y_K^*) \text{ as } N \to \infty. \tag{49}$$

Arrival process

To prove (49), we start from considering the convergence of the diffusion-scaled arrival processes. For the Nth system, we introduce $E_k^N = \frac{1}{\sqrt{N}} \sum_{i=0}^{k-1} \left(A_i^N - \Lambda^N\right)$, which can be re-written as

$$E_k^N = \frac{\sqrt{\frac{\Lambda^N}{N}} \sum_{i=1}^{kN} \zeta_i}{\sqrt{N}}, \tag{50}$$

where $\{\zeta_i\}$ represents a sequence of iid random variables with mean 0 and variance 1. We want to show

$$\left(E_0^N, E_1^N, \ldots, E_K^N\right) \Rightarrow (E_0^*, E_1^*, \ldots, E_K^*) \text{ as } N \to \infty. \tag{51}$$

Here, $E_k^* = E(k)$ is an embedding of the Brownian motion $E(\cdot)$ with drift 0 and variance Λ^*.

To prove (51), we introduce another process $E^{N'}(\cdot)$:

$$E^{N'}(t) = \frac{\sum_{i=1}^{\lfloor tN \rfloor} \zeta_i}{\sqrt{N}}. \tag{52}$$

It is easy to see that

$$E_k^N = \sqrt{\frac{\Lambda^N}{N}} E^{N'}(k) \text{ for } k = 0, 1, \ldots.$$

The convergence of $E^{N'}(\cdot)$ to a standard Brownian motion in the space of $\mathbb{D} = \mathbb{D}([0, \infty), \mathbb{R})$ endowed with the Skorohod's J_1 topology can be easily proven by applying the Donsker's theorem. Then, using Condition (43) that $\lim_{N\to\infty} \Lambda^N/N = \Lambda^*$, we can show $\sqrt{\frac{\Lambda^N}{N}} E^{N'}(\cdot) \Rightarrow E(\cdot)$. Thus, (51) naturally follows since the convergence of the stochastic processes directly implies the convergence of the finite-dimensional joint distributions.

Discharge process

Next, we consider the diffusion-scaled discrete-time discharge process. For the Nth system, we introduce $D_k^N = \frac{1}{\sqrt{N}} \sum_{i=1}^{Z_0^N + \ldots + Z_{k-1}^N} (\xi_i - \mu)$, which can be further rewritten

as

$$D_k^N = \frac{\sum_{i=1}^{T_k^N} \eta_i}{\sqrt{N}}, \tag{53}$$

where

$$T_k^N = \sum_{j=0}^{k-1} Z_j,$$

and $\{\eta_i\}$ represents a sequence of iid random variables with mean 0 and variance $\mu(1-\mu)$. We want to show on any given compact set $[0, K]$,

$$\left(D_0^N, D_1^N, \ldots, D_K^N\right) \Rightarrow (S_0^*, S_1^*, \ldots, S_K^*) \text{ as } N \to \infty. \tag{54}$$

Here, for $0 \leq k \leq K$, $S_k^* = \tilde{S}(k)$, and $\tilde{S}(\cdot)$ is a Brownian motion with drift 0 and variance $\mu(1-\mu)$.

We define

$$\bar{T}_k^N = \frac{T_k^N}{N},$$

and

$$S_n = \sum_{i=1}^{n} \eta_i.$$

Then, we can represent D_k^N as

$$D_k^N = \frac{1}{\sqrt{N}} S_{\bar{T}_k^N N},$$

and

$$\left(D_0^N, D_1^N, \ldots, D_K^N\right) = \left(\frac{1}{\sqrt{N}} S_{\bar{T}_0^N N}, \frac{1}{\sqrt{N}} S_{\bar{T}_1^N N}, \ldots, \frac{1}{\sqrt{N}} S_{\bar{T}_K^N N}\right).$$

Thus, proving (54) is equivalent to showing

$$\left(\frac{1}{\sqrt{N}} S_{\bar{T}_0^N N}, \frac{1}{\sqrt{N}} S_{\bar{T}_1^N N}, \ldots, \frac{1}{\sqrt{N}} S_{\bar{T}_K^N N}\right) \Rightarrow (S_0^*, S_1^*, \ldots, S_K^*) \text{ when } N \to \infty. \tag{55}$$

To prove (55), we again introduce a continuous-time process $\{D^{N'}(t), t \geq 0\}$, which is a composition of two processes, $\frac{1}{\sqrt{N}} S_{\lfloor \cdot N \rfloor} \circ \bar{T}_{\lfloor \cdot \rfloor}^N$. Clearly, when $t = k$, we have

$$D_k^N = \frac{1}{\sqrt{N}} S_{\bar{T}_k^N N} = D^{N'}(k),$$

since $\bar{T}_k^N N$ is always an integer. If we can show

$$\frac{1}{\sqrt{N}} S_{\lfloor \cdot N \rfloor} \Rightarrow \tilde{S}(\cdot) \tag{56}$$

in $\mathbb{D}$ and

$$\bar{T}_{\lfloor \cdot \rfloor}^N \to \bar{T}_{\lfloor \cdot \rfloor} \text{ in probability} \tag{57}$$

in $\mathbb{D}$ with $\bar{T}_{\lfloor t \rfloor} = \lfloor t \rfloor$, then applying the random time change theorem, we can prove (55). Convergence in (56) follows from the Donsker's theorem. Below, we focus on proving (57). To do that, it is sufficient for us to show for each $0 \leq k \leq K$, $\bar{T}_k^N \to k$ almost surely. Equivalently, we use induction to show for each $0 \leq k \leq K$, $Z_k^N / N \to 1$ almost surely.

We first rewrite the system equation (use the fluid scaling)

$$\bar{X}_k^N \;\; = \;\; \bar{Y}_k^N + \sum_{i=0}^{k-1} (\bar{X}_i^N)^-, \tag{58}$$

where

$$\bar{X}_k^N = \frac{X_k^N - N}{N},$$

and

$$\bar{Y}_k^N = \bar{X}_0^N + \frac{\sum_{i=1}^{kN} \zeta_i}{N} - \frac{\sum_{i=1}^{T_k^N} \eta_i}{N} + \frac{k(\rho^N - 1)}{\sqrt{N}}.$$

Assume that $X_0^N = N$, then $\bar{X}_0^N = 0$ and $Z_0^N = N$ ($\bar{X}_0^N \to 0$ is trivial). The induction is as follows:

- When $k = 1$, we have

$$\bar{Y}_1^N = \bar{X}_0^N + \frac{\sum_{i=1}^{N} \zeta_i}{N} - \frac{\sum_{i=1}^{N} \eta_i}{N} + \frac{(\rho^N - 1)}{\sqrt{N}}.$$

  Recall that ζ_i and η_i are centered random variables with mean 0. By Law of Large Numbers, it is obvious that

$$\bar{Y}_1^N \to 0 \quad a.s. \text{ when } N \to \infty.$$

  Thus, $\bar{X}_1^N = \bar{Y}_1^N \to 0$ *a.s.*, and $Z_1^N / N \to 1$ *a.s.*.

- Assume that at k, we have for all $0 \le j \le k$, $\bar{X}_j^N \to 0$ $a.s.$ and $Z_j^N/N \to 1$ $a.s.$. Then for $k+1$, we have $\bar{T}_{k+1}^N \to (k+1)$ $a.s.$ and

$$\begin{aligned} \bar{Y}_{k+1}^N &= \bar{X}_0^N + \frac{\sum_{i=1}^{(k+1)N} \zeta_i}{N} - \frac{\sum_{i=1}^{T_{k+1}^N} \eta_i}{N} + \frac{(k+1)(\rho^N - 1)}{\sqrt{N}} \\ &= \frac{\sum_{i=1}^{(k+1)N} \zeta_i}{N} - \frac{\sum_{i=1}^{T_{k+1}^N} \eta_i}{T_{k+1}^N} \cdot \frac{T_{k+1}^N}{N} + \frac{(k+1)(\rho^N - 1)}{\sqrt{N}} \qquad (59) \\ &\to 0 \quad a.s.. \qquad (60) \end{aligned}$$

  Then

$$\bar{X}_{k+1}^N = \bar{Y}_{k+1}^N + \sum_{j=0}^{k} (\bar{X}_j^N)^- \to 0 \quad a.s.,$$

  completing the proof of Theorem 1.

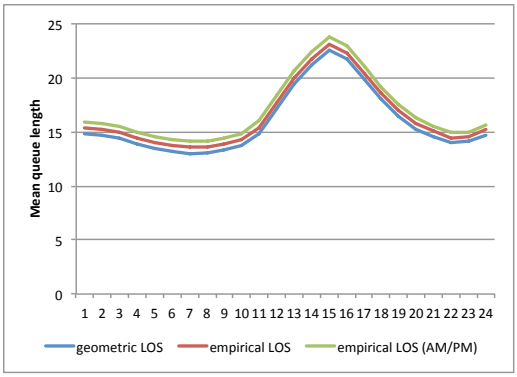
4.7 Conclusion and future work

In this chapter, we have developed an analytical framework, known as the two-time-scale analysis, to predict time-dependent performance measures for a class of time-varying queues that are motivated by modeling hospital inpatient flow management. This analytical framework overcomes many challenges that cannot be solved by existing methods for large-scale queueing systems. They include (a) the arrival process has a time-varying, periodic arrival rate, (b) the service times are no longer exogenous variables but explicitly depend on LOS, admission and discharge times, and (c) service times are extremely long compared with the time variations of the arrival rate. Using the framework, we have developed exact methods and diffusion approximations to predict time-dependent performance measures. This analytical research can greatly advance the understanding of a new class of queueing models and eventually provides insights into analyzing the high-fidelity, multi-pool stochastic processing network models. Our study also provides structural insights into eliminating the excessive long waiting times for admission to inpatient wards from ED.

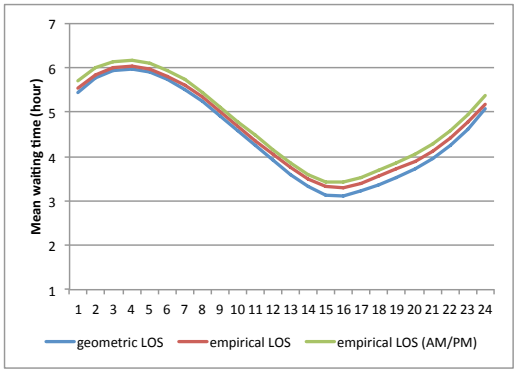
Note that a major simplified assumption we have used in the analysis of this chapter is the geometric distribution for LOS. To examine the impact of this assumption

on system's performance, we conduct simulation experiments to compare with two other LOS distributions: empirical LOS distribution estimated from all Medicine ED-GW patients (no AM/PM difference), and empirical LOS distribution with AM/PM difference. In the latter setting with AM/PM difference, each patient's LOS has a certain chance p to follow the AM distribution, and a chance $1-p$ to follow the PM distribution; see more details on AM/PM difference in Section 2.5.2. The mean LOS is kept the same as 5.3 days, while the coefficient of variation (CV) is increased as comparing to the geometric LOS distribution; the CV is the largest in the setting with AM/PM difference.

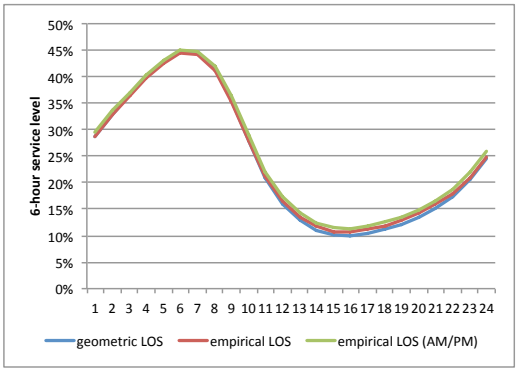
From Figures 59 and 60, we can see in general, when the LOS distribution has a larger CV, the mean waiting time and mean queue length is larger. The 6-hour service level shows little difference though. More importantly, we note that the shapes of the performance curves under different LOS distributions are similar. This indicates that the LOS distribution does not have the first-order effect in capturing the time-dependent empirical performance curves as in Figures 39a to 40a. Incorporating the multi-pool structure with the overflow mechanism is the next priority to better capture the empirical performance curves. Extending the two-time-scale analysis to a multi-pool model will be left in a future project.



(a) Mean queue length

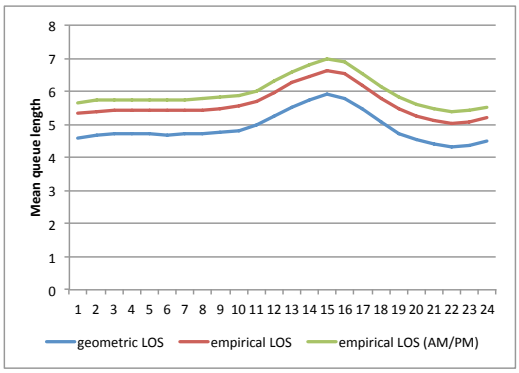


(b) Mean waiting time

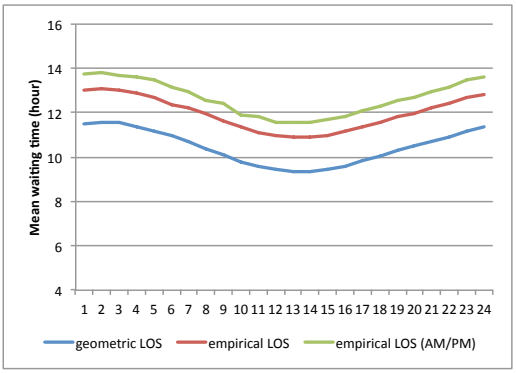


(c) 6-hour service level

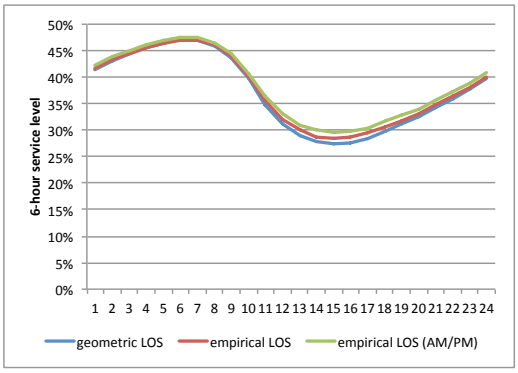
Figure 59: **Time-dependent performance measures under different LOS distributions ($N = 505$).** $N = 505$, $\Lambda = 90.95$, allocation delay follows a log-normal distribution with mean 2.5 hours and CV 1. Period 1 discharge distribution is used. Three LOS distributions are tested: geometric distribution, empirical LOS distribution estimated from all Medicine ED-GW patients (no AM/PM difference), and empirical LOS distribution with AM/PM difference. The mean LOS is 5.3 days for all three settings. The blue curves (using geometric LOS distribution) are obtained from Markov chain analysis, while other curves are from simulation estimates. In the last setting with AM/PM difference, each patient's LOS has a certain chance p to follow the AM distribution, and a chance $1 - p$ to follow the PM distribution; see more details on AM/PM difference in Section 2.5.2.



(a) Mean queue length



(b) Mean waiting time



(c) 6-hour service level

Figure 60: **Time-dependent performance measures under different LOS distributions ($N = 66$).** $N = 66$, $\Lambda = 11.37$, allocation delay follows a log-normal distribution with mean 2.5 hours and CV 1. Period 1 discharge distribution is used. Three LOS distributions are tested: geometric distribution, empirical LOS distribution estimated from all Medicine ED-GW patients (no AM/PM difference), and empirical LOS distribution with AM/PM difference. The mean LOS is 5.3 days for all three settings. The blue curves (using geometric LOS distribution) are obtained from Markov chain analysis, while other curves are from simulation estimates. In the last setting with AM/PM difference, each patient's LOS has a certain chance p to follow the AM distribution, and a chance $1 - p$ to follow the PM distribution; see more details on AM/PM difference in Section 2.5.2.

STOCHASTIC MODELING AND DECISION MAKING IN TWO HEALTHCARE APPLICATIONS: INPATIENT FLOW MANAGEMENT AND INFLUENZA PANDEMICS

PART II

Influenza Pandemic Modeling and Response

by

Pengyi Shi

CHAPTER V

OVERVIEW

5.1 Background

An influenza pandemic is defined as an epidemic of an influenza virus "occurring worldwide, or over a very wide area, crossing international boundaries and usually affecting a large number of people [119]." In contrast to the regular seasonal influenza, pandemics occur infrequently. There were three major pandemics in the 20th century, with the 1918 Spanish influenza pandemic the most severe one in recent history. Due to no previous exposure to the pandemic virus, the population usually has little or no immunity, and thus pandemics can cause much higher levels of mortality. For example, it is estimated that the 1918 Spanish flu caused approximately 50 million deaths [52].

Understanding potential changes in the spread of the disease and predicting the course of a pandemic have always been central issues for public health preparedness. They are particularly important when *multiple* waves of attack is possible. Previous influenza pandemics have shown that an outbreak can consist of multiple waves with intervening periods of relatively lower disease activity. For example, the 1918 pandemic began with an initial smaller herald outbreak in spring 1918 mostly affecting the USA and Europe, subsiding in summer 1918. A second much larger global wave occurred in autumn 1918, affecting both Northern and Southern hemispheres. After disease activity appeared to decline in January 1919, a third wave followed in late winter 1919 and early spring 1919 [113, 143]. The 1956 pandemic has also shown two waves of outbreak with the first wave attacking children and young adults and the second wave mostly attacking elderly [52]. Studies have started to notice the risk

of multiple waves of attack in a pandemic influenza outbreak and emphasized the challenges of making response plans under multiple waves [107, 109].

More recently, when the 2009 H1N1 influenza pandemic emerged in spring 2009, public health officials were prompt to forecast what would happen in fall 2009 and winter 2009-2010. When disease activity increased in September and October 2009, decision makers wondered whether another wave might occur during the winter due to the closure of schools and workplaces in the Holiday season or emergence of new strains caused by viral mutation. Public health officials tried to determine potential response strategies and surge capacity needs based on observed disease transmission characteristics during the course of the 2009 pandemic. They also tried to determine the type of data that should be collected during each stage of the epidemic to assist forecasting.

Computer simulation models can help public health officials forecast the disease spread process and make preparedness plans. In particular, *agent-based simulation* models have become one of the most popular tools to predict the spread of infectious diseases since such models can capture social contact and mixing patterns at individual ("agent") levels. In Part II of this thesis, we use agent-based simulation models with detailed data from the state of Georgia to study potential factors that could cause multiple waves and understand their impact on the entire course of an influenza pandemic. Our study aims to identify what combinations of factors would lead to multiple epidemic waves and examine the characteristics of the subsequent waves. Our eventual goal is to generate insights into intervention strategies and help public health officials make emergency response plans and decide what data to collect.

Specifically, we focus on evaluating the impact of two types of factors on the disease spread process. In Chapter 6, we study the viral characteristic aspect and explore how (i) seasonal changes in transmission dynamics and (ii) viral mutation may affect the course of an influenza pandemic. In Chapter 7, we study how changes

in human social mixing pattens would impact the disease spread process and lead to multiple waves of attack.

In the remainder of this chapter, we first summarize the contributions of our study in Section 5.2. We also give a brief literature review on commonly used disease spread models. Then, we introduce a basic version of the agent-based simulation model in Section 5.3 to provide some background information, since the simulation models we develop in Chapters 6 and 7 are based on this basic version (which is initially developed by [42, 153]). In Chapters 6 and 7, we incorporate the new features such as seasonality, viral mutation, and social mixing changes into the basic version.

5.2 Contributions and literature review

5.2.1 Contributions

Influenza viral pathogen characteristics and behavior are not necessarily static during an epidemic due to seasonal changes or viral mutation, and these non-static behavior may lead to the appearance of multiple waves. Besides, mass changes in social mixing patterns such as mass gathering and Holiday traveling have also been regarded as potential factors that can cause multiple waves of outbreak [109]. However, as far as we know, most existing disease-spread simulation models assume the viral characteristics or social mixing patters are static. These models generally consider a single wave; they did not explicitly model the possibility of multiple waves throughout the pandemic course or specifically evaluate the impact of factors that may lead to multiple waves [46, 101, 56, 153].

In our study, we explicitly incorporate the features of seasonality, viral mutation, and changes of social mixing patterns in the agent-based simulation models. Our models allow us to evaluate the impact of non-static viral characteristic and social mixing change on the entire course of an influenza pandemic. The simulation results predict various scenarios under which multiple waves of attack are possible. The

insights we have gained can alert public health officials and other decision makers the possibility of subsequent waves, and thus can help them better prepare and execute response plans and intervention strategies not only before a pandemic emerges but also throughout its entire course.

More specifically, simulating seasonality and viral mutation scenarios show that the starting month of a pandemic and the timing and degree of viral mutation can substantially alter the course of an influenza pandemic. These two factors, seasonality and mutation, can lead to two- or even three-waves of epidemic outbreaks. Our insights gained from the simulation results can help public health officials to determine what data on the viral characteristics needs to be collected to facilitate forecasting and planning. Our study also confirms the importance of active surveillance and virus typing during the entire course of an pandemic, even when the disease activity starts to subside. These findings can complement previous studies that analyze seasonality and viral mutation for other epidemics such as seasonal influenza [24] and measles [50]. More public health implications from seasonality and viral mutation scenarios are summarized in Section 6.4.

Simulating mass gathering and traveling scenarios show that traveling or gatherings that occur shortly before the epidemic peak may worsen the disease spread, resulting in a higher peak prevalence and total attack rate and in some cases generating two epidemic peaks. These findings can help public officials determine if and when to cancel large public gatherings or enforce regional travel restrictions, advisories, or surveillance during an influenza pandemic (see our recommendations on traveling and mass gatherings in Section 7.4). Although previous studies have suggested that social mixing patterns play an important role in influenza spread and social distancing measures such as school closure may be able to mitigate an epidemic [34, 36, 46, 64, 92, 13, 25, 43, 65, 73, 77, 85, 108, 127], few studies have focused on the opposite of social distancing (i.e., social gatherings) during an epidemic. There

have been studies on the potential effects of national and international travel restrictions, e.g., border closures or international air travel restrictions, but less on local or regional travel [46]. As far as we know, our study takes the first initiative to use simulation models to demonstrate the potential negative impact of Holiday traveling and mass gathering.

5.2.2 Literature review on disease spread models

Various disease spread models have been developed to predict the spread patterns and the effect of intervention strategies for different infectious diseases such as influenza, smallpox and SARS [48, 96]. Generally speaking, the following models are commonly used to model the spread of infectious disease: (i) compartmental differential equation models [53], (ii) agent-based simulation models [46, 101, 56, 153], and (iii) random graph models [23]. The most popular compartmental model is the *S-I-R model*, where every individual is in one of the disease stages: susceptible (S), infected (I), or recovered (R). The cumulative number of people in each stage have instantaneous changes over the time, which are dictated by a set of differential equations. In agent-based simulation models, the entire population is constituted by individuals (agents) and social contact networks, e.g., households and peer groups. Discrete event simulation is used to simulate the spread of the disease on the social networks. In random graph models, random graphs are used to construct the contact network, and the disease spread is predicted accordingly. Rahmandad and Sterman [123] have given a comprehensive comparison between agent-based and differential equation models.

A main feature that distinguishes these disease spread models is the mixing assumption. Homogeneous mixing (compartment models) assumes every individual has the same chance to get infected, while in heterogenous mixing (agent-based simulation models or random graph models), the chance of getting infected for an individual depends on the number of contacts he/she makes during a day and the status of the

people he/she has contacted with. Obviously, agent-based simulation models and random graph models can better represent the actual disease transmissions than the compartment models. For our study in Chapters 6 and 7, we use agent-based simulation models to get more reliable predictions on the course of an influenza pandemic. Moreover, to study the impact of social mixing patterns in Chapter 7, agent-based models naturally become one of the best choices.

5.3 Basic disease spread model

5.3.1 Baseline model settings

The basic agent-based disease spread model developed in [42, 153] is a spatially and temporally explicit agent-based simulation model that consists of a population of computer agents, with each agent representing an individual programmed with socio-demographic characteristics and behaviors. In the rest of Part II of this thesis, we use agent and individual interchangeably. The disease spread model consists of two parts: (i) the natural disease progression within an infected individual, and (ii) the contact network of each individual in the population. We now illustrate the disease progression and contact models.

Disease progression

At the beginning of each simulation run, all agents in the population are susceptible. On Day 1, three infected agents are introduced into the population. Contact with an infectious agent has a probability of transmission of the virus to the susceptible agent. A newly infected agent then progresses through the following stages: Susceptible-Exposed-Infected-Recovered (SEIR), based on the incubation and infectious periods of the disease. After being infected, each agent first progresses through the incubation period, then through the presymptomatic phase, and then has a probability p_A of remaining asymptomatic and a probability $(1 - p_A)$ of becoming symptomatic during

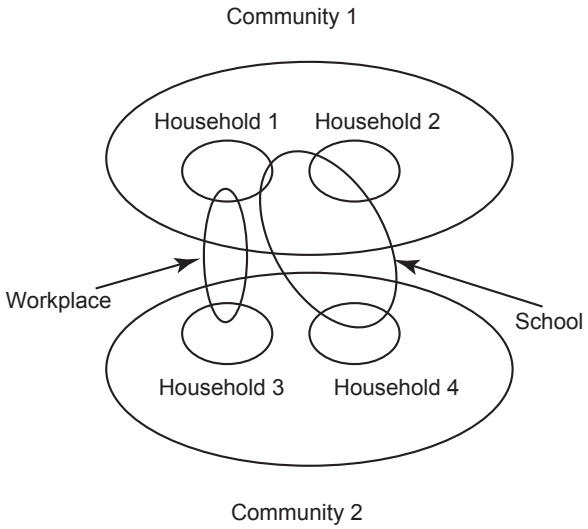


Figure 61: **An example of the contact network.**

the infectious period. Each symptomatic agent has a probability p_H of requiring hospitalization (H). Each hospitalized agent has a probability p_D of dying. Agents who survive infection eventually assume the recovered state and are immune to infection. The duration of each stage follows a certain random distribution as summarized in Table 17 below. The disease progression model is depicted in Figure 1 of [153].

Contact network

Each agent has an assigned household, a peer group (representing workplaces and schools), and a community. Each day agents move among these different locations and mix within other agents who are in the same location. Hence, a susceptible agent can get infected through contacts with his/her family members in the household, classmates/colleagues in the peergroup, or randomly meets someone when going to public places (communities) such as grocery stores, theaters, etc. Figure 61 diagrams an example of the contact network.

5.3.2 Simulation logic

We use discrete-event simulation technique to simulate the spread of the disease on the contact network. The simulation progresses with two types of events: (i) the next

infection event and (ii) disease progression event within an individual. The next infection event is generated by evaluating the instantaneous total "force of infection" [153] and drawing the corresponding event trigger time from an exponential distribution (with the mean equal to the total force of infection). The individual associated with the infection event is selected uniformly from a social group (e.g., a household or a peergroup), and this social group is randomly selected according to its contribution to the total force of infection. After infection, an individual transitions from one disease stage to another (such as become symptomatic from the pre-symptomatic stage), and the trigger time for a disease progression event is generated from the random distributions associated with each disease stage.

The instantaneous force of infection represents the rate at which susceptible persons become infected. The total force of infection is the sum of the force of infection experienced by each individual. The force of infection experienced by the ith individual during the day (λ_i^D) and during the night (λ_i^N) in the baseline model are as follows:

$$\begin{aligned} \lambda_i^D &= S_i \sum_{j=1}^{N} (\delta_{ij}^{PG} m_j \epsilon_j h_{PG} h_{X,j} \beta + \delta_{ij}^{C} \frac{m_j h_C h_{X,j} \beta}{N_i}) \\ \lambda_i^N &= S_i \sum_{j=1}^{N} (\delta_{ij}^{H} \frac{m_j h_{X,j} \beta}{n_i^{HA}} + \delta_{ij}^{C} \frac{m_j h_C h_{X,j} \beta}{N_i}) \end{aligned}$$

where S_i and m_i are the relative susceptibility and infectivity of the ith individual. N_i is the number of population in the ith individual's community and n_i^{HA} is the active household size of this individual where dead and hospitalized individuals are not counted. δ_{ij}^{Y} ($Y \in \{H,\ PG,\ C\}$) is the indicator function defined for location Y (households, peer groups or community)

$$\delta_{ij}^{Y} = \begin{cases} 1 & \text{if person i and j are in the same location } Y \\ 0 & \text{otherwise} \end{cases}$$

and ϵ_j is the indicator showing whether person j withdraws from workplace or school:

$$\epsilon_j = \begin{cases} 1 & \text{if person j mixes in the peergroup} \\ 0 & \text{if person j withdraws from the peergroup} \end{cases}$$

We assume that 100% symptomatic children and 50% symptomatic adults withdraw from their peergroups. Finally, $h_{X,j}$ is the relative hazard rate of the jth person if he/she is in the disease stage X, i.e.,

$$h_{X,j} = \begin{cases} h_{PS} & \text{if person j is in the presymptomatic stage} \\ h_{AS} & \text{if person j is in the asymptomatic stage} \\ 1 & \text{if person j is in the symptomatic stage} \\ 0 & \text{otherwise} \end{cases}$$

We can see that the total force of infection is determined by the number of infectious individuals (i.e., individuals in the presymptomatic, asymptomatic and symptomatic stages) and number of susceptible individuals at the current time, as well as the parameters β, h_{PS}, h_{AS}, h_{PG}, and h_C. The relative hazard rates (h_{PS}, h_{AS}, h_{PG}, and h_C) adjust the probability of infection between two individuals' contacts in different social groups and in different disease stages. Generally speaking, the probability of infection is the highest when contacts occur in households, medium when occur in peer groups, lowest when occur in communities. An susceptible individual is more infectious than an individual in the presymptomatic or asymptomatic stage.

5.3.3 Data and model calibration

The agent-based simulation model is age-structured. We divide the population into five age groups: 0-5, 6-11, 12-18, 19-64, $\geq$65 years. The first three groups represent children who are assumed to have higher susceptibility and infectivity compared to adults. Individuals in the fourth group are working adults. The last age group

represents the elderly. The demographic properties for each age group are populated by the 2000 U.S. Census Data.

For the contact network, we assign each agent to a household according to the household distributions obtained from the 2000 U.S. Census Data. The average peer-group (classroom) sizes are 14, 20, and 30 for the three childrens groups, respectively. The workplace size for working adults is a Poisson random variable with mean 20 (maximum 1000). We assume the elderly do not mix in peer groups. Table 17 below lists the distributions of the size of the households, peer groups and communities. It also lists the parameter values for the disease progression model.

Finally, to calibrate the disease spread through the contact network, we need to estimate the five parameters: the coefficient of transmission β and the relative hazard rates h_{PS}, h_{AS}, h_{PG}, and h_C. It is difficult to directly estimate these parameters from data. Thus, we use a similar nonlinear technique as in [153] to calculate the values of the five parameters from another five parameters that are easier to be estimated from data.

Specifically, let r_{XY} be the average number of people infected in Y by an individual who is at stage X where Y is the household (H), peer group (PG) or the community (C) and X is the presymptomatic (P), asymptomatic (A) or symptomatic (S) stage. We can write r_{XY} in terms of β, h_{PS}, h_{AS}, h_{PG}, and h_C as follows:

$$
\begin{aligned}
r_{PH} &= \sum_{n=1}^{7} p_n(n-1)(1-\phi_P(\frac{h_{PS}\beta}{2n})) \\
r_{AH} &= p_A \sum_{n=1}^{7} p_n(n-1)\phi_P(\frac{h_{PS}\beta}{2n})(1-\phi_A(\frac{h_{AS}\beta}{2n})) \\
r_{SH} &= (1-p_A) \sum_{n=1}^{7} p_n(n-1)\phi_P(\frac{h_{PS}\beta}{2n})(1-\phi_S(\frac{\beta}{2n})) \\
r_{P,PG} &= (q_1n_1+q_2n_2+q_3n_3+q_4n_4+q_5n_5)(1-\phi_P(\frac{h_{PS}h_{PG}\beta}{2})) \\
r_{A,PG} &= p_A(q_1n_1+q_2n_2+q_3n_3+q_4n_4+q_5n_5)\phi_P(\frac{h_{PS}h_{PG}\beta}{2})(1-\phi_A(\frac{h_{AS}h_{PG}\beta}{2})) \\
r_{S,PG} &= (1-p_A)[(q_1n_1+q_2n_2+q_3n_3)\phi_P(\frac{h_{PS}h_{PG}\beta}{2})(1-\phi_S(0)) + (q_4n_4+q_5n_5)\phi_P(\frac{h_{PS}h_{PG}\beta}{2})(1-\phi_S(\frac{h_{PG}\beta}{4}))] \\
r_{PC} &= N(1-\phi_P(\frac{h_{PS}h_C\beta}{2N})) \\
r_{AC} &= p_AN\phi_P(\frac{h_{PS}h_C\beta}{2N})(1-\phi_A(\frac{h_{AS}h_C\beta}{2N})) \\
r_{SC} &= (1-p_A)N\phi_P(\frac{h_{PS}h_C\beta}{2N})(1-\phi_S(\frac{h_C\beta}{2N}))
\end{aligned}
$$

Here, q_i is the proportion of population in age group i for $i = 1, \cdots, 5$, p_A is the probability that a presymptomatic individual does not develop symptoms, and n_i is the average size of peer groups for age group i. We assume the maximum household size is 7, and p_n is the probability that an individual lives in a household size of n. N is the total population size. $\phi_X(h) = E(e^{-hD_X})$ is the probability that an infection does not occur between two individuals during phase X ($X \in \{P, A, S\}$, D_X is the duration of stage X) for a constant hazard of infection h.

We can also represent r_{XY} in terms of the following disease parameters: the reproduction rate R_0 (average number of secondary cases generated by each infected individual); θ, the proportion of transmission that occurs at either presymptomatic or asymptomatic stage; ω, the proportion of infections generated by individuals who are never symptomatic; γ, the proportion of transmission that occurs outside the households; and δ, the proportion of transmission outside the home that occurs in the community.

$$
\begin{aligned}
R_0 &= r_{PH} + r_{AH} + r_{SH} + r_{P,PG} + r_{A,PG} + r_{S,PG} + r_{PC} + r_{AC} + r_{SC} \\
\theta &= \frac{r_{PH} + r_{AH} + r_{P,PG} + r_{A,PG} + r_{PC} + r_{AC}}{R_0} \\
\omega &= \frac{r_{AH} + p_A r_{PH} + r_{A,PG} + p_A r_{P,PG} + r_{AC} + p_A r_{PC}}{R_0} \\
\gamma &= \frac{r_{P,PG} + r_{A,PG} + r_{S,PG} + r_{PC} + r_{AC} + r_{SC}}{R_0} \\
\delta &= \frac{r_{PC} + r_{AC} + r_{SC}}{r_{P,PG} + r_{A,PG} + r_{S,PG} + r_{PC} + r_{AC} + r_{SC}}
\end{aligned}
$$

To derive the first equation between r_{XY} and R_0, we use the idea that the average number of secondary cases from a typical infectious individual generated in his/her social groups is equal to R_0. Similarly, we can derive the other four equations. Through the intermediate values of r_{XY}, for given values of R_0, θ, ω, γ, and δ, we can solve the above nonlinear equations and obtain the values for β, h_{PS}, h_{AS}, h_{PG}, and h_C. The values of R_0, θ, ω, γ, and δ used in the baseline simulation are given in Table 17 (also see Table 1 of [133]).

Model calibration

Calibration of the disease spread simulation model is based on previously published methods [46, 153]. For a given R_0, we target the corresponding attack rates from studies of previous pandemics, i.e., we fine-tune certain parameters until an appropriate *attack rate* (proportion of symptomatic cases out of the total population) is obtained. The baseline simulation model is calibrated to match the attack rate in the 1918 pandemic: when R_0=1.8, the clinical attack rate is 50% [153]. We have also done experiments to calibrate the model for other pandemics in the history. For example, we fine-tune the R_0 value to 1.53 to match the age-specific attack rates in the 1957 pandemic as shown in [31]. Table 18 shows the values of adjusted parameters to achieve the age-specific clinical attack rates in the 1957 pandemic. Table 19 reports the age-specific attack rates from [31] and from our simulation model, respectively. Note that $R_0 = 1.53$ is consistent with the estimates 1.5-1.7 for the 1957

Table 17: **Values of the key parameters in the baseline disease spread simulation model.**

Parameter	Explanation	Base Value	Source
p_A	Probability of infected individual being asymptomatic	0.4 for working adults, 0.25 for others	[49, 56, 101]
p_H	Probability of symptomatic individual requiring hospitalization	0.18 for age 0-5, 0.12 for 65+ and 0.06 for others	[153, 101]
p_D	Probability hospitalized individual not surviving	0.344 for age 0-5, 0.172 for others	[153, 23]
Duration of $E + I_P$	Duration of exposed and presymptomatic stage	Weibull with mean 1.48, std. 0.47, offset 0.5	[153, 47]
Duration of I_S	Duration of symptomatic stage	Exponential with mean 2.7313	[153]
Duration of I_A	Duration of asymptomatic stage	Exponential with mean 1.63878	[153]
Duration of I_H	Duration of hospitalized stage	Exponential with mean 14	[153, 47]
R_0	Reproductive rate	1.8 in baseline (sensitivity test: 1.3 and 1.5)	[153, 46, 101] [47, 65]
θ	Proportion of transmission that occurs at pre- or asymptomatic stage	0.3	[153]
ω	Proportion of infections generated by individuals who are asymptomatic	0.15	[153]
γ	Proportion of transmission that occurs outside the households	0.7	[46]
δ	Proportion of transmission outside the home that occurs in the community	0.5	[46]
m_i	Relative infectivity of the ith person	1.3158 for children, 0.8772 for adults	[23, 42]
S_i	Relative susceptibility of the ith person (0 if not susceptible)	1.1036 for children, 0.9597 for adults	[23, 42]
p_n	Probability that an individual lives in a household size of n	p_1=10.33%, p_2=23.55%, p_3=20.45%, p_4=23.00%, p_5=12.79%, p_6=5.91%, p_7=3.97%	[20]
Classroom size	Number of individuals in a classroom (children's peergroup)	Uniform(9,19) for ages 0-5, uniform(15,25) for ages 6-11, uniform(25,35) for ages 12-18	[153, 35]
Workplace size	Number of individuals in a workplace (adults' peergroup)	Poisson with mean 20 (maximum 1000)	[153, 101]
Community size	Number of individuals in a census tract (1615 tracts in total)	Maximum = 29341, minimum =218	[20]

Table 18: **Adjusted parameter values to achieve the attack rates in the 1957 Pandemic.**

Parameter	**Baseline value**	**Adjusted value**
p_A	0.4 for working adults, 0.25 for others	0.33 for age 0-18, 0.50 for age 19-64, 0.68 for age 65+
R_0	1.8	1.53
S_i	1.1036 for children, 0.9597 for adults, 0 if not susceptible	1.4236 for children, 0.8374 for adults, 0 if not susceptible

Table 19: **Age-specific attack rates from the 1957 pandemic [31] and from simulation.**

Age Group	Attack Rate of the 1957 Pandemic [31]	Attack Rate from simulation
0-5 years	32.17%	33.05%
6-11 years	35.02%	35.37%
12-18 years	38.44%	38.67%
19-64 years	22.24%	21.88%
≥65 years	10.00%	10.04%
Total	24.72%	24.53%

pandemic [31]. This also provides us a way to validate the simulation model (similar validation method has been used in other papers [46]).

CHAPTER VI

THE IMPACT OF SEASONALITY AND VIRAL MUTATION ON THE COURSE OF AN INFLUENZA PANDEMIC

6.1 Introduction

Influenza viral mutations and environmental factors can affect viral transmission dynamics, virulence, and population and individual host susceptibility and, in turn, alter the course of the epidemic and may lead to the appearance of multiple waves. Various infectious pathogens including influenza, measles, chickenpox, and pertussis have exhibited seasonality in their outbreak and epidemic patterns [2, 10, 12, 34, 51, 57, 100, 115]. Transmission of seasonal influenza tends to substantially increase from November to February in the Northern hemisphere and from May to August in the Southern hemisphere. Studies have postulated a number of possible causes of influenza seasonality, including changes in human mixing patterns, fluctuations in human immunity, and most recently environmental humidity [102, 131]. The influenza virus can also mutate, resulting in either incremental changes (antigenic drift) or more substantial changes (antigenic shift).

In this chapter, we focus on exploring how (i) seasonal changes in transmission dynamics and (ii) viral mutation may affect the course of an influenza pandemic. Based on the basic agent-base simulation model described in Section 5.3, we specify the details of modeling seasonality and viral mutation in Section 6.2. We then demonstrate simulation results in Section 6.3 to show the impact of seasonality and viral mutation with a focus on identifying scenarios with multiple waves of attack. Finally, we summarize our finds and implications for public health planning in Section 6.4.

6.2 Modeling seasonality and viral mutation

6.2.1 Modeling seasonality

To model seasonality, we assume the reproductive rate (R_0) value as a sinusoidal function of time t:

$$R_0(t) = R_0^* \cdot (1 + \epsilon \cos(2\pi t)). \tag{61}$$

Here, R_0^* is the baseline reproductive rate, and ϵ ($0 < \epsilon < 1$) characterizes the degree of seasonality. More temperate regions tend to have a higher ϵ and lower R_0^*; more tropical regions tend to have lower ϵ but higher R_0^*. Figure 62 demonstrates examples of the variation of $R_0(t)$ under different values of ϵ and the starting time when the initial seed case appears.

Equation (61) allows us to vary R_0 during the course of the pandemic. We directly set the R_0 value (rather than the transmission rate β) as a function of time t, while β and other disease parameters at time t are calculated based on the values of $R_0(t)$. For computational efficiency, we used linear interpolation to convert the continuous function in (61) into a step function with 12 discrete monthly R_0 values (see Table 1 of [134]). The discretization approximated the continuous function with sufficient accuracy.

6.2.2 Modeling viral mutation

To model viral mutation, we assume that a new viral strain is introduced at time t^* (i.e. t^* number of days after the appearance of the initial seed) in the simulation. After the introduction of the new strain, each day a fraction δ ($0 < \delta < 1$) of the recovered population lose their immunity (i.e., reverted to being susceptible). A reverted susceptible individual, if got infected by the new strain, goes over the same disease progression process as we described in the basic SEIR model. Figure 63 depicts the revised SEIR model which incorporates viral mutation. We assume that each individual can be infected at most two times (once by the original strain and

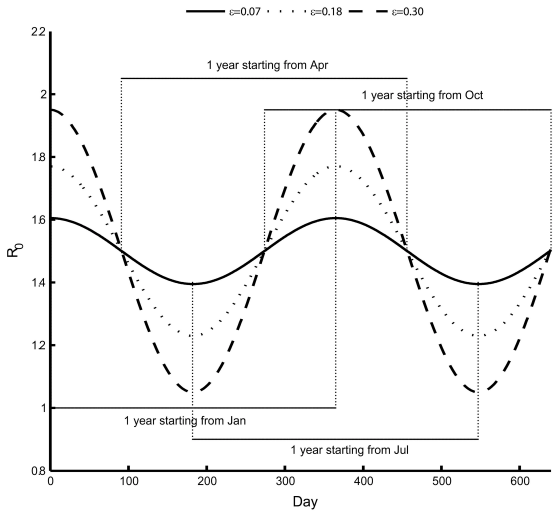


Figure 62: **Plot of R_0 value as function of time.** The figure shows the baseline reproductive rate $R_0^* = 1.5$, degree of seasonality $\epsilon = 0.07,\ 0.18$, and 0.30. The four intervals represent four different times (January, April, July, October) when the initial seed case appears. Within different intervals, the variation patterns of R_0 are different. For example, in the first interval, the value of R_0 first decreases then increases; in the third interval, the value of R_0 first increases then decreases).

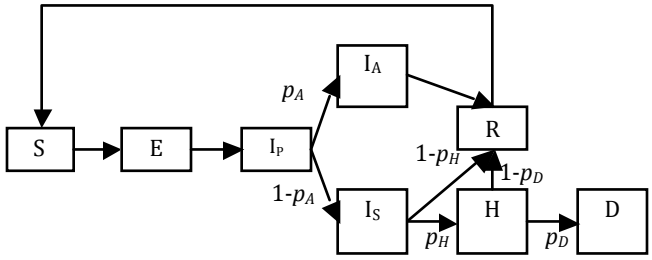


Figure 63: **Natural disease history with viral mutation.**

once by the new strain).

Our mutation model assumes that the new mutant strain had the same R_0 value as the original strain. Hence, our simulation runs track disease spread in the population without distinguishing the infections caused by the original strain from those caused by the new strain, which is consistent with the assumption employed by Ferguson et al. [49].

Note that the probability that a recovered individual loses his or her immunity and becomes susceptible on day t ($t > t^*$), $p(t)$, increases as t increases. For an

individual from the "candidate" population (recovered individuals who have not re-entered the susceptible stage), the day T when he/she re-enters the susceptible stage is a geometric random variable with probability distribution:

$$Pr(T = t) = \delta(1 - \delta)^{(t-t^*-1)}.$$

Thus, the probability

$$p(t) = 1 - (1 - \delta)^{(t-t^*)} \tag{62}$$

increases as t increases since $0 < \delta < 1$.

6.2.3 Combination of seasonality and viral mutation

In order to study the joint impact of seasonality and viral mutation on an influenza pandemic, we combine the seasonality and mutation models for some experiments. The combined model assumes that the R_0 value of the circulating strain (either the original or the mutant strain) will change over time in the same manner described by (61) and employes the resulting discrete monthly R_0 values for computational efficiency, as we described above.

6.2.4 Simulation runs and sensitivity analysis

To study the impact of seasonality, we test three values of R_0^* (1.5, 1.8, and 2.0) and three values of ϵ (0.07, 0.18, and 0.30). Under each combination of R_0^* and ϵ, we test four different starting times (January, April, July, October) of the initial seed case. We use t_0 to denote this starting time in the remaining of this chapter. Moreover, considering the time of emergence of the 2009 H1N1 pandemic influenza, we further test three more months in spring for the initial seed case (February, March, May) with each of the combinations of R_0^* and ϵ. In total, we have 63 different scenarios for seasonality, and we run 10 simulation replications for each scenario. The time horizon for each simulation replication is 365 days.

To study the impact of viral mutation, we test three R_0 values: 1.5, 1.8, and 2.1. These values are consistent with the estimates for R_0 in relevant studies [15-17]. We also explore the effects of using six different values for δ (0.5%, 1.5%, 5%, 8%, 10%, 20%) derived from previous studies [19, 25]. Using rough estimates, an antigenic drift occurs on average every 2 to 8 years, and the antigenic shift occurs approximately three times every 100 years [26]. We consider an influenza pandemic that starts from an antigenic shift (i.e. the population has low immunity to the virus), and test five values for t^* (30, 60, 90, 120, 180) to ensure a comprehensive experimental setting for the antigenic drift. If t^* is larger than 180, our simulation results show that it can be considered to be a new epidemic with a smaller susceptible population. The total number of mutation scenarios is 120 with 10 simulation replications for each scenario. The time horizon for each simulation replication is 365 days. Also see Tables 2 and 3 in [134] for the combinations of parameter values used in the seasonality and mutation scenarios.

Finally, we also consider different specific scenarios with modeling both seasonality and viral mutation to explore the combinations of factors that may lead to a third wave (e.g., simulated time horizon of 500 days and 10 replications with parameters $R_0^* = 1.5$, $\epsilon = 0.3$, $\delta = 0.5\%$, 1.5%, 5%, $t^* = 150$, 180, 250, 275, and the pandemic starting in April).

6.3 Results

We focus on the following performance measures when comparing different simulation scenarios: *daily prevalence* (the number of symptomatic and asymptomatic individuals over the total population), peak prevalence value and peak day, and whether a second or third wave emerges.

6.3.1 Seasonality scenarios

Figure 64 shows the daily prevalence curves for different combinations of (R_0^*, ϵ, t_0) from the seasonality scenarios. As Figure 64 demonstrates, a pandemic that begins in April can result in two waves (the first in spring and the second in the subsequent autumn/winter), whereas pandemics beginning in January, July, and October do not result in additional waves for the set of variables that we have tested. Our simulation also shows that a pandemic starting in March can result in two waves (the first in spring and the second in the subsequent autumn), and no additional waves appear if the pandemic begins in February or May.

As shown in Figure 64, with the degree of seasonality ϵ held constant, the peak prevalence day for the first wave of the pandemic occurs earlier for higher values of R_0^*. In situations where a second epidemic wave occurs, the second wave's peak prevalence day occurs later and the peak value is smaller for lower values of R_0^*.

Holding the baseline value R_0^* constant, a pandemic that starts in January or October has an earlier peak prevalence day and a higher peak prevalence for higher degrees of seasonality ϵ. A pandemic that starts in April or July has a later peak prevalence day and a smaller peak prevalence for higher degrees of seasonality.

6.3.2 Mutation scenarios

Figure 65 shows the daily prevalence curves for different combinations of (R_0, δ, t^*) values. The simulation results suggest that 10 days after the initial wave's peak prevalence may be a critical threshold. Viral mutations introduced before this time do not result in a second wave but can (i) increase the initial wave's peak prevalence and (ii) delay the peak prevalence day. However, a viral mutation introduced after this time could result in a second wave. Moreover, after this time threshold, the later the viral mutation emerges, then the later the peak prevalence of the second wave come, if it occurs.

Moreover, not all viral mutations introduced after this time threshold result in a second wave. A loss of immunity rate δ smaller than 1% daily seems to prevent the appearance of a second wave. Second waves appear only when δ is above 1%. Additionally, a higher δ results in an earlier peak prevalence day and a higher peak prevalence for the second wave, if one exists.

Figure 65 shows that when R_0=1.5 and δ=0.05, the peak prevalence of the second wave increases as t^* increases from 90 to 180. However, when δ increases to 0.10, the peak prevalence of the second wave decreases as t^* increases from 90 to 180. Table 20 summarizes the different impact on the second peak when increasing t^* under different values of R_0 and δ. Also see Figure 5 of [134] on the effects of varying δ, t^*, and R_0.

Table 20: **The peak prevalence value in the second wave varies as the mutant strain emerges later.** (+): the value of the second wave's peak prevalence is higher if the mutant strain emerges later. (–): the value of the second wave's peak prevalence is lower if the mutant strain emerges later.

	Loss of immunity rate (δ)		
Reproductive rate (R_0)	0.05	0.10	0.20
2.1	+	+	–
1.8	+	–	–
1.5	+	–	–

6.3.3 Seasonality and viral mutation scenarios

Certain combinations of seasonality and mutation scenarios (e.g., R_0^*=1.5, a degree of seasonality ϵ=0.3, and a loss of immunity rate δ=0.015) are able to reproduce three-wave epidemic curves similar to those seen in the 1918 pandemic. In the scenario demonstrated in Figure 66, the first case is introduced in April in the simulation and the mutant strain emerges 275 days after the initial seed infection. The simulated time horizon spans 500 days to include the third wave. Applying a constant mortality rate [15] reproduces the shapes of the observed 1918 pandemic mortality curves very closely (see figures in [1, 2]).

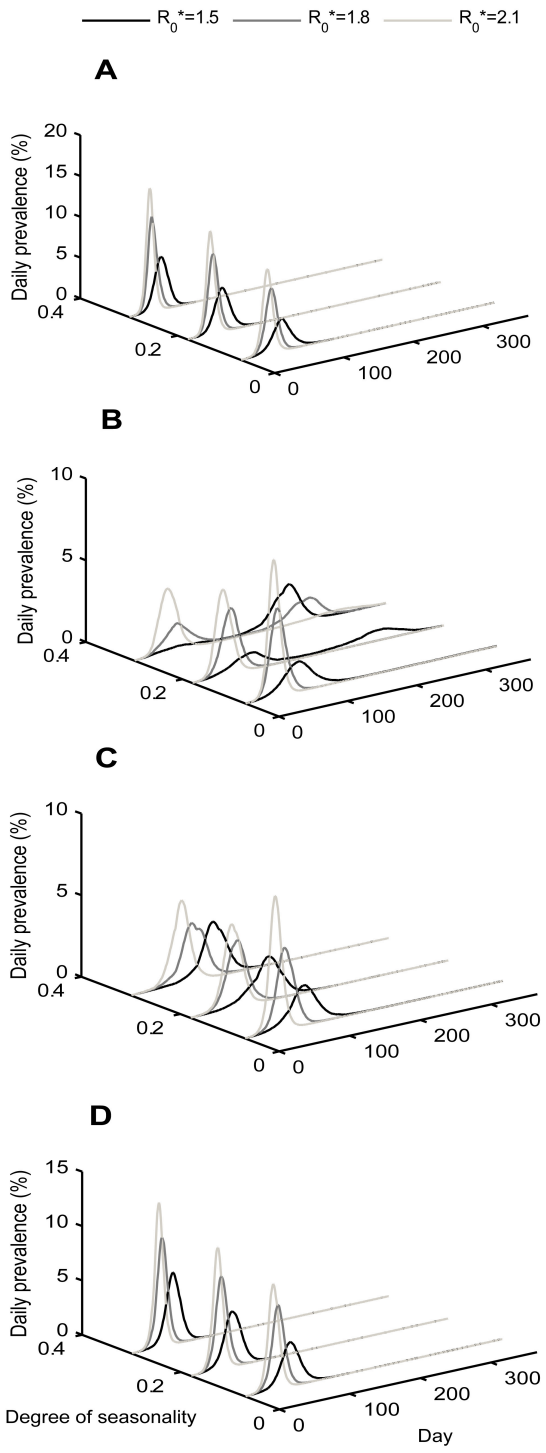


Figure 64: **Daily prevalence curves for seasonality scenarios.**. Nine curves in each panel correspond to nine pairs of (R_0^*, ϵ) values. The x axis represents the simulation day, the y axis represents the degree of seasonality ϵ, and the z axis (vertical) represents the daily prevalence of infectious cases (the number of symptomatic and asymptomatic persons over the total population). The epidemic starts in four different months: (a) January, (b) April, (c) July, (d) October.

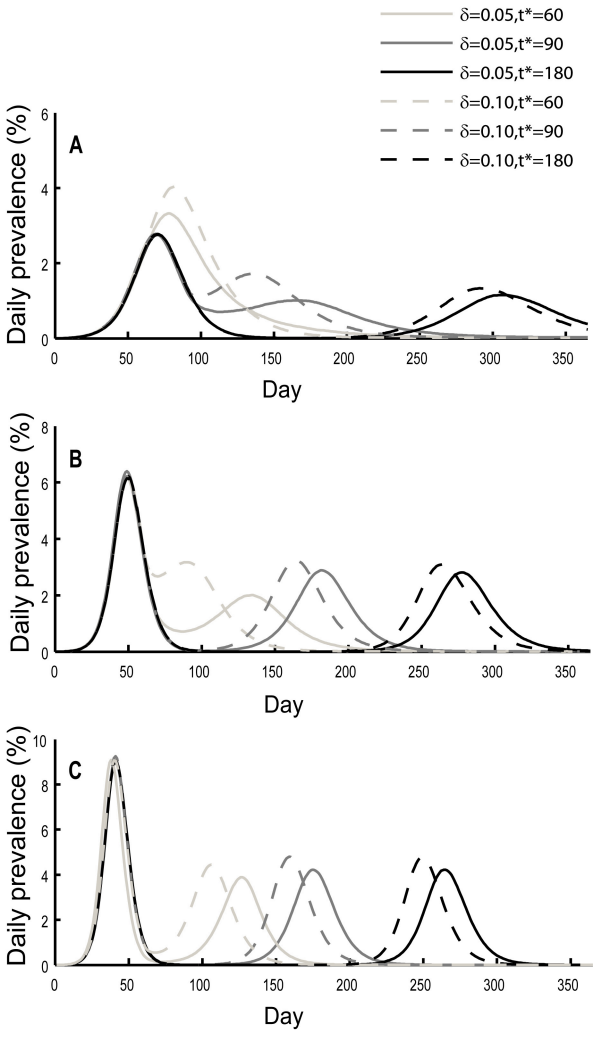


Figure 65: **Daily prevalence curves for mutation scenarios.** Each panel contains six curves corresponding to six pairs of (δ, t^*) values (δ=0.05 and 0.10, t^*=60, 90, and 180). The x axis represents the simulation day, and the y axis represents the daily prevalence of infectious cases (the number of symptomatic and asymptomatic persons over the total population). The subfigures show different reproductive rates (a) R_0=1.5, (b) R_0=1.8, (c) R_0=2.1.

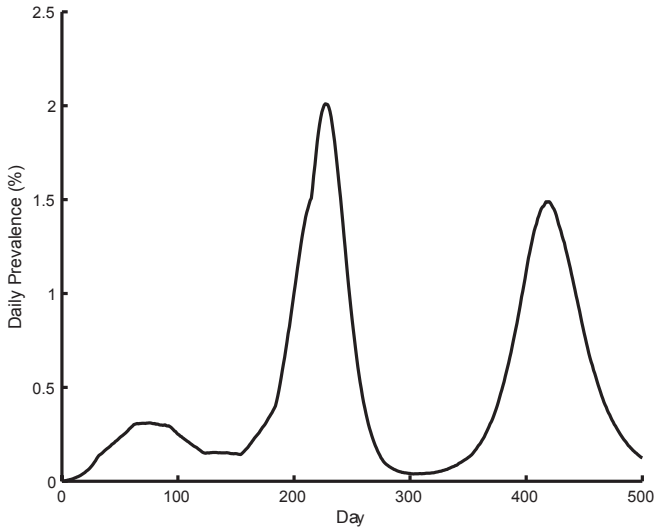


Figure 66: **Reproduced prevalence curve for the 1918 pandemics.** The x axis represents the simulation day, and the y axis represents the daily prevalence of infectious cases (the number of symptomatic and asymptomatic persons over the total population). Three prevalence peaks can occur with a baseline reproductive rate $R_0^* = 1.5$, degree of seasonality $\epsilon = 0.30$, loss of immunity rate $\delta = 0.015$. The pandemic starts in April and the mutant strain emerges at day 275.

As described earlier, without seasonality, a second wave would not appear if the loss of immunity rate is below 0.01. However, with seasonality added, we find scenarios where a third wave could occur if the loss of immunity rate is greater than 0.005, a viral mutation emerges after 180 days from the initial seed case, and the degree of seasonality is equal to 0.18. In these cases, the first two waves reflect seasonal effects, and the third wave results from the viral mutation.

6.4 Discussion

Our study shows that seasonality and viral mutation can substantially alter the course of an influenza pandemic. Both seasonality and viral mutation can lead to a second epidemic wave. The combination of seasonality and viral mutation can even lead to three epidemic waves similar to what observed in the 1918 influenza pandemic.

When modeling seasonality with a sinusoidal function, we demonstrate how different factors (e.g., degree of seasonality) can influence the daily prevalence curve. In particular, we find that the month that a pandemic appears could help determine

whether a second wave may occur. A pandemic that begins in April can result in two waves, whereas pandemics beginning in January, July, and October do not result in additional waves for the set of variables that we have tested. We give a brief explanation here. First, a pandemic starting in January (in the Northern hemisphere) with a high reproductive rate (R_0) may infect too many susceptibles (in turn, producing too many immune persons) to allow for a second wave to occur [71, 3, 78, 129]. Second, a pandemic that begins in April has only a short timeframe to infect susceptibles before summer when the R_0 value decreases. As a result, the first wave is relatively mild, leaving a large population of susceptibles remaining to be infected in autumn, when the R_0 value rises again. This situation provides a fertile ground for a second wave. Finally, a pandemic that starts in July may not have a large enough R_0 to generate an epidemic curve until autumn, while a pandemic starting in October rapidly affects a large number of persons (leaving relatively fewer susceptible persons) so that a second wave may not be possible.

Regarding viral mutation, our study shows that the time in which a viral mutation emerges may affect the peak prevalence, the timing of the peak, and whether a second wave occurs. Viral mutations that emerges more than 10 days after the peak prevalence day accompanied by a loss of immunity rate of more than 1%of the recovered population can lead to a second wave. Moreover, we find that when a viral mutation leads to a second wave, the characteristics of the second wave depend on the value of R_0, the emergence time of the mutant strain (t^*) and the loss of immunity rate (δ). The variations in the value of the peak prevalence in the second wave (Table 20) are related to the force of infection. Recall that the force of infection is the rate at which susceptible persons are infected by the virus, which is higher if there are more infectious persons. The prevalence of infections depends on the current value of the force of infection as well as the number of susceptible persons. Holding other parameters fixed, if the mutation begins earlier, then the value of the force of infection is higher

because the first wave has not completely disappeared and the number of infectious persons is higher; if the mutation begins later, the number of susceptible persons is higher because the recovered pool is larger. Thus, the peak prevalence of the second wave can be either higher or lower when the mutation emerges later. According to the simulation results, when $R_0 = 1.5$, the first factor dominates when the loss of immunity rate $\delta = 0.10$, while the second factor dominates when $\delta = 0.05$.

6.4.1 Public health implications

As the 2009 influenza pandemic (H1N1) has demonstrated, public health officials and other decision makers must plan and execute strategies not only before a pandemic emerges but also throughout its course. Within a limited time window, they also must determine what data needs to be collected to facilitate forecasting and planning. Early characterization of the ambient circumstances and the emerging viral characteristics may help predict the behavior of the pandemic and the corresponding intervention requirements. When a pandemic emerges in spring, for example, a noteworthy concern is whether the pandemic will re-emerge with greater or less severity in the following autumn. Our simulation suggests that decision-makers may want to watch for certain characteristics such as the month when the pandemic initially starts and the rate at which recovered patients are being re-infected to aid their forecasts. If a second wave is possible, then decision-makers can plan medical supplies, personnel staffing, and education of the public accordingly, and the time gained for planning may even allow for a vaccine to be developed.

Additionally, our simulation study confirms the importance of active surveillance and virus typing during the course of an pandemic or epidemic. A viral mutation that emerges during the downward slope of an initial wave may take public health officials by surprise. It is valuable to closely monitor patients who have already been infected and detect new strains as soon as they emerge. Without this additional information,

the gross epidemiological behavior of an initial wave can be very deceptive.

Finally, re-creating a three-wave epidemic curve may help shed additional light on the 1918 pandemic, which has been the source of much of the scientific and preparedness communities's understanding of influenza epidemics. Our simulation offers a profile of conditions that may have been present in 1918.

6.4.2 Conclusions and limitations

Our study demonstrates how different seasonal effects and the timing and degree of viral mutation can substantially alter the course of an influenza pandemic. Early characterization of the ambient circumstances and the emerging viral characteristics may help public health officials and other decision-makers predict the subsequent behavior of a pandemic or epidemic and the corresponding intervention requirements. Further, the advance notice of potential subsequent waves can help improve planning decisions. Future studies may look at the effectiveness of different public health interventions in many of our simulated scenarios.

We also want to mention that computer models and simulations by definition are simplifications of real life. They include a number of assumptions and cannot fully capture every possible factor or effect. Computer simulations can help delineate possible relationships and understand the importance of various questions and characteristics. Caution should be used when attempting to make definitive forecasts. The current results may not be generalizable to all locations and conditions.

CHAPTER VII

THE IMPACT OF MASS GATHERINGS AND HOLIDAY TRAVELING ON THE COURSE OF AN INFLUENZA PANDEMIC

7.1 Introduction

During the 2009 H1N1 influenza pandemic, concerns arose about the potential negative effects of mass public gatherings and travel on the course of the pandemic. For example, when the H1N1 pandemic appeared to be subsiding in December 2009, public health officials contemplated whether changes in social mixing patterns due to a combination of Holiday travel with school and workplace closures could lead to an additional wave of outbreaks similar to those seen in 1918 and 1957 [67]. The World Health Organization (WHO), the U.S. Centers for Disease Control and Prevention (CDC), and many other public health organizations published recommendations [26, 27, 151, 152, 140, 122, 121] suggesting the public defer non-essential travel to infected areas and emphasizing taking appropriate precautions (e.g., hand hygiene) during traveling, attending and/or hosting mass gathering events. However, the decisions regarding cancelling or postponing mass gatherings are left to local authorities; travel restrictions are generally not recommended [152, 34, 36, 46, 64], but some countries have introduced new travel regulations relating to the 2009 H1N1 outbreaks.

Better understanding the potential effects of changes in social mixing patterns could help public officials determine if and when to cancel large public gatherings or enforce regional travel restrictions, advisories, or surveillance during an influenza pandemic. In this chapter, we explore how various changes in social mixing and contact patterns, representing mass gatherings and Holiday traveling, may affect the

course of an influenza pandemic. We use simulation to study the impact of social mixing patterns and explore scenarios in which additional epidemic waves can appear due to social mixing changes. In Section 7.2, we specify the details of modeling social mixing changes in the agent-based simulation model. Then in Section 7.3, we show simulation results under various mass gathering (non-Holiday) scenarios and Holiday traveling scenarios. We demonstrate the impact of changing social mixing patterns on the aggregate epidemic prevalence as well as its regional impact and impact on traveler's family members. Finally in Section 7.4, we summarize our finds and implications for public health planning.

7.2 Models

Our study utilizes the basic agent-based simulation model developed in Section 5.3. Recall that the simulation model consists of a population of computer agents, with each agent representing an individual with socio-demographic characteristics and behaviors. To incorporate the changes in social mixing patterns, we have significantly modified the model, particularly for disease spread on the contact network (the individual disease progression process remains the same as in the baseline simulation model). Correspondingly, model calibration needs to be adjusted to incorporate these changes. In the following two subsection, we introduce the major changes we have made to model mass public gatherings and Holiday travel.

7.2.1 Modeling mass social mixing: public gatherings and Holiday travel

To explore the effects of mass social mixing changes (e.g., large public gatherings and Holiday traveling), we divide the year into a *regular period* and a *traveling* period. The traveling period starts at day t^* after the introduction of the initial infected case and lasts for l days; the remaining days before and after this traveling period comprise the regular period. During the regular period, agents move back and forth between households and workplaces or schools. They mix in the workplaces or schools during

the day and in their households during the night. Agents also mix in the communities during the day and night by visiting common areas such as grocery stores, churches, theaters, etc.

At the beginning of the traveling period, we select $p\%$ of the total agent population (in two different ways, see below) to change mixing patterns. They mix in a large group (i.e., *traveling/gathering group*) to model temporal mass gathering locations or events, e.g., airports, shopping malls, or the annual Georgia Tech versus University of Georgia football game. We consider the following two scenarios:

- Non-Holiday: $p\%$ of the total agent population is sampled randomly. Agents selected to mix in the traveling/gathering group only have contact with each other in the group, and no longer interact with their family members or classmates/colleagues, or mix in their usual communities. The $(1-p)\%$ agents not in the traveling group retain their usual mixing routines, e.g., mix in their workplaces or schools during the day and in their households during the night. This scenario represents mass public gatherings, e.g., a football game, road race, concert, convention, or demonstration, where one does not necessarily attend the events or travel with his/her family. The traveling/gathering group can include event attendees, visitors, and local residents.

- Holiday: A subset of households is randomly sampled so that $p\%$ of the total agent population is chosen to mix in the traveling/gathering group. The agents travel with their family members (i.e., mix in the household day and night), and also interact with other agents in the traveling group during the day. However, they no longer mix in their schools, workplaces or usual communities. The agents not selected for travel reduce their peer group mixing activities. Schools and a percentage of workplaces (baseline 50%) are closed during the traveling period (l days) so that agents no longer mix in these locations. This setting

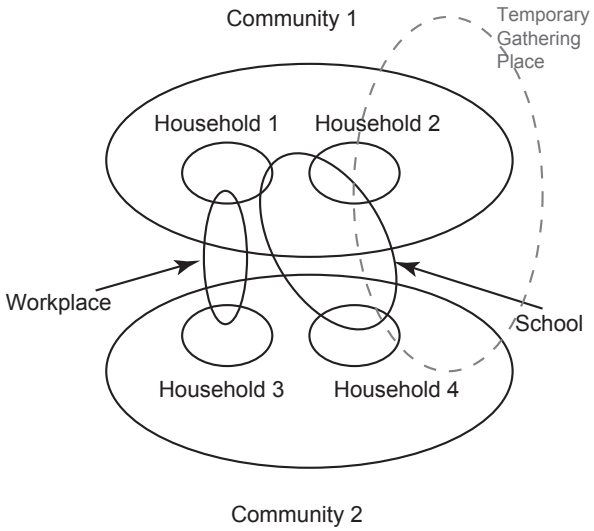


Figure 67: **An example of the contact network during the traveling period.** The figure shows an example of the contact network, i.e., how persons interact with each other in households, workplaces, schools, communities, and/or temporary mass gathering locations.

represents travel or mass gatherings during a holiday, e.g., Thanksgiving or New Year's Eve.

When the traveling period ends, all the agents return to their regular mixing routines. Figure 67 diagrams an example of the new contact network during the traveling period.

7.2.2 Force of infection and model calibration

From the description in Section 7.2.1, we can see that there are three levels of mixing for each individual during the regular period (i.e., household, peer group, and community), and four potential levels of mixing during the traveling period (i.e., household, peer group, community, and the traveling/gathering group). A susceptible individual can become infected not only through contacts with his/her family members, classmates/colleagues, or random contacts with someone when going to community places, but also through contacts with other travelers or attendees when traveling or attending mass gathering events. Therefore, we need to adjust the calculation of force of infection to reflect the new level of mixing during the traveling period, and

recalibrate the disease transmission parameters (e.g., the transmission rate β). For mixing in the regular period, the calculation of force of infection remains to be the same as in the baseline simulation model.

Traveling Period: Non-Holiday Setting

During the traveling period in the non-Holiday scenario, individuals who are traveling or gathering have different mixing patterns from those are not:

1. Individual not traveling/attending mass gatherings: retains his/her usual mixing pattern in the regular period, i.e. mix in household (night), peergroup (day) and community (day/night);

2. Individual traveling/attending mass gatherings: only mix in the traveling/mass gathering group (day/night).

Thus, the force of infection experienced by the ith person during the day (λ_i^D) and during the night (λ_i^N) in the traveling period are calculated as follows:

$$\begin{aligned}
\lambda_i^D &= S_i \cdot \left[(1-\sigma_i) \cdot \sum_{j=1}^{N} (\delta_{ij}^{PG} m_j \epsilon_j h_{PG} h_{X,j} \beta + \delta_{ij}^{C} \frac{(1-\sigma_j) m_j h_C h_{X,j} \beta}{N_i^A}) + \sigma_i \cdot \sum_{j=1}^{N} \delta_{ij}^{T} \frac{\sigma_j m_j \tilde{h}_{X,j} \tilde{\beta}}{N_T} \right] \\
\lambda_i^N &= S_i \cdot \left[(1-\sigma_i) \cdot \sum_{j=1}^{N} (\delta_{ij}^{H} \cdot \frac{(1-\sigma_j) m_j h_{X,j} \beta}{n_i^{HA}} + \delta_{ij}^{C} \cdot \frac{(1-\sigma_j) m_j h_C h_{X,j} \beta}{N_i^A}) + \sigma_i \cdot \sum_{j=1}^{N} \delta_{ij}^{T} \frac{\sigma_j m_j \tilde{h}_{X,j} \tilde{\beta}}{N_T} \right]
\end{aligned}$$

Here S_i and m_i are the relative susceptibility and infectivity of the ith person (see Table 17). N_i^A is the number of population in the ith person's community except those persons who are traveling or attending mass gatherings, and n_i^{HA} is the active household size of this person where persons who are dead, hospitalized, or traveling are not counted. N_T is the number of population in the traveling group, i.e., the number of persons who are traveling or attending mass gatherings. Moreover, δ_{ij}^{Y} ($Y \in \{H,\ PG,\ C,\ T\}$) is the indicator function defined for location Y (households,

peer groups, community, or the traveling group)

$$\delta_{ij}^{Y} = \begin{cases} 1 & \text{if person i and j are in the same location } Y, \\ 0 & \text{otherwise;} \end{cases}$$

and ϵ_j is the indicator showing whether person j withdraws from workplace or school:

$$\epsilon_j = \begin{cases} 1 & \text{if person j mixes in the peergroup,} \\ 0 & \text{if person j withdraws from the peergroup.} \end{cases}$$

As in the baseline model, we still assume that 100% symptomatic children and 50% symptomatic adults withdraw from their peergroups. Besides, individuals who are traveling or attending mass gathering events also withdraw from their peergroups. The indicator variable σ_i is to show whether person i is traveling or attending mass gatherings:

$$\sigma_i = \begin{cases} 1 & \text{if person i is traveling or attending mass gathering events,} \\ 0 & \text{otherwise.} \end{cases}$$

Finally, $h_{X,j}$ is the relative hazard rate of the jth person if he/she is in the disease stage X and not traveling, and $\tilde{h}_{X,j}$ is the relative hazard rate if this person is traveling:

$$\tilde{h}_{X,j} = \begin{cases} \tilde{h}_{PS} & \text{if person j is in the presymptomatic stage,} \\ \tilde{h}_{AS} & \text{if person j is in the asymptomatic stage,} \\ 1 & \text{if person j is in the symptomatic stage,} \\ 0 & \text{otherwise.} \end{cases}$$

We use β and $\tilde{\beta}$ to differentiate the transmission rate for individuals who are not traveling or gathering and for those who are traveling/gathering.

From the above equations for the non-Holiday setting, we can see that during the day, the new infection event (determined by λ_i^D) can occur in a peergroup, in a traveling group, or in a community; while during the night, the new infection event (determined by λ_i^N) can occur in a household, in a traveling group, or in a community.

Traveling Period: Holiday Setting

In the Holiday setting, we have three different mixing patterns during the traveling period, i.e., one pattern for individuals who travel, and two patterns for those who do not travel:

1. Working adults who still go to work: keep his/her regular mixing patterns, i.e., mix in household (night), peergroup (day) and community (day/night);

2. Children, elderly, and working adults who do not travel but stay at home: mix in household (day/night) and community (day/night);

3. Individuals traveling with their family members: mix in traveling/gathering group (day) and household (day/night).

Thus, the force of infection experienced by the ith person during the day (λ_i^D) and during the night (λ_i^N) in the traveling period are calculated as follows:

$$\lambda_i^D = \begin{cases} S_i \cdot \sum_{j=1}^{N} (\delta_{ij}^{PG} m_j \epsilon_i \epsilon_j h_{PG} h_{X,j} \beta + \delta_{ij}^{C} \frac{(1-\sigma_j) m_j h_C h_{X,j} \beta}{N_i^A}) & \text{if retaining mixing patterns } (\sigma_i = 0,\ \epsilon_i = 1) \\ S_i \cdot \sum_{j=1}^{N} (\delta_{ij}^{H} \cdot \frac{(1-\epsilon_j) m_j \tilde{h}_{X,j} \tilde{\beta}}{n_i^{HA}} + \delta_{ij}^{C} \frac{(1-\sigma_j) m_j \tilde{h}_C \tilde{h}_{X,j} \tilde{\beta}}{N_i^A}) & \text{if staying at home all day } (\sigma_i = 0,\ \epsilon_i = 0) \\ S_i \cdot \sum_{j=1}^{N} (\delta_{ij}^{H} \cdot \frac{m_j \bar{h}_{X,j} \bar{\beta}}{n_i^{HA}} + \delta_{ij}^{T} \frac{\sigma_j m_j \bar{h}_T \bar{h}_{X,j} \bar{\beta}}{N_T}) & \text{if traveling } (\sigma_i = 1,\ \epsilon_i = 0) \end{cases}$$

$$\lambda_i^N = \begin{cases} S_i \cdot \sum_{j=1}^{N} \delta_{ij}^{H} \cdot \frac{m_j h_{X,j} \beta}{n_i^{HA}} & \text{if not traveling } (\sigma_i = 0) \\ S_i \cdot \sum_{j=1}^{N} \delta_{ij}^{H} \cdot \frac{m_j \bar{h}_{X,j} \bar{\beta}}{n_i^{HA}} & \text{if traveling } (\sigma_i = 1) \end{cases}$$

Here S_i and m_i are the relative susceptibility and infectivity of the ith person. N_i^A is the number of population in the ith person's community except those persons who are traveling or attending mass gatherings, and n_i^{HA} is the active household size of this person where dead and hospitalized persons are not counted. N_T is the number of population in the traveling group, i.e., the number of persons who are traveling or attending mass gatherings. Moreover, δ_{ij}^{Y} ($Y \in \{H,\ PG,\ C,\ T\}$) is the indicator

function defined for location Y (households, peer groups, community, or the traveling group)

$$\delta_{ij}^{Y} = \begin{cases} 1 & \text{if person i and j are in the same location } Y, \\ 0 & \text{otherwise;} \end{cases}$$

and ϵ_j is the indicator showing whether person j withdraws from workplace or school

$$\epsilon_j = \begin{cases} 1 & \text{if person j mixes in the peergroup,} \\ 0 & \text{if person j withdraw from the peergroup.} \end{cases}$$

We assume that 100% children and 50% adults withdraw from their peergroups during the Holiday season. Besides, individuals who are traveling also withdraw from their peergroups. The indicator variable σ_i is to show whether person i is traveling or attending mass gatherings:

$$\sigma_i = \begin{cases} 1 & \text{if person i is traveling during Holiday,} \\ 0 & \text{otherwise.} \end{cases}$$

Finally, $h_{X,j}$ is the relative hazard rate of the jth person if he/she is in the disease stage X and retains the regular mixing patterns; $\tilde{h}_{X,j}$ is the relative hazard rate if this person is not traveling and stays at home during the day; and $\bar{h}_{X,j}$ is the relative hazard rate for traveling individuals:

$$\bar{h}_{X,j} = \begin{cases} \bar{h}_{PS} & \text{if person j is in the presymptomatic stage,} \\ \bar{h}_{AS} & \text{if person j is in the asymptomatic stage,} \\ 1 & \text{if person j is in the symptomatic stage,} \\ 0 & \text{otherwise.} \end{cases}$$

From the previous equations for the Holiday setting, we can see that during the night, the new infection event (determined by λ_i^N) only occur in a household and in a community; while during the day, the new infection event (determined by λ_i^D) can occur in a household, in a peergroup, in a traveling group, or in a community.

Since the calculation of force of infection is adjusted to reflect the changes in social mixing patters, we also need to recalibrate disease transmission parameters for individuals who have changed their mixing patterns during the traveling period (e.g., the new transmission rate $\tilde{\beta}$ and $\bar{\beta}$). We use the nonlinear techniques introduced in Section 5.3.3 to estimate these parameters from other input parameters such as the reproductive rate R_0. The details of calibrating the disease parameters are in Appendix D.1.

7.2.3 Simulation runs and sensitivity analyses

The individual disease progression model adopts the same distributional settings as in the baseline model (see Table 17). Parameters for populating the contact network, unless otherwise specified, also remain the same as we used in the baseline model. The demographic features in the model are populated from the 2000 census data for the state of Georgia and are the same as listed in Table 17.

To study the impact of traveling and mass gathering events on the course of an influenza pandemic, we test different scenarios with three *initial* reproductive rates (initial R_0): 1.3, 1.5, and 1.8, which correspond to R_0 estimates from past pandemics in 1918, 1957, 1968, and 2009 [153, 46, 65, 101, 47]. Here, the initial R_0 refers to the R_0 value before any social mixing changes occur and is a model input to calculate the disease parameters in Section 7.2.2. It is different from the *resulting* R_0 value, which is the R_0 after the mass social mixing changes are instituted and is a performance measure we obtain from simulation experiments. Separate scenarios also explore the effects of using different traveling/gathering starting dates t^* (Day 30, 60, 90, 120, 180), traveling/gathering durations l (0.5 day, 1 day, 2 days, and 3 days for the non-Holiday scenario, 3 and 5 days for the Holiday scenario), and the proportion of the population that travels/gathers during the traveling period, p (1%, 5%, 10% and 25% for the non-Holiday scenario [cite: 32-37], 25% and 50% for the Holiday scenario [cite:

43,44]).

To study the regional impact of traveling and mass gathering events, we explore different proportions of population who participate in traveling/gathering in different locations (i.e., different p values). For example, similar to the Annual Cherry Blossom Festival in Macon, Georgia, we assume in one experimental scenario that 50% of the population travels/gathers in the Bibb County and its nearest 5 counties [20], and 9.5% of the population travels/gathers in other counties (so that for the entire population p = 10%) under the non-Holiday setting.

The total number of experimental scenarios is 125 for the non-Holiday scenario and 60 for the Holiday scenario with 10 replications for each experiment unless indicated otherwise. The time horizon for each simulation replication is 365 days.

7.3 Results

For the non-Holiday scenario, we focus on the characteristics of peak prevalence and the total attack rate since only one epidemic peak appears; for the Holiday scenarios, we focus on whether two epidemic peaks are present (i.e., the influenza activity declines first and increases later). In the non-Holiday setting, we also examine the impact of transmissions to the traveler and their family and within regions where gathering occurs. Tables 21, 22, and 23 report the initial baseline R_0 (before social mixing changes are introduced) values, the peak prevalence, the total attack rate, and the resulting R_0 values after the mass social mixing changes are instituted for experiments under the non-Holiday setting.

7.3.1 The timing of mass travel/public gatherings t^*

As the results from simulating non-Holiday scenarios demonstrate, when the initial $R_0 = 1.5$, mass traveling or public gatherings that commence more than 20 days (e.g., $t^* = 90$, 120, or 180) after the epidemic peak (Day 70 in the baseline scenario with $R_0 = 1.5$) have little impact on the peak prevalence or the total attack rate. Mass

Table 21: **Results from Different Mass Gathering Scenarios (Initial R_0 = 1.5).** The table shows the total attack rate (i.e., proportion of population that has ever been infected), the peak prevalence day and value in the non-Holiday scenarios, with several combinations of values for l (duration of the traveling/mass traveling period) and p (the proportion of the population traveling/gathering) when the initial R_0 equals to 1.5. In the baseline scenario, no traveling/gathering occurs. The resulting R_0 values (after adding the traveling/mass gathering period) are obtained from the baseline scenarios to match the peak prevalence and the total attack rate showed in this table. The standard deviation is 0.04-0.09% for the peak prevalence and is 0.17-0.30% for the total attack rate.

% of traveling (p)	Traveling Period		Resulting	Peak	Peak	Total
	Start	Duration	R_0	Prevalence	Day	Attack Rate
p=1%	Day 30	0.5	1.5	2.73%	70	51.00%
	Day 30	1	1.5	2.76%	70	51.00%
	Day 30	2	1.5	2.78%	71	51.00%
	Day 30	3	1.5	2.79%	70	51.00%
	Day 60	0.5	1.5	2.74%	70	51.00%
	Day 60	1	1.5	2.76%	70	51.00%
	Day 60	2	1.5	2.74%	70	51.00%
	Day 60	3	1.5	2.75%	71	51.00%
p=5%	Day 30	0.5	1.5	2.74%	69	51.00%
	Day 30	1	1.5	2.77%	70	51.00%
	Day 30	2	1.5	2.77%	70	51.00%
	Day 30	3	1.5	2.80%	70	51.00%
	Day 60	0.5	1.5	2.74%	69	51.00%
	Day 60	1	1.5	2.81%	70	51.20%
	Day 60	2	1.51	2.83%	70	51.20%
	Day 60	3	1.5	2.78%	70	51.10%
p=10%	Day 30	0.5	1.5	2.74%	69	51.00%
	Day 30	1	1.5	2.78%	69	51.00%
	Day 30	2	1.5	2.80%	69	51.00%
	Day 30	3	1.5	2.82%	68	51.10%
	Day60	0.5	1.5	2.80%	69	51.00%
	Day 60	1	1.51	2.85%	70	51.30%
	Day 60	2	1.51	2.89%	69	51.40%
	Day 60	3	1.5	2.80%	70	51.10%
p=25%	Day 30	0.5	1.5	2.79%	69	51.00%
	Day 30	1	1.5	2.80%	68	51.10%
	Day 30	2	1.5	2.80%	68	51.10%
	Day 30	3	1.5	2.83%	70	51.00%
	Day 60	0.5	1.51	2.90%	69	51.40%
	Day 60	1	1.52	3.04%	69	51.70%
	Day 60	2	1.53	3.12%	69	52.00%
	Day 60	3	1.51	2.96%	71	51.40%
Baseline			1.5	2.73%	70	51.00%

Table 22: **Results from Different Mass Gathering Scenarios (Initial R_0 = 1.3).** The table shows the total attack rate (i.e., proportion of population that has ever been infected), the peak prevalence day and value in the non-Holiday scenarios, with several combinations of values for l (duration of the traveling/mass traveling period) and p (the proportion of the population traveling/gathering) when the initial R_0 equals to 1.3. In the baseline scenario, no traveling/gathering occurs. The standard deviation is 0.02-0.05% for the peak prevalence and is 0.22-0.41% for the total attack rate.

% of traveling (p)	Traveling Period		Resulting R_0	Peak Prevalence	Peak Day	Total Attack Rate
	Start	Duration				
p=1%	Day 60	0.5	1.3	0.96%	98	32.50%
	Day 60	1	1.3	0.97%	99	32.50%
	Day 60	2	1.3	0.96%	98	32.80%
	Day 60	3	1.3	0.98%	98	32.90%
	Day 90	0.5	1.3	0.96%	97	32.50%
	Day 90	1	1.3	0.97%	98	32.60%
	Day 90	2	1.3	0.98%	97	32.80%
	Day 90	3	1.3	0.98%	97	32.80%
p=5%	Day 60	0.5	1.3	0.97%	96	32.60%
	Day 60	1	1.3	0.98%	97	32.60%
	Day 60	2	1.3	0.98%	97	32.80%
	Day 60	3	1.3	1.00%	96	32.90%
	Day 90	0.5	1.3	0.98%	97	32.80%
	Day 90	1	1.3	0.98%	96	32.70%
	Day 90	2	1.31	1.00%	99	33.10%
	Day 90	3	1.3	0.97%	101	32.70%
p=10%	Day 60	0.5	1.3	0.98%	96	32.80%
	Day 60	1	1.3	0.98%	96	32.70%
	Day 60	2	1.3	0.99%	95	32.90%
	Day 60	3	1.3	1.01%	97	32.90%
	Day 90	0.5	1.3	0.99%	98	32.80%
	Day 90	1	1.31	1.00%	97	33.10%
	Day 90	2	1.31	1.02%	99	33.10%
	Day 90	3	1.3	0.98%	99	32.80%
p=25%	Day 60	0.5	1.3	0.98%	97	32.80%
	Day 60	1	1.3	1.00%	97	32.70%
	Day 60	2	1.31	1.05%	94	33.20%
	Day 60	3	1.3	1.03%	99	32.70%
	Day 90	0.5	1.31	1.04%	98	33.10%
	Day 90	1	1.31	1.07%	99	33.30%
	Day 90	2	1.32	1.11%	99	33.70%
	Day 90	3	1.31	1.02%	99	33.00%
Baseline			1.3	0.96%	94	32.40%

Table 23: **Results from Different Mass Gathering Scenarios (Initial R_0 = 1.8).** The table shows the total attack rate (i.e., proportion of population that has ever been infected), the peak prevalence day and value in the non-Holiday scenarios, with several combinations of values for l (duration of the traveling/mass traveling period) and p (the proportion of the population traveling/gathering) when the initial R_0 equals to 1.8. In the baseline scenario, no traveling/gathering occurs. The standard deviation is 0.08-0.15% for the peak prevalence and is 0.07-0.18% for the total attack rate.

% of traveling (p)	Traveling Period		Resulting R_0	Peak Prevalence	Peak Day	Total Attack Rate
	Start	Duration				
p=1%	Day 30	0.5	1.8	5.99%	50	68.40%
	Day 30	1	1.8	5.99%	51	68.40%
	Day 30	2	1.8	6.00%	50	68.40%
	Day 30	3	1.8	6.00%	50	68.40%
	Day 45	0.5	1.8	5.99%	51	68.40%
	Day 45	1	1.8	6.00%	51	68.40%
	Day 45	2	1.8	5.99%	51	68.40%
	Day 45	3	1.8	5.96%	51	68.40%
p=5%	Day 30	0.5	1.8	6.01%	50	68.40%
	Day 30	1	1.8	6.01%	50	68.40%
	Day 30	2	1.8	6.03%	50	68.40%
	Day 30	3	1.8	6.10%	51	68.40%
	Day 45	0.5	1.8	6.04%	51	68.40%
	Day 45	1	1.8	6.05%	50	68.60%
	Day 45	2	1.81	6.09%	51	68.70%
	Day 45	3	1.8	5.94%	51	68.40%
p=10%	Day 30	0.5	1.8	6.03%	51	68.40%
	Day 30	1	1.8	6.08%	50	68.40%
	Day 30	2	1.8	6.12%	50	68.40%
	Day 30	3	1.8	6.17%	51	68.40%
	Day 45	0.5	1.8	6.05%	51	68.50%
	Day 45	1	1.81	6.20%	50	68.60%
	Day 45	2	1.8	6.07%	51	68.50%
	Day 45	3	1.8	5.99%	51	68.40%
p=25%	Day 30	0.5	1.8	6.08%	50	68.40%
	Day 30	1	1.81	6.16%	50	68.50%
	Day 30	2	1.82	6.31%	50	68.60%
	Day 30	3	1.82	6.40%	50	68.50%
	Day 45	0.5	1.82	6.20%	51	68.80%
	Day 45	1	1.83	6.49%	50	69.30%
	Day 45	2	1.83	6.58%	51	69.50%
	Day 45	3	1.82	6.21%	53	68.60%
Baseline			1.8	5.99%	50	68.40%

traveling or public gatherings that commence well prior (i.e., more than 40 days) to the epidemic peak (e.g., $t^* = 30$) have a minor but not significant impact. For example, having 25% of the population traveling increases the peak prevalence from 2.73% (baseline) to 2.80% (around 2% relative increase in the peak percentage) but does not affect the overall attack rate much.

However, mass traveling that begins shortly before the peak prevalence day (e.g., $t^* = 60$, 10 days before the peak in the baseline case) can significantly increase the peak prevalence, e.g., 25% of the population traveling for 1 day increases the peak prevalence from 2.73% (baseline) to 3.04% (around a 11% relative increase) and increases the overall attack rate from 51.0% (baseline) to 51.7%. This translates to an additional 63,502 individuals being infected in Georgia [20]. Tables 21 to 23 show how different starting time of the traveling/gathering period affect the epidemic under non-Holiday conditions for all the initial R_0 values we have tested.

The results of simulating the Holiday scenarios show similar observations on the impact of the starting time t^*. When the initial $R_0 = 1.5$, mass traveling/gatherings that occur more than 20 days after the epidemic peak or more than 40 days before the peak do not lead to a second epidemic peak; otherwise, two explicit epidemic peaks can appear under certain scenarios as we demonstrate in the next section.

7.3.2 Impact of Holiday traveling on multiple peaks

Figure 68 shows the resulting epidemic curves (i.e., the daily prevalence of infected individuals) for the entire state of Georgia under the Holiday scenario where 25% of the population mixes in the traveling group during a 5-day traveling period. Figure 68(a) shows the scenario with the initial $R_0 = 1.5$ when the traveling period starts on Day 60; Figure 68(b) shows the scenario with the initial $R_0 = 1.3$ when the traveling period starts on Day 90. Figure 68 shows that the Holiday scenario can generate two prevalence peaks, while this is not seen in any of the non-Holiday

scenarios we tested. We have simulated a variety of scenarios with different initial R_0 values, Holiday traveling durations and proportions of the population on travel to explore scenarios that can generate two distinct epidemic peaks. Generally speaking, two peaks appear when Holiday traveling occurs within 5-20 days (depending on the initial R_0 values) before the prevalence peak day in the baseline (no traveling) scenario. The prevalence, the timings of the two peaks, and the total attack rate depend on the parameter settings in each scenario.

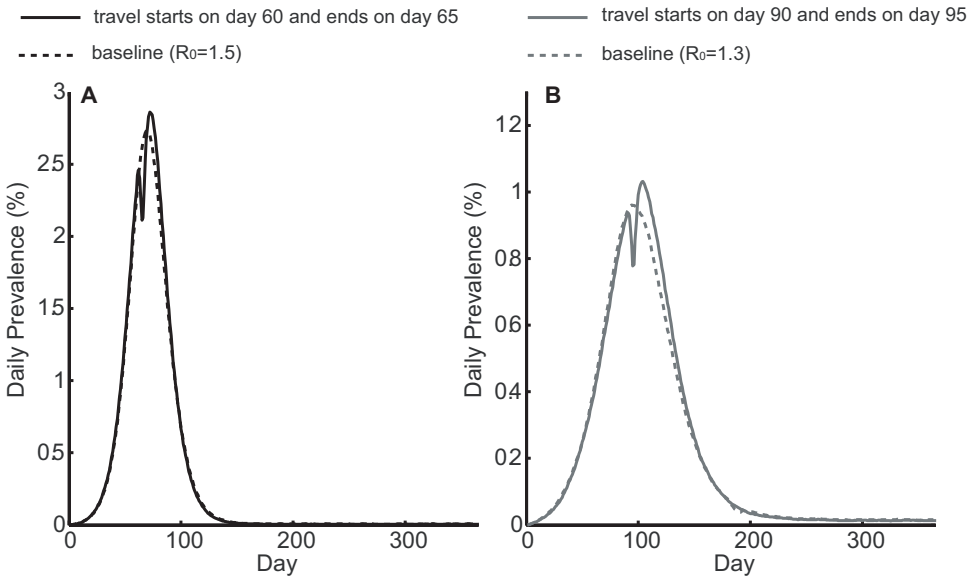


Figure 68: **Epidemic curves in the Holiday scenarios.** The figure shows the daily prevalence of infection (i.e., proportion of the symptomatic and asymptomatic persons over the entire population) for the entire state of Georgia und er the Holiday setting. Here 25% of the population travels during a 5-day traveling or mass gathering period with two initial R_0 values: a) R_0=1.5; b) R_0=1.3.

The appearance of the two epidemic peaks is due to partial social-distancing, as a large proportion of the population no longer mixes in the workplaces/schools when the Holiday (traveling) begins, causing a momentary drop in new infection cases until the Holiday is over and mixing resumes. To isolate the effects of traveling versus the reduction in peer group mixings, Figure 69 compares the epidemic curves for the entire state of Georgia in the following two scenarios using the initial $R_0 = 1.5$: (1) 25%

population on travel during a 5-day Holiday period starting on Day 60 as previously described; (2) the same number of persons reduce their peer group mixings and stay at home day and night during a 5-day period starting on Day 60. The second scenario models social distancing or household quarantine. As shown in Figure 69, there are two epidemic peaks in both scenarios; however, the prevalence of the second peak in the social-distancing scenario (2.47%) is lower than that in the traveling scenario (2.86%). The total attack rate in the former is 50.4%, and 51.7% in the latter. This is consistent with our previous observation: traveling/mass gatherings can lead to an increase in the peak prevalence and the total attack rate, but do not cause a second peak alone among the experiments we test.

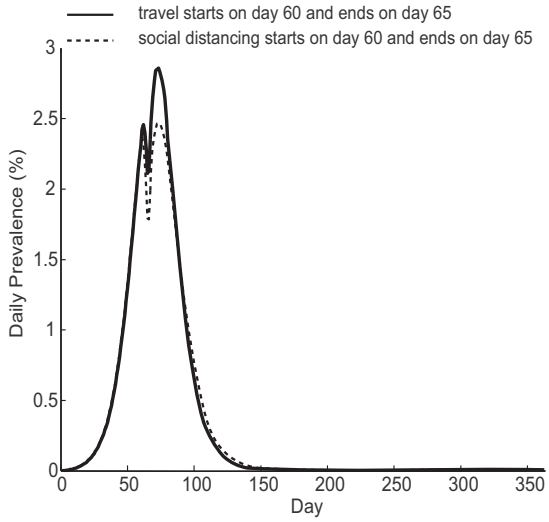


Figure 69: **Epidemic curves in the Holiday and social distancing scenarios.** The figure shows the daily prevalence of infection (i.e., proportion of the symptomatic and asymptomatic persons over the entire population) for the entire state of Georgia under the Holiday and the social distancing settings. Here the initial R_0=1.5; 25% of the population travels during a 5-day period starting on Day 60 (solid curve), or reduces their peer group mixings during the same period time (dotted curve).

7.3.3 The duration of the mass traveling period (l) and the proportion of the population traveling (p) under the non-Holiday setting

Tables 21 to 23 also compare the peak prevalence and the total attack rate in Georgia for different combinations of traveling/gathering duration l and the proportion of the population that travels/gathers when the initial $R_0 = 1.3$, 1.5, and 1.8 under non-Holiday conditions. As Tables 21 to 23 demonstrate, even a half-day event can lead to as high as an 8% increase in the peak prevalence (e.g., with 25% of the population involved in a half-day event starting on Day 90 and the initial $R_0 = 1.3$). Moreover, 1-day and 2-day traveling periods result in similar peak prevalence values to each other (a 3% maximum relative difference) and very similar total attack rates (a 1% maximum relative difference).

However, extending the event duration from 2 to 3 days reduces the peak prevalence and total attack rate somewhat (although they remain higher than if mass gathering did not occur) in some scenarios. For example, when the initial $R_0 = 1.5$ and 10% of the population is involved in a mass gathering event, the resulting peak prevalence and total attack rate are 2.89% and 51.4%, respectively, after a 2-day event starting on Day 60; however, these values are 2.80% and 51.1%, respectively, after a 3-day event starting at the same time. Note that the baseline average infectious period is 3-4 days (see Table 17). We conduct sensitivity analyses in which the infectious period is assumed to be of 7 days [93, 94]. Under the new assumption of the infectious period, when the initial $R_0 = 1.5$, the total attack rate is 49.05% and the peak prevalence is 4.05% in the baseline scenario without traveling/mass gathering. The total attack rate becomes 51.2%, 51.3%, and 51.4% when the traveling period starts at 20 days before the epidemic peak (in the baseline scenario) and lasts for 1, 2 and 3 days, respectively. The peak prevalence becomes 4.56%, 4.58%, and 4.60%, respectively. Thus, when the infectious period is 7 days, the total attack rate and peak prevalence increase when the traveling/gathering duration l increases.

The proportion of the population traveling/gathering shows a larger impact on the peak prevalence and the total attack rate. When the initial $R_0 = 1.5$ and 25% of the population starts traveling on Day 60 for 1 day, the peak prevalence increases from 2.73% (baseline) to 3.04% (approximately a 11% relative increase), significantly greater than the 4% relative peak prevalence increase (compared to baseline) when only 10% of the population travels on Day 60. Smaller mass gatherings (i.e., 1%-5% of the population) do not result in substantial increases in the peak prevalence and the total attack rate. Tables 21 to 23 show that this observation holds for other initial R_0 values as well.

7.3.4 Risk for travelers' families under the non-Holiday setting

To study the potential increase of the infection risk for the people traveling/gathering and for their family members (i.e., the impact of secondary transmissions), we compare the prevalence and the total attack rate in the non-Holiday setting to the baseline scenarios, for the population of travelers/gatherers and their family members.

When the initial $R_0 = 1.5$ and 10% of the population is on travel during a 1-day traveling period beginning at Day 60 (or Day 30), the value of the peak prevalence is 2.97% (or 2.86%, respectively) and the total attack rate is 53.5% (or 53.0%, respectively) among the population of travelers/gatherers and their family members, while the peak prevalence in the entire population is 2.85% (or 2.78%, respectively) with a total attack rate 51.3% (or 51.0%, respectively).

The peak prevalence value and the total attack rate for individuals who travel or attend mass gatherings and their family members are higher than the corresponding average values for the entire population when the traveling or mass gatherings occur before the epidemic peak. Even if the traveling period starts at Day 90 (20 days after the epidemic peak in the baseline scenario), the total attack rate for the travelers and their families is 53.0%, still higher than that for the entire state (51.0%).

7.3.5 Regional impact of traveling and mass gatherings

The aforementioned scenarios assume that the proportion of persons traveling/gathering are uniform throughout the entire state; however, mass gatherings may disproportionately involve residents of certain areas or neighborhoods (e.g., residents closer to the mass gathering event may be more likely to attend than persons remote). Therefore, an additional set of scenarios explores the impact of regional differences in traveling and mass gatherings under the non-Holiday setting. Figure 70 depicts the scenarios when the initial $R_0 = 1.5$, the traveling period is 1 day, and 50% of the population in Bibb County and its nearest 5 counties [20] are mixing in the traveling group with 9.5% of the population traveling from all other counties (resulting in 10.4% total of the entire population on travel). Figure 70 shows the maximum and minimum, the 25% and 75% percentiles, and the mean of the peak prevalence value and the peak day for Bibb County (from 50 replications) with traveling starting on Day 30, Day 60, and without traveling (baseline scenario).

As shown in Figure 70(a), when the traveling/mass gathering starts on Day 60 and lasts for 1 day, the peak prevalence in Bibb County can reach as high as 4% in some experiments (compared to 2.82% in the entire state). The average peak prevalence is 3.32%, and the average total attack rate is 50.1%, which are higher than the baseline value of peak prevalence (2.82%) and total attack rate (48.9%) for Bibb County.

Moreover, Figure 70(b) indicates that the traveling/gathering occurring before the peak prevalence day (e.g., Day 30) can synchronize the timing of the epidemic curves in a local county and in the entire state. In the baseline case, the day when the prevalence peaks in Bibb County can appear as late as Day 95, which is 25 days after the peak day in the entire state. With traveling/mass gathering occurring on Day 30, the peak day in Bibb County is mostly reached before Day 75 (with 75% chance); and furthermore, in some experiments, the peak day can occur as early as Day 56 due to the early introduction of seed infections to the local area.

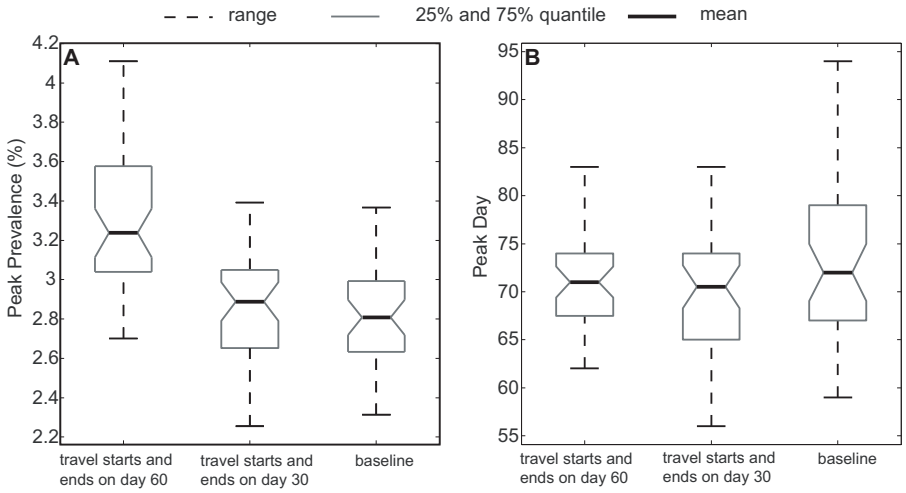


Figure 70: **Peak prevalence value and peak day in Bibb County.** The box plots show the range (with maximum and minimum, dotted line), 25% (lower gray line) and 75% (upper gray line) percentile, and the mean value (solid black line) for the peak prevalence (subfigure a) and the peak day (subfigure b) in Bibb County. Here 50% of the population from Bibb County and its nearest 5 counties travels and mixes with 9.5% of the population from other counties in the traveling group. The initial R_0=1.5, the traveling period lasts for 1 day, and it starts on Day 30, 60, or no traveling (baseline).

7.4 Discussion

Our simulation experiments have identified situations where mass traveling or gatherings that occur *shortly* before the epidemic peak may worsen or alter the course of the influenza epidemic (e.g., resulting in a higher peak prevalence and total attack rate and in some cases generating two epidemic peaks), which may substantially affect planning and potentially strain healthcare facilities and resources. This impact can be greatest on the local communities hosting the mass gatherings. Therefore, public health officials, local authorities, and other decision makers may consider closely monitoring, postponing or canceling public gatherings near the peak of an epidemic. Moreover, pandemic surveillance and other responses should not necessarily be slowed even after a large decline in influenza activity since a second epidemic peak may occur after Holiday traveling. Conversely, our experiments suggest that mass traveling or gatherings may have little effect when occurring relatively early or past the peak in

an epidemic (with high enough herd immunity achieved [71, 3, 78, 129]).

Our study emphasizes the impact of social mixing patterns and the creation and distribution of immune individuals on the progression of an epidemic. Specifically, when individuals mix in households, schools, and workplaces without major changes, they can generate pockets of adequate herd immunity to prevent additional transmission. In other words, if a large percentage of individuals at one's workplace and household are immune then one's risk of infection may be low, even though many infectious individuals are still in the population. This is because individuals tend to stick with their typical social contacts and do not mix with a majority of the population. However, a mass gathering brings together people that normally would not mix, i.e., it brings together susceptible and infectious individuals that would not have interacted otherwise, thus potentially worsening the epidemic.

Additionally, the differences between the durations of the mass gathering and the pathogen's infectious period can substantially alter the impact of mass gathering. When the mass gathering period is shorter than the pathogens infectious period (e.g., when the mass gathering is 1-2 days versus 3-4 days for the infectious period), mass gathering creates new infectious individuals who then return to their households, workplaces, and schools to infect their standard social networks, thereby worsening the epidemic. Conversely, when the mass gathering lasts as long as or longer than the pathogen's infectious period (e.g., when the mass gathering infectious period is 3-4 days and the pathogen's infectious period is 3-4 days), mass gathering can actually act as a mass immunization or mass quarantine event, keeping people in one location while they are infectious and then returning them to their social networks only after they are immune. Sensitivity analyses that increase the average infectious period to 7 days support this conclusion.

7.4.1 Public health implications

Canceling or postponing mass gatherings near the epidemic peak can be challenging. As seen during the 2009 H1N1 pandemic, it can be difficult to determine the current and anticipated future status of an ongoing epidemic. Moreover, changing a previously scheduled event can have economic and logistic consequences. In some cases, the scheduled date of a mass gathering can have significance. For example, Memish et al. [106] discussed the global religious event Hajj (pilgrimage by Muslims to Saudi Arabia, attracting more than 2.5 million pilgrims from the whole world every year), which is difficult to cancel during a pandemic.

The alternative to changing the scheduling of an event is close monitoring and enforcement of hygienic measures and precautions during the event. Memish et al. [106] and Rashid et al. [126] presented several recommendations for local governments to follow, including screening, surveillance, and most importantly, encouraging attendees from high risk groups (e.g., elderly and pregnant women) to postpone their participation in the event. Also, reducing the length and the scale of an event could be less drastic ways of reducing disease transmission. Even if an event cannot be cancelled, knowing that it may increase the overall attack rate and peak prevalence could help public health decision makers prepare (e.g., increasing health care resource availability and surge capacity).

7.4.2 Conclusion and future direction

Our study demonstrates how social mixing dynamics can be captured in a heterogeneous population, and shows the impact on prevalence, peak timing, and secondary transmissions within families or regions. Our study suggests that when mass gatherings and traveling occur close to the peak of an epidemic, they could worsen the overall attack rate and the peak prevalence. However, such changes in social mixing may have little effect when they occur earlier or later in the course of an epidemic.

Public health decision makers may use this information to help decide whether to postpone, cancel, monitor, or enforce infection control measures during a mass gathering or Holiday season.

From the modeling side, we have developed an agent-based simulation model which allows the flexibility to change mass social mixing patterns in different time periods. Such flexibility is not explicitly incorporated in other presented simulation models. We have also developed new ways to estimate transmission parameters to calibrate the disease spread model when individuals can have different mixing patterns. This model may be generalized to other settings besides Holiday traveling and mass gathering events, such as modeling patients seeking medical care and mixing with other patients in hospitals during an influenza pandemic.

Finally, similar to the previous chapter, we want to emphasize that computer simulations are simplifications of real life. Rather than make decisions, simulation can identify potentially important factors and relationships for decision makers. Our model does incorporate a number of assumptions and cannot fully capture every possible factor or effect. For example, we assume homogeneous mixing within the traveling/mass gathering group during the traveling/gathering period. In real life, people may not have contact with every attendee in a mass gathering. Also, mass gathering events are not equivalent. Some may involve closer and more extended contact than others. The type of venue and location can play a significant role. Different events can involve people of different ages, socioeconomic status, and potentially health status. Future studies could extend our model to incorporate these realistic features. We also want to mention that although we have conducted a wide-range of sensitivity analyses, it is not possible to explore every possible combination of parameters.

APPENDIX A

APPENDIX FOR CHAPTER 2

A.1 Dealing with entries with missing admission or discharge information

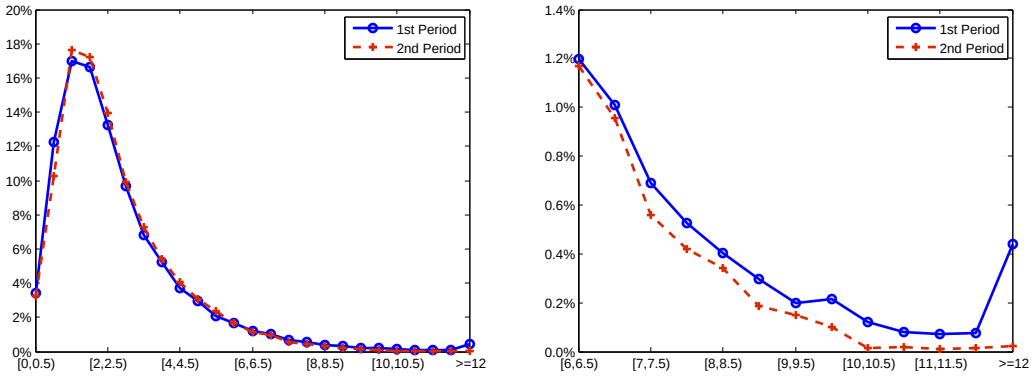
The data we have covers three years, from 2008 to 2010. Records with admission or discharge time outside the three-year period have incomplete admission or discharge information. For example, if a patient is admitted before January 1, 2008 and is discharged on January 15, 2008, then her admission information is missing. We note that there are 650 records lacking admission information and 10 lacking discharge information. To ensure consistency, we apply the following conventions to these records:

1. All records with complete admission information are included in any analysis that is related to inpatient admission (e.g., admission time, daily admission rate), no matter whether discharge information is missing or not.

2. All records with complete discharge information are included in any analysis that is related to discharge (e.g., discharge distribution in Section 2.2), no matter whether admission information is missing or not.

3. Only records with both admission and discharge information are included in the analysis for LOS and service time (e.g., Sections 2.5 and 2.6). Records with either missing admission or discharge information are excluded.

4. All records identified as "visited ED" and with complete admission information are included in analyzing ED-GW patient's waiting times and bed-request rates, no matter whether discharge information is missing or not.



(a) Distribution between 0 and 12 hours (b) Distribution between 6 and 12 hours

Figure 71: **Empirical distributions of the waiting times for ED-GW patients.** The waiting times are calculated in the conventional way, i.e., using the duration between bed-request and patient exiting ED. The bin size is 0.5 hour, and points falling beyond 12 hours are lumped together into the last bin.

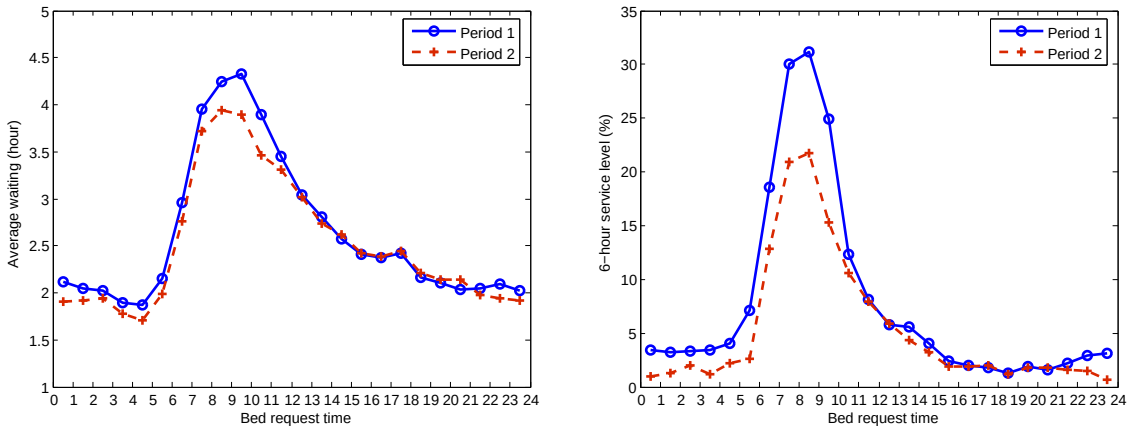
As a consequence of following these conventions, the total sample size may vary for different analyses.

A.2 Waiting time statistics calculated in the conventional way

As introduced in Section 2.1.1, waiting time reported in this thesis is calculated in a slightly different way from that used in the medical literature. Here, we report waiting time statistics calculated from the conventional way in literature, i.e., using the duration between bed-request time and when patient exiting from ED.

A.2.1 Distribution of waiting time

Figure 71a shows the empirical distributions of waiting times for all ED-GW patients in Periods 1 and 2. The bin size is 0.5 hour, and points falling beyond 12 hours are lumped together into the last bin. For hospital management purpose, we are particularly interested in those patients with excessive long waiting times. Thus in Figure 71b, we provide a detailed plot for the waiting time distribution between 6 and 12 hours. The shapes of the overall distribution curves look similar in both periods.



(a) Average waiting times

(b) 6-hour service levels

Figure 72: **Waiting time statistics for ED-GW patients by bed-request hour.** The waiting times are calculated in the conventional way, i.e., using the duration between bed-request and patient exiting ED.

The tail distributions, however, exhibit significant differences.

A.2.2 Hourly average waiting time

As reported in Table 1, the average waiting time (calculated in the conventional way) for all ED-GW patients is 2.52 hours for Period 1, and 2.46 hours for Period 2, a reduction of 3.6 minutes. Thus, there is no significant difference between the two periods.

Figure 72a plots the hourly average waiting time. Figure 72a shows a similar shape as Figure 1a(a); the difference between the two figures is that we calculate the waiting times differently. Table 24 lists the corresponding numerical values for Figure 72a.

A.2.3 Service levels

Figure 72b, which shows the 6-hour service level with respect to bed-request hour, is similar to Figure 1(b) in [135], except that we calculate the waiting times differently. Table 24 lists the corresponding numerical values. We also observe a time-dependent feature of the 6-hour service level. Patients requesting beds between 6am and 12noon

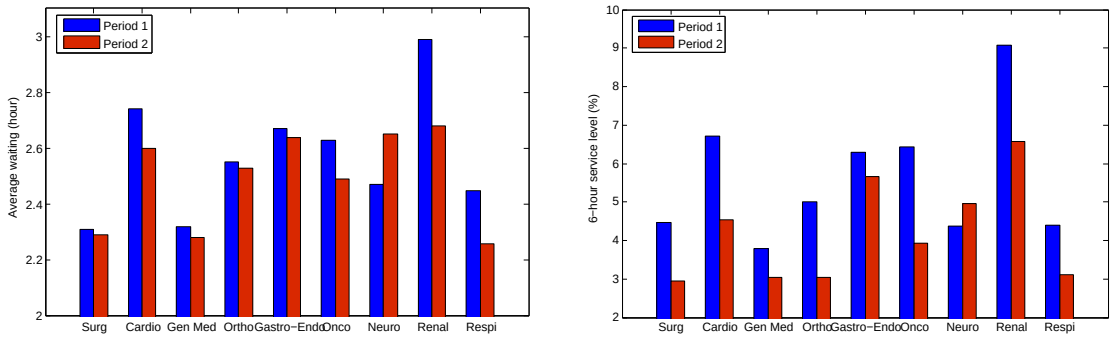
Table 24: The average waiting times and service levels by bed-request hour.

	per	bed-request hour											
		1	2	3	4	5	6	7	8	9	10	11	12
req. dist.	1	4.33	3.48	2.71	2.40	1.81	1.71	1.68	1.50	1.62	2.41	3.51	4.62
(%)	2	4.31	3.33	2.57	2.09	1.65	1.67	1.53	1.67	1.77	2.57	3.41	4.98
avg. wait	1	2.12	2.05	2.02	1.89	1.87	2.15	2.96	3.95	4.24	4.33	3.89	3.45
(h)	2	1.91	1.92	1.94	1.78	1.71	1.99	2.76	3.72	3.94	3.89	3.47	3.31
$f(W \geq 4)$	1	8.34	6.24	5.93	5.64	6.22	10.87	23.49	42.59	52.70	56.84	48.79	32.82
(%)	2	6.88	5.95	5.24	3.02	3.37	7.57	23.20	43.94	49.22	44.68	36.54	30.84
$f(W \geq 6)$	1	3.39	3.24	3.33	3.41	4.04	7.08	18.62	30.02	31.13	24.91	12.30	8.11
(%)	2	1.01	1.31	2.00	1.13	2.16	2.60	12.89	20.90	21.70	15.25	10.56	7.95
$f(W \geq 8)$	1	2.80	2.43	2.81	2.59	2.80	4.45	11.24	9.57	7.65	4.21	2.89	2.20
(%)	2	0.18	0.83	0.92	0.76	0.96	1.18	3.09	5.94	3.36	2.16	2.55	1.91
$f(W \geq 10)$	1	2.35	2.27	1.98	1.18	1.87	1.15	2.35	1.50	1.22	0.82	0.56	0.49
(%)	2	0.00	0.12	0.31	0.57	0.24	0.00	0.00	0.24	0.00	0.15	0.12	0.00
	per	bed-request hour											
		13	14	15	16	17	18	19	20	21	22	23	24
req. dist.	1	5.46	6.18	6.30	6.61	6.14	6.16	5.64	4.86	5.07	5.55	5.12	5.12
(%)	2	5.73	5.99	6.58	6.84	5.92	5.92	5.40	5.09	5.20	5.62	5.20	4.98
avg. wait	1	3.04	2.81	2.58	2.41	2.37	2.42	2.17	2.10	2.04	2.05	2.09	2.03
(h)	2	3.02	2.73	2.62	2.42	2.39	2.44	2.21	2.14	2.14	1.98	1.94	1.92
$f(W \geq 4)$	1	22.99	20.40	16.88	14.94	12.37	10.77	7.20	6.85	6.40	5.95	6.94	6.44
(%)	2	25.45	20.59	17.43	12.83	11.03	11.49	8.57	8.16	8.14	6.41	7.15	6.20
$f(W \geq 6)$	1	5.79	5.61	4.03	2.39	1.98	1.79	1.30	1.86	1.56	2.24	2.92	3.08
(%)	2	5.93	4.36	3.19	1.91	1.87	2.00	1.17	1.79	1.83	1.62	1.45	0.71
$f(W \geq 8)$	1	1.29	1.46	0.58	0.30	0.60	0.50	0.35	0.75	0.78	1.32	2.09	2.09
(%)	2	1.10	0.33	0.60	0.17	0.00	0.73	0.22	0.78	0.46	0.35	0.23	0.48
$f(W \geq 10)$	1	0.05	0.23	0.04	0.17	0.05	0.23	0.20	0.46	0.56	1.02	1.82	1.38
(%)	2	0.14	0.00	0.06	0.00	0.00	0.27	0.00	0.08	0.08	0.00	0.15	0.08

have a much higher chance of waiting more than 6 hours than patients requesting beds in other hours. In Period 1, about 1 out of 3 patients requesting beds between 8 and 9am have to wait more than 6 hours. Comparing the two periods, the peak value of the 6-hour service level (8-9am) decreases from 31% to to 22% in Period 2. Table 24 also lists the 4-, 8-, and 10-hour service levels with respect to the bed-request hour. The 8-hour and 10-hour service levels are greatly reduced in each hour in Period 2.

A.2.4 Waiting time statistics for each specialty

Figures 73a and 73b plot the average waiting times and 6-hour service levels for the nine specialties in the two periods. Table 25 shows the corresponding numerical values, and contains statistics for other service levels.



(a) Average waiting times

(b) 6-hour service level

Figure 73: **Waiting time statistics for each specialty.** The waiting times are calculated in the conventional way, i.e., using the duration between bed-request and patient exiting ED.

We make two observations. First, the nine specialties exhibit similar average waiting time and 6-hour service levels in each period (especially in Period 2). This balanced result could have been achieved through years of continual adjustment in resource allocation (e.g., bed and ward allocation) and a proper overflow policy (see Section 2.3.3). Renal and Cardiology patients show longer average waiting times than the overall average, while Surgical, General Medicine, and Respiratory patients show shorter average waiting times than the overall average. The potential reasons could be that (i) Surgery, General Medicine, and Respiratory wards have relatively low BORs (see Table 3); moreover, patients from Surgery and General Medicine can be overflowed to wards of other specialties easily since they have less specialized requirements; (ii) Renal and Cardiology wards have high BORs, and these patients need more specialized care and equipment (e.g., dialysis for Renal patients, telemetry beds for Cardiology patients) so it is more difficult for them to be overflowed.

Second, comparing the two periods, we can see that the average waiting time does not change much for each specialty except for Renal and Respiratory. Meanwhile, the 6-hour service level exhibits a significant reduction in Period 2 for each specialty except Neurology. These observations are consistent with what we observed from the hospital-level statistics (see Table 1). They all suggest that patients with long waiting

times (as noted, a very small amount) benefit more in Period 2 than most patients.

Table 25: Waiting time statistics for ED-GW patients from each specialty.

	Period	Sample size	avg. wait (h)	$f(W \geq 4)$ (%)	$f(W \geq 6)$ (%)	$f(W \geq 8)$ (%)	$f(W \geq 10)$ (%)
Surg	1	6078	2.31	13.29	4.46	1.40	0.61
	2	3926	2.29	12.66	2.95	0.76	0.05
Card	1	5437	2.74	19.22	6.71	2.21	0.74
	2	4011	2.60	18.08	4.54	0.85	0.02
Gen Med	1	7913	2.32	12.26	3.80	1.44	0.80
	2	6176	2.28	11.92	3.04	0.65	0.08
Ortho	1	4557	2.55	15.16	5.00	1.73	0.75
	2	2899	2.53	13.42	3.04	0.83	0.07
Gastro-Endo	1	3348	2.67	18.43	6.30	2.42	0.96
	2	2309	2.64	19.23	5.67	1.08	0.09
Onco	1	1586	2.63	17.40	6.43	3.22	1.32
	2	1271	2.49	16.76	3.93	0.79	0.08
Neuro	1	2669	2.47	15.36	4.38	1.57	0.52
	2	1979	2.65	18.85	4.95	0.76	0.10
Renal	1	2315	2.99	23.24	9.07	3.59	1.43
	2	1686	2.68	19.51	6.58	2.02	0.42
Respi	1	1549	2.45	14.33	4.39	1.23	0.39
	2	1028	2.26	12.16	3.11	0.58	0.00

A.3 BOR contributed by primary and overflow specialties for each ward

Besides the overflow proportion introduced in Section 2.3.3, we define *BOR share* of a specialty (or group of specialties) as the BOR of the speciality, or group, divided by the total BOR for a certain ward. To calculate the BOR of one specialty for a given ward, the numerator in Equation (1) counts the total patient days for patients from that specialty who used beds in the ward. The denominator counts the total bed days available for all beds from that ward. Thus, the sum of the BORs from each specialty equals the total BOR of that ward (as reported in Table 3). Correspondingly, the BOR share from each specialty adds up to 1 for that ward.

When modeling beds in a ward as servers, the BOR share resembles the workload share in queueing systems, i.e., out of all "busy" periods, the average proportion of

time that these beds are "working" on patients from a particular specialty. The BOR share provides us with a deeper insight into the overflow issue. Patients who are initially assigned to a wrong ward may be transferred to the right ward later (see more discussions in Section 2.8.3 on such transfers). Typically, this happens a day or two after the patient's initial admission; otherwise, the hospital usually allows the patient to remain in the wrong ward until discharge. The overflow proportion only takes patient count into consideration, without differentiating an overflow patient with a long LOS from an overflow patient with a short LOS, where the latter is always preferred for the right-siting of care. Therefore, we study this BOR share statistic, since it takes patient's LOS into consideration from the BOR calculation.

Figure 74 plots the BORs from primary and non-primary specialties for each ward in Periods 1 and 2. We also refer the two BORs as right-siting BOR and overflow BOR, respectively. Each bar in the figure represents the total BOR for each ward in the corresponding period. Even though the figure does not directly plot the BOR share (since the BOR share from primary and non-primary specialties should add up to 1 for a ward), it gives us some insight regarding the time the ward serves right-siting patients and overflow patients, as well as its "idle" time, when it is not serving patients. Table 26 contains the numerical values for the overflow BOR share for each ward in Periods 1 and 2. Using 1 minus the overflow BOR share obtains the right-siting BOR share. We observe similar features regarding the overflow BOR share and overflow proportions. For example, dedicated wards have lower BOR share from overflow patients, and Orthopedic wards and class A/B1 wards expend more time treating overflow patients. Moreover, most wards in Period 2 show a reduction in overflow BOR share.

Comparing the overflow BOR share with overflow proportion in Table 26, we can see that the overflow BOR share is generally smaller than the corresponding overflow proportion, e.g., ward 54 and 55. This has two implications. First, some overflow

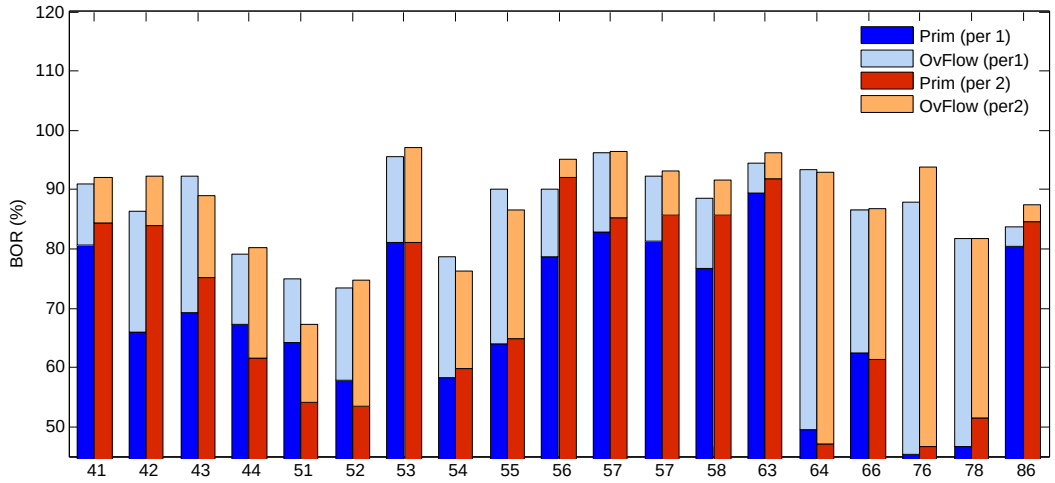


Figure 74: **BOR from primary and non-primary specialties for each ward in Period 1 and 2.** Each bar height represents the total BOR for each ward in the corresponding period. The y-axis starts from 45%.

patients only stay in the wrong wards for a day or two before transfer to a right ward. Thus, the lower overflow BOR share value (compared to overflow proportion) reflects NUH's efforts on right-siting. Second, the overflow patients have a shorter average LOS compared to the primary patients, even when they do not transfer to a right ward, e.g., ward 58 is dedicated to serve Oncology patients, and most of its overflow patients are from General Medicine with a shorter average LOS. In fact, this also explains why ward 43 shows a higher overflow BOR share value than overflow proportion in Period 1, since most of its overflow patients are from Orthopedic with a longer average LOS than its primary Surgery patients.

A.4 Additional statistics for LOS and service time

A.4.1 LOS distributions

Table 27 lists the empirical distributions of LOS in Periods 1 and 2, with the cut-off value at 30 days. Table 28 lists the tail frequencies of LOS after 30 days for the two periods; the bin size is 5 days and the cut-off value is 90 days.

Table 29 lists the total sample sizes and the LOS distributions, truncated to the first 21 values, for ED-AM and ED-PM patients in the two periods.

Table 26: **Overflow proportion and BOR share for each ward.**

Ward	OvFlow proportion (%)		OvFlow BOR share (%)	
	per 1	per 2	per 1	per 2
41	11.2	12.5	11.4	8.3
42	30.5	11.8	23.7	8.9
43	19.7	17.3	25.1	15.5
44	12.1	24.0	14.9	23.2
51	27.1	31.0	14.3	19.4
52	33.0	41.9	21.3	28.5
53	20.4	12.1	15.2	16.6
54	40.1	37.0	25.8	21.6
55	37.7	35.3	29.0	25.1
56	15.7	5.7	12.6	3.2
57	19.2	14.0	13.9	11.6
57	21.6	14.9	11.8	8.1
58	25.3	13.5	13.5	6.6
63	6.5	5.6	5.2	4.5
64	50.7	52.9	47.0	49.3
66	30.0	28.2	27.9	29.3
76	44.8	47.5	48.4	50.2
78	46.7	41.2	43.0	37.1
86	10.0	8.7	4.0	3.2
Total	27.0	25.0	21.4	19.2

Table 27: LOS distribution in Periods 1 and 2. The cut-off value is chosen at 30 days.

LOS	Period 1	Period 2	LOS	Period 1	Period 2
0	2.85%	3.30%	16	0.52%	0.43%
1	19.99%	20.40%	17	0.44%	0.32%
2	21.62%	22.05%	18	0.32%	0.30%
3	14.85%	14.95%	19	0.32%	0.25%
4	9.99%	10.20%	20	0.29%	0.26%
5	6.86%	6.88%	21	0.26%	0.21%
6	5.05%	4.98%	22	0.23%	0.20%
7	3.69%	3.43%	23	0.15%	0.17%
8	2.75%	2.69%	24	0.18%	0.21%
9	2.08%	1.96%	25	0.12%	0.11%
10	1.70%	1.51%	26	0.14%	0.13%
11	1.29%	1.05%	27	0.07%	0.08%
12	1.10%	0.93%	28	0.10%	0.08%
13	0.85%	0.82%	29	0.07%	0.08%
14	0.70%	0.67%	30	0.09%	0.05%
15	0.55%	0.56%	> 30	0.78%	0.73%

Table 28: LOS tail frequencies (start from 31 days, cut-off at 90 days).

bin	Period 1	Period 2	bin	Period 1	Period 2
(30,35]	0.25%	0.27%	(60,65]	0.03%	0.02%
(35,40]	0.16%	0.15%	(65,70]	0.02%	0.02%
(40,45]	0.10%	0.09%	(70,75]	0.01%	0.01%
(45,50]	0.08%	0.06%	(75,80]	0.01%	0.01%
(50,55]	0.04%	0.05%	(80,85]	0.01%	0.01%
(55,60]	0.04%	0.02%	(85,90]	0.01%	0.01%
			> 90	0.02%	0.02%

Table 29: **LOS distributions for ED-GW patients admitted in AM and PM.** Sample sizes only include ED-GW patients.

	ED-AM		ED-PM	
	Period 1	Period 2	Period 1	Period 2
	sample size			
	10156	7189	22897	16046
LOS	distribution (%)			
0	11.29	13.12	0.32	0.42
1	25.67	26.44	15.45	17.33
2	18.87	18.97	23.03	23.44
3	12.40	11.95	17.02	17.23
4	8.04	8.07	11.38	11.25
5	5.18	4.49	7.77	7.59
6	3.92	3.78	5.28	4.97
7	2.88	2.66	3.75	3.66
8	1.93	2.18	3.00	2.78
9	1.57	1.36	2.25	2.14
10	1.50	1.13	1.81	1.56
11	1.03	0.95	1.42	1.06
12	0.94	0.64	1.15	0.98
13	0.70	0.61	0.92	0.83
14	0.49	0.49	0.76	0.72
15	0.46	0.40	0.57	0.56
16	0.48	0.29	0.48	0.52
17	0.36	0.26	0.54	0.33
18	0.27	0.22	0.36	0.29
19	0.22	0.22	0.36	0.26
20	0.18	0.21	0.31	0.26
>20	1.63	1.54	2.08	1.82
average	3.70	3.46	4.78	4.48

A.4.2 Speculation on the one-day difference in average LOS between ED-AM and ED-PM patients

We use the following example with two hypothetical scenarios to further illustrate our speculation on the one-day difference in average LOS between ED-AM and ED-PM patients. In this example, we make three assumptions: an AM patient is admitted at 2am; a PM patient is admitted at 4pm; and both patients are discharged at 3pm. These assumptions actually represent a typical situation, since most ED-AM patients are admitted between midnight and 4 am, most ED-PM patients are admitted between 3pm and 8pm, and the discharge peak is between 2pm and 3pm.

Scenario 1

An AM-patient admitted at 2am on May 1, 2008 has a medical condition that requires 1 day for surgery and 2 days for pre/post-surgery testing and treatment. She can utilize the day of admission (May 1) to do pre-surgery tests. She receives surgery and other treatment on May 2 and May 3. She discharges at 3pm on May 4, 2008.

Scenario 2

A PM-patient admitted at 4pm on May 1, 2008 has the same medical condition as the AM-patient. But her admission time renders the day of admission wasted, and "pushes" the surgery and all pre/post testing and treatment one day later. Thus, she discharges at 3pm on May 5, 2008.

It is easy to calculate that the AM patient's entire service time is 3.54 days (85 hours) and and the LOS is 3 days, whereas the PM patient's entire service time is 3.96 days (95 hours) and the LOS is 4 days. The difference between the service time in the two scenarios is 0.42 day (10 hours), and the difference between LOS is 1 day. All these numbers match the statistics we show for ED-AM and ED-PM patients (see Section 2.6.2 for statistics on service time), which indicates our speculation could

be a reasonable explanation for the one-day difference between ED-AM and ED-PM patients. Future studies are needed to concretely identify factors causing this one-day difference.

A.5 Pre- and post-allocation delays

A.5.1 Pre-allocation delay

In Figure 28a, the hourly average for the pre-allocation delay is estimated from patients satisfying two conditions: (i) the allocated bed is available before the bed request time; and (ii) the allocated bed comes from the primary ward for the patient. We now show more empirical evidence for using condition (ii).

Figure 75 compares the empirical average durations between bed-request time and bed-allocation time for right-siting and overflow patients. Condition (i) is imposed for both groups of patients, and the two curves are plotted as functions of bed-request time. Clearly we can see that the average for overflow patients (red curve) is significantly longer than that for right-siting patients (blue curve). Moreover, we observe that for bed-request time from 1am to 8am, the differences in the average duration between overflow and right-siting patients are smaller than the differences in other hours.

We interpret Figure 75 with caution, because it cannot provide a definitive conclusion that overflow patients have a longer pre-allocation delay. In practice, BMU may wait for some time before deciding to overflow a patient if no primary bed is available upon the bed-request time. The actual search/negotiation process, which we use pre-allocation delay to capture, only starts after the overflow decision is made. Therefore, the actual pre-allocation delay for an overflow patient should equal to the duration between bed request and allocation time minus this "BMU's waiting time". However, the lack of time stamps prevents us from estimating the BMU's waiting time and thus the pre-allocation delay for overflow patients. The proposed model (see Section 4.2

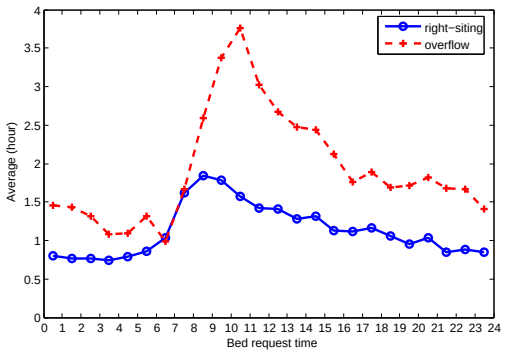


Figure 75: **Average duration between bed-request time and bed-allocation time for right-siting and overflow patients.** In both scenarios, the bed available time is earlier than the bed-request time.

of [135]) employs an overflow trigger time to mimic the BMU's waiting time, but it is only an approximation of the BMU practice and cannot be used in the allocation estimation. Thus, the stochastic model in [135] does not differentiate pre-allocation delay between right-siting and overflow patients. The model estimates the pre-allocation delay distributions from right-siting patients, and use them to approximate those of the overflow patients.

A.5.2 Post-allocation delay

Note that when estimating post-allocation delay, we do not differentiate the post-allocation delay distributions between the following two scenarios: (i) the allocated bed is available before the allocation time; and (ii) the allocated bed is available after the allocation time. Here, bed being "available" indicates that the previous patient occupying the bed has been discharged. In scenario (ii), the bed needs to be cleaned after the previous patient's discharge. This assumption on the post-allocation delay is supported by our empirical results. We separately estimate the average for the post-allocation delay under scenarios (i) and (ii). Figure 76 compares the hourly average between the two scenarios. We can see the blue curve, which represents scenario (i), is close to the red curve, which represents scenario (ii). The closeness of the two curves suggests that the bed cleaning time has almost no impact on the post-allocation delay.

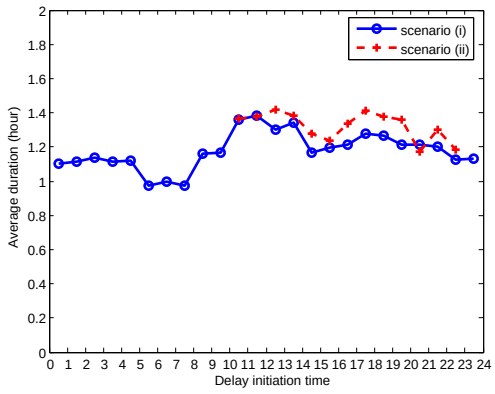


Figure 76: **Estimated average post-allocation delay with respect to the delay initiation time.** Scenario (i): the allocated bed is available before the allocation time; (ii) the allocated bed is available after the allocation time. Certain time interval of the red curve is omitted because of limited data points. Post-allocation delay equals the duration between the bed allocation time and the admission time for scenario (i); and duration between the bed available time and the admission time for scenario (ii).

The observation from Figure 76 can be partially explained as follows. NUH implements an auto countdown system for bed cleaning. After a patient is discharged, the bed tracking system marks the bed as "in cleaning" and automatically counts down for 30 minutes. After 30 minutes, no matter whether the bed is indeed cleaned or not, the system changes the bed status to "vacant", indicating it is ready to serve a new patient. The ED nurses can access the bed status information in real time. They know that the ED discharge and transfer process typically takes longer than the 30-minute cleaning time. If a patient is waiting her allocated bed to receive her, the nurses usually initiate the discharge process once the bed status changes to "in cleaning" (or shortly after the change, indicated by the fact that the red curve is slightly higher than the blue curve in Figure 76). After the bed status changes to vacant, ED can then send the patient to the allocated bed. In such a way, the auto countdown system enables the nurses to do the discharge/transfer in parallel with the bed cleaning process. This ensures that the bed cleaning time does not become a major bottleneck like those discussed in Section 2.7.1.

APPENDIX B

APPENDIX FOR CHAPTER 3

B.1 Generating iid service times from $\lfloor S \rfloor$ and residual

Following Equation (2), one can choose to model the service time S as the sum of two random variables: an integer variable corresponding to $\lfloor S \rfloor$, and a residual variable corresponding to res(S). Moreover, Figure 24b, which shows similar distributions of the residuals regardless of the values for $\lfloor S \rfloor$, suggests an independency between the integer and residual variables. For a class of patients (patient class depends on admission source, specialty, admission period, etc; see the definition in Section 3.1 of [135]), we assume that their integer and residual parts each forms an iid sequence and the two sequences are independent. Thus, the service times are also iid. This iid model is different from the non-iid service time model proposed in [135].

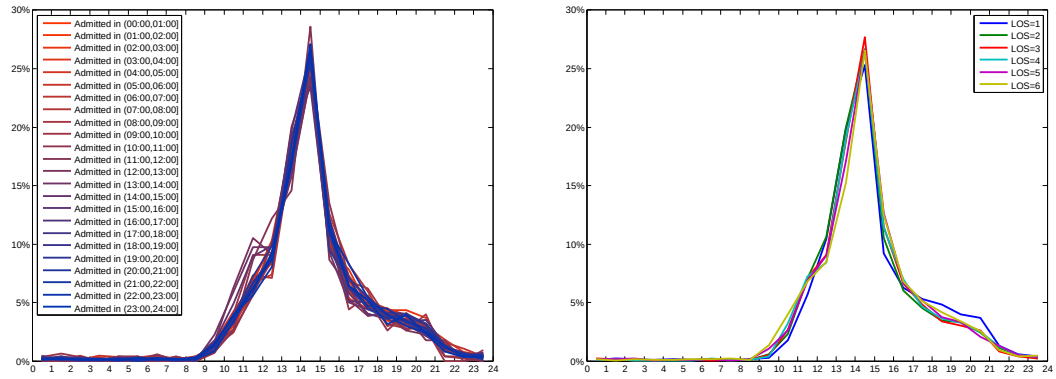
To populate this iid service time model, we empirically estimate the distributions for $\lfloor S \rfloor$ and res(S) as shown in the previous sections. For simulation, we generate the inter and residual parts independently from the appropriate empirical distributions, and use their sum as the service time.

B.2 Additional empirical results for the service time model

The proposed service time model in the main paper is in the form of (see Equation (3) in Section 4.3 of [135]):

$$S = \text{LOS} + h_{\text{dis}} - h_{\text{adm}}, \tag{63}$$

where LOS stands for the length of stay of the patient, and h_{dis} and h_{adm} represent hour of patient admission and discharge, respectively. The model assumes that h_{dis} is independent of LOS and of h_{adm} because LOS is believed to capture the amount



(a) Discharge distribution with respect to different admission hour

(b) Discharge distribution with respect to different LOS values

Figure 77: Independence between admission and discharge hours and between LOS and discharge hours using Period 1 data.

of time that a patient *needs* to spend in a ward due to medical reasons, whereas discharge hour h_{dis} clearly depends on the discharge patterns, which are mainly the results of scheduling and behaviors of medical staff. In this section, we provide some empirical evidence to support the assumption of the independency between h_{dis} and LOS and the independency between h_{dis} and h_{adm}. The dependency of LOS on the admission time has been discussed in Section 2.5.2.

Figure 77a plots the discharge distribution with respect to different admission hour, while Figure 77b plots the discharge distribution with respect to different LOS values. We note the closeness of the discharge distribution curves regardless of admission hour or LOS value. Even though we do not conduct a rigorous statistical analysis, the two figures support our assumption that the the discharge hour h_{dis} is independent of LOS and of h_{adm}.

B.3 Estimating the normal allocation probability $p(t)$

Recall that when a patient makes a bed-request at time t and there is no primary bed available at the time, we assume with probability $p(t)$ the allocation mode for the patient is the normal-allocation mode, meaning this patient will wait until a bed

is available before starting to experience the pre-allocation delay. In this section, we first demonstrate some empirical evidence to support our way of modeling the two allocation modes. Then we explain the rationale for using $p(t)$ given in (5) in Section 3.2.6. Consistent with the notation in Section 3.2.6, we use $h(t) = 24(t - \lfloor t \rfloor)$ to denote the time-of-day for the bed-request time t. We use hour as the time unit for $h(t)$, and day for t.

B.3.1 Rationale of the two allocation modes

We first empirically study the duration between bed-request time and bed-allocation time (i.e., when a bed is allocated to an incoming patient) for three different groups of patients. Figure 78a plots the empirical estimate of the average of this duration with respect to the bed-request hour for the three groups. The first group, corresponding to the blue curve, consists of ED-GW patients whose allocated bed is a primary bed and the bed is available before the bed-request time. The second group, corresponding to the red curve, consists of ED-GW patients whose allocated bed is not available at bed-request time and whose bed-allocation time is *later* than the bed-available time. For a patient in this group, the bed-allocation start time (i.e., when BMU agents start the bed searching and negotiation process) can be either before or at the bed-available time. In the latter case, the allocation is a normal allocation in our model. The third group, corresponding to the green curve, consists of ED-GW patients whose allocated bed is not available at bed-request time and whose allocation-completion time is *earlier* than the bed-available time. For a patient in this group, her bed-allocation start time is definitely before the bed-available time. Thus, this allocation cannot be a normal allocation. But we are not sure if it is a forward allocation as in the model because its allocation-start time may not start immediately at the bed-request time.

Recall that when we estimate the pre-allocation delay in the model, we use the

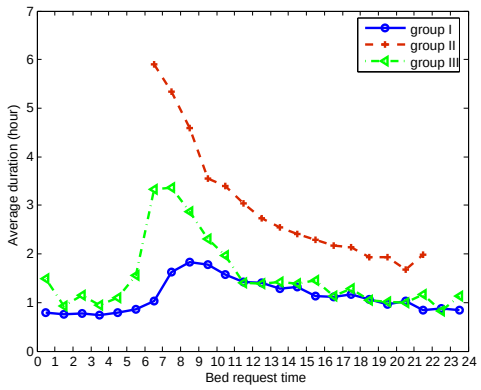
duration between bed-request time and bed-allocation time from the first group patients to reflect the minimum amount of time that BMU needs to search and allocates a bed. Thus, if the empirical data shows that

(a) the average duration between bed-available and bed-allocation time from the second group of patients is close to the average pre-allocation delay estimated from the first group of patients;

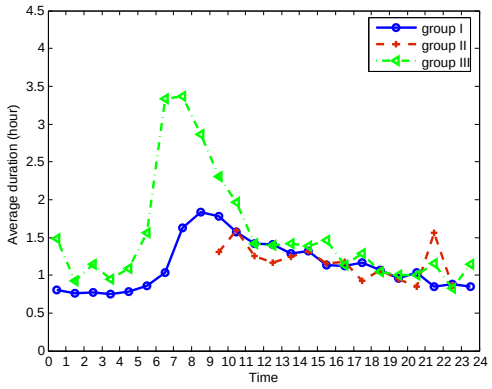
(b) the average duration between bed-request and bed-allocation time from the third group of patients is close to the average pre-allocation delay estimated from the first group of patients;

then it is a reasonable approximation of reality to (i) model the bed-allocation for the second group of patients the normal allocation, and (ii) model the bed-allocation for the third group of patients the forward allocation.

We first consider condition (b). From Figure 78a, we can see that for the majority of bed-request hours, from 11am to midnight, the blue and green curves are very close, suggesting that condition (b) approximately holds in this interval. Then we investigate condition (a). For second group of patients represented by the red curve, we plot a modified curve in Figure 78b. In the modification, we exclude the pure waiting times due to bed unavailability, and plot the average duration between their bed-available and bed-allocation times. We can see that the modified red curve is close to the blue curve between 2pm to 8pm, suggesting that condition (a) also approximately holds in the interval. Therefore, it is reasonable for us to use the two allocation modes (normal versus forward allocation) to approximate the reality. Furthermore, the second group of patients approximately correspond to patients experienced normal allocation, while the third group of patients approximately correspond to those experienced forward allocation.



(a) Average duration for three groups of patients



(b) Revised duration for the second group

Figure 78: **Estimate average duration between bed-request and bed-allocation for different groups of patients.** In sub-figure(a): average duration between bed-request time and allocation-completion time for ED-GW patients as a function of bed-request time; in the red curve, we omit certain time intervals due to the lack of data points (fewer than 15 points in each hour). In sub-figure(b): the red curve is a revision from the red one in sub-figure(a); the two other curves are kept the same. In the revision, the duration is revised to be between bed-available time and allocation-completion time, and is plotted against the bed-available time, not the bed-request time.

B.3.2 Estimating $p(t)$ in different time intervals

Now, we explain the rational of using the values in (5) for $p(t)$ in different time intervals.

First, the choice of $p(t) = 1$ for $h(t)$ between 8am and 12 noon is consistent with the current practice at NUH. In order to do a forward allocation, the planned discharge information should be available. Most wards do the morning rounds at about 9-11am, and nurses would only know which patients will be discharged after finishing the rounds. Thus, BMU typically receives the planned discharge information when the time is close to noon.

Second, for $h(t)$ between 2pm and 8pm, we use $\hat{p}(i)$ to empirically estimate $p(t)$ for each hour i between 2pm and 8pm. For all ED-GW patients (in the NUH data)

whose bed-request time falls within hour i, we define $\hat{p}(i)$ as

$$\hat{p}(i) = \frac{\#\text{ of patients whose bed-allocation time } > \text{ bed-available time}}{\#\text{ of patients whose bed is not available at bed-request time}}. \quad (64)$$

Here, the denominator consists of the patients whose allocated bed is not available at the bed-request time, i.e., the sum of the second and third groups of patients introduced in Section B.3.1. The patients included in the numerator are those from the second group of patients, which correspond approximately to normal allocations in this time interval. Therefore, it is reasonable to use $\hat{p}(i)$ to estimate $p(t)$ for $h(t)$ between 2pm and 8pm. Figure 79 shows that, between 2pm and 8pm, the ratio $\hat{p}(i)$ in (64) fluctuates near the $(40\%, 50\%)$ range. Based on these empirical estimates, we set $p(t) = .5$ between 2pm and 8pm.

Third, our empirical analysis also shows that, between 8pm and 6am the next day, there are very few (fewer than 15 each hour) normal allocations, suggesting $p(t)$ is close to zero. Therefore, we set $p(t) = 0$ for $h(t)$ between 8pm and 6am the next day.

Fourth, during each of the remaining time intervals of a day, $(6, 8]$ or $(12, 2]$, we estimate $p(t)$ by interpolating its values in the neighboring intervals to avoid sudden changes of $p(t)$. The actual values of $p(t)$ in these two intervals are obtained by trial-and-error so that the simulation estimates can approximately replicate the empirical waiting time performance.

Note that we did not use $\hat{p}(i)$ to estimate $p(t)$ in all intervals. This is because within the time interval $(6, 11)$, the three curves in Figure 78b diverge, suggesting that either condition (a) or (b) is severely violated. Therefore, we do not use $\hat{p}(i)$ in (64) to estimate $p(t)$ in this interval.

Finally, we realize that, despite our best efforts, our choice of $p(t)$ is still ad hoc. Therefore, we have conducted a sensitivity analysis of the choice of $p(t)$ in Section B.6.4.

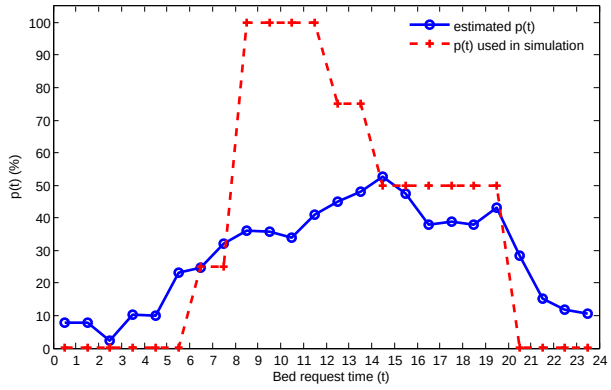


Figure 79: Estimated values of $p(t)$ from empirical data and values used in the baseline simulation.

B.4 Server pool setting

Table 13 lists 15 server pools. Each row specifies one server pool with the basic information including index, primary specialty, and number of servers. The table is based on the empirical study at NUH. We slightly adjust the number of servers in certain server pools because our proposed stochastic model does not capture all the constraints in bed assignment. For example, Orthopedic patients with open wounds cannot stay in the same room with patients who have acquired Methicillin-resistant Staphylococcus Aureus (MRSA), while our model does not differentiate MRSA patients from non-MRSA patients. Moreover, our model does not explicitly consider patients' preference for bed classes (beds in private and shared rooms are in different classes; see Section 2.3.1 for details of bed classes.) To compensate for the inefficiency caused by class mismatch in the real hospital setting, we assume pools 12, 13, and 14, which correspond to three wards that have class A or class B1 beds, to be overflow pools. These three pools only accept patients whose overflow trigger times are reached in the model. This adjustment is based on the facts that these wards usually do not admit patients who prefer class B2 or class C beds (for financial reasons) except for urgent situations. We also re-allocate some servers from the Orthopedic and Gastro-Endo pools (pools 4,7,10) into the three overflow pools. Thus, the server

numbers in these overflow pools are larger than the actual number of class A/B1 beds. This re-allocation is to capture the high overflow proportions in the Orthopedic and Gastro-Endo wards.

B.5 Simulation results for additional early discharge scenarios

In this section, we study the impact of early discharge on ED-GW patient's waiting time performance using a more comprehensive set of discharge distributions. In Section B.5.1, we introduce the hypothetical discharge distributions that will be tested. Then in Section B.5.2, we show simulation results from scenarios that use these discharge distributions. In Section B.5.3, we study a scenario that uses the Period 2 early discharge distribution and includes a capacity increase at the same time. We compare the simulation output from this scenario with the empirical performance in Period 2. In Section B.5.4, we demonstrate with an example that the Period 2 early discharge policy could have more significant benefits in reducing ED-GW patient's waiting time in other hospital settings.

B.5.1 Hypothetical discharge distributions

In our simulation experiments, we test a midnight discharge distribution and three other groups of early discharge distributions. The midnight discharge distribution simply assumes that all discharges occur at 0am each day, while the three other groups of discharge distributions are constructed as follows:

- Group (a) keeps the second discharge peak in the Period 2 discharge distribution unchanged, shifts the first discharge peak earlier by 1, 2, and 3 hours, and retains 26% discharge before noon;
- Group (b) uses a two-peak discharge distribution similar to the one in Period 2, but assumes 75% discharge before noon; the timing of the first peak is 9-10am,

10-11am, or 11am to noon;

- Group (c) shifts the entire Period 1 discharge distribution earlier by 1, 2, and 3 hours.

Figure 80 plots these three groups of discharge distributions. Note that the discharge distribution used in the Period 3 policy belongs to group (a) with the first discharge peak occurring between 8 and 9am. We differentiate the distributions within each of the three groups by their peak time, where the peak time for groups (a) and (b) refer to the time of the first discharge peak.

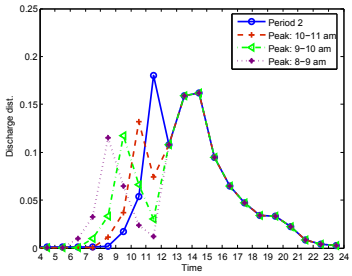
We use the midnight discharge distribution to test the maximum benefits that an early discharge policy might bring in reducing ED-GW patient's waiting time. We use groups (a) and (b) to test the impact of discharge timing and the proportion of discharge before noon on waiting time performance. Group (c) is motivated by the discharge scenarios tested in [120].

In our experiments, both the time-varying and the constant-mean allocation delay models are tested, combined with different discharge distributions as described above.

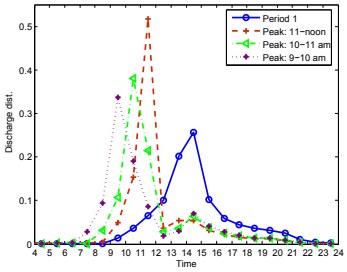
B.5.2 Selected simulation results

Simultaneous improvement is needed to flatten the waiting time curves

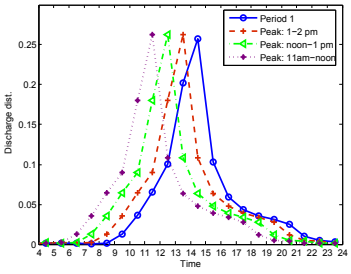
To achieve an approximately flattened waiting time performance, the hypothetical Period 3 policy proposed in Section 6.2 of the main paper [135] requires improvement in both the discharge timing and allocation delays. Here, we demonstrate that this simultaneous improvement is necessary. To show our results, we consider two scenarios. The first scenario uses the Period 2 discharge distribution and the constant-mean allocation delay model. The second scenario uses the same discharge distribution in the Period 3 policy and the time-varying allocation delay model. Each of these two scenarios differs from the Period 3 policy scenario only in one factor: either the discharge distribution or the allocation delay model.



(a) 26% patients discharge by noon



(b) 75% patients discharge by noon



(c) Shifting the entire Period 1 discharge distribution

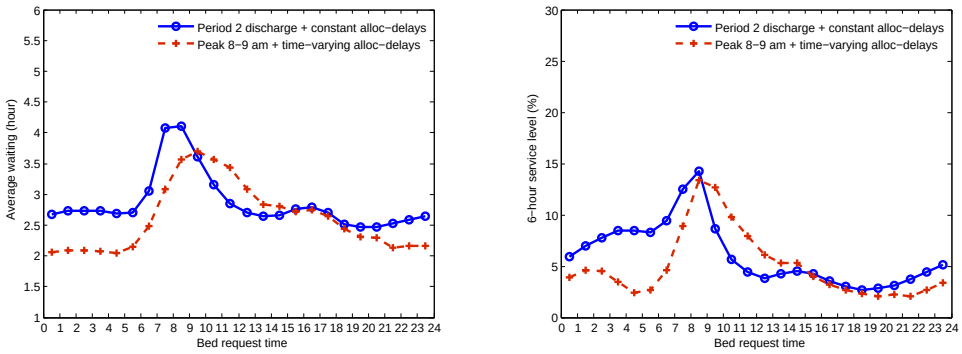
Figure 80: Three groups of hypothetical discharge distributions

Figure 81 plots the hourly waiting time statistics under these two scenarios. We see that in both scenarios, the average waiting time curve is not approximately flattened, i.e., the average waiting time for patients requesting beds between 7am and 11am is still about 1-2 hours longer than the daily average. The hourly 6-hour service level, though, appears to be more time-stable than the average waiting time for each scenario, especially considering the peak value is 30% in the baseline scenario.

Simulation experiments with other early discharge distributions that we have tested also confirm the need for simultaneous improvement in allocation delays and discharge timing to achieve time-stable waiting time performance.

Impact of the discharge timing

Figures 82 to 85 show the hourly waiting time statistics under different early discharge distributions. In each scenario, the combination of an early discharge distribution and the constant-mean allocation delay model is used; all other settings remain the same



(a) Average waiting time

(b) 6-hour service level

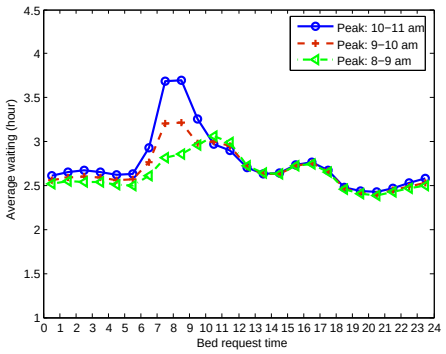
Figure 81: **Hourly waiting time statistics under two scenarios.** Scenario 1: Period 2 discharge distribution and constant mean allocation delays; Scenario 2: Period 3 discharge distribution and time-varying mean allocation delays.

as in the baseline scenario. We observe the following from the figures.

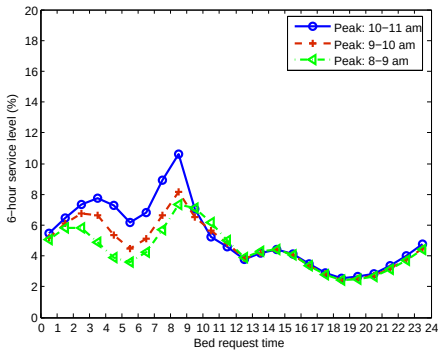
First, in the National University Hospital (NUH) setting, the combination of early discharge and stabilized allocation delays can flatten the hourly waiting time performance, but has limited impact on the daily average waiting time and overflow proportions. This is true even if every patient can be discharged as early as midnight as shown in Figure 85. Indeed, in this case the daily average waiting time can only be reduced by 24 minutes from the baseline scenario, and the overflow proportion shows a less than 3% absolute reduction from the baseline scenario.

Second, the proportion of patients discharged before noon affects the waiting time performance. Generally speaking, the waiting time is shorter if more patients are discharged before noon. Moreover, we find that the timing of the first peak is important in flattening the waiting time performance. For example, if the hospital retains the first discharge peak time to occur between 11am and noon as in the Period 2 policy, even pushing 75% of the patients to be discharged before noon and stabilizing the allocation delays cannot flatten the waiting time performance.

Third, we observe that the waiting time performance under the 9-10am discharge peak scenario in group (a) is close to the performance under the 10-11am discharge

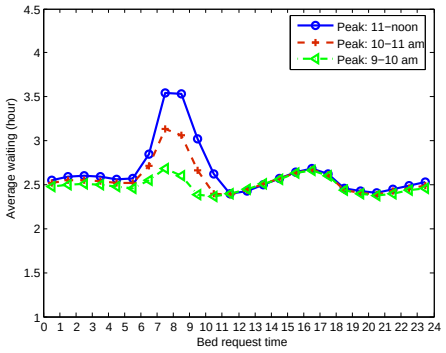


(a) Average waiting time

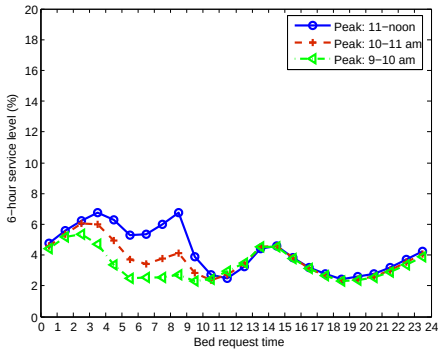


(b) 6-hour service level

Figure 82: Hourly waiting time statistics under scenarios with hypothetical discharge distributions of group (a): 26% of patients discharged before noon. A constant-mean allocation delay model is used.



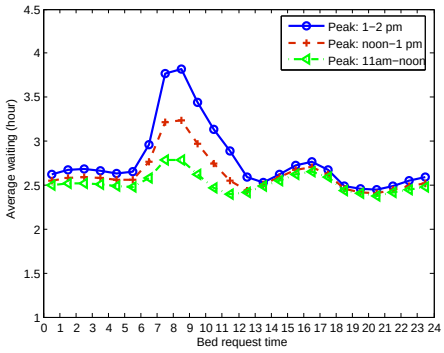
(a) Average waiting time



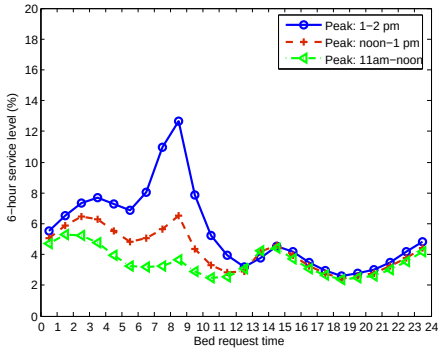
(b) 6-hour service level

Figure 83: Hourly waiting time statistics under scenarios with hypothetical discharge distributions of group (b): 75% of patients discharged before noon. A constant-mean allocation delay model is used.

peak scenario in group (b). Recall that the distributions in group (a) are based on what NUH has achieved in practice since 2010, but shift the first discharge peak to earlier time of the day. This observation indicates that if pushing 75% of the patients to be discharged before noon is too difficult, NUH (and other hospitals alike) can achieve similar waiting time performance by discharging the 26% of patients who are able to leave in the morning as early as possible.

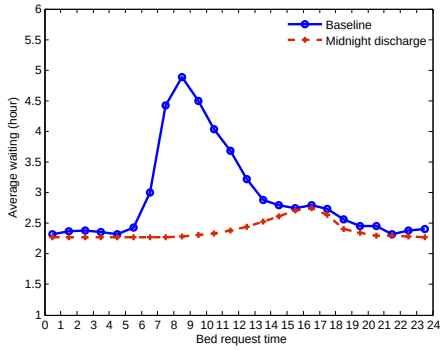


(a) Average waiting time

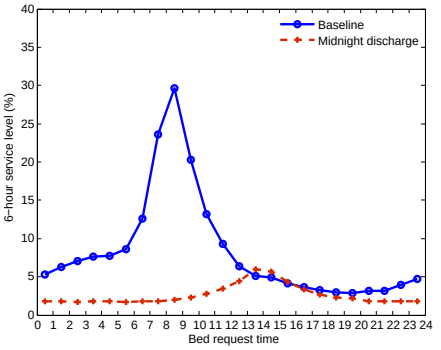


(b) 6-hour service level

Figure 84: Hourly waiting time statistics under scenarios with hypothetical discharge distributions of group (c): shift the entire Period 1 discharge distribution. A constant-mean allocation delay model is used.



(a) Average waiting time



(b) 6-hour service level

Figure 85: Hourly waiting time statistics under the midnight discharge scenario. Constant-mean allocation delay model is used.

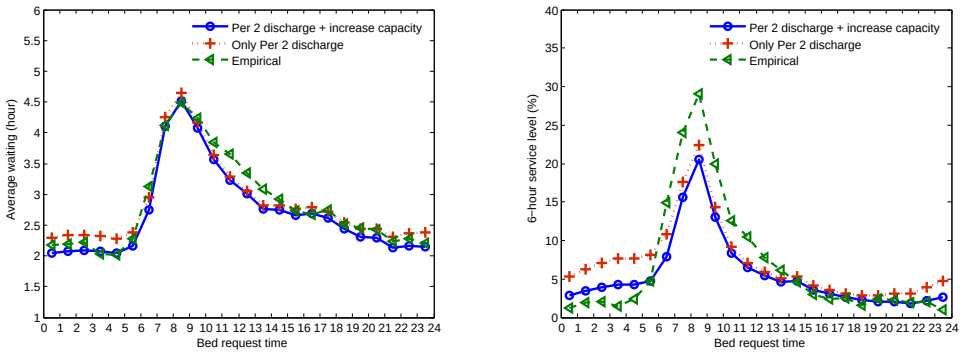
B.5.3 Comparing with Period 2 empirical statistics

In the introduction section of the main paper, we have discussed the changing operating environment from Period 1 to Period 2 at NUH. In Period 2, not only was the early discharge policy implemented, many other factors were also changed from Period 1. These factors include the arrival rates, the average length of stay (LOS), and the bed capacity. As a result, the bed occupancy rate (BOR) showed a 2.7% absolute reduction from Period 1 to Period 2, and the daily utilization showed a 1.7% absolute reduction. (Note that BOR and daily utilization are two slightly different concepts and are calculated in different ways; see Section 3.3 in the Companion paper [136].)

To compare with the empirical performance in Period 2, we simulate a scenario in which (i) the Period 2 discharge distribution is used, and (ii) the bed capacity is increased from the baseline scenario, producing a similar reduction in the BOR and daily utilization as we observed empirically in Period 2. Other settings remain the same as in the baseline scenario. Note that this new scenario is different from the Period 2 policy scenario we introduced in Section 6.1 of the main paper, since the Period 2 policy does not include an increase in bed capacity.

Figure 86 shows the simulation estimates of hourly waiting time statistics from the new scenario (Period 2 discharge + increasing capacity) and the empirical waiting time statistics in Period 2. For reference, we also plot the simulation estimates from the Period 2 policy scenario. From the figure, we can see that the hourly waiting time curves from the new scenario and the Period 2 policy scenario are close to the Period 2 empirical waiting time curves. In particular, the curves from the new scenario can better reproduce the empirical curves between 9pm and 6am (next day) than those from the Period 2 policy scenario.

Moreover, from Figure 1 in the main paper, we can see that the empirical hourly waiting time statistics, especially the 6-hour service level, show a reduction between 9pm and 6am in Period 2. This reduction does not appear in the simulated waiting



(a) Average waiting time

(b) 6-hour service level

Figure 86: Simulation output compares with empirical estimates: hourly average waiting time and 6-hour service level. The empirical estimates are from using Period 2 data.

time statistics when we change from the baseline scenario to the Period 2 policy scenario (see Figure 16 in the main paper). However, if we compare the new scenario to the baseline scenario, we observe a similar reduction in the simulated waiting time statistics between 9pm and 6am. The reason is that the new scenario includes a capacity increase, which leads to a reduction in the waiting time for patients arriving in midnight and early morning. This is also why the new scenario can better reproduce the empirical performance in Period 2, since the actual utilization in Period 2 was indeed reduced. Readers are also referred to Section 6.5 of the main paper for our discussion on how capacity increases impact waiting time statistics.

Through the observations in this section, we again see the capability of our proposed model in capturing the time-varying hourly waiting time performance and predicting the impact of various factors on the waiting time performance.

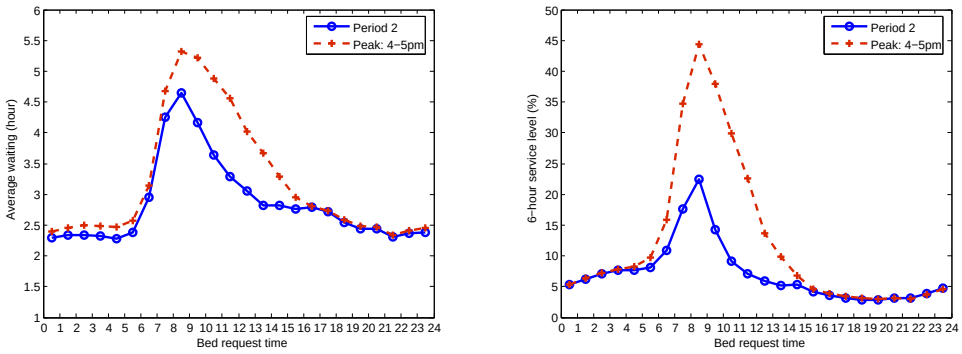
B.5.4 Period 2 policy could show more significant impact in other settings

In Section 7 of the main paper, we have mentioned two issues that readers should be aware of when interpreting our findings in Section 6. In particular, we want to point out here that, although the Period 2 early discharge policy shows limited impact on the waiting time statistics when compared to our baseline scenario, it

does not imply this early discharge policy is not beneficial in other hospital settings. Indeed, even in Period 1, NUH manages discharge planning in a more efficient way than many hospitals around the world. If NUH were not discharging patients so efficiently in Period 1 (i.e., if the baseline scenario were different), we would find that implementing a Period 2 policy could bring more significant improvements to waiting time performance. We show an example below.

Armony et al [5] report that the discharge distribution in an Israeli hospital has a peak discharge time between 4pm and 5pm, which is two hours later than the peak discharge time in Period 1 at NUH. We now evaluate the impact of the Period 2 policy in comparison with an Israeli discharge scenario, which uses a discharge distribution similar to the one at this Israeli hospital and keeps all other settings the same as in the baseline. Figure 87 plots the hourly waiting time curves under the Israeli discharge scenario and the Period 2 policy scenario. We observe a significant improvement of waiting time statistics after implementing the Period 2 early discharge policy, even though the waiting time curves are not flattened. The daily 6-hour service level reduces from 9.26% in the Israeli discharge scenario to 5.50% in the Period 2 policy scenario (with the hourly peak value reducing from 44% to 23%). The daily average waiting time also reduces from 3.08 to 2.73 hours.

The above example indicates that implementing the Period 2 early discharge policy can be very helpful to improve waiting time performance in certain settings, especially if the hospital's current discharge timing is late. Thus, other hospitals can learn from NUH's experience in implementing the Period 2 discharge policy. The Companion paper [136] documents the details on the implementation of the Period 2 discharge policy.



(a) Average waiting time

(b) 6-hour service level

Figure 87: Hourly waiting time statistics under the scenarios with the Period 2 discharge distribution and a hypothetical discharge distribution with the peak time at 4-5pm.

B.6 Sensitivity analysis of different modeling settings

In the main chapter, we evaluate the impact of five operations policies on waiting time performance and overflow proportions. These five policies are a Period 2 policy; a Period 3 policy; increasing bed capacity by 10%; reducing LOS by controlling the maximum stay being 14 days; and reducing the mean pre- and post-allocation delays by 30 minutes each.

To examine the robustness of the insights we have gained in Section 6, we test these five policies under different model settings for sensitivity analysis. These settings include using alternative arrival models (Section B.6.1), changing the priority among ICU-GW, SDA and ED-GW patients (Section B.6.2), using different distributions for the allocation delays (Section B.6.3), and choosing different values for the normal allocation probability $p(t)$ (Section B.6.4).

B.6.1 Sensitivity analysis of the arrival models

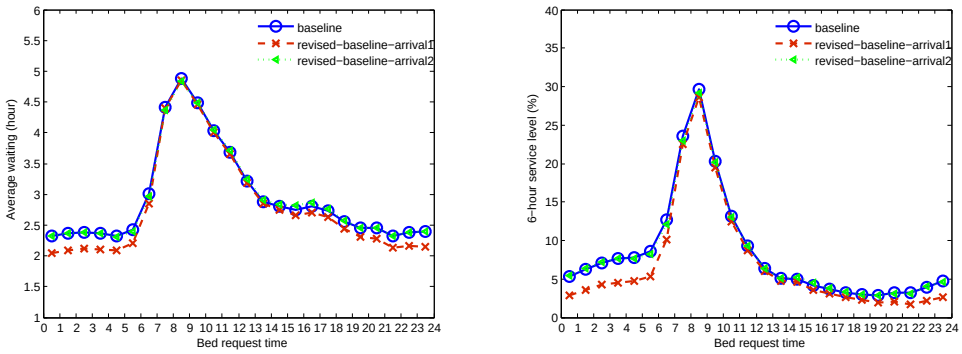
Recall that in the baseline scenario we use a time-nonhomogeneous Poisson process to model the arrivals of ED-GW patients, and non-Poisson processes to model the arrivals of other patients. (See description of the baseline setting in Sections 4.1 of

the main paper [135].) Here, we perform sensitivity analysis on the choice of the arrival process models to study its impact on the hourly waiting time performance of ED-GW patients.

We test two alternative settings for the arrival processes. In the first setting, we assume the arrival processes from the four admission sources are all time-nonhomogeneous Poisson with periods of one day. The arrival rates are plotted in Figure 7 of the main paper. In the second setting, we test a modified arrival process model for ICU-GW and SDA patients based on the one proposed in Section 4.1.2 of the main paper (the arrival processes for ED-GW and EL patients remain the same as in the baseline scenario). For the modified arrival process, after we generate the A_k^j arrivals to arrive on day k from source j, we randomly assign the first arrival to a specific time of the day according to the empirical distribution of the first bed-request time. Then we assign the arrival times of the remaining $A_k^j - 1$ arrivals sequentially, 10 minutes later than the previous one. This modified arrival model is to capture a *batching* phenomenon we have observed from the bed-request times of ICU-GW and SDA patients, i.e., the inter-bed-request time is only about 10-20 minutes for most bed-requests on the same day. See additional empirical analysis in Section 6 of the Companion paper [136].

We call the scenario using the first alternative arrival setting (all non-homogeneous Poisson) the *revised-baseline-arrival1* scenario. Similarly, we call the scenario using the second alternative arrival setting (batch model for ICU-GW and SDA patients) the *revised-baseline-arrival2* scenario. Figure 88 compares the waiting time performance under the baseline scenario, the revised-baseline-arrival1 scenario, and the revised-baseline-arrival2 scenario. From the figure we can see that the waiting time performance is not sensitive to the choice of arrival models, and in particular the performance under the revised-baseline-arrival2 scenario is almost identical to that in the baseline scenario.

Next, we evaluate the five policies in comparison to the corresponding revised

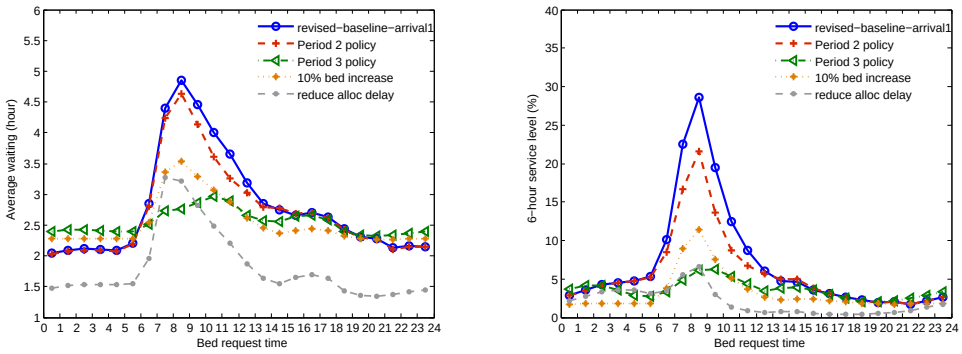


(a) Average waiting time (b) 6-hour service level

Figure 88: **Hourly waiting time statistics under the baseline scenario and scenarios with different choices of arrival models.** All simulation settings are kept the same in each scenario except the arrival models. In the *revised-baseline-arrival1* scenario, all four arrival processes are non-homogeneous Poisson. In the *revised-baseline-arrival2* scenario, a new batch arrival model is used for ICU-GW and SDA patients.

baseline scenario. For example, to evaluate the impact of the Period 2 policy, we compare the scenario using the Period 2 discharge distribution and the first alternative arrival setting with the revised-Baseline-arrival1 scenario. All other settings not specified here remain the same as in the baseline. Figures 89 and 90 plot the hourly waiting time performance for these scenarios. Note that the performance curves under the reduced LOS scenario are almost identical to those under the increased bed capacity scenario, and we do not plot them in the figures. In each figure, the choice of the arrival model is fixed.

From these figures, we can reach the following conclusions. First, the early discharge policy, implemented at the level that NUH achieved in Period 2, has limited impact on reducing or flattening the waiting time statistics for ED-GW patients. Second, the hypothetical Period 3 policy can stabilize the hourly waiting time curves but has limited impact on the daily waiting time statistics. Third, increasing capacity, reducing LOS, or reducing mean allocation delays can reduce the daily waiting time statistics and overflow proportions, but these policies alone do not necessarily stabilize the hourly waiting time performance. In other words, the insights we gained in



(a) Average waiting time

(b) 6-hour service level

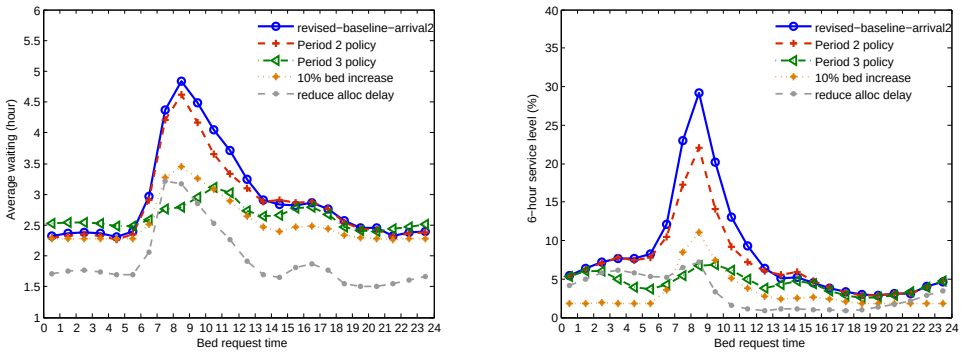
Figure 89: Hourly waiting time statistics under the *revised-baseline-arrival1* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the arrival models are the same, i.e., we assume a non-homogeneous Poisson process for each admission source. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

Section 6 of [135] are not sensitive to the choice of arrival models we have tested.

B.6.2 Sensitivity analysis of the patient priority

In the baseline simulation setting, EL patients have the highest priority, ED-GW patients the second, and ICU-GW and SDA patients have the lowest priority. We experiment with two alternative settings for patient priority. The first setting assigns ICU-GW and SDA patients a higher priority than ED-GW patients while keeping the highest priority for EL patients. The second setting assigns the highest priority to ICU-GW and SDA patients, followed by EL patients, and ED-GW patients have the lowest priority. We call the scenario using the first alternative priority setting the *revised-baseline-priority1* scenario. Similarly, we call the scenario using the second alternative priority setting the *revised-baseline-priority2* scenario.

Figure 91 compares the waiting time performance for ED-GW patients under the baseline scenario, the revised-baseline-priority1 scenario, and the revised-baseline-priority2 scenario. From the figure, we can see that the hourly waiting time curves under the two scenarios with alternative priority settings are almost identical, and



(a) Average waiting time

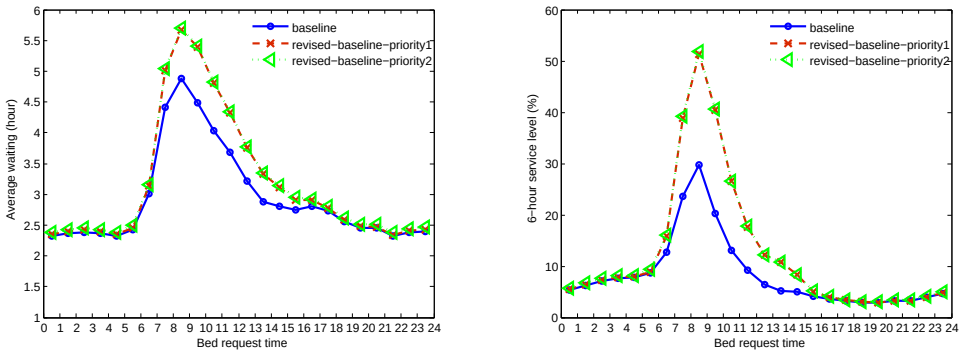
(b) 6-hour service level

Figure 90: Hourly waiting time statistics under the *revised-baseline-arrival2* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the arrival models are the same, i.e., we assume a batch arrival model for ICU-GW and SDA patients. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

they are higher than the corresponding curves from the baseline scenario. This is expected since ED-GW patients have the lowest priority in the two alternative settings, and they have to wait longer than in the baseline scenario.

We evaluate the five policies in comparison to the corresponding revised baseline scenario. Figures 92 and 93 plot the hourly waiting time performance for these scenarios. Similar to the previous section, we do not plot the performance curves under the reduced LOS scenario since they are almost identical to those under the increased bed capacity scenario. In each figure, the priority setting is fixed.

From these figures, we can see that the insights gained in Section 6 of the main paper [135] are not sensitive to the patient priority settings that we have tested. Also note that under Period 3 policy, the hourly waiting time curves in Figures 92 and 93 are not as flattened as in the baseline, though the flattening effect is still significant. This is because ICU-GW and SDA patients, who request beds mostly in the morning, now have higher priority than ED-GW patients in the revised-baseline-priority1 and revised-baseline-priority2 scenarios. As a result, the morning congestion for ED-GW patients is more severe than in the baseline. To eliminate the excessively



(a) Average waiting time

(b) 6-hour service level

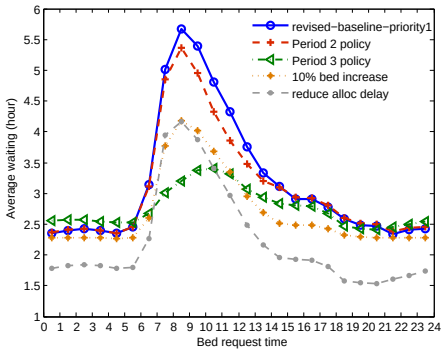
Figure 91: Hourly waiting time statistics under the baseline scenario and scenarios with different patient priority settings. All simulation settings are kept the same in each scenario except patient's priority. In the *revised-baseline-priority1* scenario, EL ¿ ICU-GW = SDA ¿ ED-GW. In the *revised-baseline-priority2* scenario, ICU-GW = SDA ¿ EL ¿ ED-GW.

long waiting times for morning ED-GW bed-requests, an early discharge policy even more aggressive than Period 3 policy needs to be implemented.

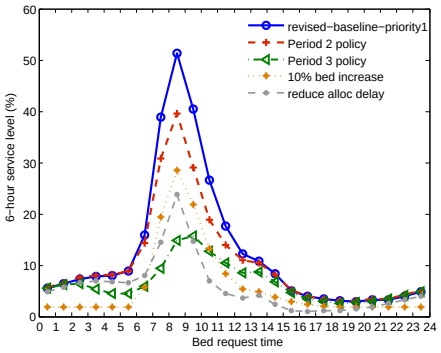
B.6.3 Sensitivity analysis of the allocation delay distributions

In the baseline setting, the pre- and post-allocation delays follow log-normal distributions with time-dependent means and coefficients of variation (CVs). For sensitivity analysis, we test two other distributions for the pre- and post-allocation delays: exponential and normal distributions. We assume the means (and the CVs for normal distributions) are still time-dependent, following the dashed lines with plus sign in Figure 10 of the main paper [135]. We call the scenario using the exponential allocation delay assumption the *revised-baseline-exponential* scenario. Similarly, we call the scenario using the normal allocation delay assumption the *revised-baseline-normal* scenario.

Figure 94 compares the waiting time performance for ED-GW patients under the baseline scenario, the revised-baseline-exponential scenario, and the revised-baseline-normal scenario. From the figures we can see that the performance measures are not very sensitive to the allocation delay distributions. In fact, the hourly average

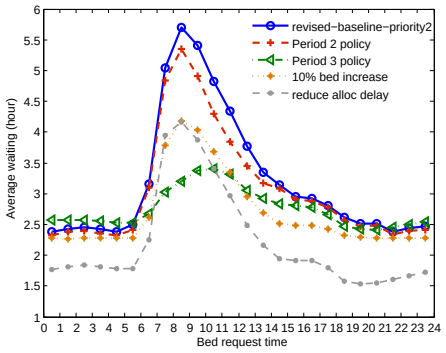


(a) Average waiting time

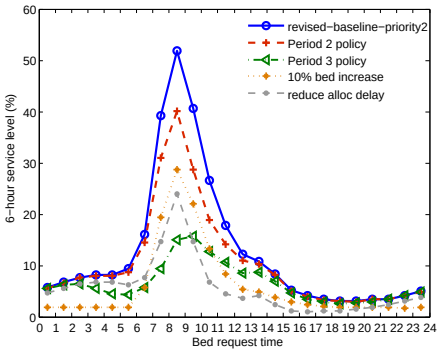


(b) 6-hour service level

Figure 92: Hourly waiting time statistics under the *revised-baseline-priority1* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the patient priority settings are the same (EL ¿ ICU-GW = SDA ¿ ED-GW). For Policy (ii) to (iv), the constant-mean allocation delay model is used.

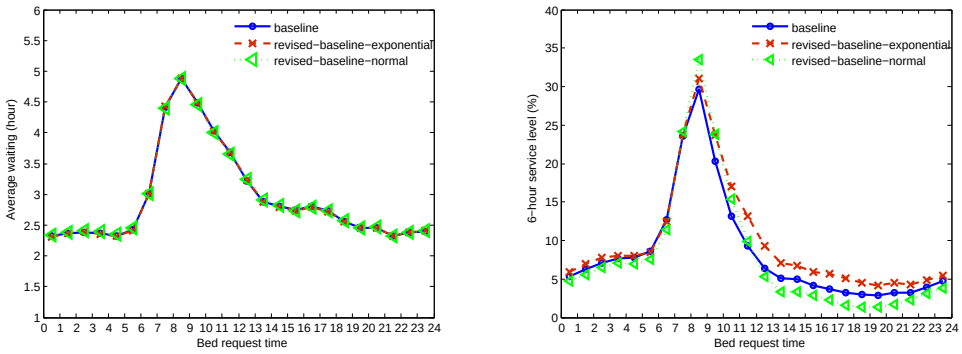


(a) Average waiting time



(b) 6-hour service level

Figure 93: Hourly waiting time statistics under the *revised-baseline-priority2* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the patient priority settings are the same (ICU-GW = SDA ¿ EL ¿ ED-GW). For Policy (ii) to (iv), the constant-mean allocation delay model is used.



(a) Average waiting time

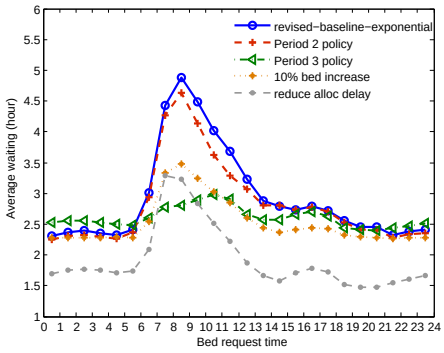
(b) 6-hour service level

Figure 94: Hourly waiting time statistics under the baseline scenario and scenarios with different allocation delay distributions. All simulation settings are kept the same in each scenario except the distributions of allocation delays. In the *revised-baseline-exponential* scenario, exponential distributions are used for the two allocation delays. In the *revised-baseline-normal* scenario, normal distributions are used.

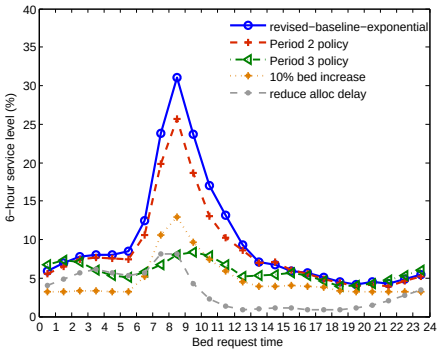
waiting time curves under the three scenarios are almost identical. This is because the average waiting time is affected by the mean allocation delays, while these mean values remain the same in all three scenarios. The differences in the allocation delay distributions are reflected through the 6-hour service level, which captures the tail distribution of the waiting times. Recall that the CV of an exponential distribution is 1, which is higher than the empirical CVs observed in Figure 10 of the main paper. Figure 94b is consistent with the common belief that higher variability contributes to longer waiting times.

We evaluate the five policies in comparison to the corresponding revised baseline scenario. Figures 95 and 96 plot the hourly waiting time performance for these scenarios. We do not plot the performance curves under the reduced LOS scenario since they are almost identical to those under the increased bed capacity scenario. In each figure, the allocation delay distributions are fixed.

Not surprisingly, the insights gained in Section 6 of the main paper are robust with respect to the tested allocation delay distributions, since the waiting time performance is not sensitive to the tested distributions.

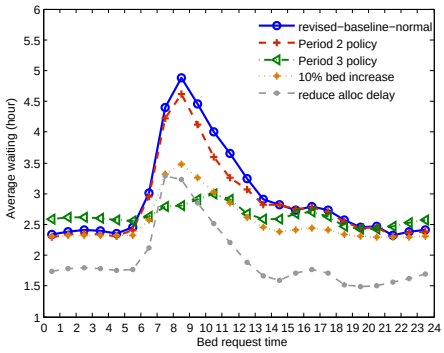


(a) Average waiting time

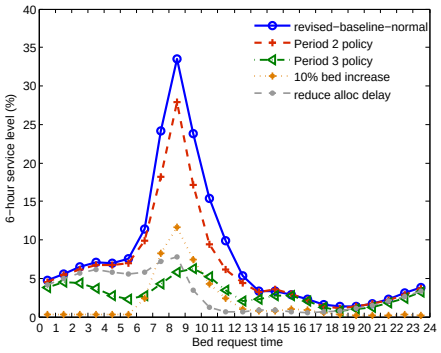


(b) 6-hour service level

Figure 95: Hourly waiting time statistics under the *revised-baseline-exponential* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the allocation delays follow exponential distributions. For Policy (ii) to (iv), the constant-mean allocation delay model is used.



(a) Average waiting time



(b) 6-hour service level

Figure 96: Hourly waiting time statistics under the *revised-baseline-normal* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the allocation delays follow normal distributions. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

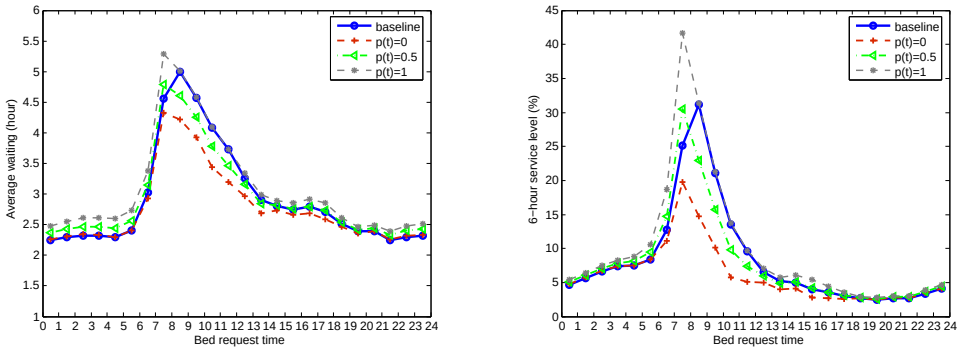
B.6.4 Sensitivity analysis of the normal allocation probability $p(t)$

In the baseline scenario, the normal allocation probability, $p(t)$ follows a step function with respect to t (see (2) in Section 4.6.2 of [135]). In this section, we perform sensitivity analysis on the value of $p(t)$ to study its impact on the hourly waiting time performance. We adopt three constant functions and assume $p(t) = 0,\ 0.5,$ or 1 for all t. Here, $p(t) = 0$ and $p(t) = 1$ serve as the lower bound and upper bound for all possible choices of $p(t)$, respectively, while $p(t) = 0.5$ is in between.

Figure 97 plots the hourly waiting time statistics under the baseline scenario and three new scenarios, which have the same settings as the baseline except for the values of $p(t)$. We call the three new scenarios the *revised-baseline-$p(t)$-j* scenario for $p(t) = j$ $(j = 0,\ 0.5,\ 1)$.

From Figure 97 we can see that the waiting time is longer when the value of $p(t)$ is larger, i.e., when normal-allocation mode is more frequently used than forward-allocation mode. This is because in the normal-allocation mode, the pre-allocation delay starts only after a bed becomes available, which is later than or the same as the bed-request time; in contrast, the pre-allocation delay always starts at the bed-request time in the forward-allocation mode. As a result, the entire waiting time for a patient in the normal-allocation mode is longer than or equal to that in the forward-allocation mode on a given sample path. Moreover, note that the value of $p(t)$ seems to have a local effect on the hourly waiting time performance. The waiting time curve for the baseline scenario coincides with one of the other three waiting time curves during certain intervals when the values of $p(t)$ are the same. For example, the baseline curve overlaps with the curve from the $p(t) = 0$ scenario between 0 and 6am since we set $p(t) = 0$ during that interval in the baseline scenario.

We evaluate the five policies in comparison to the corresponding revised baseline scenario. Figures 98 through 100 plot the hourly waiting time performance for these scenarios. We do not plot the performance curves under the reduced LOS scenario



(a) Average waiting time

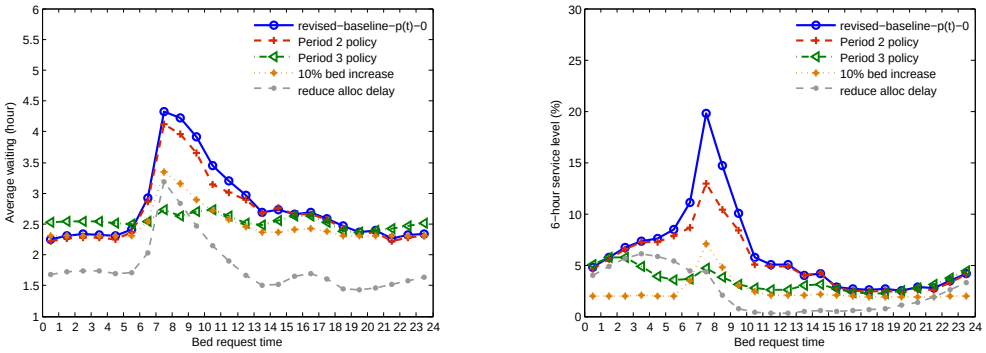
(b) 6-hour service level

Figure 97: Hourly waiting time statistics under the baseline scenario and scenarios with different choices of $p(t)$. All simulation settings are kept the same in each scenario except the values of $p(t)$.

since they are almost identical to those under the increased bed capacity scenario. In each figure, the choice of $p(t)$ is fixed. Again, we can see that the insights gained in Section 6 of the main paper are not sensitive to the tested values of $p(t)$.

B.7 Sensitivity analysis of system load

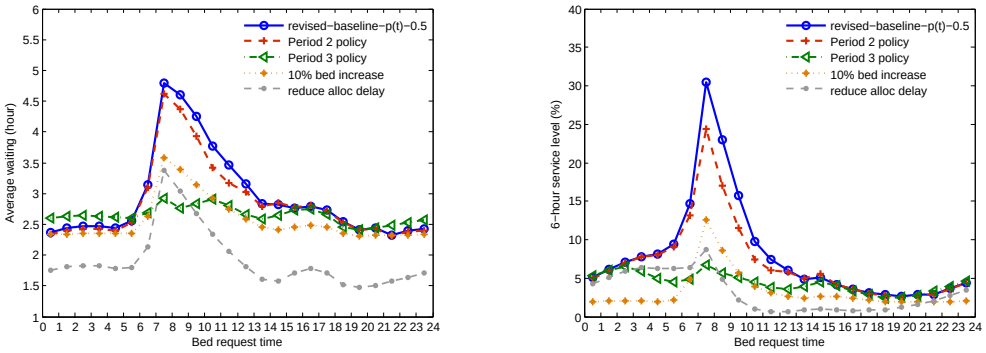
In Section 6.3 of the main paper, we find that the modeled hospital queueing system is not heavily loaded in the NUH setting. The 2-3 hours average waiting time at NUH mainly comes from secondary bottlenecks such as nurse shortages rather than bed unavailability. In this section, to examine the robustness of our gained insights in a more heavily utilized setting, we increase the system load. In Section B.7.1, we increase the daily arrival rate of ED-GW patients and evaluate the five operational policies that are tested in Section B.6. Under the increased arrival rate setting, we find that the Period 3 policy can have a great impact on the daily waiting time statistics because of its side effect in reducing LOS, while this side effect is caused by the different LOS distributions between patients admitted before noon (AM) and after noon (PM). Thus, to separate the impact of discharge timing from the impact of reducing LOS, we eliminate the difference between the LOS distributions and re-evaluate the five policies under a similar heavily-loaded environment in Section B.7.2. Finally, in



(a) Average waiting time

(b) 6-hour service level

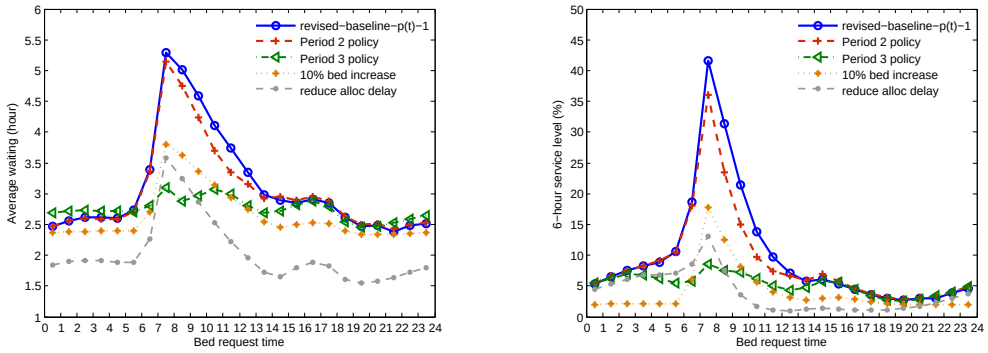
Figure 98: Hourly waiting time statistics under the *revised-baseline-p(t)*-0 scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, $p(t) = \mathbf{0}$ for all t. For Policy (ii) to (iv), the constant-mean allocation delay model is used.



(a) Average waiting time

(b) 6-hour service level

Figure 99: Hourly waiting time statistics under the *revised-baseline-p(t)*-0.5 scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, $p(t) = \mathbf{0.5}$ for all t. For Policy (ii) to (iv), the constant-mean allocation delay model is used.



(a) Average waiting time

(b) 6-hour service level

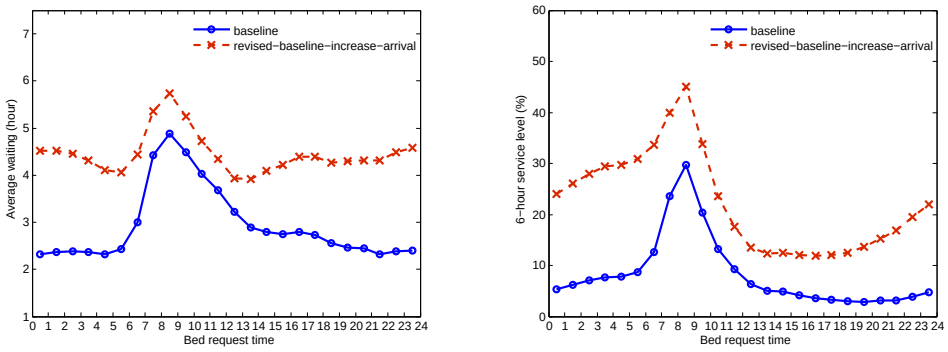
Figure 100: Hourly waiting time statistics under the *revised-baseline-p(t)*-1 scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, $p(t) = \mathbf{1}$ for all t. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

Section B.7.3 we summarize several conditions under which the early discharge policy can significantly impact the daily waiting time performance.

B.7.1 Impact of the five policies under the increased arrival rate setting

We increase the daily arrival rate of ED-GW patients by 7% from the baseline setting, similar to the increase from Period 1 to Period 2 we empirically observed. When all other settings remain the same as in the baseline, simulation shows the utilization under the increased arrival scenario becomes 93%, and the daily average waiting time and 6-hour service level become 4.37 hours and 18.60%, respectively. In other words, we create a more capacity-constrained scenario than the baseline scenario, and we call this new scenario the *revised-baseline-increase-arrival* scenario. Figure 101 compares the hourly waiting time curves between the new scenario and the baseline scenario. The curves from the new scenario have similar shapes as the curves from the baseline scenario, but are higher than the latter because of the increased system load.

We evaluate the impact of the five policies under the increased arrival rate setting and compare them with the revised-baseline-increase-arrival scenario. Figures 102



(a) Average waiting time

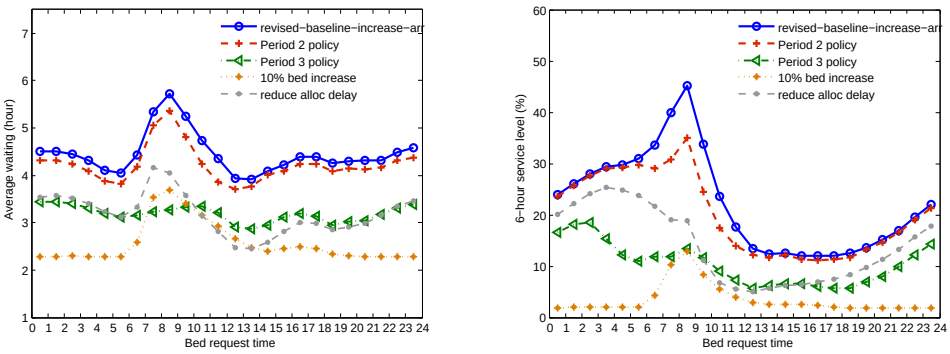
(b) 6-hour service level

Figure 101: Hourly waiting time statistics under the baseline scenario and the scenario with increased arrival rate (*revised-baseline-increase-arrival scenario*).

plots the hourly waiting time performance for these scenarios. Note that the performance curves under the reduced LOS scenario are almost identical to those under the increased bed capacity scenario, and we do not plot them in the figures.

From these figures, we can see that most conclusions we get in Section 6 of the main paper [135] still hold. First, the Period 2 early discharge policy has limited impact on reducing or flattening the waiting time statistics for ED-GW patients. Second, increasing capacity, reducing LOS, or reducing mean allocation delays can reduce the daily waiting time statistics and overflow proportions, but these policies alone cannot stabilize the hourly waiting time performance. In particular, comparing to the revised-baseline-increase-arrival scenario, increasing 10% bed capacity here reduces the daily average waiting time from 4.37 hours to 2.49 hours and the 6-hour service level from 18.60% to 2.82%, a much more significant impact on reducing the daily waiting time statistics than what we observed under the original NUH setting. This is expected because increasing capacity can greatly reduce system congestion and patient waiting time in a capacity-constrained setting, but has smaller impact if the system is not heavily loaded.

An exception is that the hypothetical Period 3 policy now not only stabilizes the hourly waiting time, but also has significant impact on the daily waiting time



(a) Average waiting time

(b) 6-hour service level

Figure 102: Hourly waiting time statistics under the *revised-baseline-increase-arrival* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the daily arrival rate for ED-GW patients is increased by 7% from the baseline setting. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

statistics. The daily average waiting time is reduced from 4.37 hours in the revised-baseline-increase-arrival scenario to 3.16 hours in the Period 3 policy scenario, and the 6-hour service level is reduced from 18.60% to 9.41%. The large reduction in the daily waiting times is mainly because of our assumption that the AM-admitted and PM-admitted ED-GW patients have different LOS distributions. This assumption is supported by our empirical study at NUH; see Section 4.3 of the main paper which shows that the mean LOS of AM-admitted ED-GW patients is about 1 day less than the mean LOS of PM-admitted patients across all specialties. After the Period 3 early discharge, more morning arrivals can be admitted before noon instead of waiting till the afternoon, and they become AM-admitted patients. As a result, the LOS is reduced and eventually the system utilization is reduced. We further verify this argument in the next section.

B.7.2 Impact of the five policies without the AM/PM difference in LOS

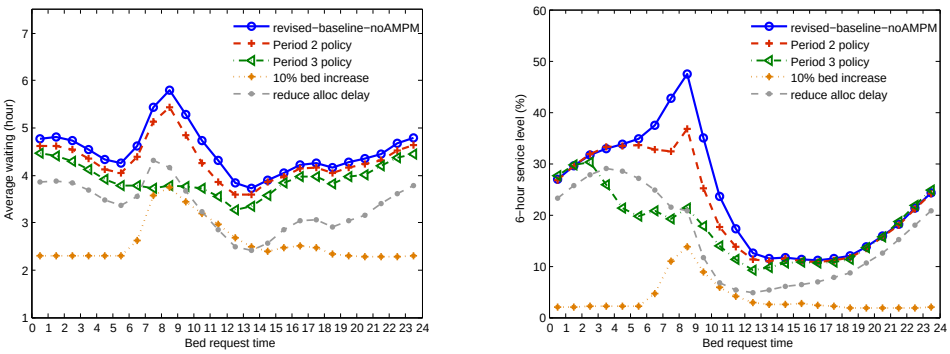
In this section, we assume that the AM-admitted ED-GW patients have the same LOS distributions as PM-admitted ED-GW patients for each specialty. We do so to eliminate the side effect of reducing LOS and to gain insights into the impact of

discharge timing when we evaluate early discharge policies.

Because the AM-admitted patients now have longer average LOS, we adjust the number of servers to create a capacity-constrained setting that has a similar system load as the revised-baseline-increase-arrival scenario introduced in the previous section. All other settings remain the same as in the baseline scenario. We call this scenario, without the difference in LOS distributions between AM- and PM-admitted patients (or AM/PM difference for short), the *revised-baseline-noAMPM* scenario. Simulation shows the system utilization under this new scenario is 94%. The daily average waiting time and 6-hour service level become 4.38 hours and 19.34%, respectively, which are similar to the values in the revised-baseline-increase-arrival scenario. The hourly waiting time curves under this new scenario are also close to those under the revised-baseline-increase-arrival scenario; see the solid lines in Figure 103.

We re-evaluate the impact of the five policies without the AM/PM difference in LOS. Figure 103 plots the hourly waiting time curves under these policies. Note that the performance curves under the reduced LOS scenario (i.e., control maximum LOS to be 14 days) are almost identical to those under the increased bed capacity scenario, so we do not plot them in the figure.

Comparing Figure 103 with Figure 102, we can see that Period 2 policy, increasing capacity, and reducing mean allocation delays show similar impact on the waiting time statistics no matter whether we consider the AM/PM difference in LOS or not. However, Period 3 policy shows a very different impact after we eliminate the AM/PM difference: it approximately flattens the hourly waiting time curves, but has limited impact on reducing the daily waiting time statistics. The daily average waiting time is reduced from 4.38 hours in the revised-baseline-noAMPM scenario to 3.92 hours in the Period 3 policy scenario, and the 6-hour service level is only reduced from 19.34% to 16.18%. This observation is consistent with what we get in Section 6.2 of the main paper.

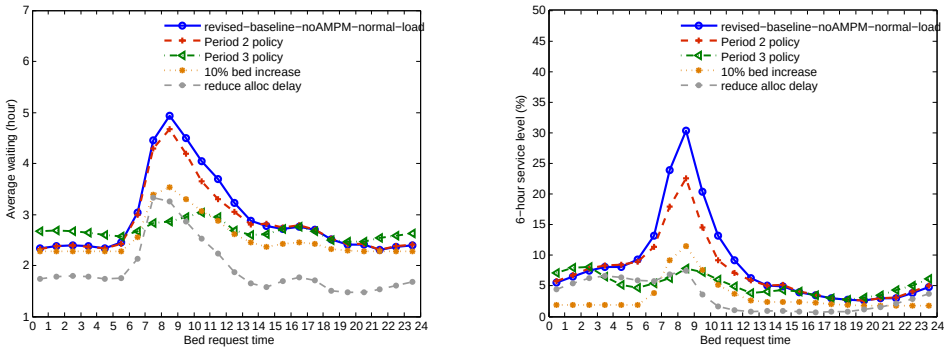


(a) Average waiting time

(b) 6-hour service level

Figure 103: Hourly waiting time statistics under the *revised-baseline-noAMPM* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the AM-admitted patients have the same LOS distributions as PM-admitted patients. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

In addition, we study the impact of the AM/PM difference in LOS when the system is not heavily loaded. We develop a *revised-baseline-noAMPM-normal-load* scenario by (i) assuming the AM-admitted patients have the same LOS distributions as PM-admitted patients and (ii) adjusting the number of servers to reach a similar system load as in the baseline scenario. Under this scenario, the daily average waiting time and 6-hour service level from simulation estimates are 2.80 hours and 6.27%, respectively, close to the values in the baseline scenario. We re-evaluate the impact of the five policies under this lower system load. Figures 104 plots the hourly waiting time performance for these scenarios. Comparing the performance curves in Figures 104 to those in Figures 16 to 18 of the main paper, we can see that the five policies show similar impact with or without considering the AM/PM difference. In particular, the side effect of reducing LOS brought by the early discharge policy dost not show much impact on the waiting time when the system is not heavily loaded.



(a) Average waiting time

(b) 6-hour service level

Figure 104: Hourly waiting time statistics under the *revised-baseline-noAMPM-normal-load* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the AM-admitted patients have the same LOS distributions as PM-admitted patients. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

B.7.3 Conditions for an early discharge policy to significantly impact the daily waiting time performance

Based on our simulation findings in Sections B.7.1 and B.7.2, we summarize here a few conditions under which an early discharge policy can show a significant impact on the daily waiting time statistics.

First, when the LOS of AM-admitted patients is shorter than the LOS of PM-admitted patients, implementing an early discharge policy can reduce LOS in addition to shifting the discharge timing. Therefore, early discharge can significantly affect the system load and reduce the daily waiting time statistics. However, when the AM- and PM-admitted patients have the same LOS distributions, the early discharge policy no longer affects the LOS but only influences the discharge timing. In this case, shifting the discharge timing can flatten the waiting time curve but has limited impact on reducing the daily waiting time statistics. In Section 6.5 of the main paper, we have provided some intuitive explanation for why reducing LOS and shifting discharge timing have different impacts on the daily and hourly waiting time performance.

Second, given that the early discharge policy shows a side effect of reducing LOS,

the impact of reducing LOS on the waiting time statistics is significant only when the system is heavily loaded. In the NUH setting, the Period 3 policy shows a limited impact on the daily performance even if we use different LOS distributions for AM- and PM-admitted patients (see Section 6.2 of the main paper). The main reason is that the system load is not high enough in the NUH setting.

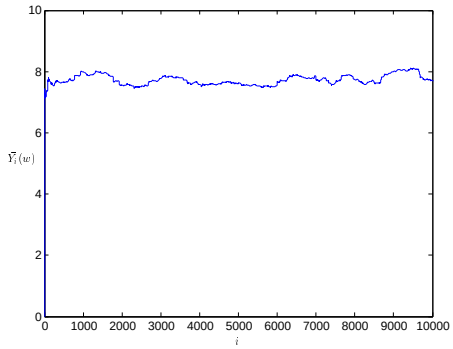
Third, in order for an early discharge policy to show a significant side effect of reducing LOS, the discharge timing needs to be early enough. Unlike the Period 3 policy, the Period 2 early discharge policy cannot reduce the daily waiting time statistics much, no matter whether we differentiate between AM- and PM-admitted patients or not. This is because, under the Period 2 policy, the first discharge peak is between 11am and noon. Even after implementing the Period 2 early discharge, most morning arrivals still have to be admitted after noon due to the allocation delays (which on average takes about 2 hours) and the LOS is not effectively reduced.

Finally, we want to point out the need of future research to identify the factors causing the AM/PM difference in LOS. This line of research can help us better understand whether the 1-day difference in the mean LOS between AM- and PM-admitted patients will still exist when more patients are admitted in the morning than what we observed so far. Eventually, this research can help us generate more comprehensive insights into the benefits of early discharge policies and other operational policies.

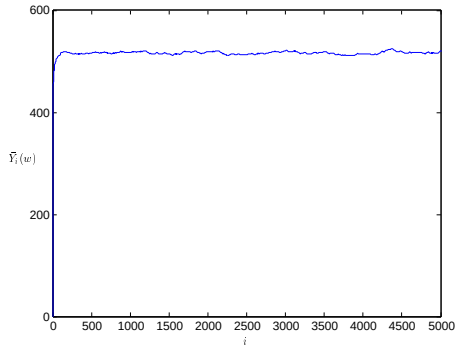
B.8 The warm-up period and the length and number of batches

This section contains supplementary details for the simulation experiment settings. We show that our choices of the warm-up period and the length and number of batches are appropriate for our simulation study.

In each simulation experiment, we simulate for a total of 10^6 days, and divide the simulation output into 10 batches. The performance measures are calculated by averaging the last 9 batches, with the first batch discarded to eliminate transient



(a) Daily average buffer size with $w = 500$



(b) Daily average number of busy servers with $w = 100$

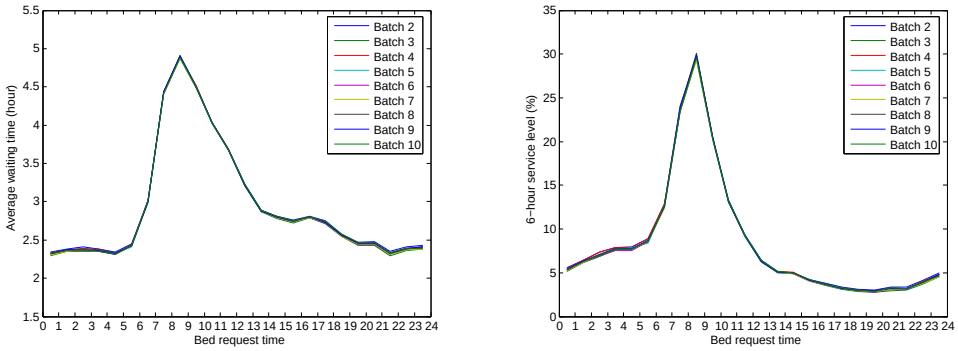
Figure 105: Moving average plots from 5 replications. Each replication contains 10^5 days.

effects. This simulation setting is justified as below.

First, we follow the standard procedures in the literature to observe the moving average plot [91] and determine the warm-up period. We run $n = 5$ replications of the baseline scenario (5-10 replications are recommended choices in the literature), with each replication running for $m = 10^5$ days. We define Y_{ji} be the ith observation from the jth replication, where the ith observation can be a chosen performance measure on day i, e.g., the daily average buffer size or average number of busy servers on day i. Let $\bar{Y}_i = \frac{1}{n}\sum_{j=1}^{n} Y_{ji}$. For a given time-window w, we define the moving average as

$$\bar{Y}_i(w) = \begin{cases} \frac{1}{2w+1}\sum_{s=-w}^{w} \bar{Y}_{i+s}, & \text{if } i = w+1, \ldots, m-w; \\ \frac{1}{2i-1}\sum_{s=-(i-1)}^{i-1} \bar{Y}_{i+s}, & \text{if } i = 1, \ldots, w. \end{cases}$$

The time-window w is chosen through experiments, so that we can both observe the initial transient effects and have a reasonably smooth plot after the system converges to steady state. Figure 105 shows the moving average plots for the daily average buffer size and average number of busy servers. We chose $w = 500$ for the buffer size plot and $w = 100$ for the busy server plot. It is clear that before 10^4 days, the sequence of $\{\bar{Y}_i(w)\}$ appears to have converged, indicating our choice of 10^5 days as the warm-up period is more than enough.



(a) Average waiting time

(b) 6-hour service level

Figure 106: Hourly waiting time statistics from each batch.

Second, Figure 106 compares the hourly waiting time statistics from each of the last nine batches under the baseline scenario (the warm-up batch is excluded), and each batch contains a total of 10^5 days. It is clear from the figure that the waiting time curves from the 9 batches are very close to each other, suggesting (i) the system is running in the steady state; and (ii) the choice of the number of batches and length of each batch is appropriate to produce a tight confidence interval for the reported batch means. Table 30 below reports the the confidence intervals for the batch means of the hourly average waiting time and the hourly 6-hour service level. For comparison, the confidence intervals of the corresponding empirical hourly waiting time statistics are also reported in the table. Table 3 in the main paper reports the confidence intervals for the batch means of the daily and specialty-level average waiting times and 6-hour service levels.

B.9 Additional discussion on the overflow proportion

Besides ED-GW patient's waiting time, the overflow proportion is one of the performance measures of interest to us and is monitored closely at NUH. This section provides more discussion on the overflow proportions. In Section B.9.1, we explain the challenges in reproducing the empirical overflow proportions with our model. In

Table 30: **Simulation and empirical estimates of the waiting time statistics for ED-GW patients requesting beds in each hour of the day.** The simulation estimates are from simulating the baseline scenario, and the empirical estimates are from Period 1 data. The numbers in the parentheses are for the 95% confidence interval of the corresponding value. The confidence intervals for the simulation output are calculated following the batch mean method [91]; the confidence intervals for the empirical statistics are calculated with the standard deviations and sample sizes from the actual data.

	average waiting time (hour)		6-hour service level (%)	
hour	simulation	empirical	simulation	empirical
1	2.32 (2.31, 2.33)	2.41 (2.30, 2.51)	5.33 (5.25, 5.41)	3.62 (2.69, 4.56)
2	2.36 (2.36, 2.37)	2.32 (2.20, 2.43)	6.28 (6.20, 6.35)	3.54 (2.51, 4.56)
3	2.38 (2.36, 2.39)	2.31 (2.18, 2.43)	7.08 (6.97, 7.19)	3.70 (2.52, 4.89)
4	2.36 (2.35, 2.37)	2.17 (2.05, 2.28)	7.68 (7.60, 7.76)	3.71 (2.45, 4.97)
5	2.32 (2.32, 2.33)	2.19 (2.04, 2.34)	7.76 (7.68, 7.83)	4.81 (3.16, 6.47)
6	2.43 (2.42, 2.43)	2.51 (2.32, 2.70)	8.64 (8.53, 8.75)	8.70 (6.46, 10.94)
7	3.01 (3.00, 3.01)	3.24 (3.01, 3.47)	12.64 (12.54, 12.75)	18.97 (15.83, 22.10)
8	4.42 (4.41, 4.43)	4.33 (4.08, 4.57)	23.60 (23.46, 23.74)	33.21 (29.22, 37.20)
9	4.89 (4.88, 4.90)	4.64 (4.41, 4.86)	29.67 (29.52, 29.81)	36.52 (32.59, 40.46)
10	4.49 (4.49, 4.50)	4.74 (4.58, 4.89)	20.32 (20.24, 20.40)	30.84 (27.75, 33.94)
11	4.03 (4.03, 4.04)	4.22 (4.11, 4.33)	13.15 (13.10, 13.21)	16.61 (14.55, 18.68)
12	3.68 (3.68, 3.68)	3.81 (3.72, 3.90)	9.25 (9.21, 9.28)	10.99 (9.48, 12.50)
13	3.22 (3.21, 3.22)	3.37 (3.30, 3.45)	6.33 (6.29, 6.38)	7.50 (6.33, 8.67)
14	2.88 (2.88, 2.89)	3.15 (3.07, 3.22)	5.09 (5.06, 5.12)	7.24 (6.16, 8.33)
15	2.80 (2.79, 2.81)	2.88 (2.81, 2.94)	4.95 (4.91, 4.99)	5.46 (4.51, 6.40)
16	2.74 (2.74, 2.75)	2.73 (2.67, 2.80)	4.11 (4.07, 4.15)	3.87 (3.09, 4.65)
17	2.80 (2.79, 2.80)	2.65 (2.59, 2.71)	3.65 (3.61, 3.68)	2.56 (1.90, 3.22)
18	2.73 (2.72, 2.74)	2.74 (2.68, 2.80)	3.22 (3.18, 3.27)	2.56 (1.90, 3.22)
19	2.56 (2.55, 2.57)	2.48 (2.42, 2.55)	2.98 (2.93, 3.02)	1.84 (1.25, 2.42)
20	2.46 (2.45, 2.47)	2.40 (2.33, 2.48)	2.88 (2.82, 2.93)	2.25 (1.55, 2.94)
21	2.45 (2.44, 2.46)	2.32 (2.25, 2.40)	3.18 (3.11, 3.25)	1.78 (1.17, 2.39)
22	2.32 (2.31, 2.33)	2.30 (2.22, 2.38)	3.18 (3.11, 3.25)	2.59 (1.89, 3.29)
23	2.38 (2.37, 2.39)	2.33 (2.23, 2.43)	3.92 (3.83, 4.00)	3.50 (2.66, 4.34)
24	2.40 (2.39, 2.41)	2.27 (2.18, 2.36)	4.73 (4.64, 4.82)	3.13 (2.33, 3.93)

Section B.9.2, we provide some intuitive explanation on why early discharge policies show limited impact on overflow proportions. We also discuss future research directions on overflow policies.

B.9.1 Challenges in calibrating the overflow proportion

Figure 12b in the main paper [135] compares the simulation estimates of the specialty-level overflow proportions from the baseline scenario with the empirical estimates in Period 1. From the figure, we observe that, for most specialties, their overflow proportions from simulation are close to the empirical estimates. The exceptions are Surgery, General Medicine, and Neurology, whose overflow proportions from simulation are much lower than the empirical values. In fact, this is the main reason why the overflow proportion across all specialties from the baseline simulation (16.35%) is lower than the empirical estimate (26.95%) in Period 1. We point out that perfectly calibrating the overflow proportions is challenging with our current model setting for two reasons.

First, our model does not capture all overflow events happened in reality. There are two kinds of overflow in practice, which are triggered by different factors: *passive overflow* is triggered to avoid excessively long waiting times, while *intentional overflow* is triggered by other reasons such as medical needs. An example for an intentional overflow is that a General Medicine patient with a potential heart problem is overflowed to a Cardiology ward for telemetry care. See similar descriptions on intentional overflow in [144]. Our model captures passive overflow but not intentional overflow, whereas the empirical estimates include both overflow events. (Note that it is difficult to differentiate between passive overflow and intentional overflow from the data we currently have.) As a result, the empirical overflow proportions can be higher than the simulation estimates from our model.

Second, our model assumes the shared server pools have *complete* flexibility, i.e.,

each bed in such a pool can serve a patient from any primary specialty. In practice, however, complete flexibility is impossible in the shared wards, as indicated by the empirical study at NUH [136]. This complete flexibility in our model can reduce the occurrence of overflow events. For example, Neurology and General Medicine specialities share a ward (a server pool in the model). From the baseline simulation, Neurology patients constitute 29% of all primary admissions to the shared server pool. However, this proportion is only 18% for the shared ward (Ward 53) in Period 1 from the empirical data, which suggests that sometimes a Neurology patient may not be able to be admitted to the shared ward even if a bed is available there. In this case, the Neurology patient could be overflowed to other wards in practice, but such an overflow does not occur in the model. As a result, the overflow proportions estimated from our model can be smaller than the empirical estimates. Future research is needed to identify better ways of modeling the shared wards.

B.9.2 Early discharge policies have limited impact on the overflow proportions

Sections 6.1 and 6.2 of the main paper [135] demonstrate that early discharge policies have a limited impact on reducing the overflow proportion, even under the extreme midnight discharge policy. We provide an intuitive explanation here. Consider patients requesting beds in the morning (7am to noon) since early discharge policies mainly affect these patients. In our simulation setting, the overflow trigger time T is long in the morning ($T = 5.0$ hours from 7am to 7pm). Thus, even in the baseline scenario, primary beds are likely to become available for morning arrivals before their waiting times exceed five hours (recall that most discharges start to occur from noon under the baseline Period 1 discharge distribution). In other words, under the given overflow policy, there are already very few morning arrivals overflowed in the baseline scenario. Therefore, although more beds become available in the morning after the early discharge, the overflow proportion will not be affected much.

In our present simulation study, we assume the overflow policy is fixed, and we focus on evaluating the early discharge policy and other policies under the given overflow policy. Clearly, the choice of overflow policy can greatly impact the waiting time performance; see Figure 11 and Figure 15 in the main paper for an example. Future research is needed to identify the impact of different overflow policies on waiting time performance and to evaluate the impact of other operational policies under different overflow policies. Moreover, the overflow policy we currently assume in the model is motivated by what NUH used in practice, and it may not be optimal. For efficient inpatient operations, it will be important to identify optimal or near-optimal overflow policies when taking multiple objectives into consideration, e.g., balancing the trade-off between the overflow proportion and waiting time.

APPENDIX C

APPENDIX FOR CHAPTER 4

C.1 The stationary distribution of discrete OU process

Similar to the continuous-time version of the Ornstein-Uhlenbeck (OU) process, we define its discrete-time version $\{X_k, k = 0, 1, \dots\}$ as:

$$X_k = Y_k - \mu \sum_{i=0}^{k-1} X_i, \quad k = 0, 1, \dots \tag{65}$$

where $\{Y_k := \sum_{i=0}^{k-1} \xi_i, k = 0, 1, \dots\}$ is a Gaussian random walk, i.e., $\{\xi_i\}$ is a sequence of iid random variables following a normal distribution with mean θ and variance σ^2. In our analysis below, we enforce a critical assumption on μ, i.e., $0 < \mu < 2$.

Note that $\{X_k, k = 0, 1, \dots\}$ is a Markov process since

$$X_{k+1} - X_k = (Y_{k+1} - Y_k) - \mu X_k.$$

The transition probability from state y to state x is

$$\mathbb{P}(X_{k+1} = x | X_k = y) = f_{\theta,\sigma^2}(x - (1 - \mu)y),$$

where $f_{\theta,\sigma^2}(s)$ denotes the probability density function associated with a normal random variable with mean θ and variance σ^2.

Let π denote a normal density with mean θ/μ and variance

$$\frac{\sigma^2}{2\mu - \mu^2},$$

i.e.,

$$\pi(y) = \frac{\sqrt{2\mu - \mu^2}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(2\mu - \mu^2)(y - \theta/\mu)^2}{2\sigma^2}\right).$$

Since we assume $0 < \mu < 2$, the variance is always positive. We now show that π is the stationary density for the discrete version of OU process. It is equivalent to showing that

$$\pi(x) = \int_{-\infty}^{\infty} \mathbb{P}(x|y)\pi(y)dy \tag{66}$$

for any given x.

We have

$$\begin{aligned}
P(x|y)\pi(y) &= \frac{\sqrt{2\mu-\mu^2}}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(2\mu-\mu^2)(y-\theta/\mu)^2}{2\sigma^2}\right)\exp\left(-\frac{(x-(1-\mu)y-\theta)^2}{2\sigma^2}\right) \\
&= \frac{\sqrt{(2\mu-\mu^2)}}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{y^2-2[(1-\mu)x+\theta]y}{2\sigma^2}\right)\exp\left(-\frac{(2\mu-\mu^2)\theta^2/\mu^2+(x-\theta)^2}{2\sigma^2}\right) \\
&= \frac{\sqrt{(2\mu-\mu^2)}}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{[y-((1-\mu)x+\theta)]^2}{2\sigma^2}\right) \\
&\quad \cdot \exp\left(-\frac{(2\mu-\mu^2)\theta^2/\mu^2+(x-\theta)^2-[(1-\mu)x+\theta]^2}{2\sigma^2}\right).
\end{aligned}$$

Among which,

$$\begin{aligned}
V(x) &= \exp\left(-\frac{(2\mu-\mu^2)\theta^2/\mu^2+(x-\theta)^2-[(1-\mu)x+\theta]^2}{2\sigma^2}\right) \\
&= \exp\left(-\frac{(2\mu-\mu^2)\theta^2/\mu^2+(2\mu-\mu^2)x^2-2(2-\mu)\theta x}{2\sigma^2}\right) \\
&= \exp\left(-\frac{(2\mu-\mu^2)(x-\theta/\mu)^2}{2\sigma^2}\right).
\end{aligned}$$

Then, we have

$$\begin{aligned}
\int_{-\infty}^{\infty}\mathbb{P}(x|y)\pi(y)dy &= \frac{\sqrt{(2\mu-\mu^2)}}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(2\mu-\mu^2)(x-\theta/\mu)^2}{2\sigma^2}\right) \\
&\quad \cdot \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{[y-((1-\mu)x+\theta)]^2}{2\sigma^2})dy \\
&= \frac{\sqrt{(2\mu-\mu^2)}}{\sqrt{2\pi}\sigma}\cdot\exp\left(-\frac{(2\mu-\mu^2)(x-\theta/\mu)^2}{2\sigma^2}\right),
\end{aligned}$$

which takes the exact form as the normal density with mean θ/μ and variance $\sigma^2/(2\mu-\mu^2)$ and thus, equals to $\pi(x)$. This completes our proof for π being the stationary density

APPENDIX D

APPENDIX FOR CHAPTER 7

D.1 Model calibration for mass gathering and Holiday traveling scenarios

D.1.1 Mass gathering (non-Holiday) scenarios

During the traveling period, the disease parameters for individuals not traveling/gathering (β, h_{PS}, h_{AS}, h_{PG}, h_C) remain the same as in the baseline model (see Section 5.3.3). For individuals who travel or attend mass gathering events, we need to calculate their infection hazard rate ($\tilde{h}_{PS}$ and $\tilde{h}_{AS}$) and the transmission rate $\tilde{\beta}$. To do that, let r_{XT} be the average number of people infected in T (traveling/gathering group) by an individual who is at stage X (X can be the presymptomatic (P), asymptomatic (A) or symptomatic (S) stage):

$$\begin{aligned} r_{PT} &= N_T(1-\phi_P(\frac{\tilde{h}_{PS}\tilde{\beta}}{N_T})) \\ r_{AT} &= p_A N_T \phi_P(\frac{\tilde{h}_{PS}\tilde{\beta}}{N_T})(1-\phi_A(\frac{\tilde{h}_{AS}\tilde{\beta}}{N_T})) \\ r_{ST} &= (1-p_A) N_T \phi_P(\frac{\tilde{h}_{PS}\tilde{\beta}}{N_T})(1-\phi_S(\frac{\tilde{\beta}}{N_T})) \end{aligned}$$

where N_T is the number of population on travel.

We then use R_0, θ, and ω to represent r_{XT}:

$$\begin{aligned} R_0 &= r_{PT} + r_{AT} + r_{ST} \\ \theta &= \frac{r_{PT} + r_{AT}}{R_0} \\ \omega &= \frac{r_{AT} + p_A r_{PT}}{R_0}. \end{aligned}$$

Given R_0, θ, and ω, we can calculate $\tilde{\beta}$, $\tilde{h}_{PS}$ and $\tilde{h}_{AS}$ through solving these non-linear equations. Note that the parameters γ and δ (see explanation in Table 17) are not

involved. It is because if a susceptible person who travels or attends gathering events gets infected during the traveling period, the source of infection is 100% from contacting with other people in the traveling/gathering group. We assume the values of R_0, θ, and ω remain the same as in the baseline model for individuals who travel/gather during the traveling period.

D.1.2 Holiday traveling scenarios

The disease parameters for individuals who retain their regular mixing patterns (β, h_{PS}, h_{AS}, h_{PG}, h_C) remain the same as in the baseline model (see Section 5.3.3). We need to recalibrate the disease parameters for individuals who do not travel but stay at home during the day ($\tilde{\beta}$, $\tilde{h}_{PS}$, $\tilde{h}_{AS}$, and $\tilde{h}_C$) and for the traveling individuals ($\bar{\beta}$, $\bar{h}_{PS}$, $\bar{h}_{AS}$, and $\bar{h}_T$). Next, we discuss how to calculate the disease parameters for these two groups of individuals.

Individuals not traveling and staying at home all day

Let $\tilde{r}_{XY}$ be the average number of people infected in Y by an individual who is at stage X (i.e., X can be the presymptomatic (P), asymptomatic (A) or symptomatic (S) stage; Y can be the household (H) or the community (C)):

$$\begin{aligned}
\tilde{r}_{PH} &= \sum_{n=1}^{7} p_n(n-1)(1-\phi_P(\frac{\tilde{h}_{PS}\tilde{\beta}}{n})) \\
\tilde{r}_{AH} &= p_A \sum_{n=1}^{7} p_n(n-1)\phi_P(\frac{\tilde{h}_{PS}\tilde{\beta}}{n})(1-\phi_A(\frac{\tilde{h}_{AS}\tilde{\beta}}{n})) \\
\tilde{r}_{SH} &= (1-p_A) \sum_{n=1}^{7} p_n(n-1)\phi_P(\frac{\tilde{h}_{PS}\tilde{\beta}}{n})(1-\phi_S(\frac{\tilde{\beta}}{n})) \\
\tilde{r}_{PC} &= N(1-\phi_P(\frac{\tilde{h}_{PS}\tilde{h}_C\tilde{\beta}}{N})) \\
\tilde{r}_{AC} &= p_A N\phi_P(\frac{\tilde{h}_{PS}\tilde{h}_C\tilde{\beta}}{N})(1-\phi_A(\frac{\tilde{h}_{AS}h_C\tilde{\beta}}{N})) \\
\tilde{r}_{SC} &= (1-p_A)N\phi_P(\frac{\tilde{h}_{PS}\tilde{h}_C\tilde{\beta}}{N})(1-\phi_S(\frac{\tilde{h}_C\tilde{\beta}}{N})).
\end{aligned}$$

We can also represent $\tilde{r}_{XY}$ in terms of R_0, θ, ω and γ:

$$\begin{aligned} R_0 &= \tilde{r}_{PH} + \tilde{r}_{AH} + \tilde{r}_{SH} + \tilde{r}_{PC} + \tilde{r}_{AC} + \tilde{r}_{SC} \\ \theta &= \frac{\tilde{r}_{PH} + \tilde{r}_{AH} + \tilde{r}_{PC} + \tilde{r}_{AC}}{R_0} \\ \omega &= \frac{\tilde{r}_{AH} + p_A\tilde{r}_{PH} + \tilde{r}_{AC} + p_A\tilde{r}_{PC}}{R_0} \\ \gamma &= \frac{\tilde{r}_{PC} + \tilde{r}_{AC} + \tilde{r}_{SC}}{R_0}. \end{aligned}$$

Given the values of R_0, θ, ω and γ, we can determine $\tilde{r}_{XT}$ and further calculate $\tilde{\beta}$, $\tilde{h}_{PS}$, $\tilde{h}_{AS}$, and $\tilde{h}_C$. We assume R_0, θ, ω, and γ take the same values as in the baseline model (see Table 17).

Traveling individuals

Let $\bar{r}_{XY}$ be the average number of people infected in Y by an individual who is at stage X (i.e., X can be the presymptomatic (P), asymptomatic (A) or symptomatic (S) stage; Y can be the household (H) or the traveling group (T)):

$$\begin{aligned} \bar{r}_{PH} &= \sum_{n=1}^{7} p_n(n-1)(1-\phi_P(\frac{\bar{h}_{PS}\bar{\beta}}{n})) \\ \bar{r}_{AH} &= p_A\sum_{n=1}^{7} p_n(n-1)\phi_P(\frac{\bar{h}_{PS}\bar{\beta}}{n})(1-\phi_A(\frac{\bar{h}_{AS}\bar{\beta}}{n})) \\ \bar{r}_{SH} &= (1-p_A)\sum_{n=1}^{7} p_n(n-1)\phi_P(\frac{\bar{h}_{PS}\bar{\beta}}{n})(1-\phi_S(\frac{\bar{\beta}}{n})) \\ \bar{r}_{PT} &= N_T(1-\phi_P(\frac{\bar{h}_{PS}\bar{h}_T\bar{\beta}}{2N_T})) \\ \bar{r}_{AT} &= p_A N_T\phi_P(\frac{\bar{h}_{PS}\bar{h}_T\bar{\beta}}{2N_T})(1-\phi_A(\frac{\bar{h}_{AS}\bar{h}_T\bar{\beta}}{2N_T})) \\ \bar{r}_{ST} &= (1-p_A)N_T\phi_P(\frac{\bar{h}_{PS}\bar{h}_T\bar{\beta}}{2N_T})(1-\phi_S(\frac{\bar{h}_T\bar{\beta}}{2N_T})) \end{aligned}$$

where N_T is the number of individuals on travel.

We can also represent $\bar{r}_{XY}$ in terms of R_0, θ, ω and γ.

$$\begin{aligned} R_0 &= \bar{r}_{PH} + \bar{r}_{AH} + \bar{r}_{SH} + \bar{r}_{PT} + \bar{r}_{AT} + \bar{r}_{ST} \\ \theta &= \frac{\bar{r}_{PH} + \bar{r}_{AH} + \bar{r}_{PT} + \bar{r}_{AT}}{R_0} \\ \omega &= \frac{\bar{r}_{AH} + p_A\bar{r}_{PH} + \bar{r}_{AT} + p_A\bar{r}_{PT}}{R_0} \\ \gamma &= \frac{\bar{r}_{PT} + \bar{r}_{AT} + \bar{r}_{ST}}{R_0}. \end{aligned}$$

Given the values of R_0, θ, ω and γ, we can determine $\bar{r}_{XY}$ and then further calculate $\bar{\beta}$, $\bar{h}_{PS}$, $\bar{h}_{AS}$, and $\bar{h}_T$. We assume R_0, θ, and ω take the same values as in the baseline model (see Table 17), and γ=0.4 (i.e., the proportion of transmission that occurs in the traveling group is 40%).

REFERENCES

[1] Allon, G., Deo, S., and Lin, W., "The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence," 2012. working paper.

[2] Altizer, S., Dobson, A., Hosseini, P., Hudson, P., Pascual, M., and Rohani, P., "Seasonality and the dynamics of infectious diseases," *Ecology Letters*, vol. 9, no. 4, pp. 467–484, 2006.

[3] Anderson, R. M. and May, R. M., *Infectious diseases of humans-dynamics and control*. Oxford, UK: Oxford Science Publications, 1991.

[4] Anthony, D., Chetty, V. K., Kartha, A., McKenna, K., DePaoli, M. R., and Jack, B., "Re-engineering the hospital discharge: An example of a multifaceted process evaluation," in *Advances in patients safety: from research to implementation* (Henriksen, K., Battles, J., Marks, E., and Lewin, D., eds.), Rockville, MD: Agency for Healthcare Research and Quality, 2005.

[5] Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y., and Yom-Tov, G., "Patient flow in hospitals: A data-based queueing perspective," 2011. working paper.

[6] Armony, M. and Zacharias, C., "Panel sizing and appointment scheduling in outpatient medical care," 2013. working paper.

[7] Bair, A., Song, W., Chen, Y., and Morris, B., "The impact of inpatient boarding on ED efficiency:a discrete-event simulation study," *J Med Syst*, vol. 34, pp. 919–929, 2010.

[8] Bertsekas, D. and Gallager, R., *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[9] Birjandi, A. and Bragg, L. M., *Discharge Planning Handbook for Healthcare: Top 10 Secrets to Unlocking a New Revenue Pipeline*. New York: Productivity Press, 2008.

[10] Bjornstad, O. N., Finkenstadt, B. F., and Grenfell, B. T., "Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series sir model," *Ecological Monographs*, vol. 72, no. 2, pp. 169–184, 2002.

[11] Blanchet, J. and Glynn., P., "Complete corrected diffusion for the maximum of the random walk.," *Annals of Applied Probability*, vol. 16, pp. 951–953, 2006.

[12] Bolker, B. and Grenfell, B., "Space, persistence and dynamics of measles epidemics," *Biological Sciences*, vol. 348, pp. 309–320, 1995.

[13] Bootsma, M. C. and Ferguson, N. M., "The effect of public health measures on the 1918 influenza pandemic in u.s. cities," *Proc Natl Acad Sci U S A*, vol. 104, no. 18, pp. 7588–93, 2007.

[14] Borghans, I., Heijink, R., Kool, T., Lagoe, R., and Westert, G., "Benchmarking and reducing length of stay in Dutch hospitals," *BMC Health Services Research*, vol. 8, no. 1, p. 220, 2008.

[15] Bramson, M., *Stability of Queueing Networks:*. Lecture Notes in Mathematics / École d'Été de Probabilités de Saint-Flour, Springer, 2008.

[16] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L., "Statistical analysis of a telephone call center," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 36–50, 2005.

[17] Browne, B. and Kuo, D., "Patients admitted through the emergency department are more profitable than patients admitted electively," *Annals of Emergency Medicine*, vol. 44, no. 4, Supplement, pp. S132 –, 2004.

[18] Browne, S. and Whitt, W., "Piecewise-linear diffusion processes," in *Advances in Queueing* (Dshalalow, J., ed.), (Boca Raton, FL), pp. 463–480, CRC Press, 1995.

[19] Bruin, A. M. d., Rossum, A. v., Visser, M. C., and Koole, G., "Modeling the emergency cardiac in-patient flow: An application of queuing theory," *Health Care Management Science*, pp. 1–13, 2006.

[20] Bureau of the Census, U. D. o. C., "Census 2000," May 1, 2008 2008. (http://www.census.gov). (Accessed May 1, 2008).

[21] Canadian Institute for Health Information, C., "Inpatient Hospitalizations and Average Length of Stay Trends in Canada, 2003-2004 and 2004-2005," 2005.

[22] Carr, B. G., Hollander, J. E., Baxt, W. G., Datner, E. M., and Pines, J. M., "Trends in boarding of admitted patients in US emergency departments 2003-2005," *Journal of Emergency Medicine*, vol. 39, no. 4, pp. 506–511, 2010.

[23] Carrat, F., Luong, J., Lao, H., Sall, A., Lajaunie, C., and Wackernagel, H., "A "small-world-like" model for comparing interventions aimed at preventing and controlling influenza pandemics," *BMC Medicine*, vol. 4, no. 26, 2006.

[24] Casagrandi, R., Bolzoni, L., Levin, S., and Andreasen, V., "The sirc model and influenza a," *Mathematical Biosciences*, vol. 200, pp. 152–169, 2006.

[25] Cauchemez, S., Valleron, A. J., Boelle, P. Y., Flahault, A., and Ferguson, N. M., "Estimating the impact of school closure on influenza transmission from sentinel data," *Nature*, vol. 452, no. 7188, pp. 750–U6, 2008.

[26] Centers for Disease Control and Prevention, USA, "2009 h1n1 flu and travel," 2009.

[27] Centers for Disease Control and Prevention, USA, "Interim cdc guidance for public gatherings in response to human infections with novel influenza a (h1n1)," 2009.

[28] Centers for Disease Control and Prevention, USA, "Health, United States," 2010.

[29] Chan, C., Yom-Tov, G., and Escobar, G. J., "When to use Speedup: An Examination of Intensive Care Units with Readmissions," 2011. working paper.

[30] Chan, N. H. and Tran, L. T., "Nonparametric tests for serial dependence," *Journal of Time Series Analysis*, vol. 13, no. 1, pp. 19–28, 1992.

[31] Chin, T. D., Foley, J. F., Doto, I. L., Gravelle, C. R., and Weston, J., "Morbidity and mortality characteristics of asian strain influenza," *Public health reports*, vol. 75, no. 2, p. 149, 1960.

[32] Choudhury, G., Lucantoni, D., and Whitt, W., "Numerical solution of piecewise-stationary $M_t/G_t/1$ queues," *Operations Research*, vol. 45, pp. 451–463, MAY-JUN 1997.

[33] Cochran, J. and Bharti, A., "Stochastic bed balancing of an obstetrics hospital," *Health Care Management Science*, vol. 9, no. 1, pp. 31–45, 2006.

[34] Colizza, V., Barrat, A., Barthelemy, M., Valleron, A., and Vespignani, A., "Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions," *PLoS Medicine*, vol. 4, no. 1, pp. 95–110, 2007.

[35] Commission, G. A., 2008.

[36] Cooper, B. S., Pitman, R. J., Edmunds, W. J., and Gay, N. J., "Delaying the international spread of pandemic influenza," *PLoS Med*, vol. 3, no. 6, p. e212, 2006.

[37] Dai, J. G., He, S., and Tezcan, T., "Many-server diffusion limits for $G/Ph/n + GI$ queues," *Annals of Applied Probability*, vol. 20, no. 5, pp. 1854–1890, 2010.

[38] Dai, J. G. and Lin, W., "Maximum pressure policies in stochastic processing networks," *Operations Research*, vol. 53, pp. 197–218, 2005.

[39] Department of Health, United Kingdom, "Achieving timely simple discharge from hospital: A toolkit for the multi-disciplinary team," 2004.

[40] Department of Health, United Kingdom, "Emergency admissions through accident and emergency," 2012.

[41] Eick, S. G., Massey, W. A., and Whitt, W. *Operations Research*, vol. 41, no. 4, pp. 731–742, 1993.

[42] Ekici, A., Keskinocak, P., and Swann, J., "Modeling influenza pandemic and planning food distribution," 2013. to appear in Manufacturing and Service Operations Management.

[43] Epstein, J. M., Goedecke, D. M., Yu, F., Morris, R. J., Wagener, D. K., and Bobashev, G. V., "Controlling pandemic flu: the value of international air travel restrictions," *PLoS One*, vol. 2, no. 5, p. e401, 2007.

[44] Feldman, Z., Mandelbaum, A., Massey, W. A., and Whitt, W., "Staffing of time-varying queues to achieve time-stable performance," *Management Science*, vol. 54, no. 2, pp. 324–338, 2008.

[45] Feldman, Z., Mandelbaum, A., Massey, W. A., and Whitt, W., "Staffing of Time-Varying Queues to Achieve Time-Stable Performance," *Management Science*, vol. 54, no. 2, pp. 324–338, 2008.

[46] Ferguson, N. M., Cummings, D. A., Fraser, C., Cajka, J. C., Cooley, P. C., and Burke, D. S., "Strategies for mitigating an influenza pandemic," *Nature*, vol. 442, no. 7101, pp. 448–52, 2006.

[47] Ferguson, N. M., Cummings, D. A., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., and Burke, D. S., "Strategies for containing an emerging influenza pandemic in southeast asia," *Nature*, vol. 437, no. 7056, pp. 209–214, 2005.

[48] Ferguson, N. M., Keeling, M. J., Edmunds, W. J., Gani, R., Grenfell, B. T., Anderson, R. M., and Leach, S.

[49] Ferguson, N. M., Mallett, S., Jackson, H., Roberts, N., and Ward, P., "A population-dynamic model for evaluating the potential spread of drug-resistant influenza virus infections during community-based use of antivirals," *Journal of Antimicrobial Chemotherapy*, vol. 51, no. 4, pp. 977–990, 2003.

[50] Ferrari, M. J., Grais, R. F., Bharti, N., Conlan, A. J., Bjornstad, O. N., Wolfson, L. J., Guerin, P. J., Djibo, A., and Grenfell, B. T., "The dynamics of measles in sub-saharan africa," *Nature*, vol. 451, no. 7179, pp. 679–84, 2008.

[51] Fine, P. and Clarkson, J., "Measles in england and wales : An analysis of factors underlying seasonal patterns," *International Journal of Epidemiology*, vol. 11, pp. 5–14, 1982.

[52] FLU.GOV, "Pandemic flu history," 2012.

[53] Fraser, C., Riley, S., Anderson, R. M., and Ferguson, N. M., "Factors that make an infectious disease outbreak controllable," *Proc. Natl. Acad. Sci.*, vol. 101, no. 16, pp. 6146–6151, 2004.

[54] Gans, N., Koole, G., and Mandelbaum, A., "Telephone call centers: Tutorial, review, and research prospects," *Manufacturing & Service Operations Management*, vol. 5, no. 2, pp. 79–141, 2003.

[55] Gans, N., Koole, G., and Mandelbaum, A., "Telephone call centers: Tutorial, review, and research prospects," *Manufacturing & Service Operations Management*, vol. 5, no. 2, pp. 79–141, 2003.

[56] Germann, T. C., Kadau, K., Longini, I. M., and Macken, C. A., "Mitigation strategies for pandemic influenza in the united states," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 15, pp. 5935–5940, 2006.

[57] Grais, R. F., Ellis, J., Glass, G., and Kress, A., "Modeling the spread of annual influenza epidemics in the u.s.: The potential role of air travel," *Health Care Management Science*, vol. 7, pp. 127–134, 2004.

[58] Green, L. V. and Kolesar, P. J., "The lagged PSA for estimating peak congestion in multiserver Markovian queues with periodic arrival rates," *Management Science*, vol. 43, pp. 80–87, JAN 1997.

[59] Green, L., "Queueing analysis in healthcare," in *Patient Flow: Reducing Delay in Healthcare Delivery* (Hall, R. W., ed.), vol. 91 of *International Series in Operations Research and Management Science*, pp. 281–307, Springer US, 2006.

[60] Green, L. and Kolesar, P. J., "The pointwise stationary approximation for queues with non-stationary arrivals," *Management Science*, vol. 37, pp. 84–97, 1991.

[61] Green, L. V., Soares, J., Giglio, J. F., and Green, R. A., "Using queueing theory to increase the effectiveness of emergency department provider staffing," *Academic Emergency Medicine*, vol. 13, no. 1, pp. 61–68, 2006.

[62] Green, L., "How many hospital beds?," *Inquiry*, vol. 39, no. 4, pp. 400–412, 2002.

[63] Griffin, J., Xia, S., Peng, S., and Keskinocak, P., "Improving patient flow in an obstetric unit," *Health Care Manag Sci*, 2011.

[64] Gustafson, R., "Pandemic influenza: Public health measures," *BC Medical Journal*, vol. 49, no. 5, pp. 254–257, 2007.

[65] Halder, N., Kelso, J. K., and Milne, G. J., "Analysis of the effectiveness of interventions used during the 2009 a/h1n1 influenza pandemic," *BMC Public Health*, vol. 10, 2010.

[66] Halfin, S. and Whitt, W., "Heavy-traffic limits for queues with many exponential servers," *Operations Research*, vol. 29, no. 3, pp. 567–588, 1981.

[67] Hall, I. M., Gani, R., Hughes, H. E., and Leach, S., "Real-time epidemic forecasting for pandemic influenza," *Epidemiology and Infection*, vol. 135, no. 3, pp. 372–385, 2007.

[68] Hall, M. J., DeFrances, C. J., Williams, S. N., Golosinskiy, A., and Schwartzman, A., "National hospital discharge survey: 2007 summary," *Natl Health Stat Report*, no. 29, pp. 1–20, 24, 2010.

[69] Hall, R., "Bed assignment and bed management," in *Handbook of Healthcare System Scheduling* (Hall, R., ed.), vol. 168 of *International Series in Operations Research & Management Science*, pp. 177–200, Springer US, 2012.

[70] Hall, R., Belson, D., Murali, P., and Dessouky, M., "Modeling patient flows through the healthcare system," in *Patient Flow: Reducing Delay in Healthcare Delivery* (Hall, R., ed.), Springer, 2006.

[71] Handel, A., Longini, I., and Antia, R., "What is the best control strategy for multiple infectious disease outbreaks?," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 274, pp. 833–837, 2007.

[72] Harrison, J. M., "Brownian models of open processing networks: canonical representation of workload," *Annals of Applied Probability*, vol. 10, pp. 75–103, 2000. Correction: **13**, 390–393 (2003).

[73] Hatchett, R. J., Mecher, C. E., and Lipsitch, M., "Public health interventions and epidemic intensity during the 1918 influenza pandemic," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 18, pp. 7582–7587, 2007.

[74] Helm, J. and Van Oyen, M., "Design and optimization methods for elective hospital admissions," 2012. Working paper.

[75] Helm, J. E., AhmadBeygi, S., and Van Oyen, M. P., "Design and analysis of hospital admission control for operational effectiveness," *Production and Operations Management*, vol. 20, no. 3, pp. 359–374, 2011.

[76] Henneman, P. L., Lemanski, M., Smithline, H. A., Tomaszewski, A., and Mayforth, J. A., "Emergency department admissions are more profitable than non-emergency department admissions," *Annals of Emergency Medicine*, vol. 53, no. 2, pp. 249 – 255.e2, 2009.

[77] Hens, N., Goeyvaerts, N., Aerts, M., Shkedy, Z., Van Damme, P., and Beutels, P., "Mining social mixing patterns for infectious disease models based on a two-day population survey in belgium," *Bmc Infectious Diseases*, vol. 9, 2009.

[78] Hill, A. N. and Longini, I. M., "The critical vaccination fraction for heterogeneous epidemic models," *Mathematical Biosciences*, vol. 181, pp. 85–106, 2003.

[79] Hoot, N. and Aronsky, D., "Systematic review of emergency department crowding: Causes, effects, and solutions.," *Ann Emerg Med*, vol. 52, pp. 126–36, 2008.

[80] Hospitalist Management Advisor, "To free beds for new admissions, triage best candidates for early discharge," 2006.

[81] Howell, E., Bessman, E., Kravet, S., Kolodner, K., Marshall, R., and Wright, S., "Active bed management by hospitalists and emergency department throughput.," *Annals of Internal Medicine*, vol. 149, no. 11, pp. 804–810, 2008.

[82] Jacobson, S. H., Hall, S. N., and Swisher, J. R., "Discrete-event simulation of health care systems," in *Patient Flow: Reducing Delay in Healthcare Delivery* (Hall, R. W., ed.), vol. 91 of *International Series in Operations Research and Management Science*, pp. 211–252, Springer US, 2006.

[83] Jennings, O., Mandelbaum, A., Massey, W., and Whitt, W., "Server staffing to meet time-varying demand," *Management Science*, vol. 42, pp. 1383–1394, OCT 1996.

[84] Kanji, G., *100 Statistical Tests*. SAGE Publications, 2006.

[85] Kelso, J. K., Milne, G. J., and Kelly, H., "Simulation suggests that rapid activation of social distancing can arrest epidemic development due to a novel strain of influenza," *Bmc Public Health*, vol. 9, 2009.

[86] Khanna, S., Boyle, J., Good, N., and Lind, J., "Impact of admission and discharge peak times on hospital overcrowding," *Health Informatics: The Transformative Power of Innovation*, pp. 82–88, 2011.

[87] Kim, S.-H., Chan, C., Olivares, M., and Escobar, G. J., "ICU admission control: An empirical study of capacity allocation and its implication on patient outcomes," 2012. working paper.

[88] Kingman, J. F. C., "Ergodic properties of continuous-time markov processes and their discrete skeletons," *Proceedings of the London Mathematical Society*, vol. s3-13, no. 1, pp. 593–604, 1963.

[89] Koizumi, N., Kuno, E., and Smith, T. E., "Modeling patient flows using a queuing network with blocking," *Health care management science*, vol. 8, no. 1, pp. 49–60, 2005.

[90] Kumar, P. R., "Re-entrant lines," *Queueing Systems*, vol. 13, pp. 87–110, 1993.

[91] Law, A. M. and Kelton, D. W., *Simulation Modelling and Analysis*. McGraw-Hill Education - Europe, 2000.

[92] Lee, B. Y., Brown, S. T., Cooley, P., Potter, M. A., Wheaton, W. D., Voorhees, R. E., Stebbins, S., Grefenstette, J. J., Zimmer, S. M., Zimmerman, R. K., Assi, T. M., Bailey, R. R., Wagener, D. K., and Burke, D. S., "Simulating school closure strategies to mitigate an influenza epidemic," *Journal of Public Health Management and Practice*, 2009.

[93] Lee, B. Y., Brown, S. T., Cooley, P. C., Zimmerman, R. K., Wheaton, W. D., Zimmer, S. M., Grefenstette, J. J., Assi, T.-M., Furphy, T. J., Wagener, D. K., and others, "A computer simulation of employee vaccination to mitigate an influenza epidemic," *American journal of preventive medicine*, vol. 38, no. 3, pp. 247–257, 2010.

[94] Lee, B. Y., Brown, S. T., Korch, G. W., Cooley, P. C., Zimmerman, R. K., Wheaton, W. D., Zimmer, S. M., Grefenstette, J. J., Bailey, R. R., Assi, T.-M., and others, "A computer simulation of vaccine prioritization, allocation, and rationing during the 2009 h1n1 influenza pandemic," *Vaccine*, vol. 28, no. 31, pp. 4875–4879, 2010.

[95] Lewis, P. A. and Shedler, G. S., "Simulation methods for Poisson processes in nonstationary systems," in *Proceedings of the 10th conference on Winter simulation - Volume 1*, WSC '78, (Piscataway, NJ, USA), pp. 155–163, IEEE Press, 1978.

[96] Lipsitch, M., Cohen, T., Cooper, B., Robins, J. M., Ma, S., James, L., Gopalakrishna, G., Chew, S. K., Tan, C. C., Samore, M. H., Fisman, D., and Murray, M., "Transmission dynamics and control of severe acute respiratory syndrome," *Science*, vol. 300, no. 5627, pp. 1966–1970, 2003.

[97] Litvak, E., Long, M. C., Cooper, A. B., and McManus, M. L., "Emergency department diversion: Causes and solutions," *Academic Emergency Medicine*, vol. 8, no. 11, pp. 1108–110, 2001.

[98] Liu, S. W., Thomas, S. H., Gordon, J. A., Hamedani, A. G., and Weissman, J. S., "A pilot study examining undesirable events among emergency department-boarded patients awaiting inpatient beds," *Annals of Emergency Medicine*, vol. 54, no. 3, pp. 381–385, 2009.

[99] Liu, Y. and Whitt, W., "Stabilizing customer abandonment in many-server queues with time-varying arrivals," 2012. working paper.

[100] London, W. and Yorke, J., "Recurrent outbreaks of measles, chickenpox and mumps seasonal variation in contact rates," *American Journal of Epidemiology*, vol. 98, pp. 453–468, 1973.

[101] Longini, I. M., J., Nizam, A., Xu, S., Ungchusak, K., Hanshaoworakul, W., Cummings, D. A., and Halloran, M. E., "Containing pandemic influenza at the source," *Science*, vol. 309, no. 5737, pp. 1083–7, 2005.

[102] Lowen, A. C., Mubareka, S., Steel, J., and Palese, P., "Influenza virus transmission is dependent on relative humidity and temperature," *PLoS Pathogens*, vol. 3, no. 10, pp. 1470–1476. doi:10.1371/journal.ppat.0030151, 2007.

[103] Maman, S., "Uncertainty in the demand for service: The case of call centers and emergency departments," July 2009.

[104] Mandelbaum, A., Momcilovic, P., and Tseytlin, Y., "On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers," *Management Science*, 2012.

[105] Massey, W. A. and Whitt, W., "An analysis of the modified offered-load approximation for the nonstationary Erlang loss model," *Annals of Applied Probability*, vol. 4, no. 4, pp. 1145–1160, 1994.

[106] Memish, Z. A., McNabb, S. J. N., Mahoney, F., Alrabiah, F., Marano, N., Ahmed, Q. A., Mahjour, J., Hajjeh, R. A., Formenty, P., Harmanci, F. H., El Bushra, H., Uyeki, T. M., Nunn, M., Isla, N., Barbeschi, M., and Jeddah Hajj Consultancy, G., "Establishment of public health security in saudi arabia for the 2009 hajj in response to pandemic influenza a h1n1," *Lancet*, vol. 374, no. 9703, pp. 1786–1791, 2009.

[107] Mills, C. E., Robins, J. M., Bergstrom, C. T., and Lipsitch, M., "Pandemic influenza: Risk of multiple introductions and the need to prepare for them," *PLoS Medicine*, vol. 3, no. 6, pp. 769–773. doi:10.1371/journal.pmed.0030135, 2006.

[108] Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., and Edmunds, W. J., "Social contacts and mixing patterns relevant to the spread of infectious diseases," *PLoS Med*, vol. 5, no. 3, p. e74, 2008.

[109] Mummert, A., Weiss, H., Long, L.-P., Amigo, J. M., and Wan, X.-F., "A perspective on multiple waves of influenza pandemics," *PloS one*, vol. 8, no. 4, p. e60343, 2013.

[110] National University Hospital, "BMU training guide: Inpatient operations," December 2011.

[111] National University Hospital, "NUH Inpatient Charges," 2012.

[112] NHS National Services, Scotland, "Average length of stay," 2010.

[113] Nicholls, H., "Pandemic influenza: The inside story," *PLoS Biology*, vol. 4, no. 2, pp. 156–160. doi:10.1371/journal.pbio.0040050, 2006.

[114] Osborn, C., *Essentials of Statistics in Health Information Technology*. USA: Jones and Bartlett Publishers, Inc., 1st ed., 2007.

[115] Pascual, M. and Dobson, A., "Seasonal patterns of infectious diseases," *PLoS Medicine*, vol. 2, no. 1, pp. 18–20. doi:10.1371/journal.pmed.0020005, 2005.

[116] Pines, J. M., Batt, R. J., Hilton, J. A., and Terwiesch, C., "The financial consequences of lost demand and reducing boarding in hospital emergency departments," *Annals of Emergency Medicine*, vol. 58, no. 4, pp. 331–340, 2011.

[117] Pines, J. M., Hilton, J. A., Weber, E. J., Alkemade, A. J., Al Shabanah, H., Anderson, P. D., Bernhard, M., Bertini, A., Gries, A., Ferrandiz, S., Kumar, V. A., Harjola, V.-P., Hogan, B., Madsen, B., Mason, S., Ohlen, G., Rainer, T., Rathlev, N., Revue, E., Richardson, D., Sattarian, M., and Schull, M. J., "International perspectives on emergency department crowding," *Academic Emergency Medicine*, vol. 18, no. 12, pp. 1358–1370, 2011.

[118] Pines, J. M., Iyer, S., Disbot, M., Hollander, J. E., Shofer, F. S., and Datner, E. M., "The effect of emergency department crowding on patient satisfaction for admitted patients," *Academic Emergency Medicine*, vol. 15, no. 9, pp. 825–831, 2008.

[119] Porta, M., *A Dictionary of Epidemiology*. Oxford University Press, USA, 2008.

[120] Powell, E. S., Khare, R. K., Venkatesh, A. K., Van Roo, B. D., Adams, J. G., and Reinhardt, G., "The relationship between inpatient discharge timing and emergency department boarding," *The Journal of Emergency Medicine*, 2011.

[121] Public Health Agency, Canada, "Public Health Guidance for the prevention and management of Influenza-like-illness (ILI), including the Pandemic (H1N1) 2009 Influenza Virus, related to mass gatherings," 2009.

[122] Public Health Agency, Canada, "Travel health notice," 2009.

[123] Rahmandad, H. and Sterman, J., "Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models," *Management Science*, vol. 54, no. 5, pp. 998–1014, 2008.

[124] Ramakrishnan, M., Sier, D., and Taylor, P., "A two-time-scale model for hospital patient flow," *IMA Journal of Management Mathematics*, vol. 16, no. 3, pp. 197–215, 2005.

[125] Ramakrishnan, M., Sier, D., and Taylor, P., "A two-time-scale model for hospital patient flow," *IMA Journal of Management Mathematics*, vol. 16, no. 3, pp. 197–215, 2005.

[126] Rashid, H., Haworth, E., Shafi, S., Memish, Z. A., and Booy, R., "Pandemic influenza: mass gatherings and mass infection," *Lancet Infectious Diseases*, vol. 8, no. 9, pp. 526–527, 2008.

[127] Read, J. M., Eames, K. T., and Edmunds, W. J., "Dynamic social networks and the implications for the spread of infectious disease," *J R Soc Interface*, vol. 5, no. 26, pp. 1001–7, 2008.

[128] Rubino, L., Stahl, L., and Chan, M., "Innovative approach to the aims for improvement: Emergency department patient throughput in an impacted urban setting," *The Journal of Ambulatory Care Management*, vol. 30, no. 4, pp. 327–337, 2007.

[129] Scherer, A. and Mclean, A., "Mathematical models of vaccination," *British Medical Bulletin*, vol. 62, no. 1, pp. 187–199, 2002.

[130] Schneider, S., Zwemer, F., Doniger, A., Dick, R., Czapranski, T., and Davis, E., "Rochester, New York: a decade of emergency department overcrowding," *Academic Emergency Medicine*, vol. 8, no. 11, pp. 1044–1050, 2001.

[131] Shaman, J. and Kohn, M., "Absolute humidity modulates influenza survival, transmission, and seasonality," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 9, pp. 3243–3248, 2009.

[132] Shapira, O. M., Gorman, J., Fitzgerald, C., Traylor, D., Hunter, C., Lazar, H., Davidson, K., Chessare, J., and Shemin, R., "Process improvement to smooth daily discharge time of patients after cardiac surgery improves patient flow and reduces operating room time.," 2004. 10th Annual Scientific Symposium at the Institute of Healthcare Improvement.

[133] Shi, P., Keskinocak, P., Swann, J. L., and Lee, B. Y., "The impact of mass gatherings and holiday traveling on the course of an influenza pandemic: a computational model," *BMC Public Health*, vol. 10, p. 778, 2010.

[134] Shi, P., Keskinocak, P., Swann, J. L., and Lee, B. Y., "Modelling seasonality and viral mutation to predict the course of an influenza pandemic," *Epidemiology & Infection*, vol. 138, no. 10, pp. 1472–81, 2010.

[135] Shi, P., Chou, M., Dai, J. G., Ding, D., and Sim, J., "Models and Insights for Hospital Inpatient Operations: Time-of-Day Congestion for ED Patients Awaiting Beds," 2012. working paper.

[136] Shi, P., Dai, J. G., Ding, D., Ang, J., Chou, M., Xin, J., and Sim, J., "Patient Flow from Emergency Department to Inpatient Wards: Empirical Observations from a Singaporean Hospital," 2012.

[137] Siegmund, D., "Corrected diffusion approximations in certain random walk problems," *Advances in Applied Probability*, vol. 11, no. 4, pp. 701–719, 1979.

[138] Sigma Breakthrough Technologies, I., "Improving inpatient discharge cycle time and patient satisfaction," 2006.

[139] Sinagpore Ministry of Health, "Waiting time for admission to ward," 2012.

[140] Singapore, M. o. H., "Update on influenza a (h1n1-2009) (6 may)," 2009. (http://www.moh.gov.sg/mohcorp/pressreleases.aspx?id=21692). (Accessed March 10, 2010).

[141] Singer, A. J., Thode, Henry C., J., Viccellio, P., and Pines, J. M., "The association between length of emergency department boarding and mortality," *Academic Emergency Medicine*, vol. 18, no. 12, pp. 1324–1329, 2011.

[142] Stahl, L., "Comprehensive emergency department and inpatient changes improve emergency department patient satisfaction, reduce bottlenecks that delay admissions," 2008.

[143] Taubenberger, J. and Morens, D., "1918 influenza: The mother of all pandemics," *Emerging Infectious Diseases*, vol. 12, no. 1, pp. 15–22, 2006.

[144] Teow, K., El-Darzi, E., Foo, C., Jin, X., and Sim, J., "Intelligent analysis of acute bed overflow in a tertiary hospital in Singapore," *Journal of Medical Systems*, pp. 1–10, Janurary 2011.

[145] Thompson, S., Nunez, M., Garfinkel, R., and Dean, M., "Efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges," *Operations Research*, vol. 57, no. 2, pp. 261 – 273, 2009.

[146] United States General Accounting Office, *Hospital emergency departments: crowded conditions vary among hospitals and communities*. Washington, D.C.: United States General Accounting Office, 2003.

[147] VA Medical Center, Washington DC, "Getting patients home sooner: DCVAMC institutes new hospital discharge system," 2009.

[148] Vericourt, F. d. and Jennings, O. B., "Nurse staffing in medical units: A queueing perspective," *Operations Research*, vol. 59, no. 6, pp. 1320–1331, 2011.

[149] Whitt, W., "The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increases," *Management Science*, vol. 37, pp. 307–314, MAR 1991.

[150] Wong, H. J., Morra, D., Caesar, M., Carter, M. W., and Abrams, H., "Understanding hospital and emergency department congestion: An examination of inpatient admission trends and bed resources," *Canadian Journal of Emergency Medicine*, vol. 34, no. 1, pp. 18–26, 2010.

[151] World Health Organization, WHO, "Interim planning considerations for mass gatherings in the context of pandemic (h1n1) 2009 influenza," 2009.

[152] World Health Organization, WHO, "No rationale for travel restrictions," 2009.

[153] Wu, J. T., Riley, S., Fraser, C., and Leung, G. M., "Reducing the impact of the next influenza pandemic using household-based public health interventions," *PLoS Medicine*, vol. 3, no. 9, pp. 1532–1540, 2006.

[154] Yancer, D. A., Foshee, D., Cole, H., Beauchamp, R., de la Pena, W., Keefe, T., Smith, W., Zimmerman, K., Lavine, M., and Toops, B., "Managing capacity to reduce emergency department overcrowding and ambulance diversions," *Jt Comm J Qual Patient Saf*, vol. 32, no. 5, pp. 239–45, 2006.

[155] Yankovic, N. and Green, L. V., "Identifying good nursing levels: A queuing approach," *Operations Research*, vol. 59, no. 4, pp. 942–955, 2011.

[156] Yao, D. D., *Stochastic Modeling and Analysis of Manufacturing Systems*. Springer Series in Operations Research, New York: Springer, 1994.

[157] Zeltyn, S., Marmor, Y. N., Mandelbaum, A., Carmeli, B., Greenshpan, O., Mesika, Y., Wasserkrug, S., Vortman, P., Shtub, A., Lauterman, T., Schwartz, D., Moskovitch, K., Tzafrir, S., and Basis, F., "Simulation-based models of emergency departments: Operational, tactical, and strategic staffing," *ACM Trans. Model. Comput. Simul.*, vol. 21, pp. 24:1–24:25, Sept. 2011.