12-2013

# Dynamic Queue Management for Hospital Emergency Room Services

Kar Way TAN
*Singapore Management University*, kwtan@smu.edu.sg

## Citation

TAN, Kar Way. Dynamic Queue Management for Hospital Emergency Room Services. (2013). 1-115.
**Available at:** https://ink.library.smu.edu.sg/etd_coll/109

# Dynamic Queue Management for Hospital Emergency Room Services

**Kar Way TAN**

Submitted to School of Information Systems in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Information Systems

**Dissertation Committee**

Hoong Chuin LAU (Supervisor / Chair)
Associate Professor of Information Systems
Singapore Management University

Venky SHANKARARAMAN (Co-Supervisor)
Associate Professor of Information Systems (Education)
Singapore Management University

Robert KAUFFMAN
Professor of Information Systems
Singapore Management University

Xiaolan XIE
Professor of Industrial Engineering
Ecole Nationale Superieure des Mines, France
and
Chair Professor and Director of Center for Healthcare Engineering
Shanghai Jiao Tong University

Singapore Management University
2013

# Abstract

The emergency room (ER) – or emergency department (ED) – is often seen as a place with long waiting times and a lack of doctors to serve the patients. However, it is one of the most important departments in a hospital, and must efficiently serve patients with critical medical needs. In the existing literature, addressing the issue of long waiting times in an ED often takes the form of single-faceted queue-management strategies that are either from a demand perspective or from a supply perspective. From the demand perspective, there is work on queue design such as priority queues, or queue control strategies such as a fast-track system and demand restriction through ambulance diversion. On the supply side, existing studies looked at the management of the supply of resources (e.g., doctors, nurses, equipment). However, they may not sufficiently leverage insights that can be derived from both historical and real-time data.

In this dissertation, we present an integrated framework that manages queues *dynamically* in the ED from both the demand and supply perspectives by leveraging historical data and real-time data. More precisely, we introduce data-driven and intelligent dynamic patient-prioritization strategies to manage the demand concurrently with dynamic resource-adjustment policies to manage supply. Our framework allows decision-makers to select both demand-side and supply-side strategies to suit the needs of their ED. We verify through simulation that strategies from both perspectives work well together in our proposed framework. The results show that such a framework improves average patient length-of-stay (LOS) in the ED without having to restrict demand (stop patients from coming to the ED).

In our dynamic patient-prioritization strategies, we propose and evaluate

three schemes to allocate patients to doctors: shortest-consultation-time-first (SCON), shortest-remaining-time-first (SREM) and a mixed strategy (MIXED). We test the strategies using simulation and our experimental results show that a dynamic priority queue is effective in reducing the LOS of patients and hence improving patient flow. This is found to be better than standard queuing solutions which are based on first-in-first-out (FIFO) or static priority queues. We present results that show a trade-off between performance and risks (in terms of implementation complexity, and starvation, a situation where a patient is deprived of the chance to consult a doctor). We show that decision-makers in healthcare institutions can use the information to choose a strategy that is most suitable for their ED.

On the supply side, we consider the problem of allocating doctors in the ambulatory area of the ED based on a set of policies. Traditional staffing policies are static and do not react well to surges in patient demand. By leveraging real-time and historical information, we provide strategies in two dimensions: (1) the ability to react to changes in demand and (2) to optimize the doctor schedule so as to satisfy the hospital's desired service quality in terms of LOS. Our main contribution is a data-driven approach that performs online reallocation of doctor resources through symbiotic simulation in real time using historical as well as current arrival rates. We build a simulation prototype to demonstrate that this can be done. The experimental results from our prototype show that our approach allows the hospital to cope with varying levels of demand and to better serve the patients within the desired service level. In addition, the prototype offers insights into the trade-off between performance and risk (in terms of implementation complexity and doctor schedule stability). As such, we provide analysis and opportunities for decision-makers to select a strategy which fits the hospital concerned.

# Contents

# List of Figures

# List of Tables

# Abbreviations, Terminology and Notations

## Abbreviations

| | |
|---|---|
| DQM | Dynamic Queue Management |
| DQMS | Dynamic Queue Management System |
| DYN | Dynamic (a dynamic resource-adjustment strategy) |
| DYN-OPT | Dynamic-Optimized (a dynamic resource-adjustment strategy) |
| ED | Emergency Department |
| ER | Emergency Room (used interchangeably with ED) |
| ERP | Enterprise Resource Planning |
| FIFO | First-In-First-Out |
| HIST | Historical (a dynamic resource-adjustment strategy) |
| HIST-OPT | Historical-Optimized (a dynamic resource-adjustment strategy) |
| IS | Information Systems |
| ISA | Infinite Server Approximation |
| IT | Information Technology |
| LOS | Length of Stay |
| LWBS | Left Without Being Seen |
| MOL | Modified Offered Load |
| PSA | Piecewise Stationary Analysis |
| QED | Quality- and Efficiency-Driven |
| QR code | Quick Response code |
| RCCP | Rough Cut Capacity Planning |
| SCON | Shortest-Consultation-Time-First |
| SIPP | Stationary Independent Period by Period |
| SREM | Shortest-Remaining-Time-First |

# Terminology

| | |
|---|---|
| Ambulatory area | A physical area in an ED handling non-emergency patients. |
| Analytics | The discovery and communication of meaningful patterns in data. |
| Arrival rate | The number of people arriving at a facility per unit of time. |
| Back room | Also known as the critical-care area. |
| Business process | A repeatable series of activities performed to deliver a service or product to a stakeholder. |
| Critical-care area | A physical area in an ED handling emergency patients. |
| Dynamic priority queue | A multi-class queuing system in which the priorities of all patients in the queue are dynamically calculated as the surrounding environment changes. |
| ERP System | A cross-functional integrated suite of software modules that supports the basic internal business processes of an organization. |
| Front room | Also known as the ambulatory area. |
| Jackson network | A queuing network consisting of multiple nodes. Each node is a FIFO queue with exponential service times, and s servers. |
| M/M/s | A birth-death queuing model with an infinite capacity, Poisson arrivals, exponential service times, and s servers. |
| Multi-class queue | A queuing system where customers (patients) are divided into K priority classes. |
| (Standard) Priority queue | A multi-class queuing system in which the priorities of customers are assigned at the point when they join the queue. The priority of a customer remains the same with respect to the other entities in the queue throughout its life-time in the queue system. |
| Queue control | A study on managing the queue to prevent it from building up above a certain threshold. |
| Queue design | A process to determine the parameters in the queue model. |

| | |
|---|---|
| Queue model | Mathematical description of a queuing system with assumptions on (i) probabilistic nature of the arrival and service time; (ii) number and type of servers; (iii) organization; (iv) queue discipline; and (v) queue capacity. |
| Re-entrants | Patients who are consulting with the doctor for the second time in a single visit to the ED. This happens after the patient has undergone his/her investigative tests and treatment. |
| Service rate | The number of customers (in this case, patients) served at a service station per unit of time. |
| Starvation | A condition in which a patient is deprived of the chance to consult with a doctor due to multiple preemption by other patients with higher priorities. |
| Sub-process | A process whose functionality is part of a larger process. |
| Symbiotic simulation system | A system consisting of a simulation model interacting with physical systems. |
| Triage | A process of determining the patients' priority based on initial assessment of their conditions. This is usually performed by a nurse. |

# Notations

| | |
|---|---|
| $\lambda_b(t)$ | Time-varying arrival rates of new patients in the back room. |
| $\lambda_f(t)$ | Time-varying arrival rates of new patients in the front room. |
| $\lambda'_f(t)$ | Real-time arrival rate of new patients in the front room. |
| $\mu_n$ | Service rate of doctors if the patient is a new patient. |
| $\mu_r$ | Service rate of doctors if the patient is a re-entrant. |
| $\mu_b$ | Service rate of doctors in the back room. |
| $\delta$ | Service rate for investigative tests and treatment. |
| $b$ | Probability of re-entrance. |
| $\overline{LOS}(t)$ | Average LOS of patients who leave the ED within a time interval of $[t, t+1)$. |
| $LOS_{max}$ | Hospital's desired service quality in terms of LOS. |

| | |
|---|---|
| $\mu$ | Service rate of doctors at any service station. |
| $room_{max}$ | Physical constraints in the front room. |
| $S_{max}(t)$ | Maximum number of doctors that can be deployed in the ED (front and back rooms combined) at time $t$. |
| $S_b(t)$ | Number of doctors required in the back room. |
| $S_f(t)$ | Number of doctors to be placed in the front room. |
| $\beta$ | Quality- and Efficiency-Driven service parameter. |
| $\lambda_x^+$ | Aggregated-arrival-rate function to node $x$ in a queue with reentrants. |
| $S_x$ | Service time at node $x$. |
| $S_{x,e}$ | Random variable representing the excess service time at node $x$. |
| $R_x(t)$ | Offered load in Station $x$ at time $t$. |
| $G_b(t)/M_b/S_b(t)$ | Describes the queuing system in the back room – Time-varying general arrival function with exponential service rates and multiple servers. |
| $G_f(t)/M_f/S_f(t)$ | Describes the queuing system in the front room – Time-varying general arrival function with exponential service rates and multiple servers. |
| $k$ | Patient $k$. |
| $c_k$ | Estimated consultation time of patient $k$. |
| $e_k$ | The amount of time patient $k$ has spent in the ED, also known as elapsed time. |
| $t_k$ | The additional time patient $k$ has spent undergoing tests and treatment. |
| $d_k$ | Total available time patient $k$ is allowed in the ED, $d_k = LOS_{max} + t_k$. |
| $r_k$ | Time remaining of patient $k$. |
| $S_i$ | Dynamic patient-prioritization strategy where $i$ is the selected strategy type. |
| $p_k^{S_i}$ | Priority assigned to patient $k$ under dynamic patient-prioritization strategy $S_i$. |

$a_n$ Weight of each factor $n$ in N-factors MIXED dynamic patient-prioritization strategy.

$w$ Length of time (in minutes) that patients have waited in the queue for a doctor by the end of simulation run.

$X_j$ Resource-adjustment policy $j$.

$C_l$ Cost of labor for deploying a doctor for a single unit of time $t$.

$C_d$ Cost of deviation per doctor. This is applicable when the number of doctors at time $t$ is different from the number of doctors at time $t-1$.

$L$ Lead time for dynamic planning.

$H$ Time horizon for dynamic planning.

# Acknowledgements

I would like to express my heartfelt gratitude to my supervisor Prof Lau Hoong Chuin who not only has been very encouraging, understanding and offered invaluable guidance, but has been a friend, a colleague and someone who gave emotional support in difficult times. Deepest gratitude also due to my co-supervisor Prof Venky Shankararaman for the guidance and opportunities he has created for me especially in the initial years of my Ph.D. journey. Thank you, Prof Robert Kauffman, who has been particularly encouraging and positive in helping me to position my work and improve the quality of the thesis. Prof Kauffman also provided invaluable linkages to various distinguished professors of other universities.

Certainly, this thesis would not have been possible without the opportunities and support of the School of Information Systems (SIS), Singapore Management University, which enabled me to take up the Ph.D. program while working full time. I thank the Dean of SIS, Professor Steven Miller, for giving me guidance, support and understanding all these years and certainly for the valuable learning and career opportunities. I also thank SIS for the wonderful, friendly and open environment, only made possible by dedicated people in the administrative offices, faculty and also my very close circle of colleagues, the instructors.

I thank my colleague Mrs Koh Lian Chee, who leads the SMU-Alexandra Health Transformation Lab, for giving me this wonderful opportunity to work with Khoo Teck Puat Hospital (KTPH) for an interesting, challenging and practical study for my dissertation. I also express my sincere appreciation to Dr Francis Lee, Mr Lau Wing Chew and Mr Wu Dan from KTPH for their domain expertise and support of the research work in this dissertation. I also thank my collaborators: Murphy Choy for helping with the data analysis; and Wang Chao and Wei Hao, for their dedication in building parts of the Dynamic Queue Management simulator. My appreciation also goes to my fellow Ph.D. classmates for their assistance and friendship throughout my journey. Especially to Fu Na, thank you for attending to my ad hoc “knocking” at your

cubicle when I was drowned in sea of mathematical formulae.

I wish to express my love and gratitude to my beloved families; for their understanding, sacrifices and help for the duration of my studies. Special thanks to my husband Eng Kit and my father-in-law for managing the children and bearing with my stressed moods on many occasions when I had to meet tight deadlines. To my siblings, for encouragement, support and taking care of our parents.

Special thanks are to be dedicated to my group of mummy friends who were always ready to offer listening ears, advice, care and concern for me and my children over the last few years of my Ph.D. journey. Especially to my best buddy, Wen Lee, heartfelt thanks for your emotional support and believing in me.

# Dedication

I am dedicating this dissertation to my three wonderful, beloved children, Kai Ying, Yong Feng and Yong Bing. For them, I have the strength to bring the thesis to completion. Especially to Kai Ying, thank you for your understanding that Mummy often “had work to do” and was not able to put you to bed for many, many nights.

# Chapter 1

# Introduction

The key challenges in the emergency department (ED) of a hospital are service quality and efficiency. One needs to ensure that the patients receive appropriate quality medical attention, and that the department remains competitive in terms of efficiency. Long waits and hence increased patient length-of-stay (LOS) in an ED are a common problem faced by many hospitals around the world. Managing wait times in an ED is challenging because the ED deals with patients without appointments and with a wide variety of illnesses with a large variance in the time required to diagnose and treat them. In order to serve the emergency medical needs of patients while maintaining quality of care, a public hospital needs to provide better service and seek ways to improve patient flow in the ED by improving processes and queue management in the department.

## 1.1 The Challenges in Emergency Departments

### 1.1.1 Complex Queue Management

Queue management in the ED is complex, making its analysis challenging. Firstly, EDs deal with demand which is not easily predictable as patients arrive without appointments. Restricting the demand, such as through ambulance diversion [39], and the channeling of non-emergency cases to general practitioners, may not reflect well on the hospital or be suitable for certain cities. The stochastic nature of the demand makes it difficult for the ED to allocate resources. Secondly, the queuing process is multi-stage, unlike many standard M/M/1 or M/M/s queuing models which are applicable to other domains such as retail banking or merchandise shops. Multi-stage queuing systems lack an analytical model that can truly mimic the real world. Thirdly, we will see that the ED process (details in Chapter 2) requires the management of re-entrants, patients who return to see the same doctor after taking some investigative tests or treatments during a single visit to the ED. These are not patients who return to the same hospital on another occasion. The re-entrants have different service distributions from the patients in their first consultation with a doctor. Standard queuing theory cannot deal with re-entrants in a tractable way. Lastly, patients with the same acuity classification [1] are not homogeneous, and experience their visit and each part of the process differently. For example, some patients may require a blood test while others require an X-ray or no tests.

Having complex queue characteristics makes managing the patient queue and planning for resources in the ED challenging. A typical fixed queuing policy such as FIFO and a fixed resource schedule is inflexible and unable meet varying demand or adapt to operational deviation from the expected service time of a patient (if a patient requires more attention than expected). Our intention in this dissertation is to combine use of the dynamic priority queue and dynamic resource adjustment. In order to ensure that our results can be applied realistically to a real-world context, we use simulation to develop workable models based on hospital setup and consider the stochastic variability of patient arrival, processing time, need for investigative tests and treatment events, and processing time. The simulation model then uses real-life data derived from a dataset obtained from a hospital in Singapore. A simulation approach allows us to model and analyze complex ED processes which are otherwise intractable with analytical queuing models.

[1] There are four levels of acuity, P1 to P4. P1 patients are in a critical condition, P2 patients have serious conditions, P3 and P4 are non-emergencies with moderate to mild conditions

## 1.2 Motivation

The main motivation of this dissertation is to explore innovative ways to minimize operational changes (e.g., process changes) for practitioners but to manage the queue in the ED with two key aims for hospital service quality:

1. No demand restriction. Patients can come to the ED at any time and in any condition. Also, ambulance diversion is not advocated.
2. To serve patients within the parameters of the desired service quality specified by the hospital.

The reason for not restricting demand is to meet the needs of the patient in a small country like Singapore where patients may have a limited choice of treatment centers. Although our proposed models are generic and should be applicable to hospitals in general, we would like the proposed methods to be directly applicable to public hospitals in Singapore. Some government agencies only recognize the medical certificate given by doctors from public hospitals. It was also reported in an article in a leading newspaper [38] that a public campaign advising members of the public not to go to hospital for every ailment had failed. Hospitals have to be able to cope with the demand. Therefore, we explore ways that will improve service quality without resorting to demand restrictions.

Desired service quality is measured by the length-of-stay (LOS) of a patient in the ED from registration until readiness for discharge from the ED (the patient may still be admitted to the wider hospital). There is a large number of possible metrics that can be used to evaluate the performance of an ED. In a comprehensive survey on optimizing ED front-end operations [56], out of 54 pieces of work presented, most used metrics for the performance of the ED in LOS. The other frequently used metrics include wait time, number of patients left without being seen (LWBS), patient satisfaction and staff satisfaction. In addition, other research has also shown that timeliness of care (wait time or LOS) has a strong correlation with patient satisfaction [7]. There are also reports that show that poor patient satisfaction leads to decreased staff satisfaction [40] and decreased physician productivity [43]. We selected LOS over wait time because we wanted to measure the amount of time a patient spent in the ED, including the time he/she takes in a test/treatment and the review session with the doctor. Wait time usually treats each entry to the queue (doctor's queue) as a separate queue instance. We modeled only the queue to the doctor's consultation as findings in Boudreaux et al. [6] shows that the wait time to be treated by a physician has the most powerful association with satisfaction.

We asked whether, with the motivation to satisfy the desired service quality of the hospital, queue management in an ED could be more innovative so that customers are better served, without the need to restrict demand.

## 1.3 Objective

Our objective in this dissertation is to improve operational responsiveness of an ED by managing the patient in the queue and providing better decision-support to determine the number of doctors required. We propose an integrated Dynamic Queue Management (DQM) Framework that contains improvements in two key aspects: (1) managing the demand by means of a dynamic priority queue and (2) managing the supply by means of dynamic resource adjustment (i.e., the supply of doctors). The overview of the framework is as shown in Figure 1.1. To achieve this, we leverage on *real-time* information on patient arrivals and apply heuristic decision-making methods for planning and scheduling, combined with queuing theory and simulation. Such methods do not require process change and are generally transparent to patients.

## 1.4 Thesis Positioning

In this thesis, our aim is to model the ED process as close to the real world as possible so results can be reflective of real life and methods directly applicable to real life. Real-world ED process is complex. Traditional analytical methods such as queuing theory are insufficient for the modeling of such complexity.

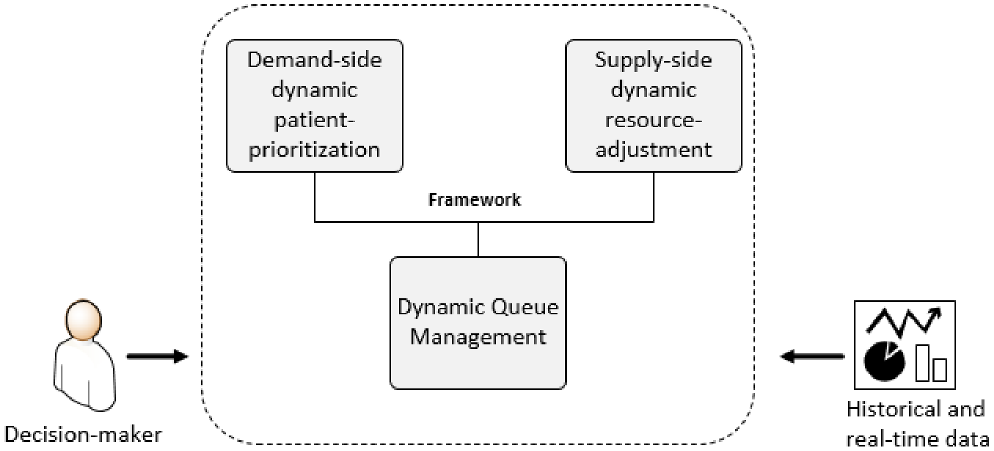

Demand-side dynamic patient-prioritization
Supply-side dynamic resource-adjustment
Framework
Dynamic Queue Management
Decision-maker
Historical and real-time data

Figure 1.1: Overview of the Dynamic Queue Management Framework

Such a challenge requires inter-disciplinary methods. In addition, we aim to analyze a hospital's data and provide system views of how a hospital can use its data to create business insights and also have a plan to implement some or all of our proposed methods. Our proposed Dynamic Queue Management Framework is an inter-disciplinary approach that draws on the disciplines as shown in Figure 1.2.

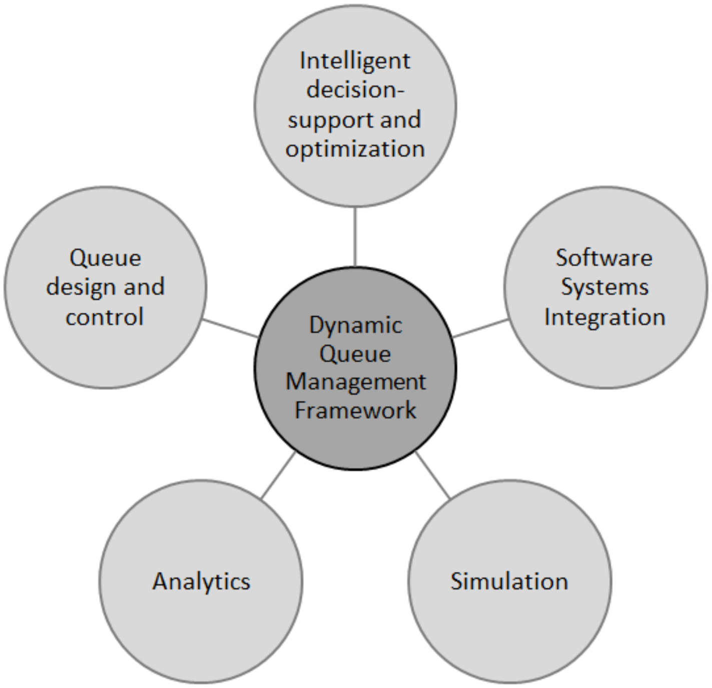

Intelligent decision-support and optimization
Queue design and control
Software Systems Integration
Dynamic Queue Management Framework
Analytics
Simulation

Figure 1.2: Multi-Disciplinary DQM Framework

The following describes how we make use of each of these disciplines in our framework:

- Queue design and control: To model the process and the queue to the doctors such that we can also benchmark the methods against some of the known analytical queuing models. The queue control considerations are to look into how we can dynamically manage and clear the queue to meet desired service levels.

- Intelligent decision-support and optimization: To formulate our problem as a constrained optimization problem that incorporates service level and other constraints. In addition, our optimization model enables decision-makers to make decisions on what parameters to use for the ED process to achieve the targets.

- Software systems integration: To evaluate the IT systems that are required to support the proposed strategies. We also provide a road map and plans to show what live systems information is required to deploy the methods and potential changes to the existing systems.

- Simulation: To provide a realistic platform to evaluate the performance of the different proposed methods.

- Analytics: To understand the trends and demographics of the patients, how the ED performs and to get the parameters that are essential to the queuing models. Also, to study the results of the simulations.

## 1.5 Contribution

Three pieces of work form the Dynamic Queue Management Framework.

### 1.5.1 Demand Perspective

We introduce the concept and a case for a dynamic priority queue model where the priorities of the entities within the queue system are recalculated when one or more resources serving the entities become available. Here, we addressed the following questions:

1. How do we prioritize the patient such that the average LOS of all patients is minimized?
2. How do we give priority to patients who will potentially have a shorter consultation with the doctor instead of having them wait for others who may occupy the doctor for an extended time?
3. How can we give priority to patients who have spent a long time in the ED, especially those waiting for periods beyond the hospital's desired service duration?
4. How can we ensure that priorities given are unbiased and that patients do not wait indefinitely (prevent starvation[2])?
5. How can we design a way to automatically calculate priorities objectively?

Our calculation of priority of a patient in the queue is based on one or more of the following factors:

1. The patient's estimated consultation time with the doctor and/or
2. The patient's remaining time in order to meet the desired service quality

We propose a queuing model that intelligently allocates patients to doctors based on the factors listed above so as to reduce the average LOS of all patients in the ED. The model is also extendable to include other factors. We found that our proposed strategies resulted in shorter LOS. We also provide analysis and results to allow healthcare decision-makers to cope with starvation, and select a strategy that best suits their implementation readiness.

[2] Starvation is a condition in which a patient is deprived of the chance to consult with a doctor due to multiple preemptions by other patients with higher priorities.

### 1.5.2 Supply Perspective

In this section, we introduce a number of strategies to dynamically adjust doctor allocation in the ED. Here, we address the following questions:

1. How do we determine the number of doctors required to meet the desired service quality based on known information about patient arrival trends in the past?

2. How do we respond to uncertainties (e.g., surges in demand) by leveraging real-time data?

3. When do we adjust the number of doctors, subject to the physical room constraints of a typical ED?

4. How do we ensure that the quality of care is not compromised, especially to the medically critical patient?

We present four staffing strategies, namely, historical (HIST), dynamic (DYN), historical with optimization (HIST-OPT) and dynamic with optimization (DYN-OPT) to address the above questions. Both HIST and HIST-OPT use only historical data, while DYN and DYN-OPT use both historical and real-time data to calculate staffing required. Our results showed that the dynamic strategies could indeed better cope with demand surges. HIST-OPT can potentially provide a doctors' schedule that meets the hospital's desired service quality with a slight increase in the number of doctors to be deployed in the ED. DYN-OPT provides opportunities to obtain more stable schedules with the ability to respond to changes. DYN performs better than DYN-OPT since it is more reactive. Similarly, we presented an analysis for healthcare decision-makers to select a strategy that is most suitable based on their quality improvement appetite and implementation readiness. Our proposed dynamic resource adjustment strategies are data-driven and provide invaluable real-time decision support to ED operations.

### 1.5.3 Integrated Dynamic Queue Management

Finally, we provided an integrated framework to combine the benefits brought by the strategies from both demand and supply perspectives. Here, we addressed the following questions:

1. What are the effects of combining strategies from both perspectives?

2. Do the strategies work together?

3. Which combinations should a hospital select?

Our key contribution in this work is the ability to seamlessly integrate strategies from both perspectives. The supply-side strategies perform well with each demand-side strategy. Similarly, we provide analysis to help decision-makers select the strategies that suit them.

## 1.6 Research Methodology

Figure 1.3 gives a schematic diagram of our research methodology and process.

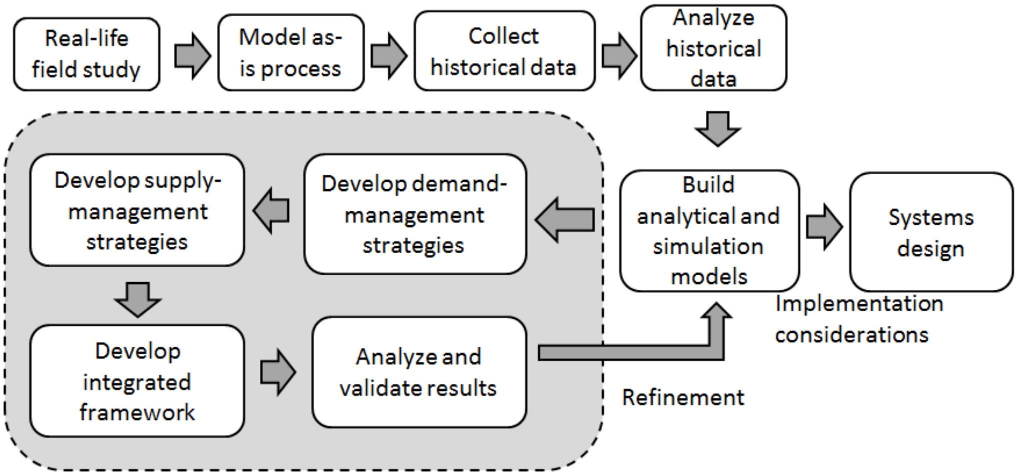

Real-life field study
Model as-is process
Collect historical data
Analyze historical data
Build analytical and simulation models
Systems design
Develop demand-management strategies
Develop supply-management strategies
Develop integrated framework
Analyze and validate results
Implementation considerations
Refinement

Figure 1.3: Research methodology

We have taken a strong practice approach to our research. We started with a field study at a selected hospital, and then mapped the physical setup of the ED and gathered inputs on ED processes, the resources (human and facilities) involved and its challenges to meet service level target. During the field study, we also interviewed ED staff and doctors. After the field study, we modeled the as-is ED process and sought verifications from the hospital. Historical data about the ED process was then collected and analyzed using an analytics tool to obtain process parameters such as arrival rates, service rates, types of investigative tests and treatment.

With the information about the ED processes and its resources, we began to build analytical and simulation models to address the hospital's intention to meet the service level target. Here, a complex ED process was simplified to the extent that it could be implemented in a simulation model but still suffice for meaningful analysis. First, we developed demand-side strategies and built a simulation prototype to analyze the performance of the strategies. Next, we developed supply-side strategies and built a simulation prototype to analyze the performance of the strategies. Finally, we designed an integrated framework, an integrated simulation prototype that executed both demand-side and supply-side strategies. The performances of the various combinations of demand-side and supply-side strategies were plotted on graphs and evaluated. We took further steps to rank the strategies by finding the statistical significance of the performances of any pair of strategies that appeared similar on graphs. Other aspects of the performances of the strategies were the cost of deploying doctors and implementation complexity. The results of these metrics were also presented in tables, charts or quadrant analysis. Based on the results, we iteratively refined the analytical and simulation models, and analyzed the results. The final step in our approach considered implementation possibilities and provided designs to implement the strategies as intelligent decision-support systems.

In addition, the methods and results were presented to the ED of two public hospitals in Singapore. Feedback was collected, considered, and (where possible) incorporated into the models in our iterations. We also learned that there were many intangible considerations in the healthcare industry.

## 1.7 Dissertation Structure

This dissertation document is structured as follows. In Chapter 2, we present the context and scope of our study by showing the setup and processes in the ED. In Chapter 3, we present a comprehensive literature review of the related work in cross-disciplinary areas of intelligent systems, decision-support, optimization, simulation and queuing theory. We study the work in the domain of healthcare, with a focus on emergency departments. In Chapter 4, we discuss the model, methods and experimental findings for our proposed demand-side strategies to dynamically prioritize the patients in the queue. In Chapter 5, we discuss the model, methods and experimental findings for our proposed supply-side strategies to dynamically adjust the resources (doctors) in accordance to the demand (arrival of patients). In Chapter 6, we show how both the demand-side and supply-side strategies can be seamlessly integrated into a single Dynamic Queue Management Framework. We discuss the framework, model, and experimental findings for the integrated Dynamic Queue Management Framework. We also provide a road map to implement the proposed strategies. Finally, in Chapter 7, we offer a summary of the contributions of the dissertation, other practical considerations in the healthcare domain and future work.

## 1.8 Chapter Summary

In this chapter, we showed the challenges of analyzing queues and managing queues in the ED due to its complexity. We presented our motivation, objectives, contribution and approach to address the issues.

# Chapter 2

# Scope of Study

The details of a real-life case in this chapter illustrate the scope of the problem addressed in this dissertation.

## 2.1 A Real-life Case

A real-life study is conducted in the ED of a selected local hospital. In this ED, based on national guidelines, the patients are classified into four acuity categories, namely P1, P2, P3 and P4, with P1 and P2 being emergency patients. P1 patients are critically (life-threateningly) ill and must be attended to immediately. P2 patients are those in great pain and must be attended to within 20 minutes. The P3 and P4 patients are considered non-emergency patients with moderate and mild illnesses. On arrival, a P1 patient is sent to the critical-care area immediately for treatment or resuscitation. Most P2 to P4 patients go through registration and triage before seeing a doctor. The acuity category of a patient is determined by a nurse during the triage sub-process.

The department (shown in Figure 2.1) is divided into two areas for patient care; the critical-care area manages P1 and P2 patients while the ambulatory area (clinic rooms) manages P3 and P4 patients. Non-emergency patients represent 70% of the workload of the ED under investigation. P3 and P4 patients are considered lesser emergencies in comparison to P1 and P2, and the relatively straightforward nature of the patients' conditions presents the opportunity for implementation of improvements and the maximizing of efficiency.

We limit our **scope of study** to the **consultation process** in the **ambulatory area**. Hospital management requires that the ED serve ambulatory area patients to a specified desired level, for example, within an LOS of 60 minutes.

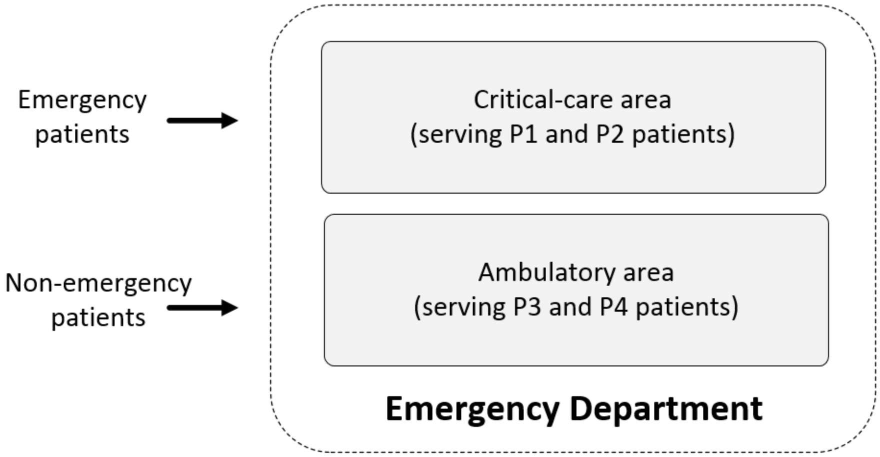

Emergency patients
Non-emergency patients
Critical-care area (serving P1 and P2 patients)
Ambulatory area (serving P3 and P4 patients)
Emergency Department

Figure 2.1: Logical segregation of work areas in ED at a local hospital

## 2.2 The ED Process in the Ambulatory Area

The patient's LOS is the time between the start of registration and the end of the case when the patient is either discharged or admitted as an in-patient. It consists of several sub-processes, namely registration, triage, consultation with a doctor, investigative tests and treatment, and discharge or admission. The registration sub-process involves the recording of a patient's personal information as well as the collection of a standard fee for use of ED services and standard medications. The patient then proceeds to the triage sub-process during which his/her condition is assessed by a nurse and is assigned an acuity category. The patient then consults a doctor. During the consultation, the doctor may order one or more investigative tests such as a blood tests, X-rays and point-of-care tests (e.g., electrocardiogram, urine, eye and hearing tests). The doctor may also order on-site treatment of the patient, such as the taking of oral medication, or the application of a bandage. Treatment could also include a period of observation. If a blood test is ordered, the doctor will draw the blood before releasing the patient to await the blood test result or to proceed to undergo other tests or treatment. The investigative tests and treatment sub-process consists of highly variable steps that differ greatly between patients. A patient could take one or more tests but receive no treatment, or could take no test but receive one or more treatments, or a combination of tests and treatments. When the test results are ready and the patient has completed his/her treatment, he/she is to be seen again by the same doctor. The patient re-enters the patient queue to await the doctor, and such patients are called **re-entrants**. During the second consultation, the doctor reviews results with the patient, reviews the patient's condition and decides whether the patient is to be discharged or admitted as an in-patient. During the discharge sub-process, some patients may require a referral to a specialist clinic for further follow-up or pay additional fees for non-standard procedures or medication.

We modeled the various sub-processes in our ED process as shown in Figure 2.2, without the details within the sub-processes, which were only useful if we were interested in exploring changes in the process sequence or the specific physical design of the ED (e.g., adding an X-ray room). However, a detailed process requires the collection of a huge amount of data for all the activities in the process. To ease the data-collection process, we found that our process model in Figure 2.2 was sufficient for capturing process information required to evaluate patient-prioritization in the patient queue and the supply requirements of doctors.

Figure 2.3, based on three months' data from the hospital, shows the types of patients who take different routes through the ED. The term "basic" refers to the steps that all non-emergency patients have to go through, namely, registration, triage, consultation and discharge or admission. For example, a patient may take path Number 9 "Basic + L + R only". This means that the patient will go through registration and triage, then the first consultation with the doctor, followed by a lab (blood) test and radiology (X-Ray) test (in any sequence). Then a review consultation with doctor will occur before the

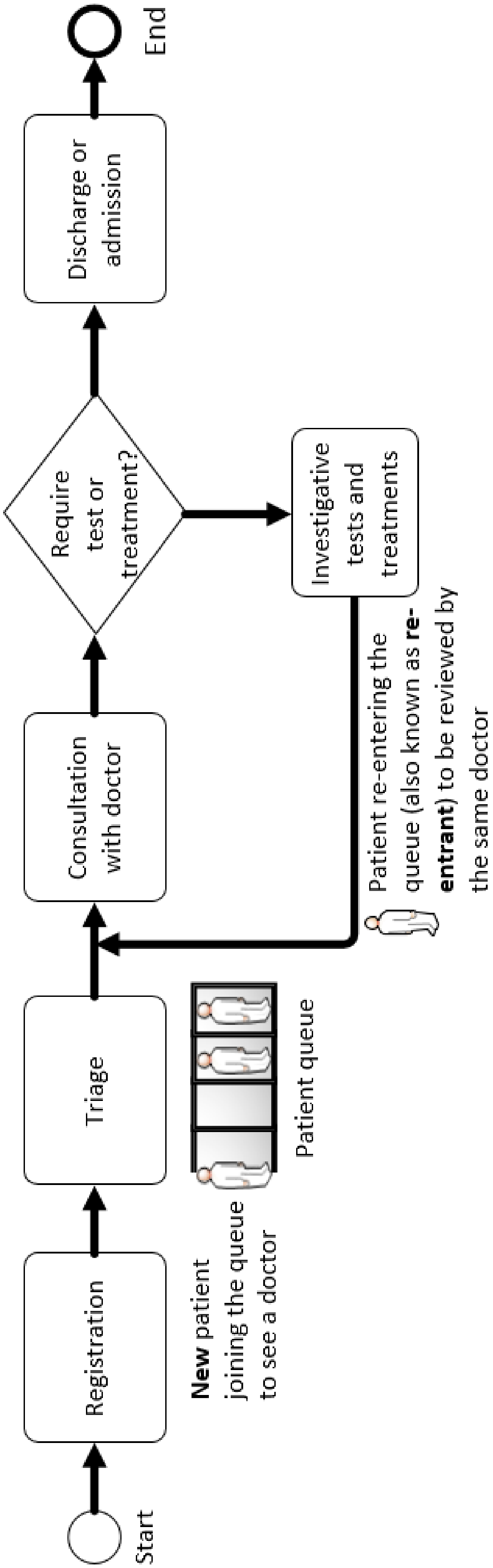
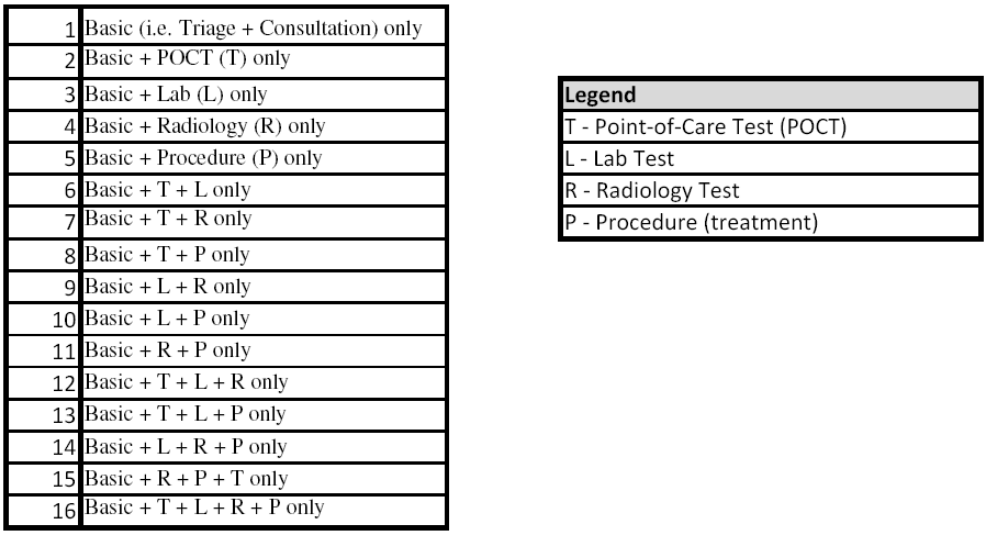

Start
Registration
Triage
Consultation with doctor
Require test or treatment?
Investigative tests and treatments
Discharge or admission
End
New patient joining the queue to see a doctor
Patient queue
Patient re-entering the queue (also known as re-entrant) to be reviewed by the same doctor

Figure 2.2: Simplified process of the ED

patient is discharged or admitted as an in-patient.



| | |
|---|---|
| 1 | Basic (i.e. Triage + Consultation) only |
| 2 | Basic + POCT (T) only |
| 3 | Basic + Lab (L) only |
| 4 | Basic + Radiology (R) only |
| 5 | Basic + Procedure (P) only |
| 6 | Basic + T + L only |
| 7 | Basic + T + R only |
| 8 | Basic + T + P only |
| 9 | Basic + L + R only |
| 10 | Basic + L + P only |
| 11 | Basic + R + P only |
| 12 | Basic + T + L + R only |
| 13 | Basic + T + L + P only |
| 14 | Basic + L + R + P only |
| 15 | Basic + R + P + T only |
| 16 | Basic + T + L + R + P only |

| Legend |
|---|
| T - Point-of-Care Test (POCT) |
| L - Lab Test |
| R - Radiology Test |
| P - Procedure (treatment) |

Figure 2.3: Patients may take different paths after the first consultation

## 2.3 The ED Queue in the Ambulatory Area

The existing patient queue (to the doctors) is managed as a single first-in-first-out (FIFO) queue for new patients but in a somewhat ad hoc manner for re-entrants. It has multiple servers (doctors). The queue capacity is infinite (no turning away of patients). The arrival rates were found (based on our analytics results) to be non-homogeneous (time-varying) Poisson processes. The hourly arrival rate followed an exponential distribution, having different arrival rates over a week's horizon. We observed that Sundays and Mondays had a higher volume of patients. Each day, the time-varying pattern was fairly similar. The low demand period was between 1am and 8am daily. The peak period was between 9am and midnight. The midnight-to-1am and 8am-to-9am periods had moderate demand. An example of the time-varying arrivals over a week is shown in Figure 2.4. The x-axis shows the day of the week starting from Sunday and the y-axis shows the number of patients arriving in the hour. Typically in queuing models, the symbol used is $\lambda$ and the symbol t in the brackets indicates the time variable. The doctors' schedule is static and is planned manually based on perceived understanding of the demand in the ED at various hours of the day, over an entire week or month. An example of a static doctors' schedule for a day is shown in Table 2.1, with the number of doctors (servers) rostered for the given time of the day.

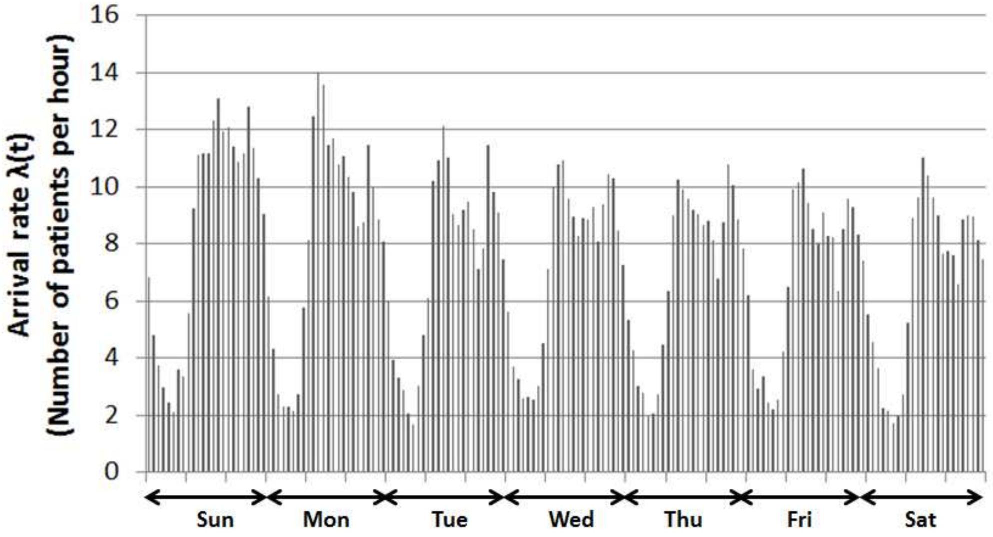


Figure 2.4: Time-varying arrival to ED

| Time | 12am - 8am | 8am - 10am | 10am - 6pm | 6pm - 12am |
|---|---|---|---|---|
| Number of doctors | 2 | 3 | 4 | 3 |

Table 2.1: Example of a static doctors' schedule for a day

We make further assumptions on the queue. The doctors are modeled to be homogeneous and have the same service rates for the same type of patient (new or re-entrants). Serving the patient is non-preemptive in nature. We consider balking (customer will not join the queue if queue length is more than a specified value) as a pre-arrival process, hence net arrival rates are used in our analysis. We do not consider reneging (customer leaving the queue if he/she has waited for more than a specified amount of time) in our model.

In our attempt to model the queue as closely as possible to a real-world queue, we use a single combined FIFO patient queue to the doctors in our model. This somewhat differs from the real-life mechanism which serves the re-entrants in an ad hoc manner. To verify that a FIFO estimate is sufficient to represent the real world, we ran a verification experiment using our simulation prototype with a FIFO patient queue and a static doctors' schedule. The outcome of the experiment showed that the differences in mean and standard deviation of the actual observed hospital data and the results of our simulation prototype were found to be less than 5% and 10% respectively. The ranges of average LOS (i.e., minimum and maximum) were also consistent. Therefore, we conclude that the results of our simulation prototype are representative of the performance of the ED process.

## 2.4 Chapter Summary

In this chapter, we showed the scope of the process being investigated, parameters to the process (e.g. trends of patient arrival) and some basic assumptions on the process that will serve as the basis of all our work in the dissertation.

# Chapter 3

# Literature Review

This dissertation draws inspiration from various cross-disciplinary areas. Here, we provide a literature review of three main areas. They are:

- Queuing (analytical) and simulation approaches to studying ED processes
- Demand-management methods
- Supply-management methods

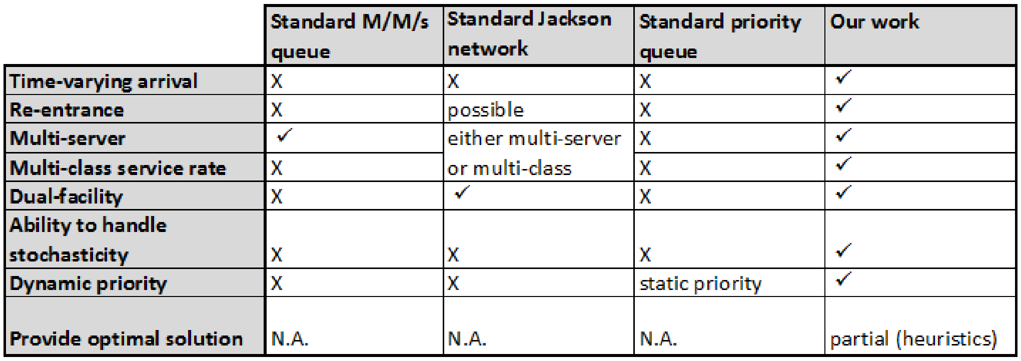
## 3.1 Queuing and Simulation Approaches to Studying ED Processes

In the extensive literature related to improvements in ED processes, we see two major approaches, namely, queuing theory and a simulation approach. Each has its advantages and limitations.

### 3.1.1 Queuing

It is generally known that queuing models are simpler, require less data, and provide more generic results than simulation, as shown in Green [18]. However, queuing models are sensitive to their parameters (e.g., arrival rates, queue discipline) and unable to capture the complexity and detail that are often required in many real-world applications.

As we have seen in the previous chapter, modeling the ED process is complex, having to deal with stochastic arrivals, a multi-stage configuration that is multi-server at each stage, re-entrants, and customers (patients) with different priorities (multi-class). There is no standard queuing theory that can address the complexities imposed by a general ED process. The standard M/M/s queue is not appropriate. The closest matches are Jackson network and priority queues. Jackson network can be used as the basis for analysis of very special cases of the process but it is still insufficient for our analysis. Jackson network is unable to handle priority queues and it is a FIFO queue at each node. The standard analytical models provide only multi-server analysis or multi-class analysis. They are unable to handle both. The standard priority queue's analytical model handles only a single server with service time as a distribution or multiple servers but require a single constant service time. See our comparison table given in Figure 3.1. None of these cater to our requirements in the ED process.



| | Standard M/M/s queue | Standard Jackson network | Standard priority queue | Our work |
|---|---|---|---|---|
| Time-varying arrival | X | X | X | ✓ |
| Re-entrance | X | possible | X | ✓ |
| Multi-server | ✓ | either multi-server or multi-class | X | ✓ |
| Multi-class service rate | X | | X | ✓ |
| Dual-facility | X | ✓ | X | ✓ |
| Ability to handle stochasticity | X | X | X | ✓ |
| Dynamic priority | X | X | static priority | ✓ |
| Provide optimal solution | N.A. | N.A. | N.A. | partial (heuristics) |

Figure 3.1: Comparison of our work with standard queuing theory

Despite the limitations of using queuing theory to model ED processes, a rich repository of academic work on EDs has used queuing theory, such as in Halfin et al. [21], Worthington [57] and Pajouh et al. [37]. Comprehensive surveys can be found in Green [18] and Fomundam et al. [14].

Pure analytical or mathematical models usually provide long-term average solutions, which do not help in a hospital's planning process. For instance, numerous researchers have attempted to model the ER using mathematical models [17]. However, the models of these attempts are often deterministic and do not account for the variability in the processes. Modeling and simulation serve as the most appropriate solutions in such circumstances.

### 3.1.2 Simulation

Simulation permits the modeling of the details of complex ED processes and their dynamics. However, simulation techniques are lacking in supporting analytical models. Simulation requires knowledge about the data, and thus a large time investment to obtain and analyze the data. The simulation approach is less sensitive to parameters. Simulation requires the development of a simulation model that is a close representation of the real system under investigation. Jacobson et al. [25] present a list of steps that must be performed carefully to model each healthcare scenario successfully using simulation, and warn about the slim margins of tolerable error and the effects of such errors being lost lives.

Mayhew and Smith [32] used queuing theory to analyze a four-hour completion time target in EDs in the UK. Setting of completion target is similar to our target LOS. We, however, have a more challenging target of one hour and hence we need intelligence to improve the process, such as dynamic prioritization to dispatch the right patient to the doctors.

Komashie et al. [28], Pajouh et al. [37], Samaha et al. [44] and Gunal et al. [20] offer examples of discrete-event simulation to enable complex problems to be analyzed. In these studies, the simulator was used as a tool to verify the proposed models and perform what-if analysis to improve the processes, or initiate more effective staff planning in the ED. The simulator, however, was not used in real time for decision support. The work of Zeltyn et al. [59] used a simulator as a tool to test which staffing method was suitable in real time. One idea is *symbiotic simulation*, first proposed by Fujimoto et al. [16]. A symbiotic simulation system consists of a simulation model interacting with physical systems. A more detailed architecture of how a symbiotic simulation systems fits into a landscape of physical systems can be found in Low et al. [29]. “The simulation system benefits from the continuous supply of the latest data and the automatic validation of its simulation outputs, whereas the physical system benefits from optimized performance obtained from the analysis of simulation experiments.” [29]. Another related work is concerned with the optimization of an objective function using simulation, generally termed *simulation optimization*. (For details, we refer the reader to the comprehensive survey on simulation optimization in Fu [15].) An example of using simulation combined with simulation optimization in an ED context is shown in April et al. [2]. This paper shows that an optimal solution can quickly create a solution to help the ED to determine the optimal number of resources (doctors, nurses) to deploy. One aspect of our work is to propose resource-allocation strategies that use symbiotic simulation with simulation optimization to determine the number of doctors required in the short-term resource plans in real time.

### 3.1.3 Combination of Simulation and Queuing

Many works have applied both simulation and queuing simultaneously. Tucker et al. [54] offer an example that uses simulation to validate, refine and complement results obtained by queuing theory. Albin et al. [1] use queuing theory to get approximate results and then use simulation models to refine them.

In this dissertation, we use queuing theory to model the ED process and we use simulation in two ways: Firstly to verify the results of our queuing model, and secondly to optimize the ED process using symbiotic simulation to perform short-term resource planning.

## 3.2 Demand-Management Methods

For the demand-management methods, we examine the existing literature that deals with managing demand in an ED. The demand perspective of the ED deals with restricting, directing, and managing the flow, or prioritizing the patients.

### 3.2.1 Restricting or Directing Patients

Ambulance diversion is a well-known method used by hospitals to spread the demand load across various hospitals. A dynamic collaborative approach to route ambulances to avoid overcrowding is shown in Barthell et al. [3]. In this method, the demand is *reduced* by having ambulances routed to other hospitals. However, such a method may not work well in an urban context where all the hospitals are equally busy.

### 3.2.2 Managing Patient Flow

Another approach is to manage patient flow by directing patients by means of *fast-track systems* (Rodi et al. [42], Sanchez et al. [45] and Roche et al. [41]). This approach creates separate staffed areas for the care of patients identified as low-acuity during triage. Our consultation with a public hospital in Singapore which used to implement the fast-track system showed that it was not a practical approach because it gives the public a false impression that non-emergency patients are given higher priority than those patients with higher acuity. As this impression prevails, more non-emergency patients who can be treated by general practitioners appear at the ED and take up the resources required to serve those patients who truly require services there.

A similar approach, *directed queuing*, assigns a doctor to the triage subprocess to make decisions to direct patients to appropriate stations [34]. Another categorization method is shown in King et al. [27], where reducing patient waiting time in the ED is done through triage systems to categorize patients into different groups and treat them differently. The work in Chakravarthy et al. [9] used a threshold for dynamically switching between two types of customers to enter the service. Directed queuing requires interactions with qualified medical doctors to achieve proper classification. We found that, in practice, in the context of our local hospitals, it is preferable to place scarce resources such as doctors in the consultation area where they can focus on diagnosis and treatment of the patient.

In addition, the fast-track system and directed queuing methods require changes to core processes and possibly the physical layout of the ED to accommodate these processes. It is our interest in this dissertation to keep such changes to a minimum as they usually require careful change management (training, organization changes and process acceptance) to ensure smooth execution.

Another approach, proposed by Shi et al. [46], was to improve the performance of an ED by reducing the *ED boarding time*, the time from an admission request to the transfer to an in-patient bed. It was found that the average waiting time was more than four hours for patients who requested a bed between 7am and 11am. This was because there were no beds immediately available in the in-patient general wards as most discharges happened between 2pm and 3pm. Hence, the ED patients had to stay in the ED, cared for by ED resources, resulting in ED overcrowding. The authors introduced an *early discharge campaign* to increase the number of in-patients discharged before noon, making beds available for bed requests from the ED in the morning. Although this solution alleviated ED overcrowding by changing the processes in other parts of the hospital, we recognize that this effort required collaboration between many departments. It would certainly involve a large hospital-wide effort on business process re-engineering and change management.

The *priority queue*, a form of managing patient flow, has been widely studied. McQuarrie [33] shows that giving priority to patients who require shorter service times can reduce waiting times. The challenge is to address the issue of starvation, the perceived unfairness (unless that class of patient is given a dedicated server) and the difficulty of estimating service times accurately. Siddhartan et al. [47] proposed a priority discipline for different categories of patients and then a FIFO discipline for each category. They found that the priority discipline reduced the average wait time for all patients; however, while the wait time for higher priority patients was reduced, lower priority patients endured a longer average waiting time. Fiems et al. [13] investigated a preemptive repeat priority queuing system in which emergency patients may interrupt scheduled patients. This work is based on a single server queue only. In Hay et al. [22], authors allowed a patient's priority to be dynamically updated based on the waiting time and the patient's underlying clinical priority. They did not consider re-entrants (patients seeing the same doctor multiple times) and also used a static priority queue. We will study dynamic priority queues and also consider re-entrants.

## 3.3 Supply-management methods

For the supply-management methods, we explored queue design, queue control, and staffing methods for handling time-varying arrivals.

### 3.3.1 Queue Design and Control

For queue design and control, there are studies on managing service operations with dual facilities, in our case a back room and front room. The main idea in queue control is to prevent a queue from building up by dynamically changing resource capacity. In another words, queue control is a technique applied to ensure that the queue satisfies some criteria, such as the wait time for all patients in the queue being at most $x$ minutes.

In a retail shopping context, Berman et al. [5] presented a scenario where the front room had shoppers and a check-out station, and the back room had other indirect work in which the customer was not directly involved. They used a staff-switching policy, where resources in both rooms were shared and back-room resources could be switched to the front room in order to control the queue so that wait-time was within an acceptable value. This is a three-dimensional queue with shoppers in the shop (not in the queue but potentially joining the queue), and the customers who are in the queue in the check-out line. The number of servers attending to the queue forms the third dimension. The paper presented a heuristic-based solution to find the optimal switching policy while having the ability to fulfill the back room minimum requirement.

Terekhov et al. [52] built on Berman et al.'s [5] work by introducing a constraint-programming approach to the solution. They claimed to be first to find proven optimality in solving the switching problem. However, they did not consider the concept of shoppers and the solution is restricted to a two-dimensional queuing model (number of people in the queue and the number of servers). Our ED process is three-dimensional as the patients who are undergoing the investigative tests and/or treatment are potential customers who will be (re-)joining the queue, as in the work of Berman et al. [5]. The other limitations of these studies that restrict us from directly applying them to ED process are: (1) strong assumption of single arrival rate and single service rate; (2) no consideration of time-varying arrival rates; (3) assumed homogeneous customer with single priority; (4) assumption that back-room demand is homogeneous and hence the resource requirement in the back room is homogeneous; and (5) no consideration of re-entrant customers.

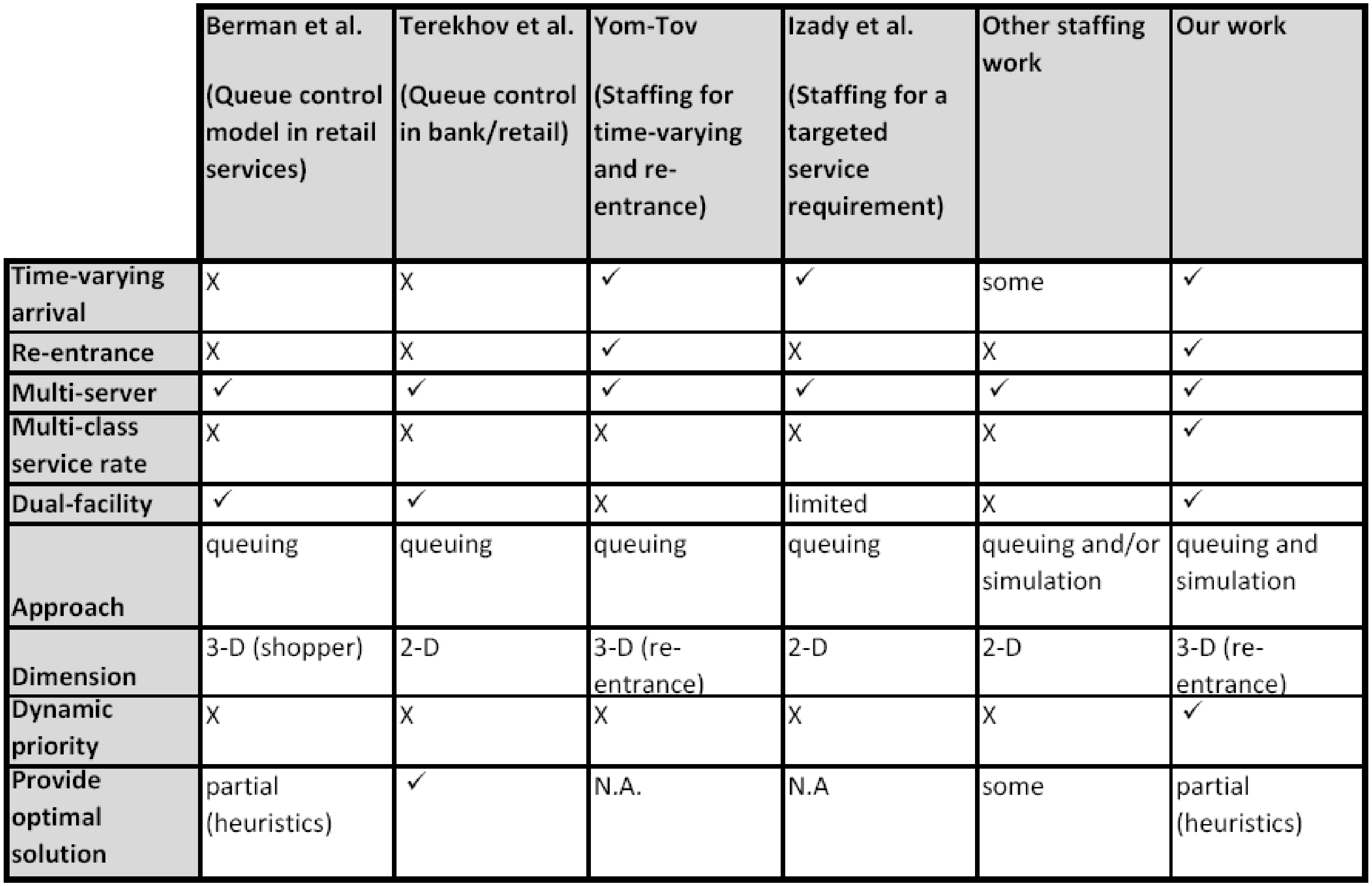
### 3.3.2 Staffing

Staffing is a well-studied problem from the supply perspective. Some approaches used simulation to design proactive staffing policies [8, 44]; some used analytical methods with considerations of time-varying arrival [12, 19, 26, 58]. In the work of Sinreich et al. [48], algorithms were presented to shift resource capacity from low-demand hours to peak hours. These are proactive methods which produce staff schedules before execution.

Real-time (dynamic) staffing has received considerable attention in recent years. Marmor et al. [30] and Thorwarth et al. [53] used simulation as a backend engine to provide decisions to real-time staffing.

There are other classical methods to address staffing problems with time-varying arrivals. There are two major approaches: one is suitable for short service-times and another for long service-times. The methods in the former approach are steady-state approximations such as PSA (Piecewise Stationary Analysis), RCCP (Rough Cut Capacity Planning) [55], Stationary Independent Period by Period (SIPP), or lag-SIPP [19, 26]. The methods for the latter approach are MOL (Modified Offered Load) [26] and ISA (Infinite Server Approximation) [12]. In a more recent study in Izady and Worthington [24] showed how staffing requirements were tailored to meet the UK's national target of 98% of ED patients completing their total consultation within four hours. The ED process was modeled as a network of K stations with time-varying Markovian arrival with general service times. The approach used was MOL and staffing requirements were calculated using square-root staffing rules for each station in the network. In this method, the first and second consultation with a doctor were modeled as different nodes and hence the staffing was not consolidated, although the authors provided a method to attempt to have consolidated staffing. Therefore, we noticed that none of above methods addressed the re-entrance needs of the ED process. Yom-Tov [58] showed that both time-varying arrival and re-entrance can be modeled using an Erlang-R queue model. In that study, the results showed that modeling the re-entrants provides a staffing policy that gives more stable performance in terms of probability of wait. Time-varying arrival was handled using the MOL approach and staffing policy was based on a square-root staffing formula. Although that study did not consider using real-time information to influence staffing, it served as a foundation and benchmark for us as we build our dynamic resource-adjustment strategies.

The switching problem with time-varying arrival and re-entering patients is a complex process that is not easily characterized using analytical models. In our approach to make supply-side staffing as close to real life as possible, we model the problem as an optimization problem and solve it via a combination of simulation and heuristics. We adopt some of the findings in Yom-Tov [58] to provide an initial solution to our local search algorithm, which we then apply to obtain an improved solution for our problem. In addition, we propose the use of simulation in real time to make predictions on future demand upon which our dynamic allocation policy is based. Figure 3.2 shows a summary of the comparison between our work and the existing literature.



| | Berman et al.<br>(Queue control model in retail services) | Terekhov et al.<br>(Queue control in bank/retail) | Yom-Tov<br>(Staffing for time-varying and re-entrance) | Izady et al.<br>(Staffing for a targeted service requirement) | Other staffing work | Our work |
|---|---|---|---|---|---|---|
| Time-varying arrival | X | X | ✓ | ✓ | some | ✓ |
| Re-entrance | X | X | ✓ | X | X | ✓ |
| Multi-server | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Multi-class service rate | X | X | X | X | X | ✓ |
| Dual-facility | ✓ | ✓ | X | limited | X | ✓ |
| Approach | queuing | queuing | queuing | queuing | queuing and/or simulation | queuing and simulation |
| Dimension | 3-D (shopper) | 2-D | 3-D (re-entrance) | 2-D | 2-D | 3-D (re-entrance) |
| Dynamic priority | X | X | X | X | X | ✓ |
| Provide optimal solution | partial (heuristics) | ✓ | N.A. | N.A | some | partial (heuristics) |

Figure 3.2: Comparison of our work with existing queue design and control literature

## 3.4 Demand and Supply Management Methods

As we saw in previous sections, the methods that attempted to address ED issues are rich from both the demand and supply perspectives. However, there is no study that integrates queue management from both perspectives in the context of EDs.

We draw inspiration from the traffic/civil engineering and economics domains, where the performance of the system is managed via controls from both demand and supply perspectives. A key model in traffic engineering is the DynaMIT system `http://mit.edu/its/dynamit.html` [4]. It is a real-time model of traffic conditions on roads (supply) and drivers' travel requirements (demand) to predict and generate strategies to guide drivers towards optimal route-choice decisions. In economics, an example of managing queues by adjusting demand and supply is shown in Mendoza et al. [35]. The methods used in both of these domains are not directly applicable to the context of ED.

In upcoming chapters we will investigate how to seamlessly integrate demand-side strategies of dynamic patient prioritization and supply-side strategies of dynamic resource adjustments in our proposed framework of Dynamic Queue Management.

## 3.5 Chapter Summary

In this chapter, we saw that the process and optimization of ED processes is complex. Knowledge from various domains, including process modeling, queuing, simulation, intelligent systems / optimization, analytics and software systems integration, are applied to address the problem of meeting targeted services levels for an ED. None of the existing work considers both demand and supply management. The existing literature also does not consider the systems integration view on how the recommended solution can be implemented in real life.

# Chapter 4

# Demand Perspectives: Dynamic Patient-Prioritization Strategies

In this chapter, we introduce the dynamic priority queue model where the priorities of the patients within the ED queue system are recalculated whenever at least one of the doctors serving the patients becomes available. The queue in our context is modeled as a multi-server, multi-class queue with time-varying arrival and re-entrants.

## 4.1 The Idea

There are several ways to manage queues. The common models are the first-in-first-out (FIFO) and standard priority queues. One application of a priority queue in the context of a hospital is that patients with more severe illnesses are generally treated as a higher priority than those with less severe illnesses. A static priority queue model calculates the priority of an entity based on some attributes of the entity (e.g., severity of illness) when it enters the queue and the priority remains the same with respect to the other entities in the queue throughout its life-time in the queue system. One of the key concerns with static queues such as FIFO or standard priority queues is that they fail to recognize that the individuals in the queue might not have the same priority as time elapses.

In our proposed *dynamic* queue prioritization strategies, priorities of all patients in the queue are calculated each time a doctor becomes available. There is a significant difference between our dynamic priority queue and the standard priority queue: the priorities of patients are dynamic in the sense they might change over time as the surrounding environment changes.

The priority of a patient is dependent on one or more of the following factors: (a) the estimated consultation time with the doctor and (b) the time remaining to meet the desired service level (e.g., a target of an average LOS of 60 minutes). Let us refer to this target as the *target LOS*. We propose a queuing model that intelligently allocates patients to doctors so as to improve patient flow and reduce the average LOS of all patients in the ED.

## 4.2 Problem Definition

The formal problem definition is as follows: Given the front room's patient queue with time-varying arrival rates and re-entrants, a fixed number of doctors in each hour over a week, the service rates of new patients and re-entrants, the service rates and probability of a patient taking a test or treatment and a fixed time horizon (e.g., two days), find the patient with the highest priority and dispatch to the next available doctor to minimize the average LOS.

We proposed three strategies in which the priorities of all patients in the queue are calculated each time a doctor becomes available: shortest-consultation-time-first (SCON), shortest-remaining-time-first (SREM) and a mixed strategy based on a combination of SCON and SREM. The remaining time of a patient is given as the difference between the target LOS and the total amount of time that the patient has spent in the sub-process of consultation with doctors. If the patient has to take any investigative test or receive a treatment, we allow the target LOS to extend beyond the original value by the amount of time spent in taking tests and treatment. This is reasonable as the patient has a need to stay in the ED for the additional activities. When the patient re-enters the queue, his/her total time spent does not start from zero, but is accumulated and hence this patient can have a higher priority than another patient who is waiting to consult the doctor for the first time. In addition, we observed in our field study that the review consultation time could be shorter than the first consultation time. As such, we consider the priorities of all the patients within the same queue to be dynamically set.

## 4.3 Dynamic Priority Queuing Model with Re-entrant Entities

An overview of our proposed dynamic priority queuing model is shown, and its parameters defined, in Figure 4.1 and Table 4.1. Our findings reveal from analytics on historical data that patients arrive according to non-homogeneous (time-varying) Poisson arrival rates. Each doctor is a server, and all servers are assumed to be identical with an exponential service time distribution whose rate is dependent on the type of patient (new patient or re-entrant). The re-entrants may have different service rates based on the results of their investigative tests and treatments. The service rate of an investigative test or treatment is also exponentially distributed and dependent on the combination of test and treatment.

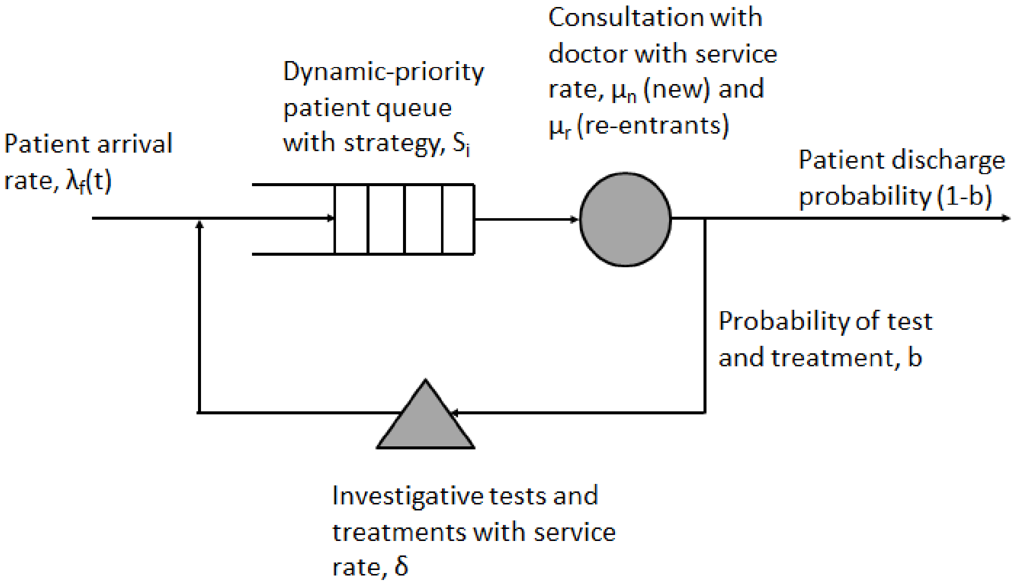

Patient arrival rate, $\lambda_f(t)$
Dynamic-priority patient queue with strategy, $S_i$
Consultation with doctor with service rate, $\mu_n$ (new) and $\mu_r$ (re-entrants)
Patient discharge probability (1-b)
Probability of test and treatment, b
Investigative tests and treatments with service rate, δ

Figure 4.1: The dynamic priority queuing model

| **Parameter** | **Definition** | **Comments** |
|---|---|---|
| $\lambda_f(t)$ | Time-varying arrival rates of new patients in the ambulatory area. | The subscript $f$ is to indicate the arrival rate for the ambulatory area, also known as the front-room operation of the ED. Each $\lambda_f(t)$ is defined per hour over a week's horizon. An example is shown in Figure 2.4. |
| $\mu_n$ | Service rate of the doctors if the patient is a new patient. | We assume a homogeneous service rate for all doctors. |
| $\mu_r$ | Service rate of the doctors if the patient is a re-entrant. | This is modeled as a set of four exponential distributions with corresponding probability of occurrence. A patient with clean test results usually takes a shorter review duration compared to a patient with complex test results. A typical example of service rates is shown in Table 4.2. |
| $\delta$ | Service rate for investigative tests or treatments. | This is modeled as a set of exponential distributions with corresponding probability occurrence. The reason is that there are different types and combinations of tests and treatments required by different patients. A typical example of service rates is shown in Table 4.3. |
| $b$ | Probability of re-entrance. | Patients who require tests and treatments are required to be reviewed by a doctor, and hence re-enter the queue. |

Table 4.1: Parameters in the demand-side queuing model

| Percentage of re-entrants | Average time-taken(mins) |
|---:|---:|
| 40% | 5 |
| 30% | 12 |
| 20% | 18 |
| 10% | 25 |

Table 4.2: Service rates of re-entrants

| Percentage of patients | Average time taken(mins) |
|---:|---:|
| 14% | 8 |
| 68% | 29 |
| 13% | 50 |
| 5% | 78 |

Table 4.3: Service rates of investigation and treatment

## 4.4 Strategies in Calculating Priorities of Patients

We have three dynamic priority strategies to calculate priorities of patients in the queue, namely shortest-consultation-time-first (SCON), shortest-remaining-time-first (SREM) and mixed strategy incorporating both consultation-time and remaining-time factors (MIXED).

The formal definitions of the variables in our model are as follows. Each patient $k$ who has registered has a targeted LOS $d_k$ and an elapsed time $e_k$ that he has spent in the department. At the end of triage (i.e. the beginning of the queuing model), $d_k$ is set to the targeted LOS (e.g., 60 minutes) and the elapsed time is set to zero. If a patient requires an investigative test or treatment, he/she incurs additional time $t_k$ based on the type of test or treatment, based on treatment service rate $\delta$. The *remaining time* for the patient, $r_k$, is then defined as $d_k + t_k - e_k$. Note that we added $t_k$ so as to allow the patient to spend more time in the ED to take the test(s) or treatment. In the consultation queue, the patient is either a new patient or a re-entrant. Each patient has an estimated *consultation time* $c_k$, computed based on the doctors' service rate $\mu_n$ or $\mu_r$ if the patient is new or a re-entrant.

### 4.4.1 Shortest-Consultation-Time-First (SCON)

In the SCON strategy $S_1$, we rank (give priority to) patients according to their estimated consultation times with the doctor. The intuition behind this strategy is based on our observation that some re-entrants have very short estimated consultation times. For example, a doctor reviews a clear blood test result with the patient in about two minutes. We suggest that such patients be cleared earlier to allow them to exit the ED. Let $c_k$ be the estimated consultation time of patient $k$. We use an exponentially decreasing function (with $\rho_1$ as the constant parameter to set the gradient of the exponential function) to assign (as shown by the arrow in the equation) the priority of a patient $p_k$:

$$p_k^{S_1} \leftarrow e^{\dfrac{\rho_1}{c_k}} \tag{4.1}$$

### 4.4.2 Shortest-Remaining-Time-First (SREM)

In the SREM strategy $S_2$, we rank patients according to their remaining times. This is because we want to inspire the patients' confidence in the hospital's ability to serve them within the targeted LOS and, as a result, maintain the hospital's reputation. We assign the patient a priority, which tends to a large number when the remaining time tends to zero, and tends to 1 when the remaining time is sufficiently large. Furthermore, since the remaining time may even become negative (i.e., when the patient is yet to be served after the targeted LOS has elapsed), his/her priority should be set to an even larger value. For this purpose, we propose an exponentially decreasing function (with $\rho_2$ as a constant parameter to set the gradient of the function) when the remaining time is a positive number. When the remaining time tends to zero, we set it to a large constant to avoid division by zero. And when the remaining time becomes more negative, it increases linearly. As such, we propose a three-segment function as shown in the following equation. Let $c$ be a small value (e.g., 0.1) to ensure priority is very high (only a few cases fall into such a category). We let $f_c = e^{\frac{\rho_2}{c}}$ when $r_k = [-c, c]$. When $r_k < 0$, we use a negative linear function with a constant slope $m > 0$.

$$p_k^{S_2} \leftarrow \begin{cases} e^{\dfrac{\rho_2}{r_k}} & \text{if } r_k > c \\ f_c & \text{if } r_k \in [-c, c] \\ f_c - m r_k & \text{if } r_k < -c \end{cases} \tag{4.2}$$

The selection of the exponentially decreasing functions in both SCON and

SREM was inspired by the work in Cheu et al. [10] in which the authors reasoned that the most computational-efficient functions for measuring dis-utility (in their case, in transportation domain) are linear and exponential functions. An analogous situation for our domain is the waiting anxieties of patients, which can be seen as a dis-utility. Since we want to give much higher priorities to patients who have a very short consultation time or remaining time, we adopted the exponential functions.

### 4.4.3 Mixed Strategy (MIXED)

We consider a MIXED strategy $S_3$ to take into consideration the multiple factors determining the priorities of patients. We also observe (see experiments below) that a pure SCON strategy may potentially result in the problem of starvation. Hence, we propose a MIXED strategy to prevent starvation. Note that although we consider only two factors in this paper, namely, consultation time and remaining time, we believe that the model can be extended to include more factors. We use a weighted scheme so that weights can be assigned to the various factors that make up the MIXED strategy. The weights are normalized such that the sum of the weights equals 1. The weights are constant parameters that are calibrated via a local search algorithm. Hence, in general, suppose we have $N$ factors, each having a weight of $a_n$ $(1 \leq n \leq N)$ contributing to the MIXED strategy, then we have the priority of patient $k$:

$$p_k^{S_3} \leftarrow \sum_n a_n . p_k^{S_n} \tag{4.3}$$

where $\sum\limits_n a_n = 1$ and $a_n \epsilon [0, 1]$

To find the weights $a_n$, we propose a multi-dimensional binary search algorithm. We begin the search with $a_n$ $(1 \leq n \leq N)$ set to $\frac{1}{N}$ which gives the centroid of the multi-dimensional search space. In each iteration of the search, we examine the $N(N-1)$ neighboring points, which are the points derived from choosing 2 out of the $N$ weight variables and shifting half the weight value from one variable to the other. We run simulations for the current point

```
Search algorithm for weights
Set $a_n$ to $\frac{1}{N}$ for each $1 \leq n \leq N$
Initialize $d_g$ to a user-defined value used to stop the search
$d \leftarrow aLargeNumber$
$a^{curr} \leftarrow a$ // current best setting
avgLOS $\leftarrow$ runSim($a$, dayOfWeek, numOfIterations)
while ($d \geq d_g$)
  for each point $a'$ in the neighborhood (of $N(N-1)$ points)
    $avgLOS' \leftarrow$ runSim($a'$, dayOfWeek, numOfIterations)
  endForEach
  if there is an improvement over the current $avgLOS$ value,
  then set $avgLOS$ and $a^{curr}$ to the values associated with the neighbor with the minimum average LOS
  Set $d$ to the difference between the current $avgLOS$ and its previous iteration value
endWhile
```

Table 4.4: Search algorithm to obtain the weight of each factor in the MIXED strategy

and those of the neighboring points. We then take the average LOS over all hours and obtain the weights that provide the best average LOS. We continue the search until the algorithm converges to a point such that difference ($d$) between the current average LOS and minimum neighboring average LOS is less than a targeted difference $d_g$. The details of the search algorithm are as shown in table 4.4.

## 4.5 Implementation Design

We propose in this section a preliminary system design to support our proposed methods. The SCON method requires a way to provide an accurate estimation of the consultation time in order for the scheme to be successful. In order not to overload the nurses or doctors, we proposed to automate the estimation by the use of IT systems. The SREM method required real-time tracking of the amount of time the patients spent in the ED. This can be fairly accurate and easily implemented in most hospitals with today's technologies.

### 4.5.1 Estimation of Consultation Time for SCON

With the example of the systems and process in the ED that we have studied, we have the patient arrival information recorded in the Patient Care System (refer to Figure 4.2). This is entered during registration. At each stage of the ED process, time-stamps and some basic information of the activities are also recorded (e.g., the time at which an X-ray is ordered and the time in which the nurse receives the result). For investigative tests, the results are recorded in the individual systems. For example, the blood test results of a patient are recorded in the Laboratory System, and the X-ray image and analysis are recorded in the X-ray System. The Enterprise Resource Planning (ERP) System records hospital-wide information including bed information. This system is used when the ED is discharging the patient to an in-patient ward or to book an appointment in another specialist's clinic within the hospital.

For new patients, we have little information other than the outcome of the triage process to estimate the consultation time. Based on interviews with the hospital's staff, we understand that the service rates for new P3 and P4 patients are relatively stable. As such, we automate the estimation of consultation time for all first consultations with a doctor using a standard service rate $\mu_n$.

The estimated consultation time for the re-entrants (second consultation) is more complex. However, with the help of the information systems, we can get a good automated estimate. The patient's test results are readily obtained and can be used in our estimate when they are available in the individual systems. As an illustration, in a typical blood test, if a patient's blood results show all of the items in the test are within the normal range, he/she is believed to have a clean blood test result. If a patient has clean test results, then the second consultation with the doctor is generally as short as two minutes and usually completed within five minutes. If the tests show few or borderline issues (readings near to borderline normal ranges), there will be some discussion between the doctor and patient during the second consultation. If the test results show many issues, the second consultation with the doctor can be longer as it either involves long discussions with the patient or the family, or

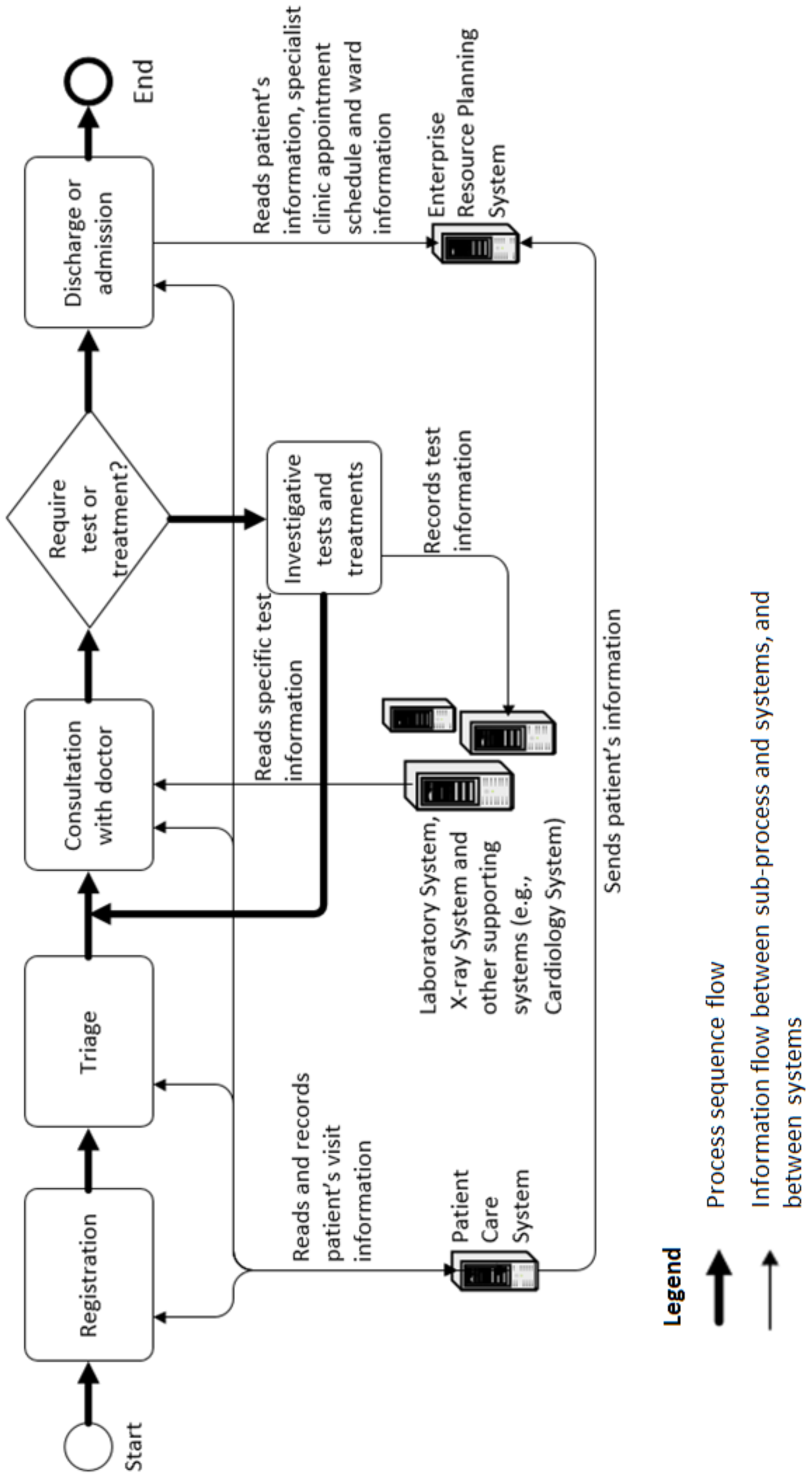

Start
Registration
Triage
Consultation with doctor
Require test or treatment?
Investigative tests and treatments
Discharge or admission
End
Reads and records patient's visit information
Patient Care System
Reads specific test information
Laboratory System, X-ray System and other supporting systems (e.g., Cardiology System)
Records test information
Reads patient's information, specialist clinic appointment schedule and ward information
Enterprise Resource Planning System
Sends patient's information
Legend
Process sequence flow
Information flow between sub-process and systems, and between systems

Figure 4.2: ED process with associated supporting systems

involves more detailed information to be indicated in the medical record or a referral memorandum before the patient is discharged to an in-patient ward or specialist clinics. There are, at times, cases when the test and treatment results are unfavorable, and a very short second consultation is observed and the patient is warded for further investigation. As a general heuristic, the doctor takes more time to see patients with observable conditions than those who have clean test results. We can form a set of re-entrant service rates $\mu_r$ such as the ones shown in Table 4.2. Using data drawn from the test results of patients, a hospital can gather statistics based on historical data to determine the service rates for each of the categories of re-entrants. We can then automate the estimation of the consultation times of re-entrants using these service rates. The process of estimating a patient's consultation time is illustrated in Figure 4.3.

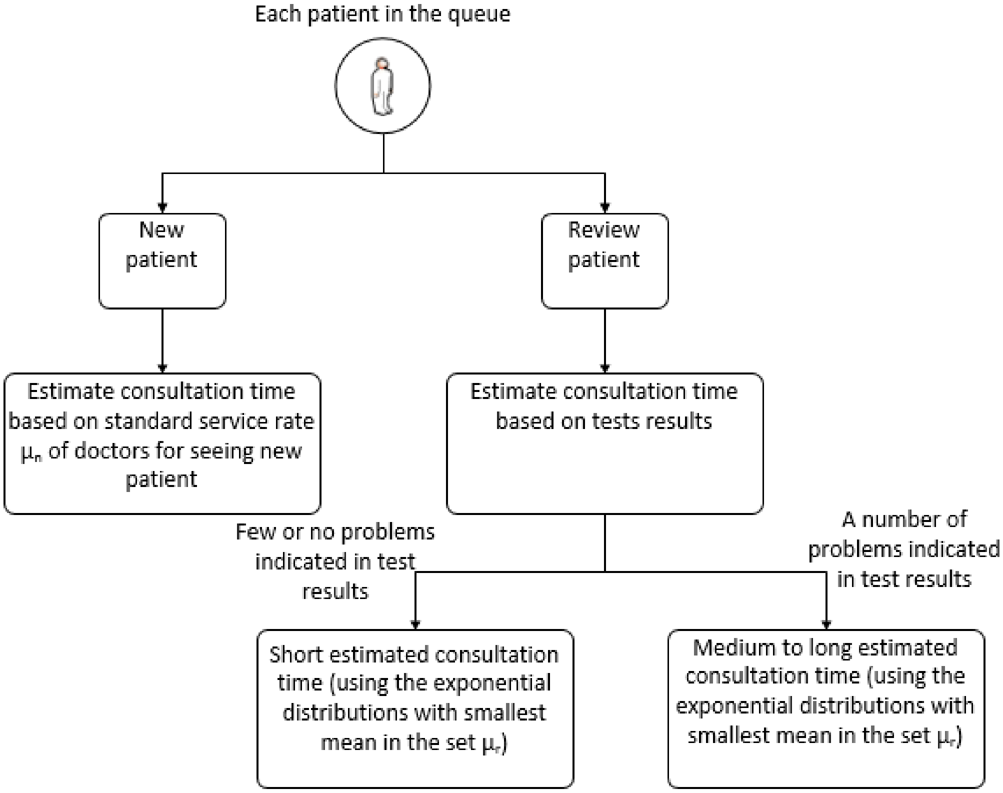

Each patient in the queue
New patient
Review patient
Estimate consultation time based on standard service rate μₙ of doctors for seeing new patient
Estimate consultation time based on tests results
Few or no problems indicated in test results
A number of problems indicated in test results
Short estimated consultation time (using the exponential distributions with smallest mean in the set μᵣ)
Medium to long estimated consultation time (using the exponential distributions with smallest mean in the set μᵣ)

Figure 4.3: Proposed calculation of estimated consultation time of patient

We recognize that the proposed method is an estimate and the realized consultation time can deviate from the estimate. The behavior of individual patients and their family members is not easily predicted using technology. As such, we need to find a more accurate method to complement SCON. In the next two sections, we will see how priorities in SREM can be more accurately calculated and how the MIXED method can reap the benefits of both SREM and SCON strategies.

### 4.5.2 Calculation of Remaining Time for SREM

The calculation of remaining time for SREM can be made more accurately. There are three components to the calculation: target LOS ($d_k$), treatment duration ($t_k$) and elapsed time ($e_k$). $d_k$ is set by the hospital. $e_k$ is the difference between the current time and the patient's registration time as recorded in the Patient Care System. $e_k$ is accurately computed. The only component that may be only partially accurate is the treatment duration $t_k$. Although most of the start and end test time-stamps captured in the individual systems supporting each investigative test (e.g. Laboratory System, X-ray System) can indicate the treatment duration, there could be slight variations due to human entry delays. For example, in some cases, nurses batch a few X-ray reports before entering them into the system. Hospitals should review the systems' ability to capture such information for accurate calculation of remaining time. We believe that capturing such information can be easily automated with simple and inexpensive solutions such as bar codes, QR-codes or any form of scanning devices.

### 4.5.3 Inclusion of Other Factors for MIXED Strategies

If a hospital wants to include other factors into the MIXED strategy, it should analyze its information systems capability and readiness to ensure that information can be captured and the relevant priorities can be calculated. An example of other factors is the patient's health conditions during his/her stay in the ED. A patient's vital statistics are collected during his/her stay in the ED and the values are used to dynamically prioritize the patient [11]. The acuity level of the patient may also change if his/her condition deteriorates greatly (up-triage). A higher priority will then be assigned to such a patient when the next doctor becomes available or the patient will be transferred to the critical-care area.

## 4.6 Experimental Evaluation

We built a prototype of the Dynamic Queue Management system as a simulator to evaluate the effects of using the strategies. We set up our experiments using six months' data from the selected local hospital from which we derived the parameters $\lambda_f(t)$, $\mu_n$, $\mu_r$, $\delta$ and $b$. We implemented the prototype using Java and ran the simulation over 10 iterations for a static FIFO queue, as well as for each of the three proposed dynamic priority queue strategies. We measured the average values over all of the iterations.

We used two sets of time-varying arrival rates, one for running the simulation for Tuesday to Saturday and another for Sunday and Monday. The latter are the two days of the week when the ED experiences a higher volume of patients, hence the arrival rates are different. In each set, exponential distributions with mean $\frac{60}{\lambda_f(t)}$ were used. In our time-varying arrival rates, the low demand period was between 1am and 8am daily. The peak period was between 9am and midnight. The midnight-to-1am and 8am-to-9am periods were moderate.

Similarly, we derived two sets of consultation times represented by exponential distributions for the services by the doctors. One set of consultation times was used for new patients (first consultation) and another for re-entrants (review consultation). The first consultation time consists of only a single exponential distribution with mean $\frac{60}{\mu_n}$ where $\mu_n$ is the doctor service rate. The review consultation time consists of a set of four exponential distributions with corresponding probabilities of occurrence as shown in Table 4.2. The service time for investigation tests and treatment is a set of four exponential distributions with corresponding probability of patients' going through tests or treatments as shown in Table 4.3. In the MIXED strategy, the patients' priorities were calculated based on a combination of SCON and SREM. The weights are determined via a local search algorithm (Table 4.4) based on a two-dimensional search. Finally, we set the probability of re-entrance $b$ to 40% as per our observation of historical data.

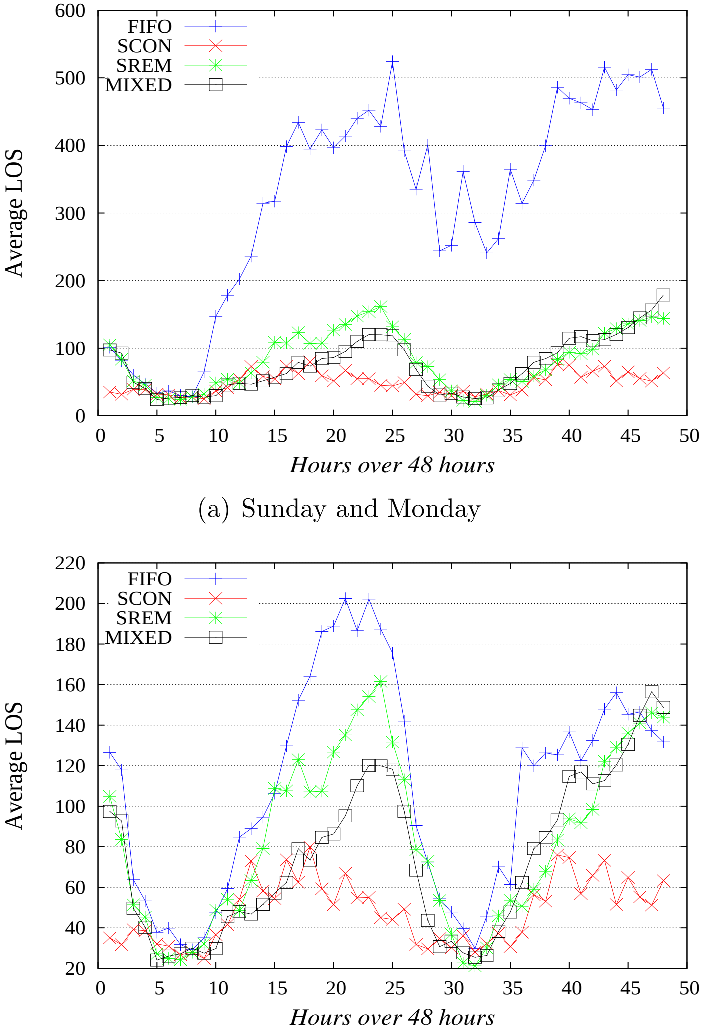
For all of the simulations (except the sensitivity test on the number of doctors), we set the number of doctors in consultation to three and we allowed one day of simulations to pass before we started collecting the results. We then ran the simulation over an additional two days and collected the results for the two days.

### 4.6.1 Experimental Results

Our first set of results is shown in Figure 4.4, with comparisons between the proposed strategies and FIFO. The number of doctors is three and the service rate of the doctors ($\mu_n$) is set to six new consultations per hour (i.e., doctors' service time is an exponential distribution with mean $(60/6) = 10$ minutes). These are the current settings in the ED of the hospital we studied. The average service rate information was provided by the hospital.

From Figure 4.4, we observed that the proposed strategies clearly outperformed the FIFO queue, both in terms of lower average LOS of all patients, as well as consistent and almost stable average LOS over peak and non-peak hours. The pattern of SREM and MIXED strategies is similar to FIFO: average LOS increased during peak hours and decreased during non-peak hours. On days on which the ED had higher demand (Sundays and Mondays), the three strategies performed far better than FIFO. On days that did not have such a high demand, the performance of SREM and MIXED were closer to FIFO. We observed that SCON coped very well with various types of demand in the ED and yielded consistent and relatively low average LOS with smaller variances compared to other strategies.

Next, we performed a what-if analysis by setting the service rate of the doctors ($\mu_n$) to five new consultations per hour (i.e., doctors' service time is an exponential distribution with mean $(60/5) = 12$ minutes). This was to test what happened if the doctors took more time to serve a patient. We show in Figure 4.5 the results of this analysis. On Sundays and Mondays, as shown

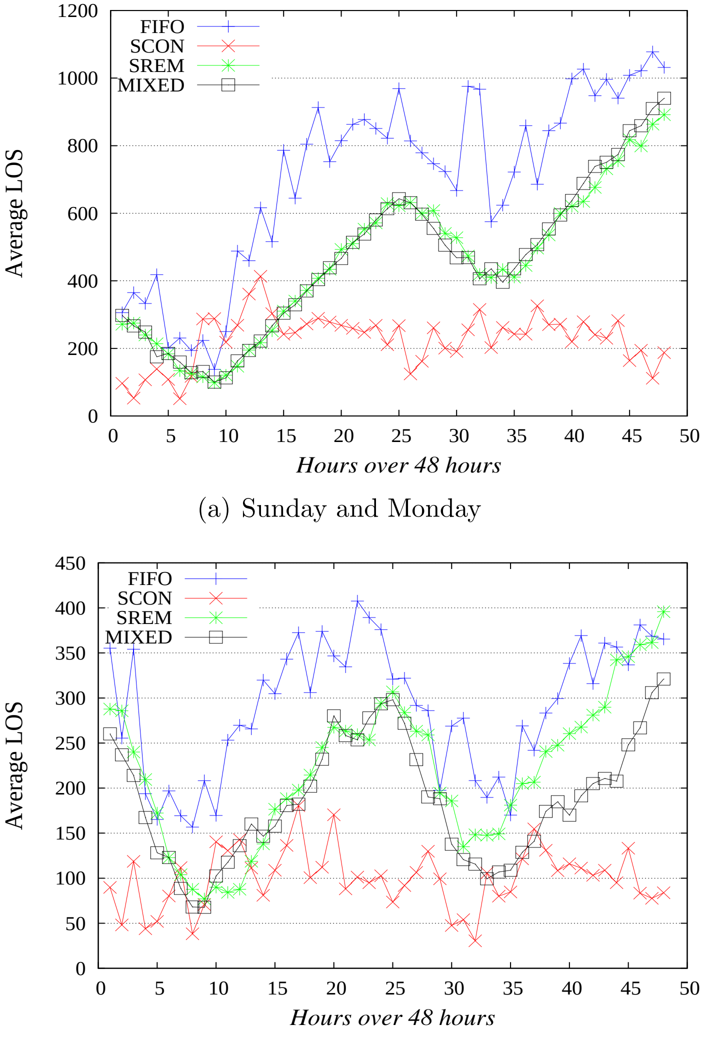


(a) Sunday and Monday

(b) Tuesday to Saturday

Figure 4.4: Comparison of proposed strategies against FIFO for three doctors with $\mu_n = 6$

in Figure 4.5(a), the queue became unstable and the average LOS became increasingly high using FIFO, SREM or MIXED strategies. Unstable queues appeared when the service was slower than the arrivals (which is consistent with standard FIFO queuing theory). In contrast, although the average LOS for SCON was also high at peak hours (reaching 400 minutes at some points), it is interesting to see that it remained fairly stable under such conditions. We draw a conclusion here that the SCON strategy scales better under high load conditions. But SREM and MIXED strategies still perform better than FIFO.

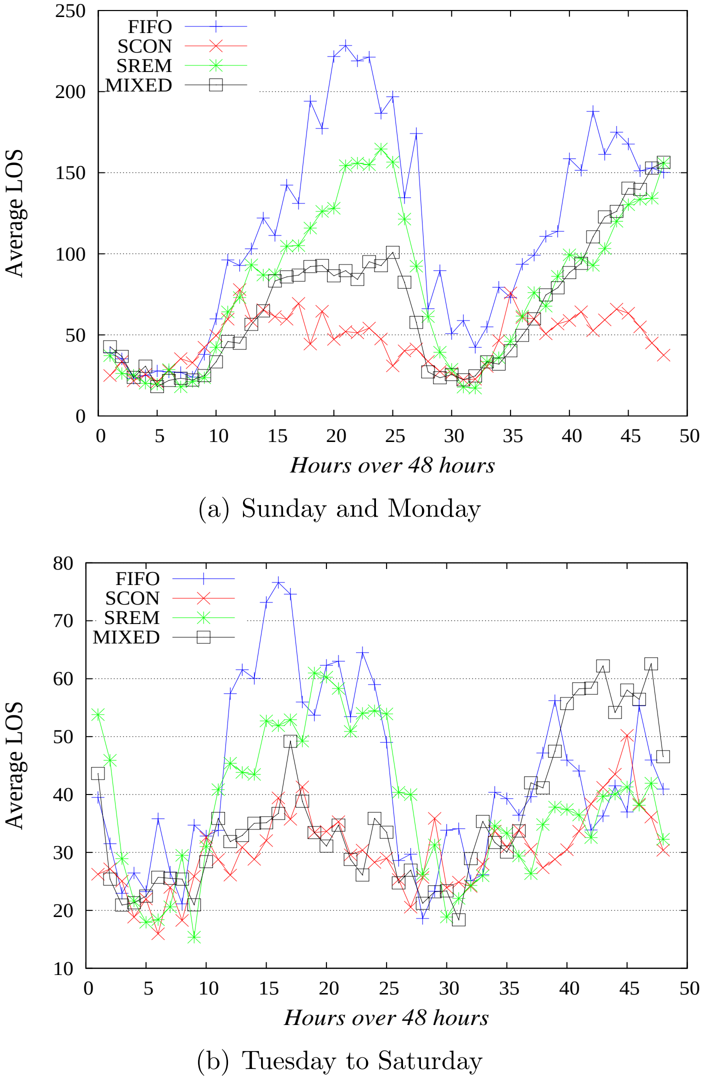


(a) Sunday and Monday

(b) Tuesday to Saturday

Figure 4.5: Comparison of proposed strategies against FIFO for three doctors with $\mu_n = 5$

Another what-if analysis we conducted was when the service rate of the doctors ($\mu_1$) was set to 7.5 new consultations per hour (i.e., doctors' service time is an exponential distribution with mean (60/7.5) = 8 minutes). We tested the situation where the hospital required increased efficiency, with doctors serving each patient more swiftly. We show in Figure 4.6 the results of our investigation. An interesting observation is in Figure 4.6(b) (results of simulation for Tuesday to Saturday). When the ED is not heavily loaded (i.e., during off-peak hours when there are sufficient resources such as doctors), the various strategies including FIFO perform similarly. The proposed strategies have less advantage over FIFO in low-peak situations.

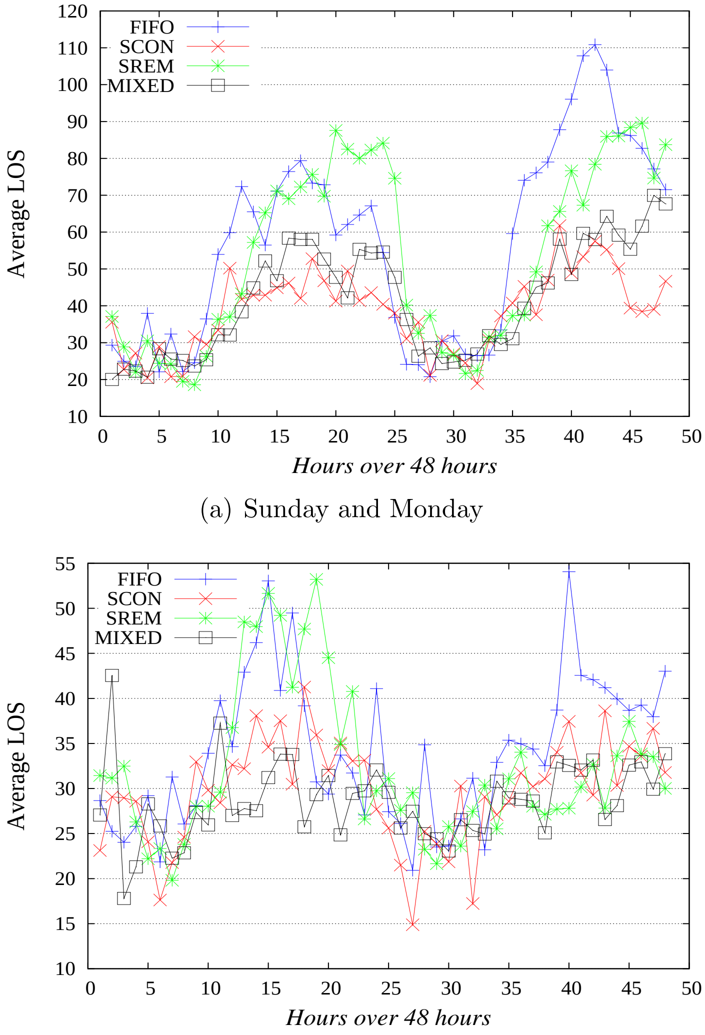


(a) Sunday and Monday

(b) Tuesday to Saturday

Figure 4.6: Comparison of proposed strategies against FIFO for three doctors with $\mu_n = 7.5$

Since a hospital must assure quality of care, increasing the service rate of doctors may not always be feasible (doctors need to thoroughly examine a patient). We examined a situation in which the hospital increases the number of doctors to four and the result is in Figure 4.7. Figure 4.7(b) further verifies our observation that, if the resources (doctors) are not extremely scarce such as on non-peak days like Tuesdays to Saturdays, any chosen strategy works similarly to any other. Our intelligent patient-allocation strategies become useful with a high volume of patients that competes for expensive resources such as doctors.



(a) Sunday and Monday

(b) Tuesday to Saturday

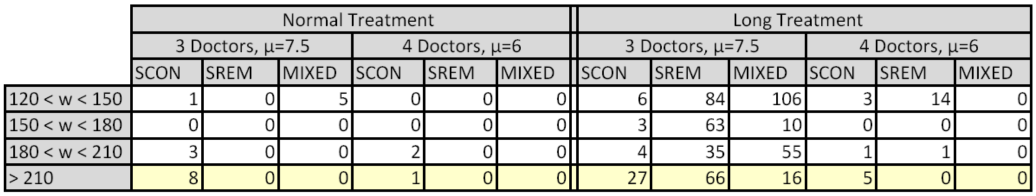
Figure 4.7: Comparison of proposed strategies against FIFO for four doctors with $\mu_n = 6$

Across all the various sensitivity tests on the doctors' service rates and number of doctors, we found that the SCON strategy yielded a very consistent and optimistic result, regardless of whether peak hours were involved. From a management perspective, this strategy not only has the best performance but also inspires the hospital confidence in ensuring the public that a certain service level can be achieved in its ED. For example, the hospital could convey to the public that x% of patients can be served within 20 to 80 minutes in the Tuesday-to-Saturday period based on the doctors' service rate of six patients per hour.

We noticed that SREM did not seem to perform well although it had been designed to meet the LOS target. In our experiments, the number of doctors did not vary but was fixed at three. As such, there were times when there were simply insufficient doctors to clear the demand. In fact, SREM behaved just like FIFO except when a patient's remaining time was very near zero or negative. The patients in such cases would still have to wait for an available doctor or the same doctor who had seen him/her previously to review him/her. As such, the LOS target could not be achieved all the time.

### 4.6.2 Starvation Analysis

Despite its good performance, we speculate that SCON has a potential limitation of patients suffering from starvation. We present the results in Figure 4.8 to illustrate this phenomenon. The symbol $w$ in the table indicates the length of time (in minutes) that patients have waited in the queue for a doctor by the end of our simulation runs. The number in each column shows the total number of such patients over the 10 iterations.



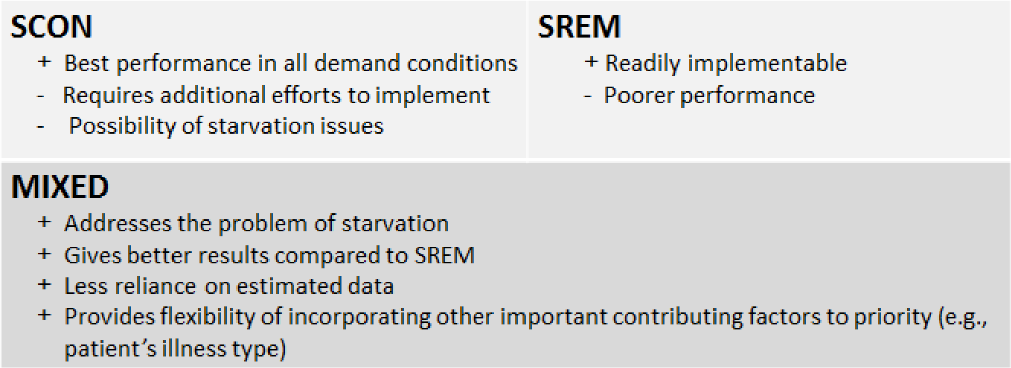
| | Normal Treatment | | | | | | Long Treatment | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 3 Doctors, μ=7.5 | | | 4 Doctors, μ=6 | | | 3 Doctors, μ=7.5 | | | 4 Doctors, μ=6 | | |
| | SCON | SREM | MIXED | SCON | SREM | MIXED | SCON | SREM | MIXED | SCON | SREM | MIXED |
| 120 < w < 150 | 1 | 0 | 5 | 0 | 0 | 0 | 6 | 84 | 106 | 3 | 14 | 0 |
| 150 < w < 180 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 63 | 10 | 0 | 0 | 0 |
| 180 < w < 210 | 3 | 0 | 0 | 2 | 0 | 0 | 4 | 35 | 55 | 1 | 1 | 0 |
| > 210 | 8 | 0 | 0 | 1 | 0 | 0 | 27 | 66 | 16 | 5 | 0 | 0 |

Figure 4.8: Comparison of starvation phenomenon between the three strategies

We ran the experiments over two types of treatment duration distributions, *normal treatment* and *long treatment*. The duration distributions for normal treatment comprise 90% of re-entrants with less than or equal to the estimated consultation time of a new patient. The duration distributions of long treatments only comprise 70% of such re-entrants. We found that even under non-peak conditions (three doctors, $\mu_n = 7.5$, normal treatment, four doctors, $\mu_n = 6$, normal treatment and four doctors, $\mu_n = 6$, long treatment), SCON has cases of starvation, while SREM and MIXED are cleared of starvation. When the ED is slightly crowded (three doctors, $\mu_n = 7.5$, long treatment), the MIXED strategy still outperformed SCON in terms of starvation cases. In this starvation study, we found that SCON had a significant disadvantage and its performance should be evaluated in conjunction with both the starvation study and the average LOS measurement. The results that were shown in Figures 4.4 to 4.7 were based on patients who had completed the ED process, but did not include the patients who were still starved when the simulation had ended (since we were not able to measure the LOS of patients who had not completed their full process). Therefore, we recommend that despite SCON's seemingly good performance, it should not be a strategy that to be used alone. In this analysis, we observed a trade-off between the performance and the risk of having starved patients.

## 4.7 Management Insights for Decision-Makers

Among the three strategies for dynamic priority queuing of patients at an ED, we found that the proposed strategies result in shorter average LOS compared to using the FIFO method and they are able to scale well over peak days and peak hours. Introduction of the SREM strategy also helps the hospital to meet its desired service level. With the advancement of IT that enables patient real-time movements to be tracked and data to be mined quickly, we believe that our proposed concept of a dynamic priority queue is implementable. We observed a trade-off between performance and risk of implementation in terms of readiness.



**SCON**

- + Best performance in all demand conditions
- - Requires additional efforts to implement
- - Possibility of starvation issues

**SREM**

- + Readily implementable
- - Poorer performance

**MIXED**

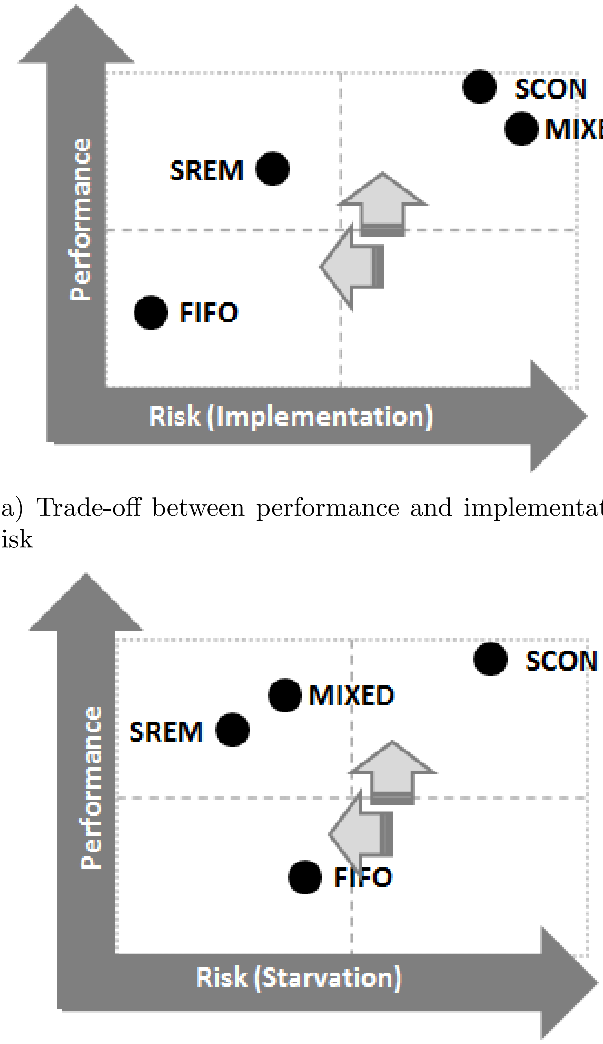
- + Addresses the problem of starvation
- + Gives better results compared to SREM
- + Less reliance on estimated data
- + Provides flexibility of incorporating other important contributing factors to priority (e.g., patient's illness type)

Figure 4.9: Summary of pros and cons of the three strategies

The advantages and disadvantages of strategies are shown in Figure 4.9. The SCON strategy yields the best performance with lower variance over the average LOS for all the scenarios, but the drawback is that the strategy requires additional effort to implement and faces the high risk of starvation. The SREM strategy is readily implementable but its performance is less promising. We see the potential of the MIXED strategy because it addresses the problem of starvation (faced by SCON) and gives better results compared to SREM. More importantly, the MIXED strategy reduces reliance on the accuracy of estimating the consultation time of patients; it also has the flexibility of incorporating other important contributing factors (e.g., patient's illness type - a patient with a fish bone in the throat may be more uncomfortable than another patient with a cold and fever) to setting the patient's priority.

In our further analysis, we see two major trade-offs that a decision-maker must consider: the trade-off between performance and risk of implementation, and the trade-off between performance and the risk of starvation. We illustrate these trade-offs in Figures 4.10(a) and 4.10(b). The arrows in the figures indicate what a decision-maker may like to achieve – high performance and low risk. Decision-makers can make a choice of the strategy to adopt based on the hospital's appetite for performance and risk. The MIXED strategy is our recommended strategy with a good trade-off between the various considerations.

In this study, we see a potential need to be able to have a combined measurement of the performance in terms of LOS and the starvation cases. We



(a) Trade-off between performance and implementation risk

(b) Trade-off between performance and starvation risk

Figure 4.10: Demand-side strategy quadrant analysis for the decision-maker on strategy selection

recognize that the performance measure of SCON (only in terms of hourly average LOS) may be overly optimistic because the hourly average LOS cannot account for the starvation cases. Since the patients who are starved will exit some hours later and only be measured in later hours, they are not taken into consideration in computing the average LOS. We anticipate that different concluding results may be observed if longer simulation runs (such as over a week or two) are used to ensure that starved patients are cleared from the system and are measured. Moving forward, another possibility for extension is to consider measurement of performance using other forms of evaluation such as agent-based simulation where each patient is differentiated and an LOS measurement may be taken for each patient, including the patient who has experienced starvation in the ED process.

## 4.8 Summary

We conclude that a dynamic priority queue has many advantages over the standard FIFO queue to improve patient flow in an emergency department. We see two trade-offs in considerations of a good strategy: performance versus risk of implementation and performance versus risk of starvation. Before deciding on the strategy to adopt, the decision-maker must understand the available state-of-the-art of technology and the hospital's ability to implement new technologies and systems.

# Chapter 5

# Supply Perspectives: Dynamic Resource-Adjustment Strategies

In this chapter, we introduce our dynamic resource-adjustment strategies for managing the supply of doctors in the ambulatory area. In doing so, we need to bear in mind the resource requirements of the critical-care area and ensure that the more critical patients (P1 and P2) have sufficient doctors to serve them before the P3 and P4 patients in ambulatory area. This is the first study to consider queue design and real-time queue control under time-varying demand with re-entering customers for a dual-facility service center. The queue model in our context is modeled as dual time-varying ($G_b(t)/M_b/S_b(t)$, $G_f(t)/M_f/S_f(t)$) queues for the critical-care area (the back room with subscript $b$) and the ambulatory area (the front room with subscript $f$). The front room is modeled as a queue with re-entering patients.

## 5.1 The Idea

A doctors' schedule for both front and back rooms is planned manually based on the hospital's estimate of demand in the ED at various times of the day, over an entire week or month. Such schedules generally do not react well to uncertainties such as surges in patient arrivals, and often the hospital is unable to react to meet its desired service level (in terms of target length-of-stay). The doctors are at times moved between the front room and back room. In current practice, this is typically done in a reactive and unplanned manner without the ability to know when there is a need to move a doctor from the critical-care area and when to return the doctor. If the ED experiences surges in demand (usually in the front room), additional doctors may be called in but there is a lack of information about how long the additional doctors will be required to serve the patients.

We designed dynamic resource-allocation strategies with the aim of achieving the following objectives: (1) ability to respond to uncertainties (e.g., surges in demand) by leveraging on real-time data; (2) ability to meet the desired service level in the front room while ensuring that the quality of the related facility (i.e., the ED's back room) is maintained. Our proposed dynamic resource-adjustment strategies are data-driven and provide real-time decision support to the ED's operations.

## 5.2 Preliminaries

To support our dynamic resource-adjustment policies, we considered the queue model with re-entrants proposed by Yom-Tov [58]. A queue model with re-entrants (known as an Erlang-R model as shown in Figure 5.1) was selected. It is most relevant to the ED process and the results have shown that the Erlang-R model provides more stable performance as measured by the probability of wait ($P(W > 0)$) of the patients under the balanced Quality- and Efficiency-Driven (QED) regime. In Yom-Tov's work, the goal was to identify the staffing required at the ED that provides a stable probability of wait regardless of the time of arrival of the patient. However, physical constraints (e.g., the total number of doctor's rooms available in the ED) and a dual-facility (critical-care area) were not considered in her work. As such, we extended one of her models to build in our other strategies. We also used the method as a benchmark for comparisons between the proposed strategies. We present the mathematical equations from her work that are essential and related to our work.

In Figure 5.1, we show two stations, Station 1 and Station 2. The numbers

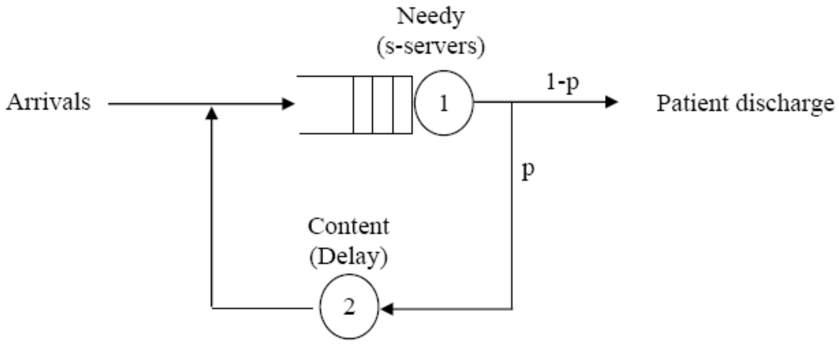


Figure 5.1: The Erlang-R Model

in the circles indicate the station numbers, not the number of servers available. Based on the modified offered load (MOL) approach, the time-varying offered load $R_x$ at Station $x$ is given by the following equation [31]:

$$R_x(t) = E[\lambda_i^+(t - S_{x,e})]E[S_x] \tag{5.1}$$

where $\lambda_x^+$ is the aggregated arrival-rate function to node $x$, $S_x$ represents the service time at node $x$, and $S_{x,e}$ is a random variable representing the excess service time at node $x$.

The doctor staffing at Station 1 where patients await consultation is then determined by substituting the time-varying offered load formula into the following staffing formula [12, 26] where $\beta$ is chosen to take on one of the values which provide a stable expected wait-time according to findings in Yom-Tov[58]. We recognize that there are other ways of setting $\beta$, such as to parameterize it according to the desired service quality. The methods can be found in Halfin and Whitt [21].

$$s_1(t) = R_1(t) + \beta\sqrt{R_1(t)}; \quad \forall t > 0 \tag{5.2}$$

Using the arrival function of our selected local hospital as in the previous chapter, the arrivals are modeled as non-homogeneous time-varying Poisson processes. The arrival pattern is found in Figure 2.4 earlier. For the purpose of comparison of our proposed strategies with existing work with an analytical model in [58], the service rates at Stations 1 and 2 are simplified to single exponential distributions of $\mu$ and $\delta$. Based on the findings in Yom-Tov [58], for general time-varying arrival and exponential service rates, we can solve via numerical approximation. The time-varying offered load with re-entrants using the following ordinary differential equations is given by:

$$\begin{aligned} \frac{d}{dt}R_1(t) &= \lambda_t + \delta R_2(t) - \mu R_1(t) \\ \frac{d}{dt}R_2(t) &= p\mu R_1(t) - \delta R_2(t) \end{aligned} \tag{5.3}$$

## 5.3 Problem Definition

The problem consists of two variants: (1) the constraint-satisfaction problem where the goal is to find the allocation of doctors (doctors' schedule) that satisfies all constraints; and (2) the optimization problem where the goal is to find an allocation of doctors that optimizes the cost of using them, and treats the service level constraint as a *soft constraint*. The term *soft constraint* is used to refer to a constraint which can be relaxed during the search for the selection of the best solution. A solution with fewer occurrences of violations to the soft constraint is a better solution than one which has higher occurrences of violations.

The cost of using doctors consists a primary cost – the hourly labor cost of a doctor multiplied by the number of doctors allocated in the hour; and a secondary cost – the deviation cost of the allocation for measuring schedule stability. Deviation refers to the total difference in the number of doctors between the two consecutive time periods, and the deviation cost is computed as the square of deviation multiplied by a numeric cost factor.

The constraint-satisfaction problem is defined as: Given the front-room's patient queue time-varying arrival rates and re-entrants, a pre-determined hourly doctors' requirement for serving the patients in the back room, hourly total number of available doctors in the whole ED, a service rate of the doctors, the service rate and probability of a patient taking a test or receiving treatment, calculate the number of doctors allocated in the front room per hour for a planning horizon such that the following constraints are satisfied:

(a) the front and back rooms' allocations do not exceed the hourly total number of doctors (back-room service level guarantee),

(b) the front room's requirements do not exceed the front room's physical capacity.

The optimized problem involves optimizing the cost of doctors, and is defined as: Given the front-room's patient queue time-varying arrival rates and re-entrants, a pre-determined hourly doctors' requirement for serving the patients in the back room, hourly total number of available doctors in the whole ED, a service rate of the doctors, the service rate and the probability of a patient taking a test or receiving treatment, the goal is to find the number of doctors allocated in the front room per hour for a planning horizon that minimizes the cost of using doctors, such that the following constraints are satisfied:

(a) the front and back rooms' allocations do not exceed the hourly total number of doctors,

(b) the front room's requirements do not exceed the front room's physical capacity,

(c) the service level target is achieved in each hour (soft constraint).

## 5.4 The Dynamic Resource-Adjustment Queuing Model with Re-entrants

Table 5.1 defines the parameters to our queuing model as shown in Figure 5.2. Our aim is to intelligently determine or forecast the optimal number of doctors required at Station 1 (the front room of the ED) in order to meet the front-room service quality desired by the hospital while maintaining back-room quality. We focus on Station 1 because we learnt that it is the bottleneck of the ED process (through our on-site observations). With reference to Figure 5.2, we define the following notation used in our model:

| Parameter | Definition | Comments |
|---|---|---|
| $\lambda_b(t)$ | Time-varying arrival rates of new patients in the back room. | Each $\lambda_b(t)$ is defined per hour over a week's horizon. |
| $\lambda_f(t)$ | Time-varying arrival rates of new patients in the front room. | Each $\lambda_f(t)$ is defined per hour over a week's horizon as shown in Figure 2.4. |
| $\overline{LOS}(t)$ | Average LOS of patients who leave the ED within a time interval of $[t, t+1)$. | |
| $LOS_{max}$ | Hospital's desired service quality in terms of LOS. | |
| $\mu_b$ | Service rate of the doctors in the back room. | Assumes an exponential service rate and homogeneous doctors. |
| $\mu$ | Service rate of the doctors at Station 1. | Assumes a homogeneous and exponential service rate for all doctors. |
| $\delta$ | Service rate for investigative tests and treatments at Station 2. | Assumes an exponential service rate. |
| $room_{max}$ | Physical constraints in the ED's front room. | This corresponds to the maximum number of consultation rooms in the set-up of the ED. |
| $S_{max}(t)$ | Maximum number of doctors that can be deployed in the ED (front and back rooms combined) at time $t$. | This corresponds to the resource capacity in real life. |
| $S_b(t)$ | Number of doctors required in the back room to serve the demand. | Estimated using steady-state offered load. |
| $S_f(t)$ | Number of doctors to be placed in the front room. | |
| $b$ | Probability that a patient will go for investigative tests and treatments. | |

Table 5.1: Parameters in the supply-side queuing model

We observed that the demand in the back room was fairly stable from day to day although it varied depending on the time of day. Assuming that it is possible to get to steady state, and using that average arrival rate of a

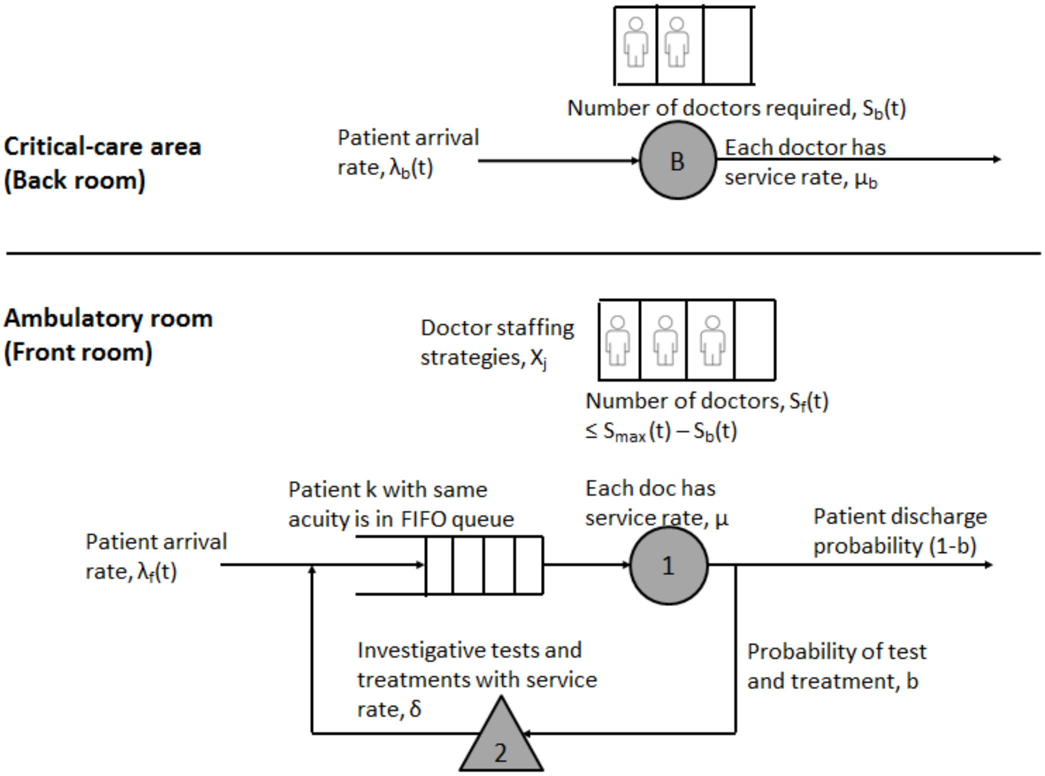

Number of doctors required, $S_b(t)$
Critical-care area (Back room)
Patient arrival rate, $\lambda_b(t)$
B
Each doctor has service rate, $\mu_b$
Ambulatory room (Front room)
Doctor staffing strategies, $X_j$
Number of doctors, $S_f(t)$ $\leq S_{max}(t) - S_b(t)$
Patient k with same acuity is in FIFO queue
Each doc has service rate, μ
Patient discharge probability (1-b)
Patient arrival rate, $\lambda_f(t)$
1
Investigative tests and treatments with service rate, δ
Probability of test and treatment, b
2

Figure 5.2: The queuing model for dynamic resource adjustment

time-interval (e.g., an hour), we calculated the offered load for each interval $R(t) \approx \frac{\lambda_b(t)}{\mu_b}$. $\frac{1}{\mu_b}$ determines the mean service time in the back room. We then found the staffing requirements for the back room using the staffing rule as per Equation 5.2 under the QED regime. The hourly time-varying back room requirements were pre-computed and then used as an input for the front-room staffing.

The following are the constraints for our problem:

$$S_f(t) \leq S_{max}(t) - S_b(t) \tag{5.4}$$

$$S_f(t) \leq room_{max} \tag{5.5}$$

$$\overline{LOS}(t) \leq LOS_{max} \quad (soft \quad constraint) \tag{5.6}$$

Constraint (5.4) specifies the back-room service level guarantee, (5.5) specifies the physical front-room constraint and (5.6) specifies the service level constraint. The service level soft constraint can be implemented as a penalty function in the objective function. For example, we can count the number of violations to the constraint and provide a penalty to the measurement of the objective function.

The queue model used in the dynamic-resource adjustment strategies is similar to the model in the previous chapter of dynamic patient-prioritization strategies. However, the model is different in this chapter in the following ways:

1. The needs of the back room are considered.
2. The queue discipline used for the patient queue is first-in-first-out (FIFO).
3. Homogeneous rates are used for the service rate of doctors (single $\mu$) and the service rate of investigative tests and treatments (single $\delta$). The reason for using homogeneous rates is to be able to allow our proposed strategies to be compared to a benchmark as given by the work in Yom-Tov [58], which is supported by an analytical model.
4. The number of doctors in each hour is not constant.

In the following chapter, we will show that we can combine the demand-side and supply-side queue models in the integrated dynamic queue-management framework.

## 5.5 Resource-Adjustment Strategies

The resource-allocation strategies allow a doctors' schedule in the front room for Station 1 to be determined either in a proactive manner or dynamically, depending on the actual arrival conditions. The doctors' schedule is designed to adjust at the minimum of hourly intervals. This is to prevent a schedule that has many changes within a small time interval. We also assume that the back-room requirement is fairly predictable based on historical data and hence back-room staffing is pre-processed and is available before the allocation of doctors to the front room.

The design of the strategies (shown in Figure 5.3) aims to achieve these objectives: to be able to react to demand changes, and to meet the target

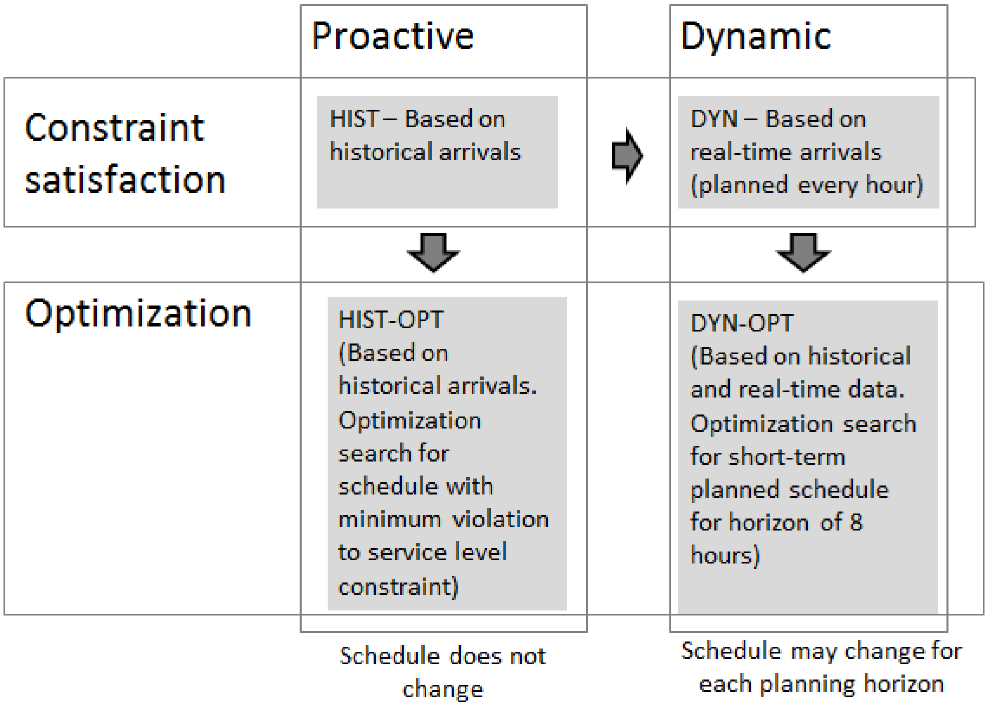

Proactive
Dynamic
Constraint satisfaction
Optimization
HIST – Based on historical arrivals
DYN – Based on real-time arrivals (planned every hour)
HIST-OPT (Based on historical arrivals. Optimization search for schedule with minimum violation to service level constraint)
DYN-OPT (Based on historical and real-time data. Optimization search for short-term planned schedule for horizon of 8 hours)
Schedule does not change
Schedule may change for each planning horizon

Figure 5.3: Resource allocation strategies

service level. We group the strategies into *proactive* strategies and *dynamic* strategies in one dimension; and into *constraint satisfaction* and *optimization* in a second dimension. The proactive strategies are those which are planned in advance by the hospital and do not change during real-life execution of the ED process. The dynamic strategies make use of real-time information about a patient's arrival and produce staffing requirements that change according to demand observed in the ED. The strategies used to address the constraint satisfaction problems use the staffing rule to find the staffing requirements, and the optimization strategies use a local search heuristic to find a schedule with the least violation of the service level. The HIST strategy is to be used as the benchmark for comparison. The HIST-OPT strategy aims to satisfy the aim of meeting the target service level. The DYN strategy is a simple way to react to demand changes. However, we notice that it is too reactive (planning every hour) and may cause the ED to be too nervous. We then have DYN-OPT whose objective is to perform short-term planning over a time horizon (e.g., eight hours) while retaining the ability to react to demand. In DYN-OPT, a symbiotic simulation system is required to be implemented to generate the short-term doctors' schedule.

### 5.5.1 Constraint Satisfaction Strategies

#### Proactive Strategy HIST

This strategy considers resource allocation using historical trends taken from the actual data for patient arrivals from a real-world hospital. Using this data, we applied the Erlang-R method [58] for the front-room model. We first estimated the requirements in the back room, then solved the ordinary differential equations (Equation 5.3) and applied the square-root staffing rule (Equation 5.2) using numerical approximation for the front-room requirements. The front-room staffing was subject to the *back-room service level guarantee* in Equation 5.4 and *physical front-room constraints* in Equation 5.5. Without the back room and physical constraints, this was proven to provide a stable queue wait-time (not LOS) over time as shown in Yom-Tov [58]. However, with the addition of the constraints, where the number of doctors that could be assigned to the front room might be less than the output from the staffing rule, the wait-time stability could no longer be guaranteed. This provides the opportunity to develop better strategies, as shown below.

#### Dynamic Strategy DYN

This strategy considers dynamic resource allocation using real-time data. With the support of enterprise systems and real-time monitoring devices, it is possible to track the time-varying arrival rates in real time. With these arrival rates, we can then calculate the *actual* time-varying offered load based on Equation 5.3 for the previous time intervals. One challenge to applying this method is that the arrival rates must be recorded in a significantly large time frame, such as hourly. However, to approximate the offered loads, one must use a small time interval such as a per-minute time frame in order to compute reasonably sensible offered load for the next time period.

The DYN strategy is designed to be reactive to real-time observation of arrivals and perform short-term forecasts (such as on an hourly basis). Let $R'_1(t)$ and $R'_2(t)$ be the real-time offered load based on *actual* real-time arrival rates at time $t$. In this method, we use one reading of the historical arrival rate as this is the only information we know about the arrival in time $t+1$ at time $t$. We recognize that if the forecast is hourly, then we have at most a one-hour delay in terms of reacting to the demand changes. The algorithm for DYN is as shown in Table 5.2:

| | *DYN Algorithm* |
|---|---|
| 01. | Use the *actual* arrival rate in the previous hour to calculate $R'_1(t)$ and $R'_2(t)$ up to the minute before current time $t$. |
| 02. | Use the *historical* arrival rate for the next hour (t+1) to find the offered load for each minute in the next hour. |
| 03. | Calculate the theoretical doctors' requirements $S'_f(t+1)$ using the square-root staffing rule (per Equation 5.2) based on the offered load for t+1. |
| 04. | $S_f(t+1) \leftarrow min[S'_f(t+1), S_{max}(t+1) - S_b(t+1), room_{max}]$ |
| 05. | Repeat steps 1 to 4 every hour in real time. |

Table 5.2: The DYN Algorithm

### 5.5.2 Optimization Strategies

Both the HIST and DYN strategies rely heavily on the staffing rule. One major limitation with the use of the staffing rule in our problem is that the only way to satisfy the back-room quality constraint or physical front-room constraint is to limit the number of doctors deployed in the front room to the minimum of $[S_{max}(t) - S_b(t+1), room_{max}]$. This may cause violations to the service level quality constraint though. To overcome this limitation and with intuition that one could allocate more doctors in the previous hour(s) before any service level violation, we propose our next two strategies – HIST-OPT and DYN-OPT – which will use heuristic search methods to find optimized resource allocations.

**Proactive Optimized Strategy HIST-OPT**

The idea for the HIST-OPT strategy is to provide a pre-planned optimized resource allocation by using historical data in case real-time arrival information is not available. We propose the use of a local search algorithm.

Let $C_l$ be the cost of labor for deploying a doctor for a single unit of time $t$ and $C_d$ be the deviation cost factor when the number doctors in time $t$ deviates from $t-1$. The deviation cost is included in the model because we do not wish to have a schedule where the number of doctors changes too frequently from hour to hour. In real life, when a doctor comes to work in the front room or when he is removed from the back room, he will incur a set-up cost and lose some productivity. Hence we factor in the deviation cost to minimize hour-to-hour changes in the doctors' schedule.

We have the following objective function:

$$min \quad C_l \sum_t S_f(t) + C_d \sum_t [S_f(t) - S_f(t-1)]^2 \tag{5.7}$$

subject to constraints (5.6) to (5.5) of the model. If no feasible solution is found, the solution with the least violation is returned.

Our local search algorithm is as shown in Table 5.3. Note that a simulator has to be used to evaluate each schedule that we try in the search algorithm. We will see later that this simulator is also our symbiotic simulation system. In practice, if a hospital does not choose a dynamic optimized strategy, the simulator used in HIST-OPT can be implemented as a normal discrete-event simulator without any interaction with the physical systems.

### Dynamic Optimized Strategy DYN-OPT

DYN-OPT is the dynamic variant of HIST-OPT using real-time data. The design of DYN-OPT attempts to address the limitations of DYN which has a short planning time (hence being too nervous to react to changes every hour) and HIST-OPT which is not unable to react to demand changes. The intuition behind DYN-OPT is to make use of existing known real-time information as well as the forecast of the future (of at least more than one hour) based on historical information about the arrivals. The aim is to plan a number of hours ahead such that resources in the ED can be better informed or additional

```
HIST-OPT Algorithm
Using the HIST schedule, run the schedule in a simulator once with x replications.
While a service level violation exists in the results or maximum number of searches has not been reached
  Find earliest time $t$ where the service quality constraint is violated.
  While violated
    Find a time $t_1$ nearest to $t$ when $(0 \leq t_1 \leq t)$ when increase in the resource will not violate constraints (5.4) and (5.5).
    Increase the resource by 1 unit at $t_1$.
    If the solution removes the violation, select this solution.
    Else,
      Find a time $t_2$ nearest to $t$ when $(t < t_2 \leq$ length of simulation$)$ when increase in the resource will not violate constraints (5.4) and (5.5).
      Increase the resources by 1 unit at $t_2$.
      If the solution removes the violation, select this solution.
    endElse.
    If none of the two solutions remove the violation, select the solution with the lower cost.
  endWhile.
endWhile.
Return the first schedule without a violation or an infeasible schedule with the least violation.
```

Table 5.3: The HIST-OPT Algorithm

resources can be better arranged.

We define a vector $(L, H)$, where $L$ denotes the lead time and $H$ denotes the time horizon. Between the current time $t$ and lead time $L$, there will be no change in staffing. This is to cater for the case when the ED cannot add or remove a doctor instantaneously from the ED. The variable $H$ defines the horizon of re-planning based on what is known currently (e.g., plan for the horizon of eight hours). The next planning period is then $t + H$.

Our optimization model becomes:

$$min \quad C_l \sum_{t=t+L}^{t+L+H} S(t) + C_d \sum_{t=t+L}^{t+L+H} [S(t) - S(t-1)]^2 \tag{5.8}$$

The DYN-OPT algorithm is as shown in Table 5.4. Again, in the DYN-OPT algorithm, a simulator is used to evaluate each schedule that we tried in the search algorithm. As DYN-OPT's optimization is done using real-time data, a symbiotic simulation system is required to support this algorithm. We show the design in the next section.

```
DYN-OPT Algorithm
Use the actual arrival rate of previous hours to calculate $R'_1(t)$ and $R'_2(t)$ up to the minute before current time $t$.
Use the historical arrival rates for hours (t+L) to (t+L+H) to find the offered loads for each of the hours in planning horizon.
Calculate the theoretical doctors requirements for the planning horizon using the square-root staffing rule subject to back room quality and physical constraints.
Run the schedule once with x replications.
For the hours from $t + L + H$
  While a service level violation exists in the results or maximum number of searches has not been reached
    Find time t' $(t+L \leq t' \leq t+L+H)$ when the service quality constraint is violated.
    While violated
      Find a time $t_1$ nearest to $t'$ when $(t + L \leq t_1 \leq t')$ when increase in the resource will not violate constraints (5.4) and (5.5).
      Increase the resource by 1 unit at $t_1$.
      If the solution removes the violation, select this solution.
      Else,
        Find a time $t_2$ nearest to $t$ when $(t < t_2 \leq t + L + H)$ when increase in the resource will not violate constraints (5.4) and (5.5).
        Increase the resource by 1 unit at $t_2$.
        If the solution removes the violation, select this solution.
      endElse.
      If none of the two solutions remove the violation, select the solution with the lower cost.
    endWhile.
  endWhile.
endFor.
Return the first schedule without a violation or an infeasible schedule with the least violation.
```

Table 5.4: The DYN-OPT Algorithm

## 5.6 Implementation Design

We present a design of a possible implementation of Dynamic Queue Management (DQM) as a system. We assume that this system is capable of receiving both historical data and real-time data from the physical systems as shown in Figure 5.4. Examples of physical systems are the Patient Care System, Laboratory System, X-Ray System and Enterprise Resource Planning (ERP) system (see Figure 4.2). Data from the physical systems can be extracted from the databases and be analyzed to form the ED's historical process data such as the time-varying arrival rates, service rates of various activities (consultation for new patient and re-entrants, various investigative tests and treatments) in the process and the probability of re-entrants. Experts' inputs on the process sequence are also collected to model the ED process and derive our queuing model.

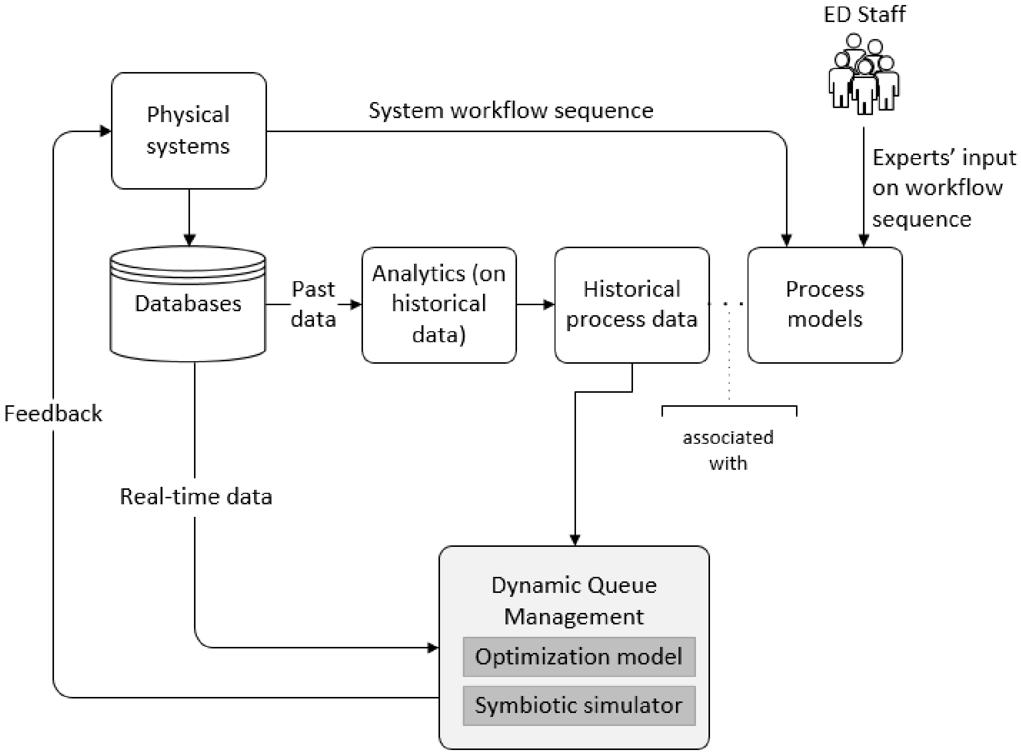


Figure 5.4: Implementation design of DQM

In terms of implementation complexity, HIST requires only historical data, while DYN also requires real-time data. HIST-OPT requires historical data and the optimization model, while DYN-OPT requires historical, real-time data and the symbiotic simulator.

We built a prototype of DQM as a discrete-event simulator using Java (see Figure 5.5). The decision-maker selects a desired strategy as an input to the prototype. In our prototype, the discrete-event simulator simulates the ED process of patients' arrival, queuing, consultation with doctor, investigative tests and treatments. This is to emulate the real-life interactions with the hospital's physical systems where information about the patients' arrival, time of consultation, investigative tests and treatments can be received. The DQM prototype reads the historical process data and doctors' schedule from external files. The doctors' schedule is used to determine the hourly number of available doctors in HIST and HIST-OPT strategies. These schedules are generated before the simulation runs. In an actual deployment, we can expect a physical system such as the ERP to store and provide the doctors' schedule.

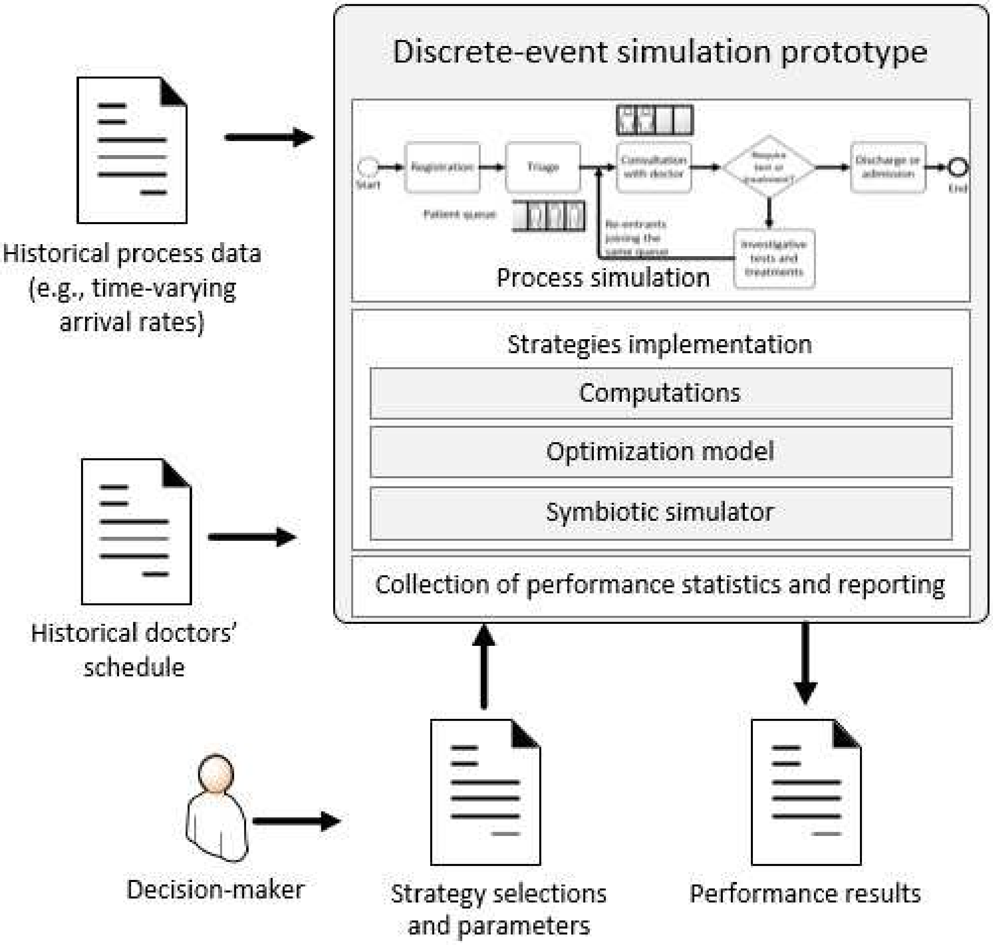

Discrete-event simulation prototype
Start
Registration
Triage
Consultation with doctor
Require test or treatment?
Discharge or admission
End
Patient queue
Re-entrants joining the same queue
Investigative tests and treatments
Process simulation
Strategies implementation
Computations
Optimization model
Symbiotic simulator
Collection of performance statistics and reporting
Historical process data (e.g., time-varying arrival rates)
Historical doctors' schedule
Decision-maker
Strategy selections and parameters
Performance results

Figure 5.5: The DQM prototype

The core component of the DQM prototype is the strategies implementation. This component comprises of *computations*, an *optimization model* and a *symbiotic simulator*. The computations component implements the logic of calculating the number of doctors required for dynamic strategies. The optimization model implements the algorithms for HIST-OPT and DYN-OPT as specified in Table 5.3 and 5.4. The symbiotic simulator is a symbiotic simulation system [16] that evaluates the performance of staffing schedules produced by the optimization model by interacting with the simulated ED process (emulating real-life interactions with physical systems). For more details about the design of a symbiotic simulation system, the reader may refer to Low et al. [29]. Our DQM prototype then collects and reports the performance measurements of each of the strategies.

## 5.7 Experimental Evaluation

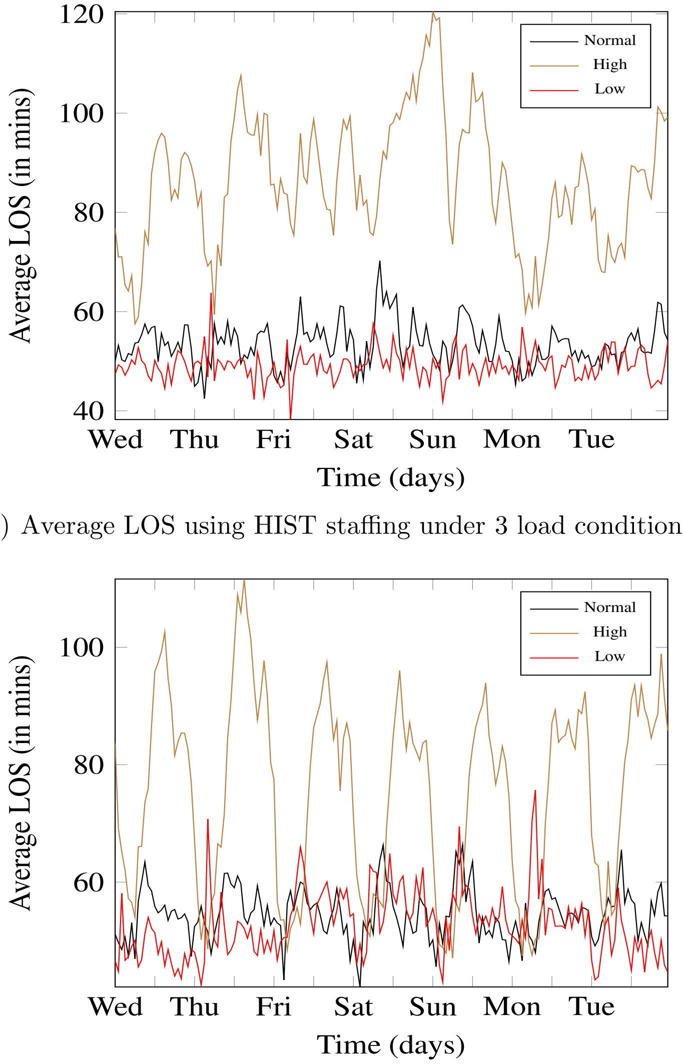
### 5.7.1 Experimental Setup

Similar to the previous chapter, we set up our experiments using six months' data from the selected local hospital. Each experiment was run over 100 replications and the result was an average of the replications. In each optimization for HIST-OPT and DYN-OPT, 50 replications were used in the symbiotic simulator and the average was taken to evaluate the solution. The maximum search iteration was set to 300. In DYN-OPT, the lead time was set to 0 (to plan for the next hour) and the planning horizon was set to 8 hours. We derived the parameters $\lambda_b(t)$, $\lambda_f(t)$, $\mu$, $\delta$ and $b$ from historical data using a commercial business analytics software, SAS. The time-varying arrival rates ($\lambda_b(t)$, $\lambda_f(t)$) were recorded over a week, and each day had its own time-varying arrival rates that changed every hour, which we refer to as the *historical arrival rates* (Figure 2.4). The probability of re-entrance $b$ was 0.4. The average service rate of doctors $\mu$ was 4 per hour. The QED service parameter $\beta$ was set to 0.5, one of the values which yields more stable expected wait for small systems (which is similar to our context) [58] and minimizes over-buffering of doctors. The registration and triage service times were assumed to be exponentially distributed with a mean of 14.2 minutes. The hospital's desired service quality, $LOS_{max}$ was set to 60 minutes. The hourly back-room requirements were pre-determined by the hospital and were fed into the simulator as inputs.

For each set of experiments (for all strategies), a simulation over 9 days was executed and the first and last days were discarded in order to remove inaccurate results from the simulation start-up and completion. The remaining 7 days represented the 7 days of the week with different patterns of arrivals as observed in real life. To simulate various real-life scenarios, we ran the experiments under three conditions: normal, low and high load. The normal load condition assumes the arrival rates to be exactly the same as historical rates. In the low load condition, the arrival rates on Thursdays, Fridays, Saturdays and Sundays are halved. In the high load condition, the arrival rates of the same affected days are doubled.

### 5.7.2 Experimental Results

The first experiment was to compare HIST and DYN (which use the staffing rule) to illustrate the value of real time information. As shown in Figure 5.6, we observe that DYN is capable of reacting to surges. The doctor staffing schedule generated by DYN was able to keep the average LOS stable under the conditions of low and normal loads. Although the average LOS was higher under the high load condition, the doctors' schedule generated by DYN allowed the queue to clear. In the case of HIST, if the load was high, the average LOS remained higher than the desired service quality over the period of demand surges (Thursday to Sunday).



(a) Average LOS using HIST staffing under 3 load conditions

(b) Average LOS using DYN staffing under 3 load conditions

Figure 5.6: Results of varying demand for the strategies using staffing rule

Table 5.5 (and its pictorial representation in Figure 5.7) shows the average number of doctor hours required in a week under the given strategy and demand conditions. Under the high-load condition, the percentage increase in the number of doctors was only about 9%. However, if the demand dropped, the number of doctors required dropped by 22%. We feel that the HIST method led to slight over-staffing. Supposed that an ED, over time, has an equal probability of experiencing high, normal or low loads, then the average cost of using a dynamic strategy (computed to be an average of 650 doctor hours) was lower than that of a static one at 681 doctor hours based on historical data.

|  | Normal | High | Low |
|---|---|---|---|
| HIST | 681 | 681 | 681 |
| DYN | 678 | 743 | 530 |
| % Change | -0.4 | 9 | -22 |

Table 5.5: Comparison of number of doctor hours required per week between the HIST and DYN strategies

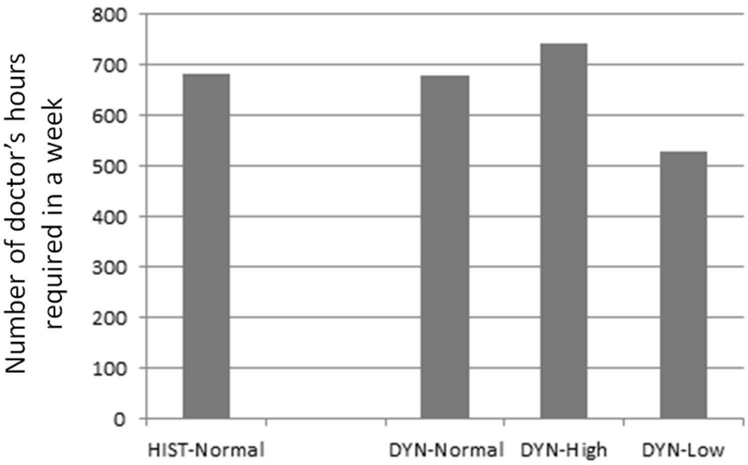


Figure 5.7: Number of doctors required

We can see that the staffing rule provided reasonably good solutions. However, neither HIST nor DYN guarantee that the solution can satisfy the service level constraint given in Equation 5.6. We provide an example of use of the optimization model in DQM to find a solution that satisfies the service level constraint. Figure 5.8 shows an instance of a simulation evaluation when the ED can achieve a performance within the desired service level using an optimization model such as HIST-OPT. This is observable under normal- and low-load conditions. The HIST-OPT strategy requires an additional increase of only eight doctor hours per week compared to HIST. However, we need to understand that this performance is not guaranteed because some patients may require more time to be monitored and be treated with quality care in a real-life situation. The results in this figure also show that HIST-OPT is a static method and is not able to react to high-load conditions.

The results from DYN-OPT provide interesting insights. We found that DYN-OPT can react to surges but its performance is limited by the length of its planning horizon. Figures 5.9(a) and 5.9(b) show the results of DYN-OPT

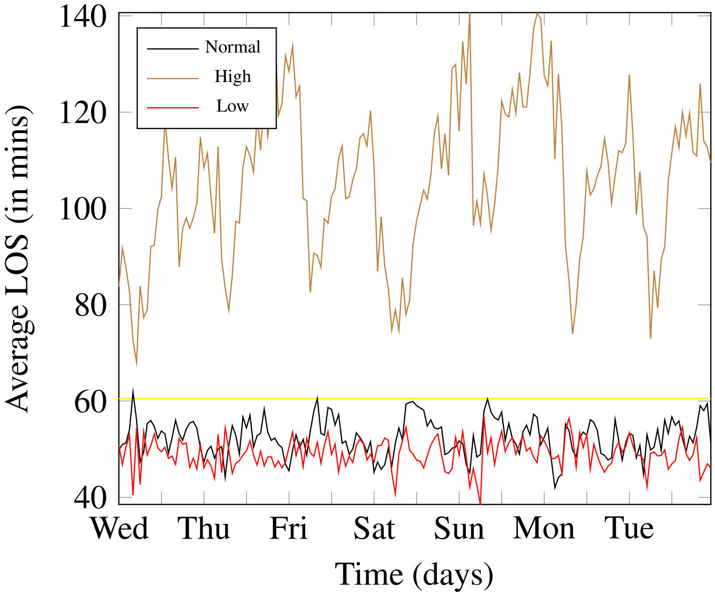
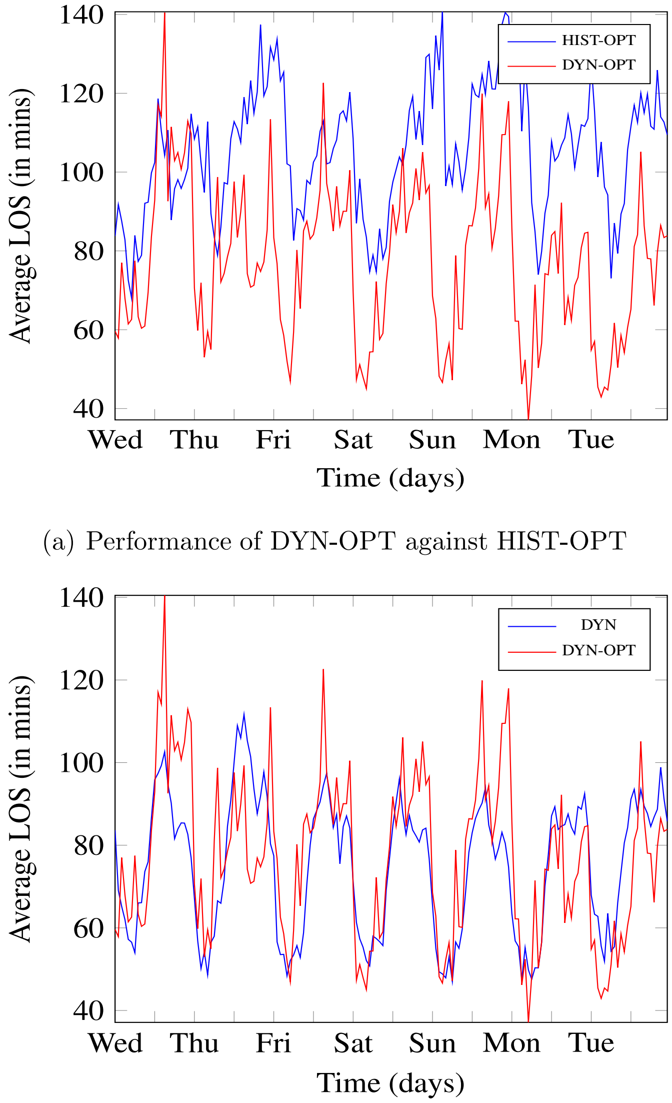


Figure 5.8: Performance of HIST-OPT under three load conditions

versus the optimized counterpart of HIST-OPT and the dynamic counterpart of DYN under the high-load condition. We found that DYN-OPT performs better than HIST-OPT as it can better respond to surges in demand. On the other hand, as the planning horizon uses the historical arrival rates (but the actual arrival rates are higher), the performance degrades over this horizon. From the DYN-OPT curve in either of the figures, we can see that at every x-axis tick (which is the eight-hourly DYN-OPT's horizon), there is a drop in the average LOS at the start of each horizon. This is because the method will indicate that more doctors are required to serve the patients under the unexpected high load. The entire horizon is then planned based on historical arrivals which are only half of the actual arrivals. As such, the queue builds up again during this period when there is no resource planning until the next horizon. DYN-OPT is able to react to demand surges by lowering the LOS at the start of each horizon, and we see that strategy DYN is better at handling surges. This is because DYN has the opportunity to adjust the resources every hour. However, one may find DYN to be a strategy that is too reactive and so the hospital may not be able to find doctors at short notice. We see more opportunity to further improve the DYN-OPT strategy or investigate how the DYN-OPT parameters, such as the horizon, can affect or improve the



(a) Performance of DYN-OPT against HIST-OPT

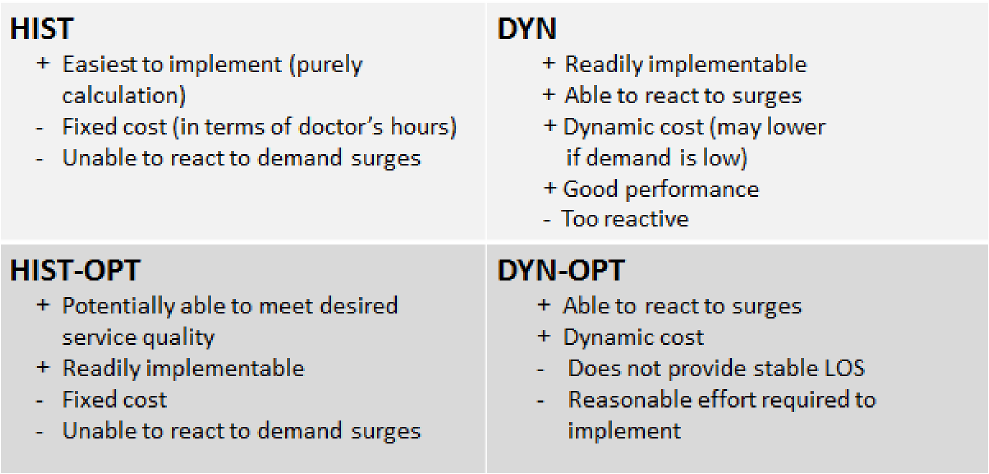
(b) Performance of DYN-OPT against DYN

Figure 5.9: Results of DYN-OPT against its optimized and dynamic counterparts

performance. DYN-OPT is potentially a good short-term planning strategy. As we recognize through the DYN and HIST-OPT strategies though, there are benefits in using real-time information and an optimization method.

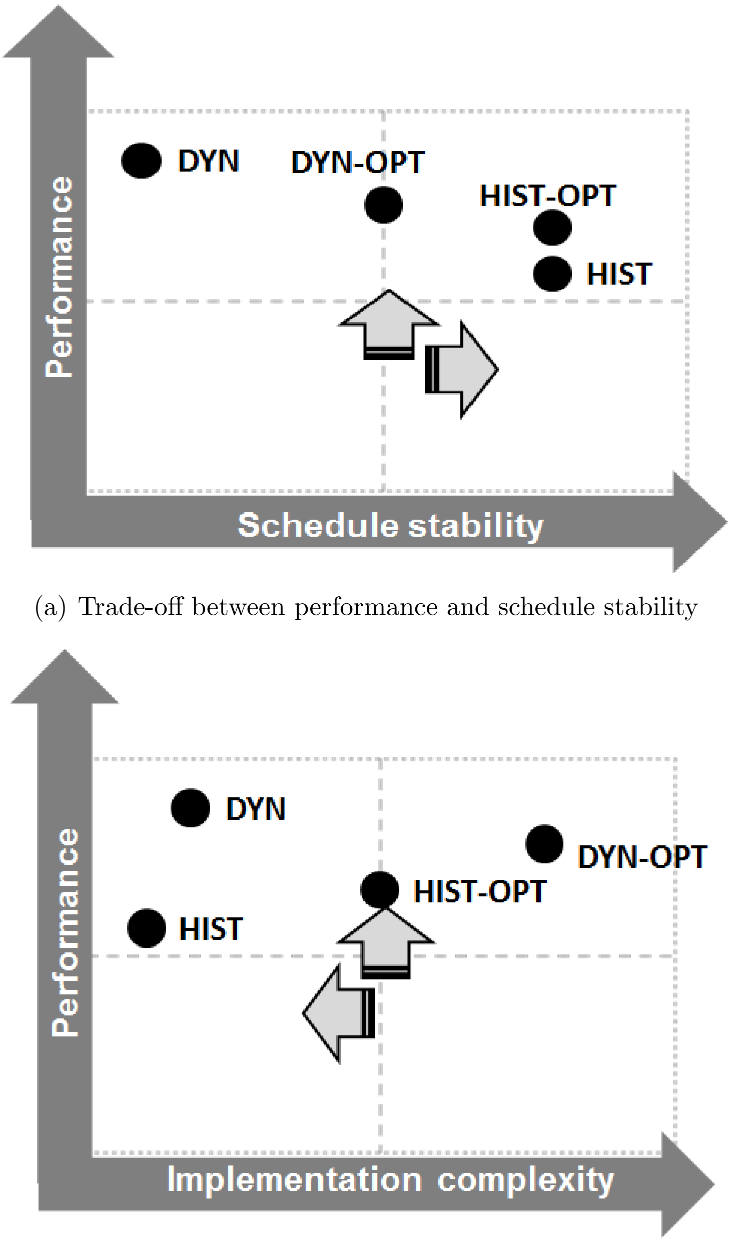
## 5.8 Management Insights for Decision-Makers

The advantages and disadvantages of our strategies are shown in Figure 5.10. Among the four strategies for resource adjustment, we found that there were two major trade-offs. We show the trade-offs in Figures 5.11(a) and 5.11(b). The arrows on the figures indicate the directions that a decision-maker may like to take. For stability, we should try to achieve high performance and high stability. For implementation complexity though, we should try to achieve high performance with low implementation complexity. On the other dimension, we want the strategies to be able to respond to demand surges. We summarize this in Figure 5.12. In this figure, we know that HIST is not able to respond to surges. HIST-OPT is also unable to react but it is slightly better because the HIST-OPT schedule contains more staff (doctors) than HIST. Hence, there is a higher buffer to serve the patient if there is an increase in demand. DYN is on the other extreme. DYN-OPT is a good balance and its ability to react to surges is dependent on the length of the planning horizon.



| | |
|---|---|
| **HIST**<br>+ Easiest to implement (purely calculation)<br>- Fixed cost (in terms of doctor's hours)<br>- Unable to react to demand surges | **DYN**<br>+ Readily implementable<br>+ Able to react to surges<br>+ Dynamic cost (may lower if demand is low)<br>+ Good performance<br>- Too reactive |
| **HIST-OPT**<br>+ Potentially able to meet desired service quality<br>+ Readily implementable<br>- Fixed cost<br>- Unable to react to demand surges | **DYN-OPT**<br>+ Able to react to surges<br>+ Dynamic cost<br>- Does not provide stable LOS<br>- Reasonable effort required to implement |

Figure 5.10: Summary of pros and cons of the four supply-side strategies



(a) Trade-off between performance and schedule stability

(b) Trade-off between performance and implementation complexity

Figure 5.11: Supply-side strategy matrices for decision-maker on strategy selection

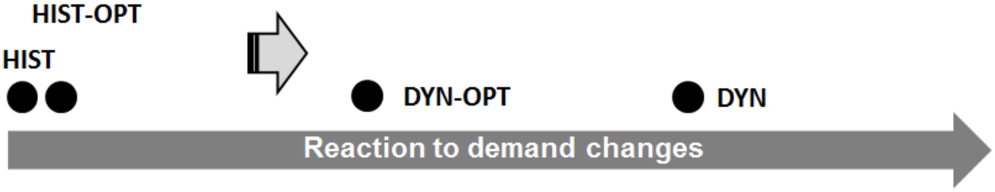


Figure 5.12: Ability to react to demand surges for various strategies

## 5.9 Chapter Summary

In this chapter, we showed strategies for resource allocation in the front room of the ED. We proposed proactive strategies (with and without optimization) and dynamic strategies, while maintaining the service quality in the back room. We found in our experiments that our methods yielded better average length-of-stay for all patients compared to a typical method that purely uses historical data and the staffing rule (HIST). Since real-time data is easily accessible, the benefit of incorporating symbiotic simulation to optimize resource allocation in a dynamic situation is valuable. We implemented a prototype with the use of symbiotic simulation via the DYN-OPT strategy. With the advances in enterprise systems that enable the real-time movements of patients to be tracked, data can be fed quickly into a real-time simulator such as our symbiotic simulator. So a dynamic resource allocation is implementable. Although it has a higher requirement for doctors than the non-optimized strategies, it yields significant benefits in reducing the LOS of patients for various arrival conditions. The other strategies that we proposed provide alternatives for hospital decision-makers to select other effective methods that do not require symbiotic simulation (DYN) or pre-planned optimized schedule (HIST-OPT).

# Chapter 6

# The Integrated Dynamic Queue Management Framework

This chapter offers an integrated framework to manage queues *dynamically* in the ED from both the demand and supply perspectives by deploying intelligent dynamic patient-prioritization strategies and dynamic resource-adjustment strategies developed in the previous two chapters (Figure 6.1). We explore the effects of such integration on the performance of the various combinations of demand-side and supply-side strategies.

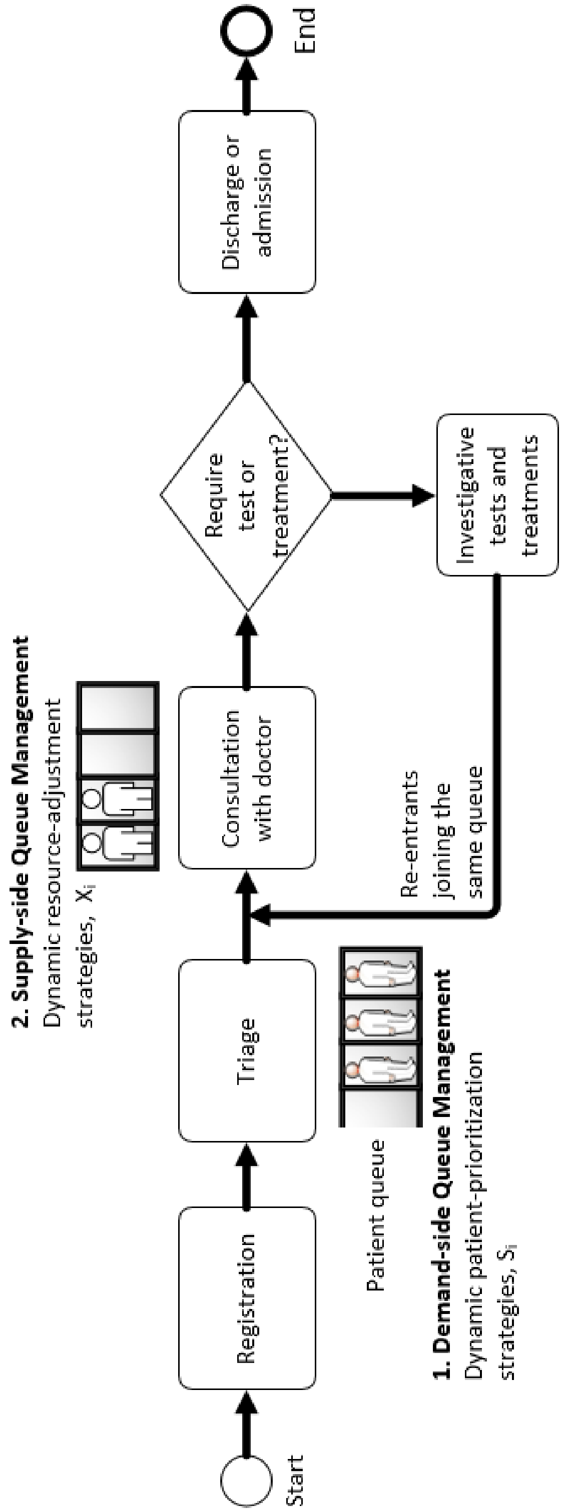

Start
Registration
Triage
Consultation with doctor
Require test or treatment?
Discharge or admission
End
Investigative tests and treatments
Re-entrants joining the same queue
Patient queue
1. Demand-side Queue Management
Dynamic patient-prioritization strategies, $S_i$
2. Supply-side Queue Management
Dynamic resource-adjustment strategies, $X_i$

Figure 6.1: Deploying dynamic patient-prioritization and dynamic resource-adjustment strategies to ED process.

## 6.1 The Dynamic Queue Management Framework

There are three main components in the framework, namely Live System and Data, Analytical Model and Decision-Support Model. A diagram illustrating the contents of the three components is shown in Figure 6.2.

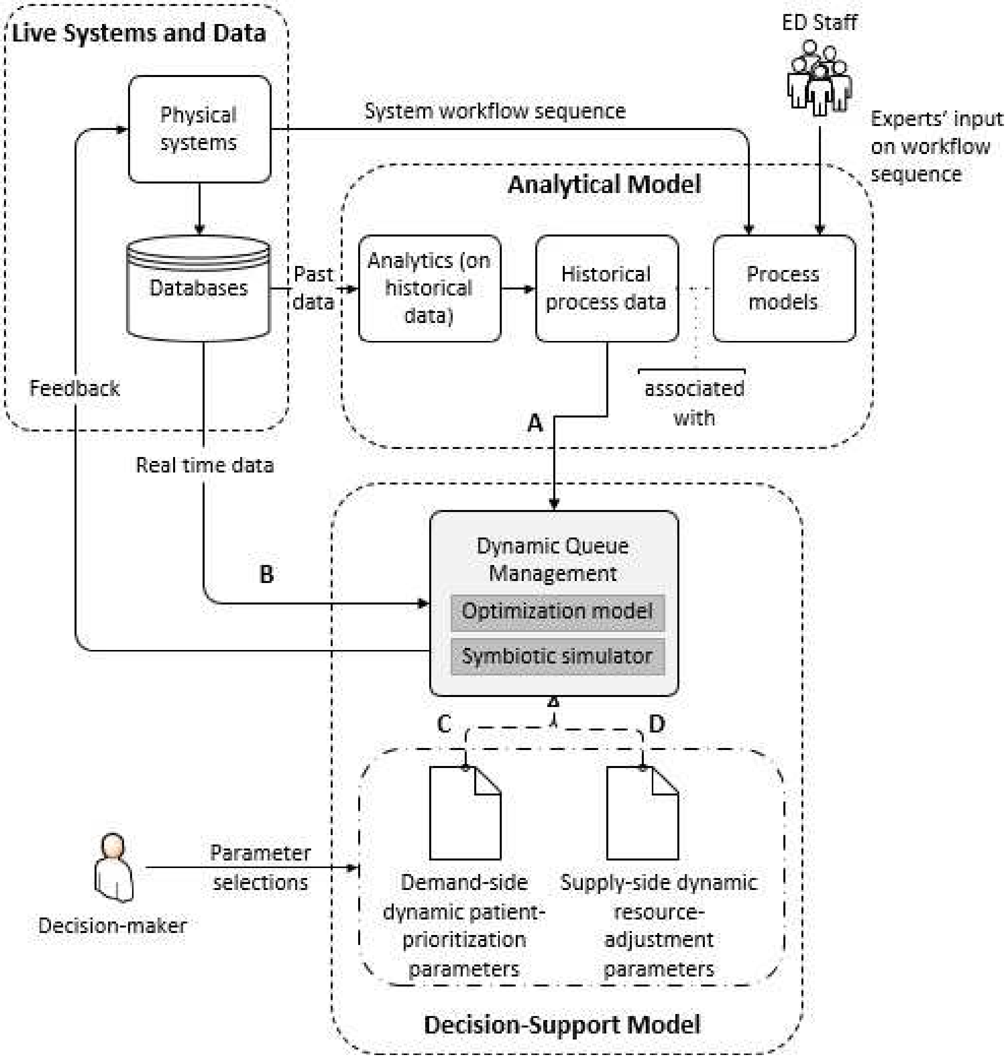


Figure 6.2: The Integrated Dynamic Queue Management Framework

In this framework, we unify the demand-side and supply-side queue models. Referring to Table 6.1, the integrated model now considers demand in the critical-care area. It has the ability to incorporate demand-side dynamic priority (or the decision-maker can also choose FIFO) in the patient queue discipline. It uses a set of exponential distributions with corresponding probabilities of occurrence for the service rates of doctors. And it has a varying number of doctors for different hours of the day and days of the week. The service rates of investigative tests and treatment remain the same as the supply-side model because we approximate the set of exponential distributions using a single exponential distribution and our initial experiments showed that there were no significant differences in the results. In order to ensure that the analytical and simulation models remain valid with these parameters, we will show how we can make use of the properties of the exponential distributions to estimate a single service rate of the doctor and a single service rate of tests and treatment so they can be used in the calculation of the staffing requirements.

| | Demand-side model | Supply-side model | Integrated model |
|---|---|---|---|
| Back room considerations | No | Yes | Yes |
| Queue discipline | Dynamic-priority | FIFO | FIFO and Dynamic-priority |
| Service rates of doctors ($\mu$) | Set | Single | Set |
| Service rates of tests and treatments($\delta$) | Set | Single | Single |
| Hourly number of doctors | Constant | Varying | Varying |

Table 6.1: Unifying demand-side and supply-side queue models into integrated model

### 6.1.1 Live Systems and Data

This consists of the various information technology (IT) physical systems and databases that support the live operations of the ED and its supporting departments. For example, the ED processes interact with processes in laboratory (e.g., blood tests) and X-ray departments. The processes in laboratory and X-ray departments use the Laboratory System and X-ray System respectively. The Laboratory System contains the blood test results and the X-ray System contains X-ray test results of a patient who has undergone each test. The live systems provide data that serve as the input to the analytical model and the decision-support model. An example of the live systems supporting the ED process is shown in Figure 4.2.

### 6.1.2 Analytical Model

We build our initial process models and queuing models based on historical data from the live systems. The components in the analytical model are:

- Analytics (on historical data): A set of actions performed on historical data using commercial business analytics software (such as SAS). Historical data represents a snapshot of the live data at the point when it is taken. Output is *Historical Process Data*.
- Historical Process Data: ED process parameters such as time-varying arrival rates, service rates of triage, service rates of doctors, service rates of investigative tests and treatment, and probability of re-entrance.
- Process Models: Depict the real-life processes of the ED. An example of a simplified process flow is as in Figure 2.2.

The Analytical Model provides a set of outputs (reflected as Input Set A in Figure 6.2) that are required by the Dynamic Queue Management module in order to carry out decision-support functions. This set of outputs consists of:

- $\lambda_b(t)$ - Time-varying arrival rate of new patients in the back room. Each $\lambda_b(t)$ is defined per hour over a week's horizon.
- $\lambda_f(t)$ - Time-varying arrival rates of new patients in the ambulatory area. Each $\lambda_f(t)$ is defined per hour over a week's horizon.
- $\mu_b$ - Service rate of the doctors in the back room.
- $\mu_n$ - Service rate of the doctors for new patients. We assume a homogeneous service rate for all doctors.

- $\mu_r$ - Service rate of the doctors if the patient is a re-entrant. The review consultation time consists of a set of four exponential distributions with corresponding probability of occurrence. A patient with clean test results usually takes a shorter duration compared to a patient with complex test results.

- $\delta$ - Service rate for investigative tests or treatment, assumed to be a set of exponential distributions with corresponding probability occurrence. This can be simplified to a single exponential distribution using an estimate.

- $b$ - Probability of re-entrance.

### 6.1.3 Decision-Support Model

The main component of the decision-support model is the Dynamic Queue Management (DQM) module. We enhanced this module (from the supply-side implementation) to also handle the demand-side dynamic priority queuing strategies. Hence, in our prototype, DQM can now generate plans on the supply of doctors based on the selected supply-side strategy and also dynamically prioritize patients in the patient queue with selected demand-side prioritization strategies. We mimic the real world by having the DQM prototype contain a discrete-event simulator that generates events such as patient arrivals, distributions of registration, triage, consultation, treatment and investigative tests. As mentioned, DQM contains an *optimization model* and a *symbiotic simulation system* (named as the symbiotic simulator).

A decision-maker makes decisions to select the dynamic queue-prioritization parameters and the dynamic resource-adjustment parameters. Strategies selection, associated parameters and real-time data are fed into DQM collectively. DQM generates output (e.g., which patient to serve next, how many doctors are required) and the information is fed back into the live systems for real-time execution. On the supply side, if the decision-maker chooses the HIST-OPT or DYN-OPT strategy, the optimization model will be used to perform the schedule optimization. If the supply-side strategy DYN-OPT is selected, the symbiotic simulator is used in real time to perform optimization for the specified planning horizon during the course of simulation. If HIST-OPT is selected, the symbiotic simulator is used as a pre-processing tool to obtain the optimized doctors' schedule. For the purpose of evaluating the HIST-OPT schedule, DQM is used for the second time to obtain the performance.

DQM receives real-time data (as reflected as Input Set B in Figure 6.2) from the live systems to support the demand and supply strategies. Input Set B requires the real-time time-varying arrival rates of new patients $\lambda'_f(t)$ in the ambulatory area.

In DQM, the symbiotic simulator is used to evaluate which doctors' schedule is to be used for the planning horizon such that there is either no LOS violation or the least violation. We use the concept of a snapshot. A snapshot contains the current queue conditions, doctor availability, patients' statuses and the realized arrival rates. At the start of each planning period, a snapshot is taken and is used with the historical arrival rates for the planning horizon. A heuristic local search algorithm is then applied to find the best schedule. When the schedule has been found, the snapshot is restored and the best schedule is used in the DQM as the schedule for the next horizon.

A decision-maker must provide two sets of parameters (as reflected as Input Set C and D in Figure 6.2) to allow DQM to work.

The decision-maker must provide the following parameters (as reflected as Input Set C in Figure 6.2) to the demand-side dynamic patient queue-prioritization configuration. Input Set C contains:

- $S_i$ - Dynamic patient-prioritization strategy where $i$ is the selected strategy type.
- $LOS_{max}$ - Hospital's desired service quality in terms of LOS.

Input Set D contains:

- $X_j$: Resource-adjustment policy $j$.
- $LOS_{max}$: Hospital's desired service quality in terms of LOS. This should be consistent with the value set in Input Set C.

- $room_{max}$: Physical constraints in the ED's ambulatory area, which correspond with the maximum number of consultation rooms in the real-life set up of the ED.

- $S_{max}(t)$: Maximum number of doctors that can be deployed in the ED (both areas of the ED combined) at time $t$.

- $C_l$: Cost of labor for deploying a doctor for a single unit of time $t$.

- $C_d$: Cost of deviation per doctor. This is applicable when the number doctors at time $t$ is different from the number of doctors at time $t-1$.

- $L$: Lead time for dynamic planning.

- $H$: Time horizon for dynamic planning.

## 6.2 Implementation Design

In a real-life implementation, the discrete-event simulator of the DQM prototype should be replaced and implemented as a physical system that interacts with the optimization model and the symbiotic simulator. Hence, we propose a design of the Dynamic Queue Management System (DQMS) with symbiotic simulation to support *implementation* of the DQM strategies, the entire suite of demand-side and supply-side strategies. An example of a possible design is presented in Figure 6.3. In this figure, we show how various systems can be responsible for feeding the DQMS with the information that it requires to enable the proposed strategies to be executed in real time. The outputs of the DQMS are (1) the next patient to serve, and (2) the short-term dynamic doctors' schedule. Each of the outputs is fed back into the system supporting the parts of the ED process.

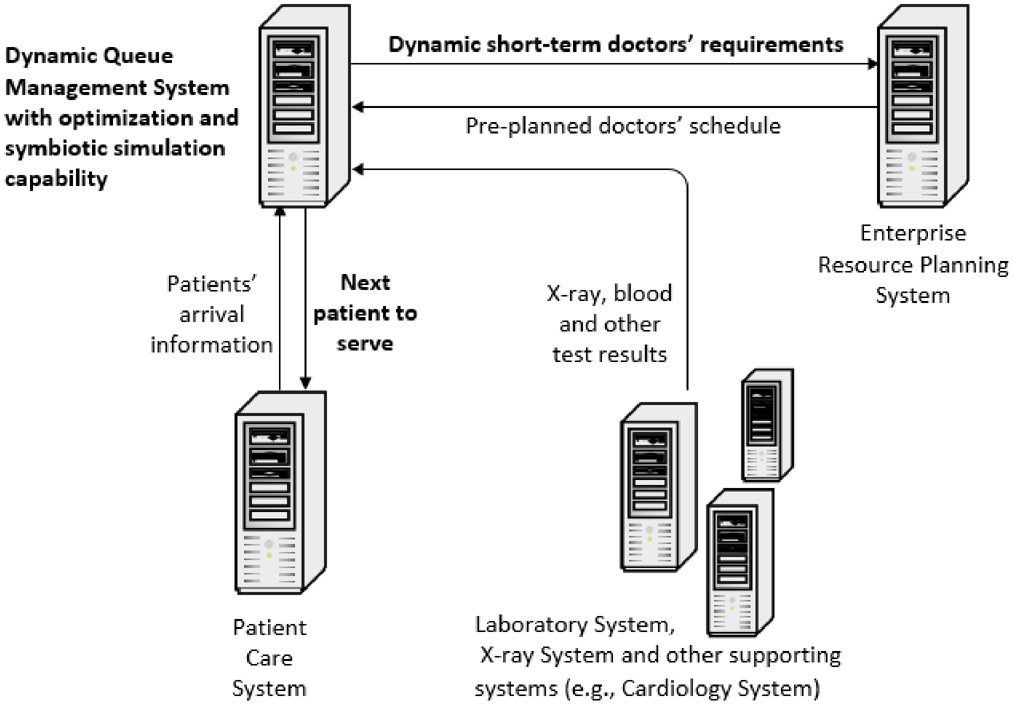

Dynamic Queue Management System with optimization and symbiotic simulation capability
Dynamic short-term doctors' requirements
Pre-planned doctors' schedule
Enterprise Resource Planning System
Patients' arrival information
Next patient to serve
X-ray, blood and other test results
Patient Care System
Laboratory System, X-ray System and other supporting systems (e.g., Cardiology System)

Figure 6.3: An example of implementation design

## 6.3 Experimental Evaluation

For evaluation of how the various demand-side and supply-side strategies work together, we used the DQM prototype as shown in Figure 5.5. We implemented the three demand-side strategies (SCON, SREM and MIXED) as part of the *computations* component of the prototype. We also allowed the decision-maker to make selections of the demand-side and supply-side strategies.

### 6.3.1 Experimental Setup

Similar to previous experiments, our experiments were run using six months' data from the selected hospital. Each experiment was run over 100 replications and the results were averaged over the replications. As for HIST-OPT and DYN-OPT, which require the symbiotic simulation system, each symbiotic simulation was run over 50 replications and the average was taken over the symbiotic simulation replications. The maximum search iteration was set to 300. In DYN-OPT, the lead time was set to 0 (plan for next hour) and the planning horizon was set to eight hours.

In order to verify that our simulator was sufficiently close to real-world performance, we used a FIFO patient queue and a static doctors' schedule in a verification experiment. The outcome of the experiment showed that the differences in the means and standard deviations of the actual hospital data and the results from the DQM simulator were less than 5% and 10%. The ranges of LOS (minimum and maximum) were also consistent. We conclude that the results from DQM simulator are representative of the performance of the ED process.

For each set of experiments (for all strategies), a simulation over nine days was run, and the first and last days were discarded in order to remove inaccurate results from simulation start-up and completion. The remaining seven days represent the seven days of the week with time-varying arrivals, as observed in real life. Through the Analytical Model, the probability of re-entrance $b$ was found to be 0.4. The average service rate of doctors for new patients, $\mu_n$, was four per hour. The review consultation time consisted of a set of four exponential distributions with corresponding probabilities of occurrence as shown in Table 4.2. The registration and triage service times were exponentially distributed with a mean of 14.2 minutes. The hospital's desired service quality, $LOS_{max}$ was set to 60 minutes.

### 6.3.2 Service Rates Estimates for Staffing Requirement Calculations

Although the service rate of the consultation station is represented by a set of exponential distributions with associated probabilities in the experiments, the staffing requirement calculations require a single service rate. The set of exponential distributions yield a hyperexponential distribution. Statistically, we know that the mean of this distribution is equal to the weighted average of the means of the underlying set of exponential distributions. Experimentally, we found from our dataset that, with the inclusion of the service rate for new patients $\mu_n$, the resulting hyperexponential distribution could be approximated closely by an exponential distribution whose mean was equal to the mean of the hyperexponential distribution. The evidence (see Figure 6.4) is derived from an experiment that simulates the service times provided by both the approximated exponential distribution and the service rates represented by the original set of exponential distributions. As such, the staffing requirements for HIST and DYN in our experiment can be computed by solving the ordinary differential equations in Equation 5.3 using the mean of the resulting hyperexponential distribution as the single service rate $\mu$. Using a similar estimation method, the rate of investigative tests ($\delta$) is computed and can be set to 2.3 per hour for the supply-side staffing requirements calculations.

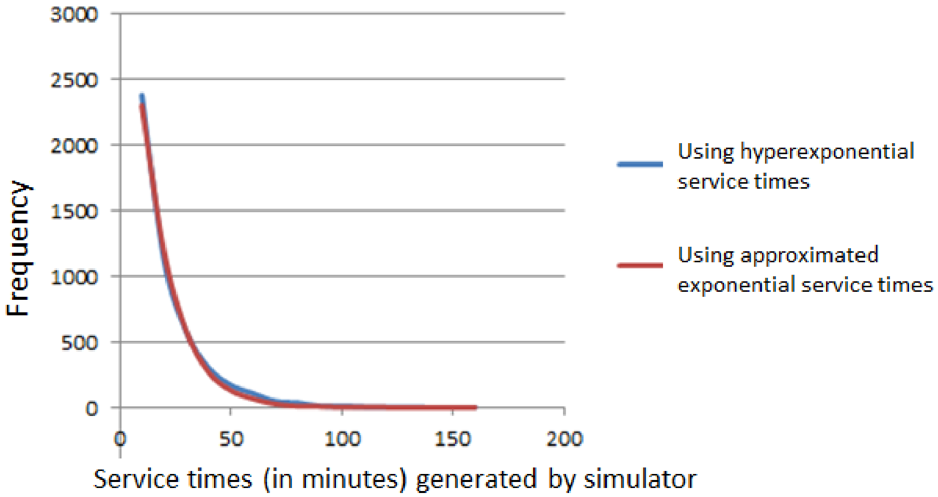


Figure 6.4: Approximation of hyperexponential distribution with exponential distribution.

### 6.3.3 Statistical Test Setup

In most cases, we charted the results on the graphs so that we could visually see which strategy was better. However, in some cases, the results were very similar, hence further steps were required to rank the strategies. We needed a statistical method to compare the result sets. The purpose of a statistical test was to compare the two sets of data to provide evidence of whether they were significantly different. This was used for the what-if analysis. For example, a decision-maker wants to evaluate two choices: (A) a MIXED strategy on the demand side and a DYN strategy on the supply side, or (B) an SREM strategy on the demand side and a DYN strategy on the supply side. Since the supply-side strategy remains constant (DYN), we are effectively comparing two demand-side strategies: S1 MIXED and S2 SREM.

Our hypotheses are as follows:

- Null Hypothesis (H0): Strategy S2 has no improvement over Strategy S1.
- Alternative Hypothesis (H1): Strategy S2 has a significant improvement over Strategy S1.

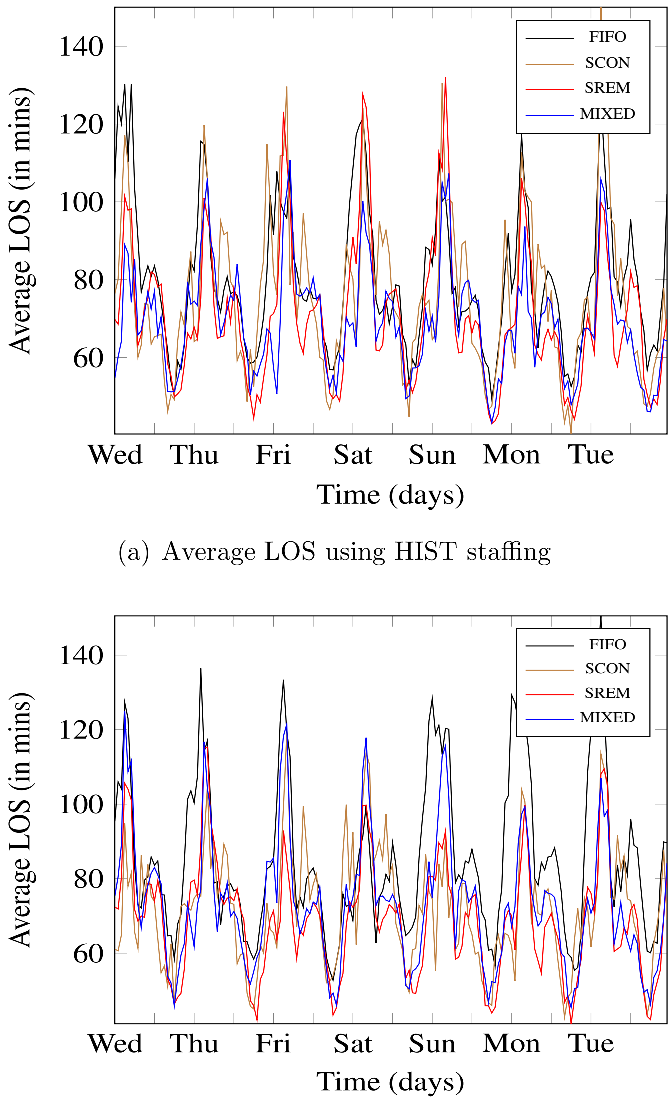
In our attempt to select a suitable statistical test, we evaluated the distribution of the differences for the two strategy selections. We found that the differences did not follow any distribution. As such, we rejected any of the parametric tests such as the Student's t-test and ANOVA, which assume the differences must follow normal distribution. The Kruskal-Wallis test does not assume that the data is normally distributed but it does assume that the observations in each group come from populations with the same shape of distribution. This is again not suitable for our case.

We selected the Wilcoxon Signed-Rank test because it is a non-parametric test which does not assume that the differences must be normal or difference samples must be from the same distribution. We determine whether we reject H0 by the computation of the $p$-value. If the $p$-value is lower than 0.05, we can conclude that the null hypothesis does not hold (i.e., S2 is better than S1).

Using this method, we presented statistically-proven rankings of the various strategies so that a decision-maker can make informed decisions about the demand-side and supply-side strategy pairs that he/she would take.

### 6.3.4 Experimental Results

Two sets of experiments were set up. First, we simulated the real-life situation when each hour had a maximum number of doctors that could be deployed. That means, $S_{max}(t)$ varies hourly. This experiment allowed us to show the effects of the demand-side strategies on the supply-side strategies. We show in Figure 6.5 the results for HIST and DYN as a representation of the static and dynamic supply-side strategies.

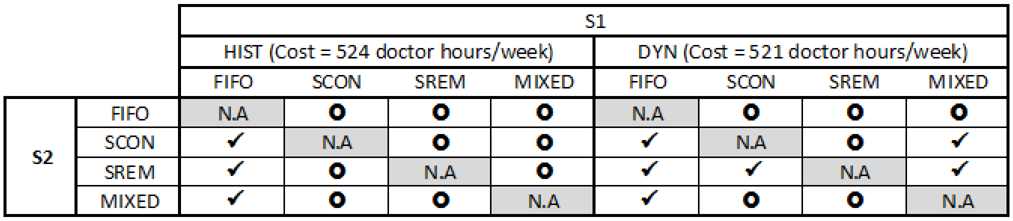


(a) Average LOS using HIST staffing

(b) Average LOS using DYN staffing

Figure 6.5: Results of using demand-side strategies with a selected supply-side strategy.

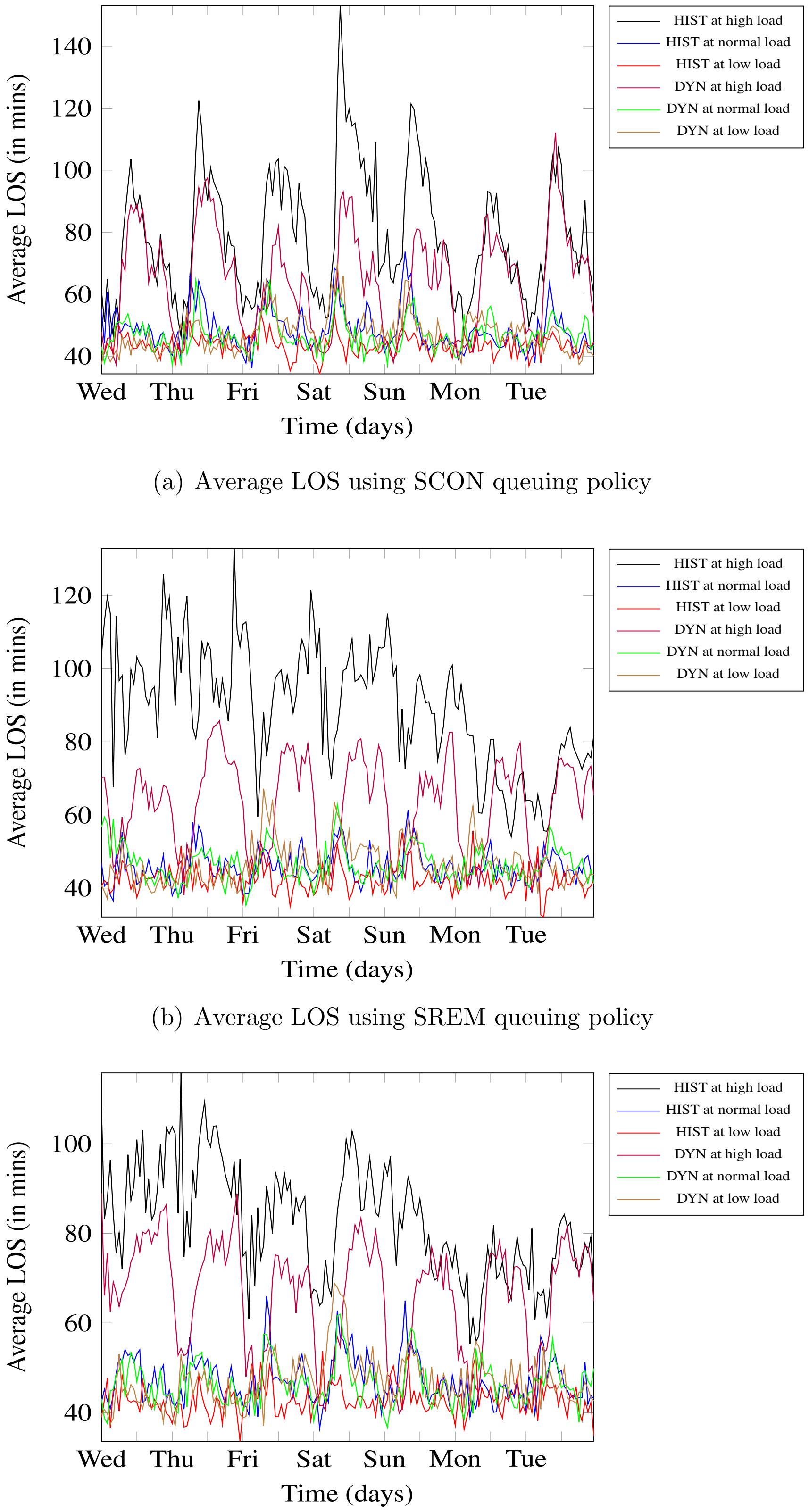
We can see that the demand-side strategies (SCON, SREM, MIXED) have similar performances to FIFO in the HIST and DYN supply-side strategies. One reason for similar performance is that all the supply-side strategies performed well. This is a consistent observation when evaluating the demand-side strategies. When the ED is not as crowded, the improvements provided by the demand-side strategies are not as great compared to FIFO. Although all the strategies (including FIFO) performed well, we discovered from Wilcoxon Signed-Rank tests that the SCON, SREM and MIXED strategies still provide significant improvements over FIFO in HIST and DYN experiments. We present the outcome of Wilcoxon Signed-Rank tests in Figure 6.6. For simplicity, we use a tick to represent values less than 0.05 and a circle otherwise.



| | | S1 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | HIST (Cost = 524 doctor hours/week) | | | | DYN (Cost = 521 doctor hours/week) | | | |
| | | FIFO | SCON | SREM | MIXED | FIFO | SCON | SREM | MIXED |
| S2 | FIFO | N.A | ○ | ○ | ○ | N.A | ○ | ○ | ○ |
| | SCON | ✓ | N.A | ○ | ○ | ✓ | N.A | ○ | ✓ |
| | SREM | ✓ | ○ | N.A | ○ | ✓ | ✓ | N.A | ✓ |
| | MIXED | ✓ | ○ | ○ | N.A | ✓ | ○ | ○ | N.A |

Figure 6.6: The results of Wilcoxon Signed-Rank test.

Our second set of experiments aimed to determine how the supply-side strategies are useful in working together with the demand-side strategies. We now allow the ED to increase or decrease doctors to adapt to demand changes as long as the number of doctors at any time $t$ is still below the physical constraints of the ED. The physical constraint $room_{max}$ in this case is set to five. To test the demand changes, we provided the DQM prototype with three different types of arrival rates: high, normal and low. In the case of a high load, the arrival rates were doubled for Thursdays, Fridays, Saturdays and Sundays. Likewise for the low load, the arrival rates were halved for the same days of the week. Using HIST and DYN as representations of static and dynamic strategies, Figure 6.7 shows how HIST and DYN perform under each demand-side strategy and with high-, normal- or low-load conditions.



(a) Average LOS using SCON queuing policy

(b) Average LOS using SREM queuing policy

(c) Average LOS using MIXED queuing policy

Figure 6.7: Results of using supply-side strategies with a selected demand-side strategy.

As we can see from Figure 6.7(a)-(c), the dynamic method DYN did better with demand surges when compared to HIST. This came with the price of a slight increase in the number of doctor hours deployed at the ED (in a week) to handle the additional load. This is shown in Figure 6.8. The more interesting result is that the corresponding decrease in demand in fact yielded a larger decrease in the number of doctor hours to be deployed, yet the performance is similar to that of HIST which is over-staffed under normal- and low-load conditions. Hence, we conclude that use of a dynamic staffing method is effective in its ability to cope with demand surges and also reduces the amount of doctor effort when the demand is low.

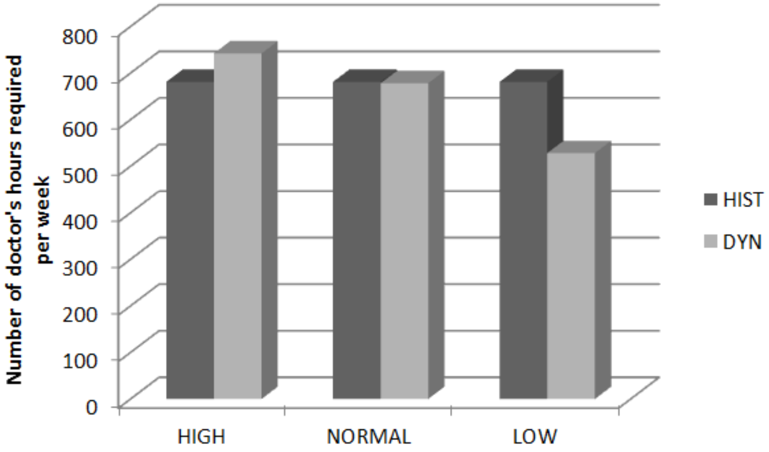


Figure 6.8: The number of doctor hours required for deployment in a week

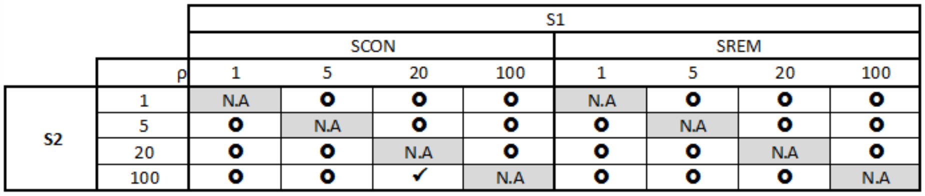
### 6.3.5 Sensitivity analysis of parameters in patient-prioritization functions

The next set of experiments we carried out was the sensitivity analysis of the parameters $\rho_1$ and $\rho_2$ of the demand-side patient-prioritization functions in Equations 4.1 and 4.2. More interestingly, we wanted to evaluate if the results of the MIXED strategy (comprising both SCON and SREM functions) were sensitive to the parameters. As the results were very similar, we used Wilcoxon Signed-Rank tests to test significant differences. The set of values used to set the parameters is (1, 5, 20, 100).

The hypotheses are modified slightly and are set up as follows:

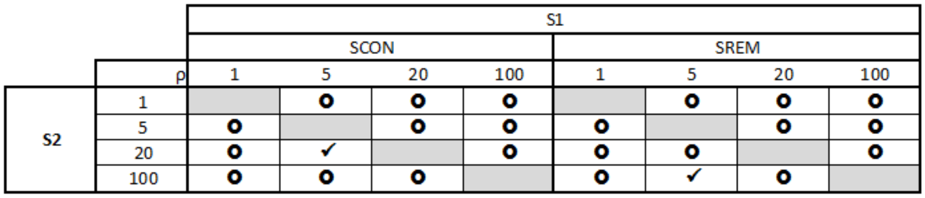
- Null Hypothesis (H0): S1 parameter has no significant improvement over S2 parameter in the same strategy.
- Alternative Hypothesis (H1): S1 parameter has a significant improvement over S2 parameter in the same strategy.

The first set of results showed the outcome of changing a single parameter, e.g., $\rho_1$ or $\rho_2$ in each of the individual strategies. For example, in SCON, we evaluated if the results were significantly different when we changed the value of $\rho_1$. Similarly, we did this for SREM. We ran tests by varying the parameters for HIST strategy as the proactive representative of the supply-side strategy. The results of the Wilcoxon Signed-Ranked tests were shown in Figure 6.9. The symbol $\rho$ shown in the table simply means either $\rho_1$ or $\rho_2$ depending on whether it was for SCON or SREM. The same test for DYN was also done and tabulated in Figure 6.10 for SCON and SREM. For simplicity, we used a tick to represent the situation when the null hypothesis H0 was rejected, i.e., there was a significant difference. We can see from both the diagrams that in most cases, there was no significant difference when varying the patient-prioritization parameters.



| | | S1 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | SCON | | | | SREM | | | |
| | ρ | 1 | 5 | 20 | 100 | 1 | 5 | 20 | 100 |
| S2 | 1 | N.A | ○ | ○ | ○ | N.A | ○ | ○ | ○ |
| | 5 | ○ | N.A | ○ | ○ | ○ | N.A | ○ | ○ |
| | 20 | ○ | ○ | N.A | ○ | ○ | ○ | N.A | ○ |
| | 100 | ○ | ○ | ✓ | N.A | ○ | ○ | ○ | N.A |

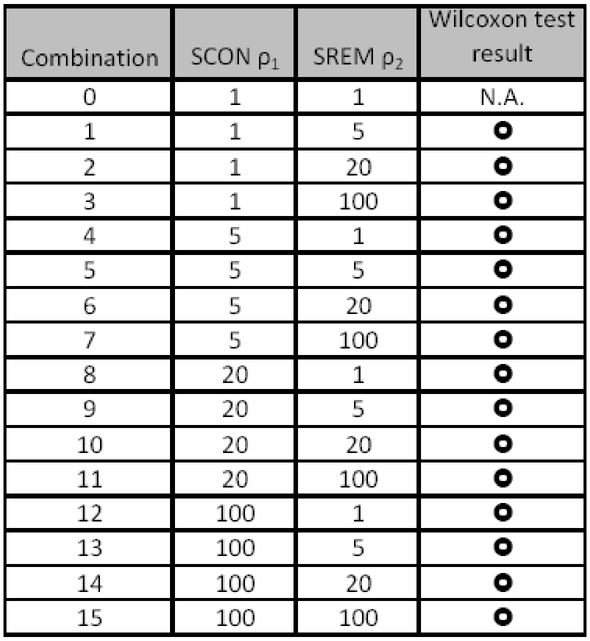
Figure 6.9: Sensitivity test of parameters in SCON and SREM for the HIST strategy



| | | S1 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | SCON | | | | SREM | | | |
| | ρ | 1 | 5 | 20 | 100 | 1 | 5 | 20 | 100 |
| S2 | 1 | | ○ | ○ | ○ | | ○ | ○ | ○ |
| | 5 | ○ | | ○ | ○ | ○ | | ○ | ○ |
| | 20 | ○ | ✓ | | ○ | ○ | ○ | | ○ |
| | 100 | ○ | ○ | ○ | | ○ | ✓ | ○ | |

Figure 6.10: Sensitivity test of parameters in SCON and SREM for the DYN strategy

In our investigation of the effects on the MIXED strategy, we used a single value of $\rho_1$ (e.g., $\rho_1 = 1$) and varied the value of $\rho_2$ ($\rho_2 = 1, 5, 20, 100$). We compared the results against the case when we set $\rho_1 = 1$ and $\rho_2 = 1$, as these were the values we used for all of our other experiments. We used the DYN strategy as the representative supply-side strategy. Wilcoxon Signed-Rank tests were used to establish if the results of using different parameters yielded results which would be significantly different. The results were shown in Figure 6.11. We concluded that the changes in parameters in each contributing strategy produced no significant difference in the results of the MIXED strategy. The results of the demand-side strategies were not sensitive to the parameters of the patient-prioritization functions.

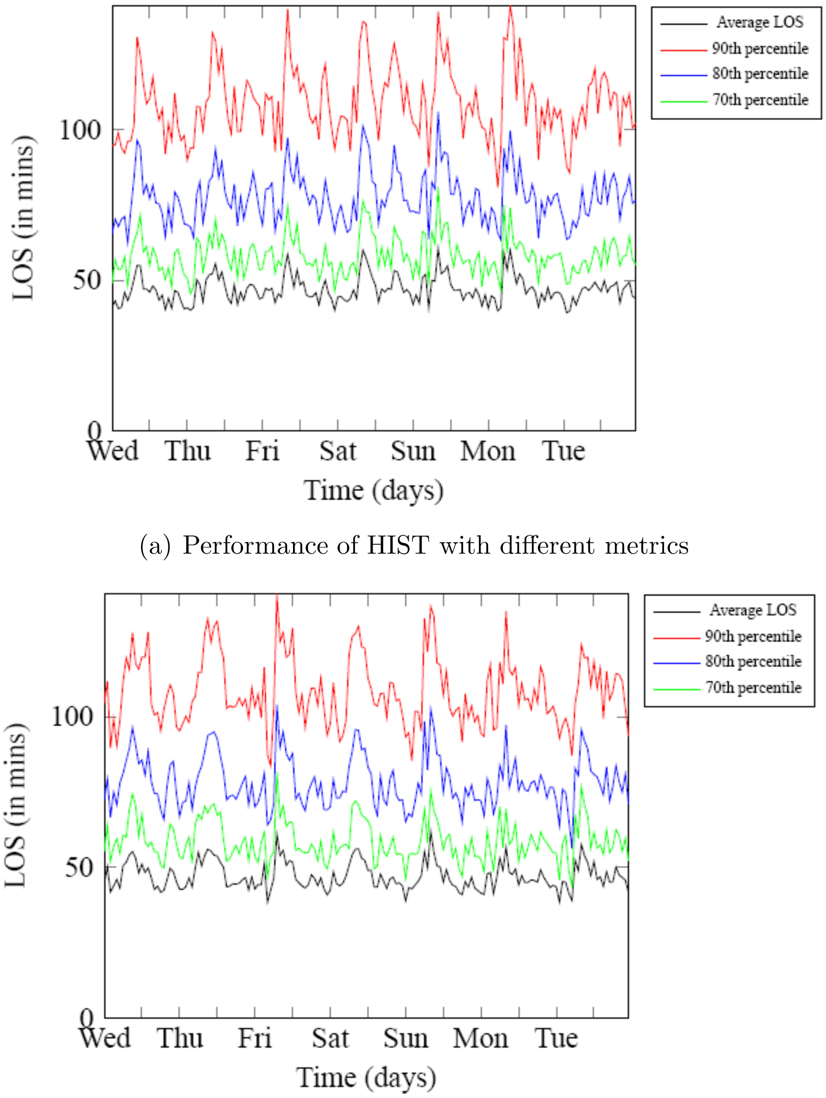


| Combination | SCON $\rho_1$ | SREM $\rho_2$ | Wilcoxon test result |
|---|---|---|---|
| 0 | 1 | 1 | N.A. |
| 1 | 1 | 5 | ○ |
| 2 | 1 | 20 | ○ |
| 3 | 1 | 100 | ○ |
| 4 | 5 | 1 | ○ |
| 5 | 5 | 5 | ○ |
| 6 | 5 | 20 | ○ |
| 7 | 5 | 100 | ○ |
| 8 | 20 | 1 | ○ |
| 9 | 20 | 5 | ○ |
| 10 | 20 | 20 | ○ |
| 11 | 20 | 100 | ○ |
| 12 | 100 | 1 | ○ |
| 13 | 100 | 5 | ○ |
| 14 | 100 | 20 | ○ |
| 15 | 100 | 100 | ○ |

Figure 6.11: Sensitivity test of parameter $\rho_1$ in the MIXED strategy

### 6.3.6 Sensitivity analysis of performance metric

In our next sensitivity analysis, we aimed to investigate if the other measurements such as the $90^{th}$ percentile of the hourly LOS (instead of average LOS) might give a different conclusion to the performances of the various strategies. Supposing we could deploy as many doctors as the physical capacity of the front room allowed, we ran experiments to output the various performance metrics – $90^{th}$ percentile, $80^{th}$ percentile, $70^{th}$ percentile and average LOS. Using the MIXED strategy on the demand side and HIST and DYN on the supply side (HIST to represent a proactive strategy and DYN to represent a dynamic strategy), we show the results in Figures 6.12(a) and 6.12(b).



(a) Performance of HIST with different metrics

(b) Performance of DYN with different metrics

Figure 6.12: Results of HIST and DYN as measured by different performance metrics

From the figures, we observed that the different percentile measurements were different scales of average LOS. We plotted a graph (Figure 6.13) based on the $90^{th}$ percentile measurement when subjected to different demand conditions (high load, normal load and low load) for the MIXED strategy on the demand-side and HIST and DYN strategies on supply side. We observed that the patterns from using other measurements were consistent with the case when we plotted using average LOS. With this, we noticed that the conclusions drawn about the performance of the various strategies were consistent whether we used the average LOS or the $90^{th}$ percentile measurements.

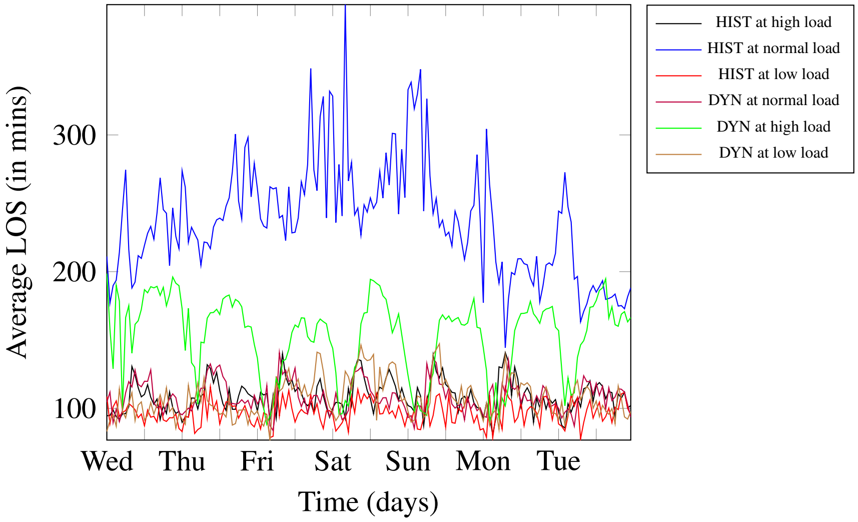


Figure 6.13: Performance of the MIXED strategy with HIST and DYN strategies as measured by $90^{th}$ percentile LOS

### 6.3.7 Sensitivity analysis of randomness

To minimize the effects due to randomness on the results of the simulation runs, besides the 100 replications (and 50 replications for each symbiotic simulations) used in each simulation run, we further investigated if multiple runs of the simulation (of 100 replications each) would yield the same conclusion. In this test, we set the demand-side strategy to the MIXED strategy and we investigated how HIST and DYN responded to various demand conditions. For each of the supply-side strategies (HIST and DYN) and each demand condition, we ran the simulation 10 times. For each simulation run, we collected the average LOS for each hour. We took an average for each hour for the 10 simulation runs. The results of this test are plotted in Figure 6.14. We observed the same pattern as that observed when using a single simulation run as reported in Figure 6.7(c). For example, the DYN strategy was able to cope with demand surges as compared to the HIST strategy. We also determined if the two sets of results have any significant difference for each demand conditions with the use of Wilcoxon Signed-Rank tests. We found that there was no statistically significant difference between the single simulation run and the average of 10 simulation runs. The results of the Wilcoxon Signed-Rank test are shown in Figure 6.15. As such, we report that despite with some randomness, the observations and conclusions drawn were consistent.

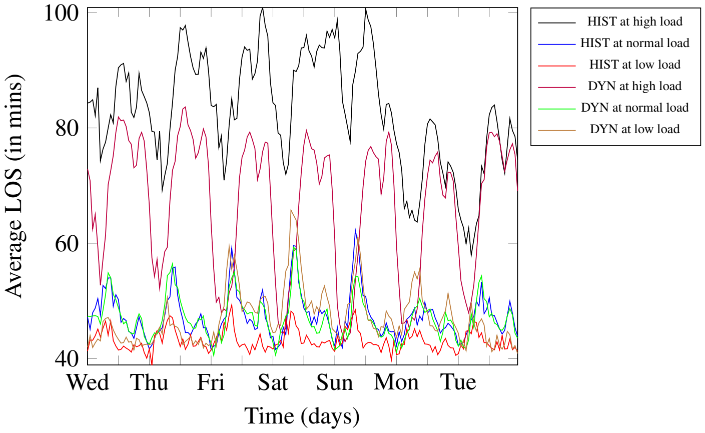
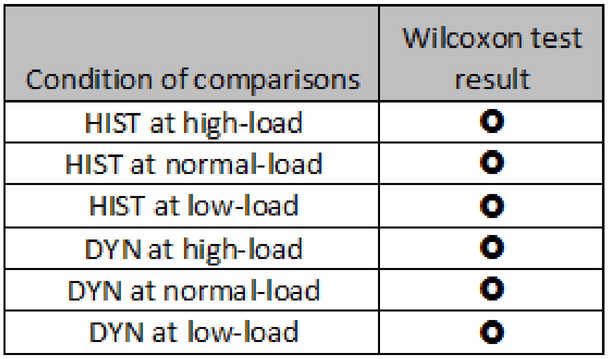


Figure 6.14: Performance of the MIXED strategy on supply-side strategies using average of 10 simulation runs



| Condition of comparisons | Wilcoxon test result |
|---|---|
| HIST at high-load | ○ |
| HIST at normal-load | ○ |
| HIST at low-load | ○ |
| DYN at high-load | ○ |
| DYN at normal-load | ○ |
| DYN at low-load | ○ |

Figure 6.15: Results of Wilcoxon Signed-Rank tests to compare the performances of a single simulation run and the average of 10 simulation runs

## 6.4 Visualization Tool for Decision-Makers

In order to help decision-makers visualize which combination of demand-side and supply-side strategies performs best, we built a simulation-playback visualization. The tool shows a playback of when the patients join the ED, when they go to consultation, undergo tests or treatment, and when they return to the doctors and are discharged.

Many commercial simulation tools in the market offer an animated view of the simulation while it is running. Unlike these, our visualization tool offers the animated view only after a simulation is completed. This is because the simulation is run so quickly that the execution of many strategies will only occur in seconds. We need our simulator to perform quickly due to the large number of replications, especially the supply-side strategy that requires optimization in real time. Our simulation is run too fast to be visualized during execution. As such, we enabled our simulation prototype to export all the simulation events, the number of doctors for each simulation hour and the average LOS for each simulation hour. We then presented the playback using a custom-made visualizer built in Java. The design is shown in Figure 6.16 and an example of the visualizer is shown in Figure 6.17.

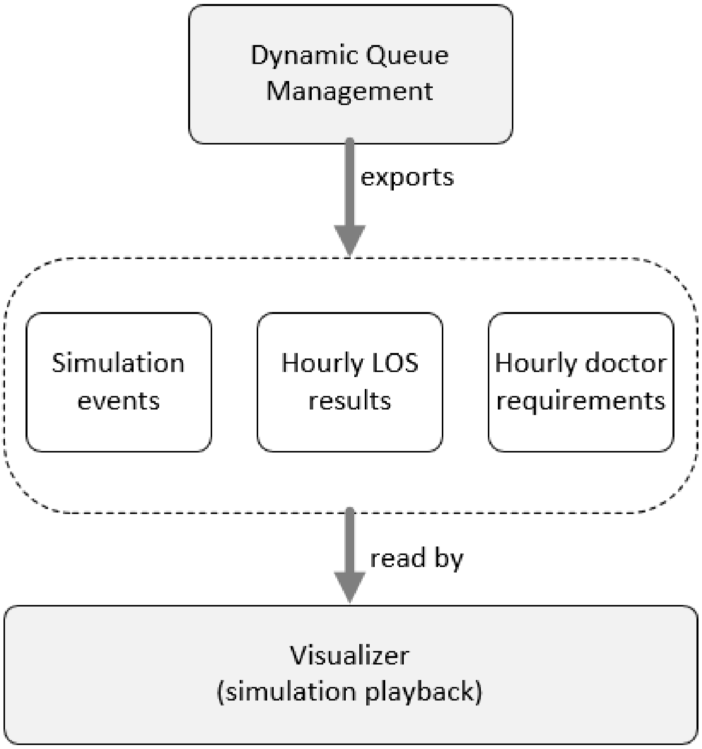


Figure 6.16: Design of Visualization Tool

In our design, the integrated DQM simulator exports the simulation events into a file. The exported events file, the results (hourly LOS) from the simulation results, and the hourly doctor requirements are then fed into the visualizer for a playback of what happened during the simulation.

As shown in Figure 6.17, each dot on the visualizer is a patient. If a patient has waited past the target LOS, the dot is displayed in red, otherwise, it is green. Although the current version of the visualization tool is simplistic, it serves its purpose to provide validation for non-IT personnel (or healthcare personnel) to see the effects on the ED of applying the strategies.

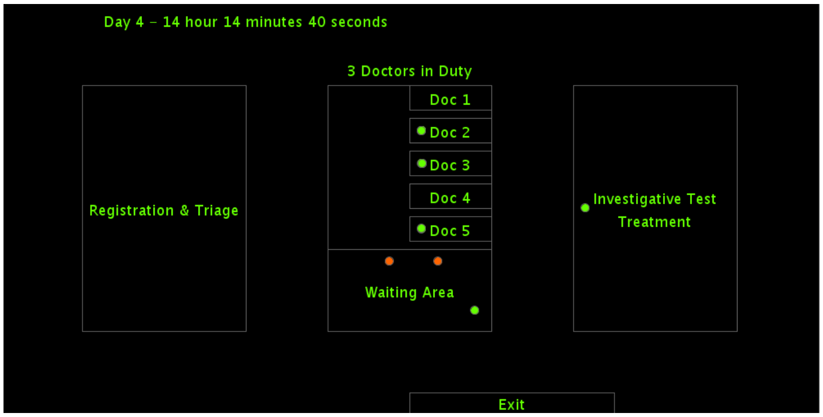


Figure 6.17: A snapshot of the visualization tool during a playback

## 6.5 Implementation Road Map

We provide an implementation road map as shown in Figure 6.18, to help hospital planners and the IT team take the proposed strategies and framework from design to implementation. The design of the road map is in increasing order of implementation complexity.

In this road map, we recommend that a hospital should first analyze its ED process and apply any best practices. An example can be found in Miller et al. [36] where six sigma is applied to the ED process. Simulation is then used to evaluate the performance. In real life, in the field of Business Process Management, applying best practices to processes can also be done without the use of simulation. The next step in the road map is to derive intelligence from historical data. With the process model and historical data, we can observe trends such as arrival rates, service rates of doctors, and service rates of tests

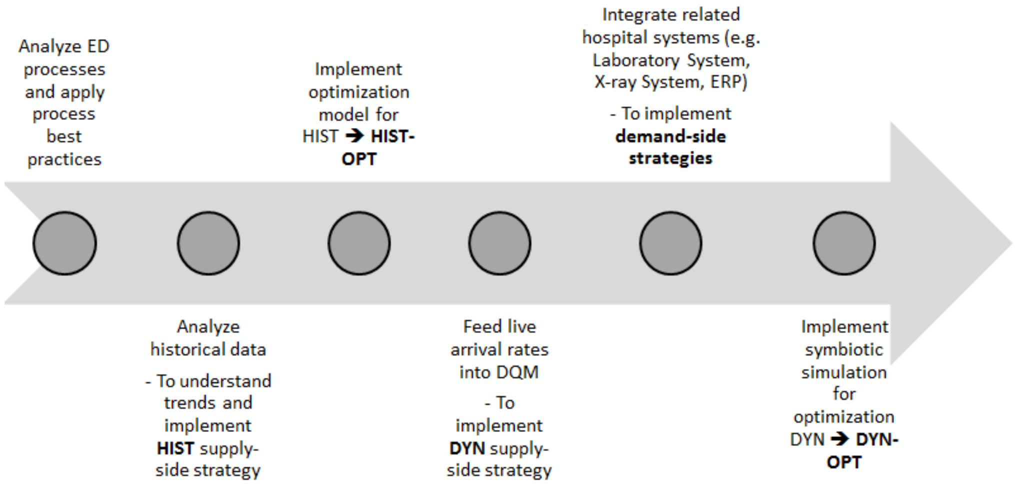

Analyze ED processes and apply process best practices
Analyze historical data
- To understand trends and implement **HIST** supply-side strategy
Implement optimization model for HIST ➔ **HIST-OPT**
Feed live arrival rates into DQM
- To implement **DYN** supply-side strategy
Integrate related hospital systems (e.g. Laboratory System, X-ray System, ERP)
- To implement **demand-side strategies**
Implement symbiotic simulation for optimization DYN ➔ **DYN-OPT**

Figure 6.18: A suggested implementation road map

and treatment. The supply-side strategy of the HIST staffing method can also be applied and staffing requirements can be computed. Following this, we suggest implementation of the optimization model in our proposed Dynamic Queue Management System that reads the historical statistics and contains the optimization model. With this, we can produce a HIST-OPT schedule.

Next we come to the more complex and higher risk part of the implementation. The next step is to be able to observe and understand real-time information. To do this, arrival rates can be easily captured and computed. It simply means the number of patients visiting the ED over a time period. We may need to integrate the DQMS with an existing system that records the patients' registrations. Now, we can dynamically produce a DYN schedule. The next step is a big leap. We suggest evaluating the hospital's IT systems and understanding where the pieces of information are stored that provide the intelligence to estimate consultation time and to compute the remaining time of a patient in the ED. This information is required to implement the demand-side strategies to dynamically prioritize patients in the queue. Finally, the hospital should also consider building a symbiotic simulator in the DQMS to provide real-time optimization and evaluation of the doctors' schedule based on historical as well as real-time information.

## 6.6 Chapter Summary

In this chapter, we presented an integrated framework for Dynamic Queue Management from both demand and supply perspectives. From our experimental analysis, we concluded that single-faceted queue management strategies (either demand side or supply side) are not cost-effective or not sufficient to provide a holistic approach to alleviate long length-of-stay in the ED. We showed via simulation that our integrated framework can synergistically combine intelligent dynamic patient queue-prioritization and dynamic resource-adjustment strategies to yield improvements in providing quality services in an ED. Our framework allows healthcare decision-makers to play a role in achieving the target service quality and select from a list of possible strategies that suit the operational needs of the ED. To reap the benefits of deploying the strategies, healthcare decision-makers must make careful plans and selections of the strategies. We provided a visualization tool and an implementation road map to assist this effort.

# Chapter 7

# Summary of Conclusion

## 7.1 Summary of Contribution

This dissertation makes valuable contributions to applied and interdisciplinary healthcare research in many ways. In all our proposed models, we took special care to go beyond standard theoretical queuing models and have our models represent the real-life ED processes. We also provided validations at logical points by comparing against real-life observed data. We then followed up by implementing prototypes of the models and evaluated the various strategies using simulations. The performances of the strategies were analyzed and compared with one another. If performances were similar visually on graphs, statistical hypothesis tests were used to further establish their significance differences. The comparisons allowed us to derive interesting managerial insights.

In our first piece of work, we modeled the ED process in the ambulatory area and showed that a hospital ED can improve the average LOS of patients by managing demand through the use of dynamic patient prioritization, leveraging both historical and real-time information. This work was published in the Proceedings of the 8th IEEE International Conference on Automation Science and Engineering (CASE 2012) [51].

In our second piece of work, we enhanced the ED process to model both the ambulatory area and the critical-care area. We showed that the hospital ED can improve the average LOS, potentially meet its desired LOS, and react to demand surges by optimizing and dynamically changing the supply of doctors based on real-time data. We applied both queue design and queue control techniques with the use of offline and online (symbiotic) simulation in our staffing strategies. This work was published in the Proceedings of the 9th IEEE International Conference on Automation Science and Engineering (CASE 2013) [50].

In our third piece of work, we provided a framework for integrating both demand-side strategies and supply-side strategies. We showed that a decision-maker can select any combination of demand-side strategies or supply-side strategies depending on the hospital's appetite for performance (in terms of average patient LOS) and other factors, such as risk of implementation, the doctors' schedule stability and ability to react to demand surges. This is the first work in the domain of healthcare to provide practical optimization capabilities from both the demand and supply perspectives. This work was published in the Proceedings of the 2013 Winter Simulation Conference (WSC2013) [49].

## 7.2 Tangible Optimization versus Intangible Considerations

During the course of completing this dissertation, we noticed the importance of intangible considerations among the operational optimization considerations in the healthcare industry. One challenge of optimization in healthcare is that it is not enough to simply find an optimal solution like a minimum LOS or a minimum wait-time. Healthcare's essential human element – patients, doctors, nurses, family members, lab technicians and many more - means there is a need to balance optimization with intangible, human considerations, such as quality of care, patient satisfaction and staff satisfaction. For example, as we consider deploying dynamic prioritization of the patients, we have to consider how a hospital should manage the patients who get preempted by other patients. This is an issue not investigated in this dissertation, but something that hospital decision-makers must explore.

One solution to balancing the two considerations could be exploring how the hospital can display information to patients to reassure them that the system is patient-centered. Suppose that Patient A is fifth in the queue and the expected waiting time is 20 minutes. When re-prioritization occurs, the display shows that he is sixth in the queue but the expected waiting time has dropped to 15 minutes. The patient is likely to be happy. Other efforts can be made to improve patient comfort, for example by providing an option for the patient to go for a snack or drink at the hospital's cafeteria during the waiting period (since we are dealing with non-emergency patients).

Likewise, there are motivational issues with doctors, and doctors (and nurses) should not be overworked. Hospitals need to consider how to balance the workload among doctors/staff, and ensure that doctors/staff have sufficient rest, even as efforts are made to improve the ED's performance.

In terms of operational quality, hospitals also need consider the quality of care in both the ambulatory area as well as other related facilities such as the critical-care and treatment areas where doctors are also required. Although there are service level targets to serve customers, hospitals need to also consider that quality of care (of both the ambulatory and related facilities) cannot be compromised. For example, it is unacceptable if doctors reduce the time spent with the patient because there is a need to meet a challenging targeted length of stay. Hospitals need to ensure that patients receive proper care during their stay in the ED.

Hospital targets may backfire resulting in quality of care being compromised, as reported in an article by the Australian Healthcare and Hospitals Association [23]. They explained that hospital targets to serve patients within a target LOS do not directly address the concerns of "improved patient access to timely and safe ED services". In a survey, they found that 80% of doctors felt that the hospital targets compromised their capacity to deliver "proper patient care". In our correspondences with Singapore hospitals, a concern was raised that an improved LOS for non-emergency patients may send a false message to the public that one can go to the ED for minor ailments which should be attended to at GP clinics. Hospitals need to consider additional campaigns such as public education on who should visit an ED in conjunction with operational improvement targets.

We illustrate the need to balance optimization with intangible considerations in Figure 7.1. Intangible considerations remain a challenge in healthcare. They are not part of the objective functions in this dissertation but certainly could be a vital focus of future research.

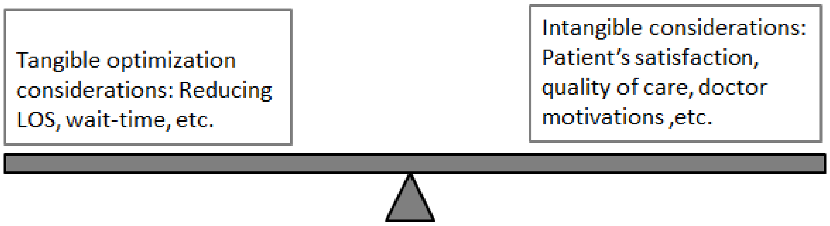

Tangible optimization considerations: Reducing LOS, wait-time, etc.
Intangible considerations: Patient's satisfaction, quality of care, doctor motivations ,etc.

Figure 7.1: The need to balance optimization and intangible considerations

## 7.3 Further Work

As there are so many possible ways to improve the ED's ability to react to demand, this work can be further developed to include other forms of intelligence for decision support. One which we have considered is *dynamic collaboration*. The idea is to allow hospitals to dynamically collaborate with other entities such as general practitioners (GPs) or ground transportation operators. In the case of collaboration with GPs, a hospital will have a view of the waiting time at a few nearby clinics. In this way, the hospital may divert patients with mild illnesses to nearby clinics which may serve them faster than the hospital. In some cases, when investigative tests or on-site treatment is required, a hospital may allow the diverted patients to take the tests or treatment at the hospital (since it is nearby). This will only be possible if the hospital has a close group of collaborators and has visibility into the queue and functions of the GP.

In the case of collaboration with transportation companies, such as taxi companies, a hospital can be informed of an incoming patient if the patient boards a taxi and indicates that he/she is visiting the ED for treatment. In that way, the hospital has more visibility into the incoming demand and better decisions can be made for staffing the ED.

We see opportunities for taking this work in various directions. In the optimization field, we can further evaluate how hospitals can leverage opportunities when demand is low to release some doctors to perform administrative or operational activities. From the perspective of information systems management, we hope to provide more managerial insights and evaluate the costs and benefits of dynamic strategies (either demand-side or supply-side or both) on the quality of patient care, doctor satisfaction and quality of care of patients in the critical-care area. In the domain of enterprise systems, we would like to see a real implementation of the strategies and evaluate the impact and complexity of dynamic strategies on real-world systems.

# Bibliography

[1] Susan L Albin, Jeffrey Barrett, David Ito, and John E Mueller. A queueing network analysis of a health center. *Queueing Systems*, 7(1):51–61, 1990.

[2] Jay April, Marco Better, Fred Glover, James Kelly, and Manuel Laguna. Enhancing business process management with simulation optimization. In *Proceedings of the 2006 Winter Simulation Conference*, pages 642–649. Institute of Electrical and Electronics Engineers, Inc., 2006.

[3] Edward Barthell, Christopher W Felton, Jasmin Jijina, and Barbara Thornburg. Getting the data and getting it straight: the Frontlines Project and similar initiatives. *Advanced Emergency Nursing Journal*, 26(2):166–175, 2004.

[4] Moshe Ben-Akiva, Michel Bierlaire, Haris Koutsopoulos, and Rabi Mishalani. DynaMIT: a simulation-based system for traffic prediction. In *DACCORS Short Term Forecasting Workshop, The Netherlands*. Citeseer, 1998.

[5] Oded Berman and Richard C. Larson. A queueing control model for retail services having back room operations and cross-trained workers. *Computers & Operations Research*, 31(2):201–222, 2004.

[6] Edwin D Boudreaux, Sarah d'Autremont, Karen Wood, and Glenn N Jones. Predictors of emergency department patient satisfaction: stability over 17 months. *Academic Emergency Medicine*, 11(1):51–58, 2004.

[7] Brenda Bursch, Joseph Beezy, and Ruth Shaw. Emergency department satisfaction: what matters most? *Annals of Emergency Medicine*, 22(3):586–591, 1993.

[8] Eduardo Cabrera, Emilio Luque, Manel Taboada, Francisco Epelde, and Ma Luisa Iglesias. ABMS optimization for emergency departments. In *Proceedings of the 2012 Winter Simulation Conference*, pages 89:1–89:12. Institute of Electrical and Electronics Engineers, Inc., 2012.

[9] S Chakravarthy. A finite capacity dynamic priority queuing model. *Computers & Industrial Engineering*, 22(4):369–385, 1992.

[10] Ruey L Cheu and Vladik Kreinovich. Exponential disutility functions in transportation problems: a new theoretical justification. *Technical Report*, University of Texas at El Paso, 2007.

[11] David Claudio. Dynamic vitals monitoring for patient prioritization in the emergency department: A technology enabled utility approach. *The Pennsylvania State University Ph.D. Thesis*, 2010.

[12] Zohar Feldman, Avishai Mandelbaum, William A. Massey, and Ward Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54(2):324–338, 2008.

[13] Dieter Fiems, Ger Koole, and Philippe Nain. Waiting times of scheduled patients in the presence of emergency requests. *Online, August, http://www.math.vu.nl/ koole/publications/2005report1/art.pdf*, 2007.

[14] Samuel Fomundam and Jeffrey W Herrmann. A survey of queuing theory applications in healthcare. *Technical Report*, University of Maryland, 2007.

[15] Michael C Fu. Optimization for simulation: Theory vs. practice. *INFORMS Journal on Computing*, 14(3):192–215, 2002.

[16] Richard Fujimoto, Dell Lunceford, Ernest Page, and Adelinde M. Uhrmacher. Technical report of the Dagstuhl-Seminar Grand Challenges for

Modelling and Simulation. Available via www.dagstuhl.de/02351/Report [accessed June 13, 2007], 2002.

[17] Tejas R Gandhi, Kaustubh Nagarkar, Monice DeGennaro, and K Srihari. *Reducing "patient Turnaround Times" at an Emergency Room*. PhD thesis, State University of New York at Binghamton, Department of Systems Science and Industrial Engineering, 2003.

[18] Linda Green. Queueing analysis in healthcare. In *Patient Flow: Reducing Delay in Healthcare Delivery*, pages 281–307. Springer, 2006.

[19] Linda Green, Peter J Kolesar, and Ward Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39, 2007.

[20] Murat M Gunal and Michael Pidd. Understanding accident and emergency department performance using simulation. In *Proceedings of the 2006 Winter Simulation Conference*, pages 446–452. Institute of Electrical and Electronics Engineers, Inc., 2006.

[21] Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[22] Andrew M Hay, Edwin C Valentin, and Rienk A Bijlsma. Modeling emergency care in hospitals: a paradox - the patient should not drive the process. In *Proceedings of the 2006 Winter Simulation Conference*, pages 439–445. Institute of Electrical and Electronics Engineers, Inc., 2006.

[23] Australian Healthcare and Hospitals Association (AHHA). "4 hour" hospital rule may backfire – harvard expert warns. *Health Media Centre, http://ahha.asn.au/news/4-hour-hospital-rule-may-backfire-harvard-expert-warns*, September 2012.

[24] Navid Izady and Dave Worthington. Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments. *European Journal of Operational Research*, 219(3):531–540, 2012.

[25] Sheldon H Jacobson, Shane N Hall, and James R Swisher. Discrete-event simulation of health care systems. In *Patient Flow: Reducing Delay in Healthcare Delivery*, pages 211–252. Springer, 2006.

[26] Otis B Jennings, Avishai Mandelbaum, William A Massey, and Ward Whitt. Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394, 1996.

[27] Diane L King, David I Ben-Tovim, and Jane Bassham. Redesigning emergency department patient flows: application of lean thinking to health care. *Emergency Medicine Australasia*, 18(4):391–397, 2006.

[28] Alexander Komashie and Ali Mousavi. Modeling emergency departments using discrete event simulation techniques. In *Proceedings of the 2005 Winter Simulation Conference*, pages 2681–2685. Institute of Electrical and Electronics Engineers, Inc., 2005.

[29] Malcolm YH Low, Stephen J Turner, Ding Ling, Hai L Peng, P Lendermann, LP Chan, and Steve Buckley. Symbiotic simulation for business process re-engineering in high-tech manufacturing and service networks. In *Proceedings of the 2007 Winter Simulation Conference*, pages 568–576. Institute of Electrical and Electronics Engineers, Inc., 2007.

[30] Yariv N Marmor, Segev Wasserkrug, Sergey Zeltyn, Yossi Mesika, Ohad Greenshpan, Boaz Carmeli, Avraham Shtub, and Avishai Mandelbaum. Toward simulation-based real-time decision-support systems for emergency departments. In *Proceedings of the 2009 Winter Simulation Conference*, pages 2042–2053. Institute of Electrical and Electronics Engineers, Inc., 2009.

[31] William A Massey and Ward Whitt. Networks of infinite-server queues with nonstationary poisson input. *Queueing Systems*, 13:183–250, 1993.

[32] L Mayhew and D Smith. Using queuing theory to analyse the governments 4-h completion time target in accident and emergency departments. *Health Care Management Science*, 11(1):11–21, 2008.

[33] DG McQuarrie. Hospitalization utilization levels. The application of queuing. Theory to a controversial medical economic problem. *Minnesota Medicine*, 66(11):679, 1983.

[34] DJ Medeiros, Eric Swenson, and Christopher DeFlitch. Improving patient flow in a hospital emergency department. In *Proceedings of 2008 Winter Simulation Conference*, pages 1526–1531. Institute of Electrical and Electronics Engineers, Inc., 2008.

[35] Gastón Mendoza, Mohammad Sedaghat, and K Paul Yoon. Queuing models to balance systems with excess supply. *International Business & Economics Research Journal (IBER)*, 8(1), 2011.

[36] Martin J Miller, David M Ferrin, and Jill M Szymanski. Simulating six sigma improvement ideas for a hospital emergency department. In *Proceedings of the 2003 Winter Simulation Conference*, pages 1926–1929. Institute of Electrical and Electronics Engineers, Inc., 2003.

[37] Foad Mahdavi Pajouh and Manjunath Kamath. Applications of queueing models in hospitals. In *Proceedings of the Midwest Association for Information Systems*, page 23, 2010.

[38] Melissa Pang. A & E units flooded with non-emergency cases. *The Straits Times, http://yourhealth.asiaone.com/content/ae-units-flooded-non-emergency-cases*, April 02, 2013.

[39] J.C. Pham, R. Patel, M.G. Millin, T.D. Kirsch, and A. Chanmugam. The effects of ambulance diversion: a comprehensive review. *Academic Emergency Medicine*, 13(11):1220–7, 2006.

[40] John R Richards, Misty L Navarro, and Robert W Derlet. Survey of directors of emergency departments in California on overcrowding. *Western Journal of Medicine*, 172(6):385, 2000.

[41] KT Roche, JK Cochran, and IA Fulton. Improving patient safety by maximizing fast-track benefits in the emergency department – a queuing network approach. In *Proceedings of the 2007 Industrial Engineering Research Conference*, pages 619–624, 2007.

[42] Scott W Rodi, Maria V Grau, and Caroline M Orsini. Evaluation of a fast track unit: alignment of resources and demand results in improved satisfaction and decreased length of stay for emergency department patients. *Quality Management in Healthcare*, 15(3):163–170, 2006.

[43] Kent V Rondeau, Louis H Francescutti, and JJ Zanardelli. Emergency department overcrowding: the impact of resource scarcity on physician job satisfaction. *Journal of Healthcare Management*, 50(5):327, 2005.

[44] Simon Samaha, Wendy S Armel, and Darrell W Starks. Emergency departments I: The use of simulation to reduce the length of stay in an emergency department. In *Proceedings of the 2003 Winter Simulation Conference*, pages 1907–1911. Institute of Electrical and Electronics Engineers, Inc., 2003.

[45] Miquel Sanchez, Alan J Smally, Robert J Grant, and Lenworth M Jacobs. Effects of a fast-track area on emergency department performance. *The Journal of Emergency Medicine*, 31(1):117–120, 2006.

[46] Pengyi Shi, Mabel C. Chou, JG Dai, Ding Ding, and Joe Sim. Hospital inpatient operations: mathematical models and managerial insights. Technical report, Working paper, 2012.

[47] K. Siddhartan, W.J. Jones, and J.A. Johnson. A priority queuing model to reduce waiting times in emergency care. *International Journal of Health Care Quality Assurance*, 9(5):10–16, 1996.

[48] David Sinreich, Ola Jabali, and Nico P Dellaert. Reducing emergency department waiting times by adjusting work shifts considering patient visits to multiple care providers. *IIE Transactions*, 44(3):163–180, 2012.

[49] Kar Way Tan, Hoong Chuin Lau, and Francis Lee. Improving patient length-of-stay in emergency department through dynamic queue management. In *Proceedings of the 2013 Winter Simulation Conference*. Institute of Electrical and Electronics Engineers, Inc., 2013.

[50] Kar Way Tan, Wei Hao Tan, and Hoong Chuin Lau. Improving patient length-of-stay in emergency department through dynamic resource allocation policies. In *Proceedings of the IEEE International Conference on Automation Science and Engineering (CASE)*, 2013.

[51] Kar Way Tan, Chao Wang, and Hoong Chuin Lau. Improving patient flow in emergency department through dynamic priority queue. In *Proceedings of the IEEE International Conference on Automation Science and Engineering (CASE)*, pages 125–130, 2012.

[52] Daria Terekhov and J. Christopher Beck. A constraint programming approach for solving a queueing control problem. *Journal of Artificial Intelligence Research*, 32(1):123–167, 2008.

[53] Michael Thorwarth and Amr Arisha. A simulation-based decision support system to model complex demand driven healthcare facilities. In *Proceedings of the 2012 Winter Simulation Conference*, pages 1–12. Institute of Electrical and Electronics Engineers, Inc., 2012.

[54] James B Tucker, James E Barone, Joseph Cecere, Robert G Blabey, and Chan-Kook Rha. Using queueing theory to determine operating room staffing needs. *The Journal of Trauma and Acute Care Surgery*, 46(1):71–79, 1999.

[55] T.E. Vollmann. Capacity planning: The missing link. *Production and Inventory Management (1st Qtr.)*, pages 61–74, 1973.

[56] Jennifer L Wiler, Christopher Gentle, James M Halfpenny, Alan Heins, Abhi Mehrotra, Michael G Mikhail, and Diana Fite. Optimizing emergency department front-end operations. *Annals of Emergency Medicine*, 55(2):142–160, 2010.

[57] David Worthington. Hospital waiting list management models. *Journal of the Operational Research Society*, 42(10):833–843, 1991.

[58] Galit Bracha Yom-Tov. *Queues in Hospitals: Queueing Networks with ReEntering Customers in the QED Regime.* PhD thesis, Technion - Israel Institute of Technology, 2010.

[59] Sergey Zeltyn, Yariv N Marmor, Avishai Mandelbaum, Boaz Carmeli, Ohad Greenshpan, Yossi Mesika, Sergev Wasserkrug, Pnina Vortman, Avraham Shtub, Tirza Lauterman, Dagan Schwartz, Kobi Moskovitch, Sara Tzafrir, and Fuad Basis. Simulation-based models of emergency departments:: Operational, tactical, and strategic staffing. *ACM Transactions on Modeling and Computer Simulation*, 21(4):24:1–24:25, 2011.