PATIENT FLOW INTERVENTIONS AND PRIORITIZATION IN
EMERGENCY DEPARTMENT

A Thesis in
Industrial Engineering

by
Yunfei He

© 2013 Yunfei He

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2013
The thesis of Yunfei He was read and approved* by the following:

Dr. Gül E. Okudan Kremer  
Associate Professor of Engineering Design and Industrial Engineering  
Thesis Adviser

Dr. Guodong Pang  
Assistant Professor of The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering

Dr. Marija Jankovic  
Assistant Professor of Industrial Engineering  
Ecole Centrale Paris

Dr. Paul Griffin  
Peter and Angela Dal Pezzo Department Chair  
Head of The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering

*Signatures on file in the Graduate School.
Abstract

Emergency Department overcrowding has been a major problem in the United States for almost two decades. Accordingly, numerous efforts have been put forth into eliminating the adverse effects that come along with it, such as long waiting times, patients leaving without being seen and deterioration of quality of care. From the published patient flow improvement interventions, we focus on strategies that are widely discussed in the relevant literature: fast track, physician triage and team triage. The effectiveness of these interventions is shown through mostly observational and a limited number of relatively controlled studies, where only a specific intervention is studied. However, these “success stories” help little when alternatives are not compared properly. To overcome this problem, we propose a simulation approach to examine these interventions in a controlled environment and to enable the sensitivity analyses of critical ED parameters such as resources or patients’ inputs, and performance measures. The impact on patients of different urgency level is also addressed so that generalized conclusions on the effectiveness of interventions can be made to guide future implementations.
# Table of Contents

List of Tables ................................. vi
List of Figures ................................. vii
Acknowledgments .............................. viii

Chapter 1. Introduction ........................ 1
  1.1 Motivation .................................. 1
  1.2 Research Framework ....................... 3
  1.3 Thesis Organization ....................... 5

Chapter 2. Literature Review .................. 7
  2.1 Patient Flow Improvements and ED Overcrowding .......................... 7
    2.1.1 Interventions through Resource Re-arrangement ..................... 7
    2.1.2 Patient Prioritization .................................. 13
  2.2 Risk of Waiting and Dynamic Priority ........................ 18
  2.3 Simulation Study of Patient Flow .................... 20

Chapter 3. Methodology ........................ 25
  3.1 Problem Definition .......................... 25
  3.2 Conceptual Intervention Models ............ 26
  3.3 Performance Measurement ................... 29
    3.3.1 Vulnerability Adjusted Waiting Time ......................... 30
    3.3.2 Waiting-at-risk .................................. 31
  3.4 Dynamic Prioritization ...................... 33
  3.5 Models Assumptions and Parameters .......... 36
    3.5.1 Structural Assumptions .......................... 36
    3.5.2 Numerical Assumptions ........................ 38
    3.5.3 Model Input and Output .......................... 39
  3.6 Analytical Approximation of the Baseline Model .................... 40
  3.7 Computerized Model and Validation ........ 42

Chapter 4. Experimentation .................... 47
  4.1 Preparation .................................. 47
    4.1.1 Model Parameterization ...................... 47
    4.1.2 Testing Scenarios .......................... 48
    4.1.3 Comparison of Multiple Alternatives ............ 51
  4.2 Comparing Interventions .................... 52
    4.2.1 Regular-Hour Test .......................... 52
    4.2.2 Stress Test .................................. 54
    4.2.3 Result of Waiting-at-Risk ...................... 57
List of Tables

2.1 Several Triage Acuity Systems by Number of Levels ................................. 14
2.2 International Triage Systems ................................................................. 14
2.3 ESI Triage System .................................................................................... 15
2.4 Re-assessment Time in CTAS ................................................................. 19

3.1 Processing Time Distributions ................................................................. 38
3.2 System Input and Output .......................................................................... 39

4.1 Resources Initialization ............................................................................. 47
4.2 Time to See Doctors (Regular Hours) ...................................................... 52
4.3 Time in ED (Regular Hours) .................................................................... 55
4.4 Time to See Doctors (Stress Test) ........................................................... 55
4.5 Time in ED (Stress Test) .......................................................................... 56
4.6 Cumulative VAWT for Non-stationary Arrival ........................................ 63
4.7 Cumulative VAWT of Scheduling Policies in Regular Hours ................. 65
List of Figures

1.1 Research Framework ................................................. 4
2.1 ESI Triage Algorithm, v4 ........................................... 16
3.1 Visualization of Selected ED Performance Measures .............. 26
3.2 Process Flow of Baseline Model .................................... 28
3.3 Process Flow of Fast Track ......................................... 29
3.4 Process Flow of Physician Triage .................................... 29
3.5 Process Flow of Triage Team ........................................ 30
3.6 Waiting at Risk ...................................................... 33
3.7 Dynamic Prioritization .............................................. 34
3.8 Mutiple Class Many Server Queue ................................ 41
3.9 Control Logic to ED ................................................ 44
3.10 Routing Logic to ED ................................................ 45
3.11 Using Link Weight to Routing ESI 5 patients .................... 45
4.1 Approximated Waiting Time for Selecting Scenarios ............ 49
4.2 Testing Scenarios Scaled to ED Capacity .......................... 50
4.3 Time to See Doctors in Regular Hours ............................ 53
4.4 Time in ED in Regular Hours ....................................... 54
4.5 Time to See Doctors in Peak Hours ............................... 56
4.6 Time in ED in Peak Hours ......................................... 57
4.7 WaR Plot of Baseline Model ........................................ 58
4.8 WaR Plot of FT Model ............................................. 59
4.9 WaR Plot of PT Model ............................................. 59
4.10 WaR Plot of TT Model ............................................. 60
4.11 Peak Hour WaRs ................................................... 61
4.12 Cumulative VAWT for Non-stationary Arrival ................. 62
4.13 Comparison of Scheduling Policies on Mean .................... 64
4.14 Comparison of Scheduling Policies on WaR ..................... 65
I would like to express the deepest appreciation to my thesis advisor Dr. Gül E. Okudan Kremer. Without the guidance and help of her, I could never have been able to finish this work. Her support and encouragement motivated me to work harder in achieving my goals.

I would also like to thank Dr. Guodong Pang and Dr. Marija Jankovic for agreeing to join in my thesis committee. They have been a great source of knowledge for me. Special thanks go to Dr. Guodong Pang, who helped me develop my background in stochastic models and patiently corrected my writing for the thesis.

Many thanks to all the professors at Penn State who have given me the knowledge that helped me in ascending the echelons of success. I would also like to thank my lab mate Ting, who was always willing to help and give her best suggestions.

Finally, I would thank to all the staff and friends, who have made my stay at Penn State so enjoyable and left me many wonderful memories.
Unfortunately, the clock is ticking, the hours are going by. The past increases, the future recedes. Possibilities decreasing, regrets mounting.

–Haruki Murakami
Chapter 1

Introduction

1.1 Motivation

Emergency department crowding has been identified as an increasing crisis for nearly twenty years. According to the results of the National Hospital Ambulatory Medical Care Survey, from 1997 through 2007, annual number of emergency department visits rose from 94.9 million to 116.8 million; meanwhile the number of EDs decreased about 10 percent in the same period (Niska et al., 2010). As a consequence, nearly 70% of urban hospital EDs and 30% rural hospital EDs are operating at or over capacity (Wiler et al., 2010).

Patient flow improvement is a low-cost solution to handle the increasing overcrowding in Emergency Departments (EDs) (McHugh et al., 2011). The challenge of overcrowding in ED necessitates effective release of ED capacity by reducing “boarding” or holding of admitted patients in ED, and optimal allocation of resources to meet diversified demand of medical care (Bernstein & Asplin, 2006). The front-end ED processes are defined as the set of all the operations before the patient receives her or his treatment, and include initial patient presentation, registration, triage, bed placement and medical evaluation (Wiler et al., 2010). Among them, triage is a key element in the process for its role in allocating scarce medical resources to demand (FitzGerald et al., 2010). In the triage process, patients are assigned priorities of receiving treatment based on their
acuity and urgency of illness. In this sense, the differentiation among patients is realized at the triage, and the following steps are carried out based on the triage information. Major patient flow improvements can also improve operations after the triage.

Presumably, different interventions may not show consistent performance across all empirical situations such as different patient arrival patterns or ED capacity mix. From managerial perspective, it is important to know the timing to implement patient flow interventions. It would be helpful for making decisions on when to open fast track and whether to implement triage team when ED is running over capacity.

However, subject to a lack of controlled environment and data availability, it is difficult for existing EDs to study patient flow interventions, especially through empirical ones, as particular demand makes it impossible examine the interventions under different scenarios.

Another noteworthy issue is the prioritization of patients. Traditionally hospital EDs adopt a static 3-level or 5-level triage system (FitzGerald et al., 2010). The drawback of static priority is that a patient with higher level of acuity always takes precedence over patients with lower priority. This assignment ignores changes of patients’ condition during waiting. Imagine that an ESI level-3 patient who has waited for two or more hours has to give the priority to an ESI level-2 patient who just arrives in ED.

This phenomenon is called ‘starvation’ in scheduling literature. Traditional triage systems assign patients’ priorities permanently, which means the recorded condition of patients does not change after triage station. Although frequent reassessment is needed to ensure the patient is not reaching a dangerous state, this process does not explicitly include the possibilities of state deterioration in the triage prioritization. Moreover, in an
over-crowded ED waiting room, it is difficult for a limited number of nursing resources to keep the timely re-assessment. To cope with this undesired situation, a more adaptable ranking system is necessary. We propose a dynamic prioritization model that can avoid starvation and keep track of patients’ acuity conditions.

1.2 Research Framework

The research is developed around the patient flow interventions and dynamic prioritization policies. Intervention and ranking models are tested under empirical testing scenarios for performance measurements, which are based on the vulnerability and risk of patients.

Four types of patient flow interventions (including one baseline case) are studied to examine their impact on key patient-centric performance indicators such as waiting time (WT), average length of stay (LOS) and leaving without being seen (LWBS), as well as statistics from the provider’s prospective like resource utilization and total throughput. The four patient flow strategies we are interested in are: baseline process without intervention (BL), Fast Track (FT), Physician Triage (PT) and Triage Team (TT). All of these strategies are academically studied, implemented and empirically analyzed in various settings. However, their comprehensive comparisons are not done thus far; hence we fill this gap in literature.

Patients with more serious acuity levels are more sensitive to the delay of treatment. In order to fairly describe the patients waiting time with impact, we incorporate the vulnerability in measurement. The vulnerability can refer to the triage groups like
ESI levels, or if ED can keep track of a patient’s detailed condition, it can also be
determined with respective to individuals.

As discussed earlier, ED performance can be measured by average Length of Stay, average Waiting Time. The problem with mean statistic is the difficulty to show the variation of population. However with regards to ED problems, it is the worst cases rather than mean that are more valuable information to compare interventions. If everyone with similar acuity level waits for the same amount of time, then no one will be in danger of deterioration. The most critical patients are those who have waited for too long.

Computer simulation provides a totally controlled experimentation environment; therefore it enables us to closely review the performance of interventions under various
testing scenarios. Several scenarios are carefully selected to mimic the empirical patient flow arrivals, including regular time, peak hour and a non-stationary process with empirical intensity. In addition to the patient flows, the diversity of capacity between hospital EDs are also considered. Moreover, an important contributor to ED overcrowding, boarding phenomenon, is also considered a testing scenario.

The major motivation of the thesis is to deliver a collective study of the patient flow interventions and patient prioritization in ED. In order to fill the gap in patient flow intervention literature, computer simulation is selected as a platform to perform comparisons under various settings and empirical scenarios. As the other part of the objective, patient prioritization rules are tested for both traditional performance measures and risk-adjusted measures developed in the study as well.

1.3 Thesis Organization

The thesis is organized as follows: Chapter 2 contains a literature review on both the problem domain of ED patient flow and methodology of simulation approach. Especially, through the literature review, the research gap in patient flow interventions providing motivation of the thesis is identified.

Chapter 3 establishes the framework of the research, including the definition of conceptual model, computerized model and dynamic scheduling policy. The performance measurement is designed with consideration of ED throughput and patients’ vulnerability for waiting. In addition, a queueing analysis is conducted for the baseline case to approximate the ED performance and give some insight to initialize the simulation model.
Chapter 4 provides the experiments and discussion of the interventions under different empirical scenarios. The ED performance is reviewed with respect to new measurement and traditional measurement as well. Dynamic routing is also tested.

Chapter 5 summarized the major conclusions obtained from chapter 4 and the future direction of simulation study on patient flow.
Chapter 2

Literature Review

This chapter presents the key concepts and main methodologies in the research. Section 2.1 introduces the current research on patient flow interventions in ED and how overcrowding is relevant to patient flows. Section 2.2 discusses the role of triage in ED and scheduling research related to dynamic patient acuity ranking. Section 2.3 explains the negative effect of patients’ of delayed access to urgent care, and the rationale for the vulnerability-adjusted measurement. Section 2.4 summarizes the study of using simulation for patient flow.

2.1 Patient Flow Improvements and ED Overcrowding

Reflecting the physical movement of patients in Emergency Departments as well as sequences of necessary medical procedures, patient flow is considered central to understanding and optimizing ED performance, and thus ensuring good care. Four types of patient flow interventions (including one baseline case) are studied in the research.

2.1.1 Interventions through Resource Re-arrangement

Baseline Strategy Baseline strategy refers to the most commonly practiced flow system in emergency departments: patients wait for a sequence of processes before receiving treatment in order of priority (usually based on urgency) and resources, including
providers and beds pooling together in one big ED. This strategy has been widely implemented since triage was introduced into ED (Slater, 1970). It is generally treated as a control group in various empirical studies of competing interventions (Kilic et al., 1998). Baseline strategy guarantees that the urgent patients are always first to receive medical care and patients with low level acuity can be served only when no patients with higher priority is waiting. This policy works fairly well in settings of medium or light traffic of patient flow. In the presence of heavy traffic, however, the patients with low acuity suffer from excessively long waiting times before receiving a simple and fast treatment (Saunders, 1987). This phenomenon, along with the average Length-of-Stay and high rate of Leave-Without-Being-Seen, becomes more intolerable with increasing medical demand.

**Fast Track** Here we adopt the definition of fast track as “a lane dedicated to serve a particular type of patients with the sole intent of reducing their waiting time; thus, reducing their total time in the system” (García et al., 1995). This definition clarifies that the sole purpose of setting up a fast track is to reduce the Waiting-Time, which is an operation-based indicator of one type of patients rather than clinical advantages for all. It also means that the separation of fixed amount of resources into two units, main ED and fast track, rather than increasing capacity. Moreover, queuing theory explains that partitioning usually deteriorates the performance in comparison to pooling (Hall, 1991). This situation follows when treatment time is random since partitioning increases the variance of each unit.
Some observations state that fast track not only decreases the Waiting-Time, but also promotes the diagnosis process to become faster. In one of the most representative cases (Hampers et al., 1999), the patients who have been seen in the Fast-Track had fewer tests ordered and had shorter Length-of-Stay. Theoretically, this is possible. In manufacturing, the widely studied phenomenon that a worker’s performance improves by repeating similar activities over time is defined as the “learning effect” (Mosheiov, 2001). It is reasonable to infer that learning effect and set up time also exist in health care processes. But it is hard to say this is a good or bad; in manufacturing environments, job types are known; however, in EDs, triage information is not always correct about patient acuity. If tests are necessary, then reducing them may risk the urgent patients with imperceptible symptoms. On the other hand, if tests are not necessary, we believe these redundancies can be reduced even without a fast track. In other words, fast track strategy can hardly take the credit of reducing treatment time (Bernstein & Asplin, 2006).

Since fast track is basically re-allocation of existing resources, we are also curious about the results that “fast track patients wait less without making others worse off” in many empirical analyses. For example, O’Brien et al. (2006) showed results of reducing Waiting-Time and Length-of-Stay without impacting the flow of other acute patients. However, most relevant reviews ignored the minimal level of Waiting Time reduction, for which the confidence interval was [-4,8] minutes, and there was increased capacity during intervention. Considine et al. (2008) provided similar conclusions with a case-control study; yet, their result was rendered by letting nurses perform treatments independently in Fast-Track. Ardagh et al. (2002) concluded that no negative effect was observed
on other patients even when they observed an increase in other patients’ waiting time, merely because their small sample size produced a large p value. However, we are concerned about the significance of the increased Waiting-Time for a larger sample size. Intuitively, when non-urgent patients have relatively shorter processing times, providing service to them instead of treating one urgent patient might be worth more. Therefore, in terms of “average Length-of-Stay” or “average Waiting-Time”, the number of patients affected adversely will decrease. Kwa & Blake (2008) erroneously observed similar results with extra resources, and concluded that “no negative effect” was observed.

We want to stress that with fixed resources, fast track does not necessarily bring advantages. The American College of Emergency Physicians (ACEP) reported an opinion that is consistent with our hypothesis on the impact of fast track (Bernstein & Asplin, 2006). Conducting a flawless experiment on the effectiveness of Fast-Track in an empirical setting is extremely difficult. For a comprehensive review on quality and credibility of empirical studies on fast track, please refer to Wiler et al. (2010).

We think, however, the “zero-sum effect” doesn’t nullify the fast track strategy. By implementing a proper fast track strategy, ED can achieve a desired balance between priority of urgency and throughput: for instance, we can adjust the fast track capacity to achieve the maximum throughput without violating the maximum waiting time constraints for each level of ESI. This kind of “optimal” decision can only be made with a thorough understanding of sensitivity between ED parameters, like the ratio between resources for the main ED and resource for Fast-Track. The difficulties to control the samples and the complex settings for physical experiments make it impossible to conduct sensitivity analysis; therefore, we turn to computer simulation for a general test bed.
Physician Triage  Physician Triage intervention, also known as “triage liaison physician”, “provider-level triage” or “faculty triage” etc., was first introduced to meet the need of more accurate identification of patients’ acuity level by placing a medical provider at the triage process. The provider enhances the triage process by performing a brief assessment, initiating necessary testing and handling other questions beyond the capacity of a triage nurse (Holroyd et al., 2007). Although the routine of “nurse requested X-ray” is piloted in several hospitals (Oredsson et al., 2011), evidence shows that the quality of decision on test ordering is questionable when it is performed by a practitioner without enough medical training: triage nurse over-ordering and under-ordering occurs significantly (Wiler et al., 2010).

The rationale behind this intervention is similar to the effect of “front-loading” in product development process. Front-loading is the art of shifting problem-identification and problem-solving to earlier stages of product development (Thomke & Fujimoto, 2000). The difference is that physician triage reduces the cost occurring along the process in terms of patient’s waiting time. By ordering the test in advance of entering ED, parallel processing is achieved while the patients are still waiting in queues. Another strategy also taking advantage of time savings from parallel processing is bed-side registration (Takakuwa et al., 2007). However, since this strategy is not exclusive with triage based interventions, it will not be included in the scope of this study.

Physician triage does not affect urgent patients since they barely wait because of high priority. For less urgent patients, their test results will be evaluated by a different care provider in ED. A potential problem arises about the effectiveness of the physician triage-based brief assessment: if the physician at triage ordered fewer tests than needed
by ED doctors, the effect of physician triage will be diminished. Another issue is how to effectively pass the information of initial assessment to physician in ED to avoid time spent on secondary assessment. Unfortunately, no study was found addressing these two problems.

**Team Triage**  Triage Team, also called “Provider directed queuing”, is an extension of physician triage intervention by “See and Treat” philosophy (Partovi et al., 2001). In addition to ordering tests for patients with complex resource needs, the comprehensive triage team performs simple treatments for less urgent patients with low level resource needs, after a brief assessment; this is called “see and treat”. The components of a Team Triage vary across different implementations. For example, PDQ implemented in Hershey Medical Center (HMC) comprises of one physician, one charge nurse, one triage nurse and one ED technician (Medeiros et al., 2008). But in theory, it requires a “union” of the set of fast track resources and the set of physician triage resources. Studies on team triage report substantial reduction of Waiting-Time. Choi et al. (2006) showed 38% decrease in Waiting-Time and 23% decrease in Length-of-Stay. Travers & Lee (2006) observed decreased mean time to physician evaluation for non-acute (35.3 to 19 min; $P > 0.05$) and serious but not life-threatening patients (28 to 14 min). The sources of reduction lie in the definition of Waiting-Time: the time between arrival and first time of seeing a medical provider. An initial brief assessment greatly advances the time of being seen by a doctor. On the other hand, by utilizing patients’ waiting time to conduct medical evaluation and order testing, many patients can be treated and discharged without using a room. This dramatically decreases patients’ waiting time,
especially for patients with lower acuity, due to the fact that waiting in a value-adding fashion, as opposed to waiting just for a room before treatment is initiated. This type of waiting time reduction also appears in pure physician triage intervention.

We are interested in the sensitivity and robustness of each intervention. As a combination of fast track and physician triage intervention, the impact of team triage should be presumably similar in characteristics of previous two methods. For tests under different scenarios, the combination may outperform its parents, but it’s also possible that the team triage inherits the drawbacks. Answering this type of what-if questions is also inapplicable in empirical analysis, which is another motivation for this study.

2.1.2 Patient Prioritization

Emergency department triage serves as a prioritization system for incoming patients. This is accomplished by evaluating the patient’s acuity or severity. Acuity is the degree to which the patient’s illness or injury is life-threatening and whether immediate treatment is necessary to maintain the patient’s condition (Baren & Rothrock, 2007).

There are various triage acuity systems ranging from two to five levels (See Table 2.1). According to an Emergency Nurses Association (ENA) survey done in 2001 in US, a three level triage system was most commonly used 68% yet there was a rising trend of five level triage system (MacLean, 2002). As for validated and implemented triage systems, various ones have been used across the globe: the Australian Triage Scale (ATS) and the Manchester Triage Scale (MTS), the Canadian Triage and Acuity Scale (CTAS) and the Emergency Severity Index (ESI). These systems share some commonalities: they are all
five scale systems and have similar required or recommended response time. Please refer to Table 2.2 for details.

Table 2.1. Several Triage Acuity Systems by Number of Levels

<table>
<thead>
<tr>
<th>Triage Acuity Systems by Level</th>
<th>2 Levels</th>
<th>3 Levels</th>
<th>4 Levels</th>
<th>5 Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emergent</td>
<td>Emergent</td>
<td>Life threatening</td>
<td>Resuscitation</td>
<td></td>
</tr>
<tr>
<td>Non-emergent</td>
<td>Urgent</td>
<td>Emergent</td>
<td>Emergent</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-urgent</td>
<td>Urgent</td>
<td>Urgent</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-urgent</td>
<td>Non-urgent</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2. International Triage Systems (Reproduced from Baren & Rothrock (2007))

<table>
<thead>
<tr>
<th>Australasian</th>
<th>Manchester(UK)</th>
<th>Canadian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>Response Time(min)</td>
<td>Level</td>
</tr>
<tr>
<td>1=Resuscitation</td>
<td>0</td>
<td>1=Immediate</td>
</tr>
<tr>
<td>2=Emergency</td>
<td>≤ 10</td>
<td>2=Very Urgent</td>
</tr>
<tr>
<td>3=Urgent</td>
<td>≤ 30</td>
<td>3=Urgent</td>
</tr>
<tr>
<td>4=Semi-urgent</td>
<td>≤ 60</td>
<td>4=Standard</td>
</tr>
<tr>
<td>5=Non-urgent</td>
<td>≤ 120</td>
<td>5=Non-urgent</td>
</tr>
</tbody>
</table>

In addition to patient acuity, the ESI system also includes the expected resource and timeliness to define a patient’s severity index (see Table 2.3). In addition to answering “who should be seen first?” but also “what will the patient need?” (Ashour & Okudan Kremer, 2013).

Patients with ESI level of 3, 4 and 5 can be generally categorized in a non-urgent group. What differentiates them is the complexity of expected diagnostic procedure and expected time to disposition (Gilboy et al., 2005). ESI follows a flow chart based algorithm:
## Table 2.3. ESI Triage System (Reproduced from Baren & Rothrock (2007))

<table>
<thead>
<tr>
<th></th>
<th>ESI 1</th>
<th>ESI 2</th>
<th>ESI 3</th>
<th>ESI 4</th>
<th>ESI 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stability of vital functions (ABCs)</strong></td>
<td>Unstable</td>
<td>Threatened</td>
<td>Stable</td>
<td>Stable</td>
<td>Stable</td>
</tr>
<tr>
<td><strong>Life threat or organ threat</strong></td>
<td>Obvious</td>
<td>Reasonably likely</td>
<td>Unlikely</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Require resuscitation</strong></td>
<td>Immediately</td>
<td>Sometimes</td>
<td>Seldom</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Severe pain or distress</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Expected resource intensity</strong></td>
<td>Maximum: staff at bedside continuously mobilization of outside resources</td>
<td>High: multiple, often complex diagnostic studies; frequent consultation remote monitoring</td>
<td>Medium: Multiple diagnostic studies; or brief period of observation; or complex procedure</td>
<td>Low: one simple diagnostic study; or one simple procedure</td>
<td>Low: Examination only</td>
</tr>
<tr>
<td><strong>Physician/staff response</strong></td>
<td>Immediate team effort</td>
<td>Minutes</td>
<td>Up to 1 hr</td>
<td>Could be delayed</td>
<td>Could be delayed</td>
</tr>
<tr>
<td><strong>Expected time to disposition</strong></td>
<td>1.5 hr</td>
<td>4 hr</td>
<td>6 hr</td>
<td>2 hr</td>
<td>1 hr</td>
</tr>
<tr>
<td><strong>Examples</strong></td>
<td>Cardiac arrest, intubated trauma patient, severe drug overdose</td>
<td>Most chest pain, stable trauma, elderly pneumonia patient, altered mental status, behavioral disturbance (potential violence)</td>
<td>Most abdominal pain, dehydration, esophageal food impaction, hip fracture</td>
<td>Closed extremity trauma, simple laceration, cystitis, typical migraine</td>
<td>Sore throat, minor burn, recheck</td>
</tr>
</tbody>
</table>
The usage of triage in the Emergency Room of hospitals in the US can trace back to 1960s. At the early stage of triage history, many hospitals did not have a formal system or triage rules (Purnell, 1991). Yet some effects of this patient prioritization system have already been identified, such as the controversial reduction in waiting time and resulting patient satisfaction (McDonald et al., 1995). In a major study in Sweden, Hansagi et al. (1992) concluded that patients’ satisfaction in ED increases in pace with the urgency of their condition.

The implementation of 5 level triage systems to replace three level systems reflects the interest of a more accurate description of patients’ condition. But in the practice of ESI, there are still limitations such as the prioritization for patients within each ESI level. Imagine a scenario that “there are six level 3 patients in the waiting room and

---

**Fig. 2.1. ESI Triage Algorithm, v4 (Reproduced from Gilboy et al. (2005))**

![ESI Triage Algorithm](image-url)
they have similar waiting time”; then who should be seen first? Many efforts have been generated to address this problem. For instance, Claudio & Okudan (2010) applied the utility theory to prioritize patients within same level of acuity level.

Another drawback of ESI is that, like other flow-chart based algorithms, ESI has difficulty to evaluate the relative importance between symptoms (Sinuff et al., 2004). Hierarchical models are capable of dealing with this problem. Ashour & Okudan (2010) utilized the Fuzzy-AHP to generate scores for patient chief complaint such as temperature, blood pressure, pulse and $SaO_2$; with the scores as the input, they ranked the patients based on a multi-attribute utility function. Their FAHP-MAUT method incorporated the complaints as well as descriptive variables such as gender age and pain level in the prioritization.

On the other hand, one may argue that a flow-chart based algorithm is more friendly for nurses to follow and does not require the assistance of computerized tools. However, triage decisions by nurses are not consistent: they largely depend on the training and experience at individual level (FitzGerald et al., 2010). Even many advanced protocols have been suggested to help triage decision makers, effectiveness and compliance were both worrying problems in experiments or pilot studies (Oredsson et al., 2011). Moreover, triage is a stressful duty. Researchers have suggested either extra staff or shorter spans of triage duty (Rock & Pledge, 1991; Nuttall, 1986). If computerized assistance is reliable and accessible, it would be a valuable improvement.

A few computerized triage systems have been evaluated previously, such as Soterion Rapid Triage system and e-TRIAGE (Grafstein et al., 2003; Maningas et al., 2006; Dong et al., 2005). Studies showed higher scores of inter-rater agreement (discrepancies
among nurses) (Maningas et al., 2006; Dong et al., 2005). A typical procedure of computerized triage is that firstly a nurse inputs the chief complaint and then the computer asks the key information to assign a triage level. The nurse can finalize the triage level after the computer give the suggestion by accepting it or overriding it (Gravel et al., 2007). This type of interaction gives nurses total control of triage level assignment, moreover, it is compatible with electronic health records (EHR). Computers can retrieve information from historical data to differentiate patients with the same syndrome. A study for patients with syncope showed that clinical data available at presentation to the emergency department can effectively be used for the risk stratification (Colivicchi et al., 2003).

2.2 Risk of Waiting and Dynamic Priority

The current focus of triage study is to assign an accurate and proper priority to incoming patients. Although frequent reassessment of patients is necessary, in an overcrowded ED, it is difficult to perform on-time periodical review (Fields et al., 2013). Through the assignment, patients obtain a static priority. The underlying assumption of a static priority is that the patients’ condition does not change during waiting.

Now imagine the same scenario discussed earlier, “there are six ESI level 3 patients in the waiting room and they have similar waiting-times”. But this time, “there is also a ESI level 4 patient who has waited for over 3 hours”, then which one of them should be treated first? The static assignment neglects the waiting time in ED and subsequent change of patients’ status. Very few papers addressed the dynamic priority problems with regard to triage. Murray et al. (2004) suggested the CTAS guidelines should include a
recommended reassessment time for each level of patients to keep the triage process dynamic. The reassessment time is the same with recommended response time.

Table 2.4. Re-assessment Time in CTAS

<table>
<thead>
<tr>
<th>Level</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>every 15 min</td>
</tr>
<tr>
<td>II</td>
<td>every 30 min</td>
</tr>
<tr>
<td>III</td>
<td>every 60 min</td>
</tr>
<tr>
<td>IV</td>
<td>every 120 min</td>
</tr>
<tr>
<td>V</td>
<td>every 120 min</td>
</tr>
</tbody>
</table>

Murray et al. (2004) also pointed out that “the patient’s status may change because of changing modifiers associated with the presenting complaint or because the presenting complaint has actually changed”. Tan et al. (2012) studied a dynamic priority queue in ED and adopted general scheduling rules from computer science literature to minimize length-of-stay. However, their study lacked in consideration of patient urgency levels and empirical ED setting.

Down & Lewis (2010) introduced upgrades for mitigating long waiting time in an N-Network model. They allowed class 1 customers to be “upgraded” to class 2 after they have been in queue for some time. In most cases they reported the optimal control policy reduced waiting times of both customer classes.

Although some studies identified the effect of ED crowding, due to the difficulty of data collection, very few papers assess the subsequent patient risk of prolonged waiting time quantitatively. Mortality rate is commonly used in studies of adverse outcomes of patient waiting.
Miro et al. (1999) found a significant correlation between mortality rate with weekly ED volume. Begley et al. (2004) showed a statistically significant trend for higher mortality rate for trauma patients during ambulance diversion. A study focusing on the patients who leave without being seen found out that LWBS patients were twice as likely to report worsened health problems (Bindman et al., 1991).

From a provider’s perspective, a study showed a $204 revenue loss for each chest pain patient who had waited for over 3 hours in ED because of boarding (Bayley et al., 2005). They also found that ED Length of Stay is not relevant with total hospital revenues or total hospital Length of Stay, but boarding imposed substantial ED opportunity costs, which decreased potential revenue.

2.3 Simulation Study of Patient Flow

Various types of computer simulation approaches have been applied in modeling patient flow related activities in Emergency Departments and other hospital units. The simulation approaches can by roughly categorized into two groups according to whether they have static model structures or not (Sobolev et al., 2011).

Static models refer to numerical experiments by sampling from a fixed model repetitively, such as Markov Chain Monte Carlo. Côté & Stein (2000) presented a stochastic model to reproduce the flow of discharged patients over time. Their Erlang-based model utilized a transition matrix to represent the evolving phases of the illness. Static models are useful to examine patient flows in larger population groups or public health administration, where Markov assumptions can be made (Davies & Davies, 1994). The limitation of these types of model is the difficulty to address the constraints on the
resources and interactions between entities like the cooperation between physician and nurses.

On the other hand, in contrast, dynamic models can include more complex logics, conditions and evolving processes. System Dynamics (SD), Discrete-Event-Simulation (DES) and Agent-Based Models (ABM) are popular dynamic simulation approaches in patient flow studies (Fone et al., 2003).

System dynamics models the patients’ behavior as a whole and usually in a deterministic and continuous way (Sobolev et al., 2011). Because system dynamics only represents aggregated patient flow, SD models are convenient to build and easy to understand. But as a consequence, high level assumptions compromise the validity of the model. It is often hard to derive a quantitative conclusion from it. There have been arguments on the strategic perspective of SD models, however (Taylor & Lane, 1998).

Discrete Event models is the most frequently used tool in patient flow simulation (Sobolev et al., 2011). As implied by the name, DES describes the systems by execution of events. In the context of hospital, events are a patient’s arrival, triage, diagnosis, departure and waiting, etc. A patient’s stay in an ED consists of events, and patients themselves are usually modeled as passive entities who will seize resources such as physicians and nurses.

The credit of the Discrete Event Simulation’s success should be given to its capacity to capture the essence of human activity. Most service systems operated in a discontinuous way can be abstracted into event-driven models. Depending on the languages or packages, DES models in patient flow study were constructed in different ways:
event scheduling, process interaction or objective orientation, yet the underlying logic is the same (Pegden, 2007).

In comparison to event driven models, Agent Based Model views the world from the entities’ perspective. It does not passively schedule the interactions between patients and physicians but rather specifies the behavioral rules of entities such as patients and let them interact autonomously. This novel approach has shown some potential in modeling complex interactions and concurrent activities. Some efforts have been exerted in this area; Hutzschenreuter et al. (2008) utilized ABM to select an optimal patient mix for admission scheduling. However, the utility of this approach with constrained resource environment is not well known (Sobolev et al., 2011).

One thing need to notice is that one study claimed that the phenomenon of ‘starvation’ (low acuity patients will often be starved from receiving services) is caused by the ‘factory view’ of DES models (Raunak et al., 2009). The ‘factory view’ refers to modeling patients as passive entities. It should be clarified that starvation is not relevant to modeling approaches but only the ranking (or dispatching) rules of patients at triage.

There are two main areas of focus of simulation based patient flow studies:

- Testing process improvement alternatives

- Staff scheduling

A large portion of literature on patient flow simulation used the model as a platform to conduct ‘what-if’ analysis for the proposed changes in ED processes changes. In one of the early work, McGuire (1994) presented a study of several process alternatives
and identified a significant reduction in patient LoS. For the purpose of process improvement, many researchers noticed the differences between EDs were not drastically large; therefore an ED simulation model with proper flexibility can be used for EDs to solve many generalized problems. A DES model (EDSim) developed by Miller et al. (2004) was intended to be reusable for general ED analysis. Based on the software, Ferrin et al. (2007) conducted analysis on processes to increase throughput in an ED, including the introduction of a discharge area and bypassing triage. Sinreich & Marmor (2004) studied the differences and similarities between five EDs and concluded a generic process ED model can be used by EDs to investigate alternatives. They showed that ‘simple and intuitive’ model is capable of analyzing ED operations.

Another area of focus, staff scheduling, requires a more detailed model than testing process alternatives. Many models built for this purpose are hospital specific, such as the one by Duguay & Chetouane (2007). These studies include very detailed information such as physical layouts of the EDs, staff shifts and empirical service times. Rossetti et al. (1999) utilized a detailed model to test alternative ED attending physician/staffing schedules and to analyze the corresponding impacts on patient throughput and resource utilization. Many works on staffing purposes can be found in proceedings of Winter Simulation Conference especially in those of 1990s (Kumar & Kapur, 1989; Draeger, 1992).

A few studies used the DES model for other purposes, such as real-time forecasting or monitoring of ED. Hoot et al. (2008) developed a discrete event simulation of ED patient flow for the purpose of forecasting near-future operating conditions and validated the forecasts with several measures. As a very interesting application of DES model, its
main problem is the computational cost when making a forecast. A potentially better solution is to run the simulation off-line and utilize output analysis techniques such as statistical meta-models to generate estimations.

Some existing surveys of simulation of patient flow suggested that although simulation in health care was not new, only few EDs had actually used this method for re-designing their patient flow (Sobolev et al., 2011). Two main factors contribute to this problem, one is limited understanding of the utility of simulation approaches; the other one is the doubts in model validation (Ham et al., 2003). According to the successful cases of using simulation to help re-design the ED processes, better engagement of care providers and a pilot implementation may be a solution to it (Medeiros et al., 2008).
Chapter 3

Methodology

3.1 Problem Definition

The patient flow optimization comprises of two objectives:

1. Minimize patients’ risk while waiting for treatment

2. Increase the throughput of ED on the premise of first objective.

The first objective explicitly represents the interest of patients to receive proper care as soon as possible. From a patient’s perspective, it is accomplished by minimizing the expected waiting time (WT) before seeing a physician. However, since we have noticed the cost of delay is distinct for patients with different levels of acuity, it is biased to use ‘average waiting time’ as a sole indicator of performance measurement. We assert that a vulnerability-adjusted measurement for patients’ waiting time would be more appropriate.

On the other hand, hospitals looks at the patient flow from the angle of increasing throughput and therefore revenue. Because the information of how much revenue per patient for each acuity level can bring is neither available nor in the scope of this research, the throughput maximization problem can be treated as an equivalent one of minimizing patient length of stay (LOS). Length-of-Stay consists of Waiting Time and Time-in-ED. We use the term ‘Time-in-ED’ to include all periods after the patient is seen by a doctor.
We also implicitly assume that after the point of seeing a doctor, the patient’s condition will be stable and as a cost of it, the patient will consume resources such as beds, nurses’ and/or physicians’ time. Please see Figure 3.1 to review a visual representation of these measures. As a result, we only need to focus on ‘time in ED’ because hospitals presumably are more interested in utilization of current capacity and waiting does not necessarily need resources. But we also notice that hospitals may have a preference to a certain patient flow interventions since it will have a different impact on the throughput.

![Fig. 3.1. Visualization of Selected ED Performance Measures](image)

As mentioned earlier, for the waiting time study, a vulnerability-adjusted time will be applied for measuring time in ED, and comparisons will be based on real time. Three interventions and a baseline model will be tested and compared for both objectives.

### 3.2 Conceptual Intervention Models

Due to the difficulties of empirical analysis and the high cost of trying new changes in ED conditions, Discrete Event Simulation (DES) is chosen as an effective tool to compare different patient flow interventions under various scenarios. We first outline the problem and the methodology to investigate the study objectives.
We aim to address two questions in the patient flow study:

1. What are the impacts of the different patient flow interventions, from both patients’ and medical providers’ perspectives?

2. How can the benefits be realized with the same amount of resources?

To answer these questions, a computerized simulation model is built as the test bed for different intervention approaches. The simulation model is calibrated with real (clinical) data and structural assumptions are validated by comparing case studies of real ED process flow. Based on the validated baseline model, the proposed patient flow interventions will be tested under different scenarios. Before the introduction of the computerized model, the conceptual process flows of alternative interventions (fast track, physician triage and triage team) are presented.

**Baseline Strategy** In the baseline model, the incoming patients first enter the ED via the front door entrance and wait in the waiting area. The triage nurse does a preliminary examination of the patient, assigns the ESI level as the priority and sends the patient to the waiting area. As soon as a bed becomes available, and using a priority queue discipline (First-In-First-Out within each priority group), a nurse takes the patient inside the ED room where the actual medical treatment begins. When patient arrives at the bed, she or he waits in the bed for a physician, who is accompanied by a nurse. They give the patient proper diagnosis and determine whether the patient needs lab tests or radiology tests or both. Then, patient receives treatment according to the results of the tests. Treatment will be performed by nurses or physicians depending on the needs. After patients experience one or more of these processes based on their needs, then
they are either discharged to home or admitted to the inpatient department. Figure 3.2 describes the process flow diagram of baseline strategy.

Fast Track  Fast track intervention differs from the baseline model with its additional fast track lane. The fast track lane, comprised of physicians, nurses and beds, is dedicated to serve non-urgent patients who do not need much resource and complex treatments, such as ESI Level 4 and ESI Level 5 patients. To avoid loss of generality, the number of each type of resources allocated in fast track will be treated as control variables in the experimentation.

Physician Triage  Physician triage patient flow intervention differs from the baseline strategy by “front loading” the diagnosis decision, which enables ordering all the tests needed right at the triage process. The physician resource needed at triage process will also be treated as a control variable. After being diagnosed during triage process, patients won’t need to be diagnosed again in ED; instead, they will receive treatment directly when test results are available.
Fig. 3.3. Process Flow of Fast Track

Fig. 3.4. Process Flow of Physician Triage

**Triage Team**  Triage Team is a combination of fast track and physician triage interventions, where the Triage team helps order patients’ tests when ED is busy, and gives treatment to patients with low severity levels (e.g., ESI Level 5). The number of allocated resources in Triage Team will be treated as a control variable.

### 3.3 Performance Measurement

In addition to traditional performance measurement like Length-of-Stay and Waiting Time, some novel measurements are developed for a better description of ED performance.
3.3.1 Vulnerability Adjusted Waiting Time

Patients with more serious acuity levels are more sensitive to the delay of treatment. In order to fairly describe the patients waiting time with this impact, we incorporate the vulnerability in measurement. It can refer to the triage systems like ESI or CATS; patients’ vulnerability are assigned according to their ESI level or CATS level, the higher level the more vulnerable. If more accurate description of patients’ acuity can be made with detailed information of symptoms, vulnerability can also be determined with respective to individual situation.

Our model of Vulnerability-adjusted waiting time (VAWT) is inspired by the notion of Severity of Failure. It is used in a quality control procedure called Failure Modes and Effects Analysis (FMEA), which is an inductive failure analysis vastly used in product development and reliability engineering (Gilchrist, 1993).

In addition to being an unbiased description, another benefit of VAWT is that it makes the trade-off between different patients comparable. The naive measures such as waiting time are multi-dimensional with respective to patient acuity levels. We propose
this measure to unify the multi-dimensional responses into a one-dimensional indicator; thus, the comparisons of the impact of different interventions can be easily made and visualized. VAWT can be computed according to the formula:

\[
\text{VAWT} = \text{Measure of Time to Treatment} \times \text{Vulnerability of Patient} \tag{3.1}
\]

where

Measure of Time to Treatment: We consider waiting time is responsible for patient’s deterioration in health status;

Vulnerability of Patient: To make all ESI levels of patients’ vulnerability comparable, we employ the recommended waiting time of CTAS system as a reference. Since a smaller waiting time indicates a smaller likelihood of waiting caused deterioration, the vulnerability is measured as the inverse of the recommended waiting time.

When we use the inverse of recommended waiting time to indicate vulnerability, we make an implicit assumption that deterioration of patient’s condition follows a constant rate. VAWT measure potentially allows a more accurate prioritization system. It is even possible to assign it on a case-by-case basis such as MAUT-AHP of (Ashour & Okudan Kremer, 2013). On the other hand, it can adopt a utilization function to better describe the value of elapsed time to patients.

3.3.2 Waiting-at-risk

As discussed earlier, ED performance can be measured by average Length of Stay, average Waiting Time. They are all ‘average’s. The problem with a mean statistic is that
it can not show the diversification of population. However regarding to ED problems, it is the worst cases rather than mean that are more valuable while comparing interventions. If everyone with similar acuity level waits for the same amount of time, then no one will be in danger of deterioration. The most critical patients are those who have waited for too long.

The Critical-Patient-First ranking rule we advanced earlier can obviously minimize danger of deterioration, but it is also necessary to develop a criteria to compare the variance of waiting time (or Vulnerability adjusted waiting time). In addition, we are more interested in the impact of ‘positive variance’ (larger than average). Therefore, inspired by the Value-at-Risk (VaR) in finance, we propose the notion Waiting-at-risk (WaR) to indicate the cases with extremely long waiting times. VaR is a summary statistical measure of possible portfolio losses resulting from “normally” distributed market movement (Linsmeier & Pearson, 2000). For instance, if a portfolio has a one-day 5% VaR of $1 million, there is a 5% probability that the value of portfolio will fall below $1 million over a one day period.

Similar to VaR, WaR is also defined by the 5th percentile but for upper half. A one-day 5% WaR of 2 hours means during the day, there is 5% of patients who will wait for 2 hours or more. Consequentially we can define peak-hour WaR and WaR for each level of acuity.

The notion of WaR gives us an alternative to compare the interventions and prioritization rules while incorporating risk.
3.4 Dynamic Prioritization

The drawback of static priority is that a patient with a higher level of acuity always takes precedence over patients with lower priorities. We propose a dynamic prioritization model that can avoid starvation and keep track of patients’ acuity conditions. This model differs from traditional triage systems where at the end of the triage process, the patient will be assigned a static priority. However, in a dynamic model, instead of being assigned a priority, a patient will be given a Critical Time and a Vulnerability for waiting. The Critical Time describes the maximum or the recommended time this patient can wait and Vulnerability calculates the vulnerability adjusted waiting time (VAWT). This patient will be added to the queue if ED is full, otherwise it will be sent directly into ED. When there’s an available bed in ED, the system will compute the patients’ priorities according to their current state. The calculation will follow a certain scheduling rule, which will be discussed in detail later. The patient with highest priority will be dispatched to ED. Please see Figure 3.7 for detail.
Triage
Assign a Critical Time
and Vulnerability for new patient
Add new patient to Queue
When resource available, compute
patients’ priorities according to the rule*
Send the patient with highest priority
to ED

Fig. 3.7. Dynamic Prioritization

Naturally with this setting, dynamic priority gives rise to ‘optimal’ control problem, so general scheduling rules can be applied in this model. With regard to the focus on simulation study, the thesis will not go into the detail of literature on stochastic control and asymptotic optimality of each scheduling rule.

If we use ‘Static priority non-preemptive scheduling’, this model will be exactly the old triage system. We propose four alternatives, namely Cumulative-Waiting-Time(CWT), Critical-Patients-First (CPF), Shortest-Treatment-First (STF) and a Balanced-Strategy (BS) which tries to strike a trade-off between CPF and STF.
**Cumulative-Waiting-Time (CWT)** rule dispatches the patients with largest cumulative waiting time. The waiting time is adjusted by patient’s vulnerability.

This type of dynamic priority has been studied in stochastic control literature before: using a similar linear increasing priority function, Bagchi & Sullivan (1985) developed expression of bounds on expected waiting time. But their analytical work was based on a single server queue.

In the ED environment, time is usually recorded in minutes, it is possible to have a draw when comparing the patients. If so, we pick the patient with the larger vulnerability value. The motivation behind this is to prevent patients with lower acuity level from waiting too long (starvation).

**Critical-Patients-First (CPF)** rule gives the highest priority to the patient who passes his or her critical time with the highest vulnerability value. If no one passes her or his critical time, then we apply CWT rule. This type of rules, called ‘due date’ scheduling has been widely studied in the scheduling and queueing literature (Wein, 1991). However, its effect on patients of different ESI levels and in multi-server ED environment is not tested yet. Similar to the calculation of VAWT, it can also adopt a utility function to calculate the critical value of time to each patient.

**Shortest-Treatment-First (STF)** gives the patients highest priority when his or her expected treatment time is shortest. This rule will inevitably favor patients with low acuity levels since they usually need shorter treatment time. This strategy is not a ‘fair’ policy, since it ignores the condition of patients. But the merit of it is its potential of increasing system throughput. We understand it is impossible to implement this policy
in reality, but it can give us an upper bound of the system capacity and is a fundamental component of the balanced strategy.

**Balanced-Strategy (BS)** is a mixture of previous strategies considering multiple factors in determining the priorities of the patients. 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 more factors.

There are two possible ways to balance different rules. One is to use a weighted scheme so that weights can be assigned to the various factors that make up to the Balanced-Strategy. Another way, possibly better way regarding to patient safety, is to apply Shortest-Treatment-First values within Critical-Patients-First rule. For instance, if there are patients who wait longer than their critical time, then dispatch the one with largest vulnerability value; if not, then apply STF rule to find the patient with shortest treatment time. Ideally this mixture may optimize the system output without harming the patients, but it needs to be tested in experiments under different scenarios.

### 3.5 Models Assumptions and Parameters

Model assumptions made in the conceptual model are categorized in two groups, structural (logic) assumptions and numerical (data) assumptions.

#### 3.5.1 Structural Assumptions

1. Static triage applies a modified ESI system.
For static prioritization, we adopted a modified ESI system. The Emergency Severity Index (ESI) is a five-level triage scale developed by ED physicians Richard Wuerz and David Eitel according to patients’ acuity and resource needed (Gilboy et al., 2005). Please note currently there’s no recommended standard for waiting time in ESI, we adopted the recommended waiting time from CTAS, which is also a five-level triage system.

2. Only physicians can order tests.

This assumption rules out the possibility that nurse ordering test because of the reliability issue and the regulatory requirements of the physician ordering in some states.

3. Diagnosis and treatment are decomposed into a three-stage process

Stage 1, first contact, including initial diagnosis and other actions performed together by a physician and a nurse before ordering tests; stage 2, second contact, including assessing test results and other decision making involving both physicians and nurses; stage 3, final treatment; the rest of the treatment performed by only nurses. The time length of each stage varies between patients of different ESI levels.

4. Perfect information sharing between physicians.

Doctors can perform treatment based on other’s diagnosis, which means patients won’t need to be diagnosed again in ED; instead, he will receive treatment directly whenever tests results are available.
5. Tests are simplified into two types.

Radiology: Patients need to visit testing center personally to be tested.

Lab tests: Patients do not need personal visit; instead specimens are collected beside the bed.

6. There is no interruption inside the ED.

3.5.2 Numerical Assumptions

We assume the patients’ arrival time is a non-stationary Poisson process. Without loss of generality, the input variables are modeled as random variables with fitted probability distributions from an empirical study (Peck, 2008), and other numerical assumptions in simulation model come from the recommendations of ESI Implementation Handbook and interviewing with practitioners, as listed in Table 3.1.

Table 3.1. Processing Time Distributions (in unit of hours; U=Uniform, E=Exponential, W=Weibull)

<table>
<thead>
<tr>
<th>ESI</th>
<th>Mix</th>
<th>Triage time</th>
<th>1st Contact</th>
<th>2nd Contact</th>
<th>Treatment</th>
<th>Lab test</th>
<th>Radiology</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5%</td>
<td>U(0.02,0.05)</td>
<td>E (0.28)</td>
<td>E (0.15)</td>
<td>E(0.33)</td>
<td>W(1.78, 0.5)</td>
<td>U(0.2,0.3)</td>
</tr>
<tr>
<td>2</td>
<td>25%</td>
<td>U(0.02,0.05)</td>
<td>E (0.35)</td>
<td>E (0.23)</td>
<td>E(0.25)</td>
<td>W(1.78, 0.5)</td>
<td>U(0.2,0.3)</td>
</tr>
<tr>
<td>3</td>
<td>30%</td>
<td>U(0.02,0.05)</td>
<td>E (0.16)</td>
<td>E (0.15)</td>
<td>E(0.15)</td>
<td>W(1.78, 0.5)</td>
<td>U(0.2,0.3)</td>
</tr>
<tr>
<td>4</td>
<td>20%</td>
<td>U(0.02,0.05)</td>
<td>E (0.15)</td>
<td>E (0.05)</td>
<td>E(0.1)</td>
<td>W(1.78, 0.5)</td>
<td>U(0.2,0.3)</td>
</tr>
<tr>
<td>5</td>
<td>20%</td>
<td>U(0.02,0.05)</td>
<td>E (0.10)</td>
<td>0</td>
<td>E(0.1)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
3.5.3 Model Input and Output

The input variables used in the simulation model include control variables and model parameters, as listed in Table (3.2). Table (3.2) also defines the system outputs that are regarded as traditional performance measures:

1. Length of Stay (LOS): the total time a patient spends in the system;

2. Time in ED (TED): the average length of time from the moment a patient enters the ED after being seen by a triage nurse to when he/she is discharged from the ED;

3. Waiting time (WT): defines the average waiting time from arrival to main ED (or treatment);

4. Utilization.

Table 3.2. System Input and Output

<table>
<thead>
<tr>
<th>Control Variables</th>
<th>Model Parameters</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allocation</td>
<td>Capacity</td>
<td>Length-of-Stay</td>
</tr>
<tr>
<td>Physician Allocation</td>
<td>Physician</td>
<td>Processing Time</td>
</tr>
<tr>
<td>Nurse Allocation</td>
<td>Nurse</td>
<td>Triage Time</td>
</tr>
<tr>
<td>Bed Allocation</td>
<td>Beds</td>
<td>Treatment Time</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>Diagnosis Time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test Time</td>
</tr>
</tbody>
</table>
3.6 Analytical Approximation of the Baseline Model

Before heading into setting up the simulation model and running experiments, some insights from queueing models is helpful to understand the feedback process between model inputs and outputs. From a queueing prospective, our model is a non-preemptive priority queueing with many servers ($M/G/N with priorities$). In a non-preemptive setting, once the ED procedure starts, it must be completed without any interruption. The major assumptions in the $M/G/N with priorities$ model are:

- There are $k$ classes of patients indexed by their priority such that class 1 is the highest and class $K$, the lowest;
- Arrival process of each class follows Poisson with intensity of $\lambda_k$, $k \in 1, \ldots, K$;
- First-In-First-Out within priority class;
- The capacity of Emergency Department is decomposed into smallest unit of servers – beds, with physician and nurse capacity allocated (therefore $N$ servers);
- Simplify all ED procedures into one step of service with generally distributed service times;
- The service rates of the $k$ class patients are denoted $\mu_1, \mu_2, \ldots, \mu_K$, respectively.

The model can be depicted in figure (3.8).

We are interested in the steady-state performance of the model, such as expected waiting time, queue length. However, in order to get concise approximation of the system state, we need to relax the assumption of class specified service time.
Let $\mu_k \equiv \frac{1}{\mu}$ equal for all classes. The overall utilization of ED is given by $\rho$ and the steady-state condition is denoted as

$$\rho = \sum_{i=1}^{K} \rho_k \leq 1, \text{where } \rho_k = \frac{\lambda_k}{N\mu}$$ (3.2)

Kella & Yechiali (1985) gave the expressions of first two moments of waiting time by deriving the Laplace-Stieltjes transformation using a vacations(idle time) approach. Let $E_{2,N}$ be the probability of delay in the $M/M/N$ queue (Erlang-C formula):

$$E_{2,N} = \frac{(N\rho)^N}{N!(1-\rho)} \left[ \sum_{k=0}^{N-1} \frac{(N\rho)^k}{k!} + \frac{(N\rho)^N}{N!(1-\rho)} \right]^{-1}$$ (3.3)

Then the expected waiting time of the $k^{th}$ class is:

$$E(W_q^k) = \frac{1}{N\mu} \frac{E_{2,N}}{(1-\sigma_k)(1-\sigma_{k-1})}$$ (3.4)

where, $\sigma_k = \sum_{i=1}^{k} \rho_i$, $\sigma_0 = 0$.
and the second moment of expected waiting time is:

\[
M_{W^k_q}^2(0) = \frac{2E_{2, N}}{(N\mu)^2} \frac{(1 - \sigma_k \sigma_{k-1})}{(1 - \sigma_k)^2(1 - \sigma_{k-1})^3} 
\] (3.5)

With the help of (3.4) and (3.5), we can get a reasonable approximation of mean waiting time of patients at each ESI level in baseline model. This will further help us to parameterize the simulation model and select some test scenarios.

3.7 Computerized Model and Validation

We use Simio version 4 (Pegden, 2007), a discrete event simulation software package to construct the simulation models presenting four types of patient flow interventions in ED. The software is capable of simulating time-dependent processes, where entities, such as patients, interact with resources like physicians and nurses. The model has one source to create patient entities, which carries information based on a data table including patients’ mix, types and priorities. The waiting area, triage area, radiology room and beds are modeled as servers.

Because of structural differences and control logic in four patient flow methods, we built them as four separate models in Simio. In the following experiment section, we synchronized the input and parameters, and manually calculated the results.

**Baseline**  Model description is introduced following the order from the creation of a patient through entering the sink.

- Model Entity Referenced State Variables
Patients, as entities in our model, need to carry information with them, so we add several state variables in “ModelEntity” to implement them.

• Entity Referenced Variables

Data table is exploited to control the patients’ mix, types, priorities and respective diagnosing time and treating time. Once the patient is born, it will get priority and type in his or her state variables. When he or she is processed by the triage center, integers (0 or 1) will be assigned to the state variables “labtested” and “radiology” respectively depending on the patient’s type.

• Waiting Area

Waiting area are modeled as servers, so we can apply ranking rules (smallest value first by priority) on them. We have two waiting areas, where patients are sorted differently; one before triage with ranking rule FIFO, the other one after triage by priority. This follows that patients’ priority are realized after triage.

• Control Logic

A NodeList is built to routing the patients, and routing logic is realized in a transfer node.

When patients (in ED) are sent to radiology tests, their beds are still considered as occupied. So in addition to routing patients by bed status, we need to also control total number of patients in ED (including patients in radiology department). In our finalized model, we subclassed our Transfernode and set the Routingoutlogic publicly accessible. Thus we can suspend it when the ED is full, and resume it
Fig. 3.9. Control Logic to ED

when ED is not. The process to control suspending is triggered when modelentity passes through a specific node representing the gate of ED; the one to control resuming is triggered when a patient arrives to billing station.

- Diagnosis & Treatment

After entering in ED, the patient will be sent to an empty bed by nurse. Before processing by a physician and a nurse, here we need to check whether the patients need treatment or diagnosis: if he or she is the first time entering in ED (patients from waiting area), he or she will be processed with diagnosing time. If the patient is from radiology department, then a treatment time will be assigned to the entity’s state variable “Processtime”.

If a patient is processed with diagnosing time, then the doctor will order lab tests and sent him or her to radiology department if necessary. Thus, we need a process to realize this judgment. Moreover, if it’s a patient’s second time to come to a bed, before starts treatment, we also need a process to check whether his or her lab test result has come out yet. If it’s not, then the patient will be delayed until lab test result comes.
• Radiology Department The patients come out of radiology department also use a node list the routing back to an empty bed.

**Fast track** In fast track methods, we set a bed as a fast track area. The major structural modification here is to add a special waiting area for ESI 5 patients. Since the fast track is dedicated to treat ESI 5 patients, ESI 5 patients will be diverted at the ‘outputnode@Triage’ using link weight. Though there is a fast track dedicated to serve ESI 5 patients, if the fast track is full, ESI patients can also join the lines of waiting in ED. We use link weight to implement this logic.
**Physician Triage & Triage Team**  In Physician Triage and Triage Team model, patients go to triage team (or Triage Team area) when ED is full, where they will receive diagnosis and sent to radiology if necessary. Thus we need a server to contain the patients coming back from radiology department. On the other hand, if ED is not full, patients can go directly into ED. We realized this logic using link weight. Also using link weight, we routed patients coming out of the radiology department.
Chapter 4

Experimentation

4.1 Preparation

4.1.1 Model Parameterization

There are three types of resources at our ED: physicians, nurses and beds. We arbitrarily set the ED size (which is denoted by number of beds) to 40, a typical number of mid-size EDs. To avoid any misspecification of resources, an experiment was conducted for baseline model to make sure the number of physicians and nurses are optimal in the current setting, with respect to total throughput and average waiting time. Also by doing so, we can prevent the alternative interventions from exploiting possible under-utilization to achieve better performance and thus, the generality can be kept.

Since the statistics we are interested in are steady-state performances, a warm-up period is set to 5 hours to rule out the initialization effect. The total run length is 40 hours, which is ten times the warm-up period. Results show that under current

<table>
<thead>
<tr>
<th>Intervention</th>
<th># of physicians</th>
<th># of nurses</th>
<th># of beds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast Track</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Physician Triage</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Triage Team</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>
assumptions, the minimum required numbers of physicians and nurses are 12 and 22, respectively. The physician-to-bed ratio is 0.3, and the nurse-to-bed ratio is 0.55. The results are also in compliance with empirical observations and state regulations. For instance, CA requires that the nurse-to-patient ratio at emergency service should be at least 1:4. When number of physicians and nurses are fixed, the capacity of ED is computed by approximation, and the benchmark arrival rate for test scenarios is 34.5 patients per hour.

In comparison to main ED, the resource allocated to interventions is relatively small. We set the number of physicians allocated to each alternative as 1 and then computed required number of nurses and beds by maximizing the utilization of beds and nurses. The effect of more dynamic staffing between interventions and main ED will be discussed later.

### 4.1.2 Testing Scenarios

Testing scenarios are carefully selected to cover a variety of empirical situations.

With the help of the analytical approximation (3.4) and (3.5) for the baseline model, we are able to find out proper arrival rates with respect to patient’s waiting time. The limitation of the analytical model is the assumption that the service rates (diagnosis and treatment rates) are equal for every type of patients. Yet, the results can still give us some insight in the system performance.

From figure 4.1, we found the performance in terms of waiting time is steady for patients of first four ESI levels. Only for ESI 5 patients, the waiting time grows significantly as ED utilization approaches one.
Fig. 4.1. Approximated Waiting Time for Selecting Scenarios
We pick three different patient arrival processes as our testing scenarios: two stationary and one non-stationary Poisson processes. Among two stationary ones, one’s arrival rate is less than the capacity of ED so that a steady performance for patients of all ESI levels can be achieved; the other one’s rate is very close to Emergency Department’s capacity in order to perform a stress test. From figure (4.1), we set the regular hour rate to 85% of ED capacity, and the stress test’s rate to 95%. For stress test, we focus on its 5-hour transient behavior.

The non-homogeneous case has an empirically distributed shape of arrival rate. Figure 4.2 shows the rates of testing processes scaled to capacity.

![Fig. 4.2. Testing Scenarios Scaled to ED Capacity](image-url)
4.1.3 Comparison of Multiple Alternatives

Bonferroni correction is applied to compare confidence intervals of differences in performance measures. Each statement $S_i$ represents an alternative in the comparison.

$$P(\text{all statement } S_i \text{ are true, } i = 1, \ldots, C)$$

To construct 95% overall confidence level, $C = 3$ (for three comparisons with baseline model), the overall error probability is $\alpha_E = 0.05$, then $\alpha_i = 0.05/3 = 0.0167$ for $i = 2, 3, 4$. Using independent sampling with unequal variances, the confidence interval is:

$$\bar{Y}_1 - \bar{Y}_2 \pm t_{\frac{0.05}{2}} \cdot \text{s.e.}(\bar{Y}_1 - \bar{Y}_2)$$

with the standard error

$$\text{s.e.}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{S_1^2}{R_1} + \frac{S_2^2}{R_2}}$$

and degrees of freedom

$$V = \frac{(\frac{S_1^2}{R_1} + \frac{S_2^2}{R_2})^2}{\frac{(S_1^2)^2}{R_1-1} + \frac{(S_2^2)^2}{R_2-1}}$$
4.2 Comparing Interventions

4.2.1 Regular-Hour Test

The experiment was carried out with 100 trials of baseline model and each intervention to achieve significant differences. The results of comparisons are shown in Table 4.2.

<table>
<thead>
<tr>
<th>ESI</th>
<th>Baseline</th>
<th>FT vs. Baseline</th>
<th>PT vs. Baseline</th>
<th>TT vs. Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.585</td>
<td>[-0.004, 0.147]</td>
<td>[-0.981, -0.882]</td>
<td>[-0.524, -0.407]</td>
</tr>
<tr>
<td>2</td>
<td>5.157</td>
<td>[-0.027, 0.144]</td>
<td>[-1.273, -1.155]</td>
<td>[-0.734, -0.595]</td>
</tr>
<tr>
<td>3</td>
<td>7.495</td>
<td>[0.944, 1.447]</td>
<td>[-2.775, -2.501]</td>
<td>[-1.563, -1.253]</td>
</tr>
</tbody>
</table>

According to Table 4.2, we observed the two interventions with “front-loading” the test-ordering decisions significantly improve the time to see the doctor of all ESI levels of patients. Although not specifically designed to benefit ESI level 5, the physician triage model also have a surprisingly large reduction (about 60% percent) in waiting time.

As for fast track approach, its targeted patients (ESI levels 4 and 5) receive substantial benefits: ESI levels 4 and 5 patients enjoy 8 and 21 minutes less waiting time, respectively. On the other hand, in contrast to some empirical studies which claimed that fast track improved performance in treating all kinds of patients, we find that fast track always benefits some types at the cost of making other patients worse off (shown in Figure 4.3). The patients most likely to be sacrificed are those whose priorities are
According to Table 5, we observe the two interventions with “front-loading” the test-ordering decisions significantly improve the time to see the doctor of all ESI levels of patients. Although not specifically designed to benefit ESI 5, the physician triage model also have a surprisingly large reduction (~60% per cent) in waiting time. As for fast track approach, its targeted patients (ESI 4, 5) receive substantial benefits: ESI 4 and 5 patients enjoy 8 and 21 minutes less waiting time, respectively. On the other hand, in contrast to some empirical studies which claimed that fast track improved performance in treating all kinds of patients, we find that fast track always benefits some types at the cost of making other patients worse off (shown in Figure 7). The patients most likely to be sacrificed are those whose priorities are slightly higher than fast-track patients. In our experiment, patients with ESI levels 1 and 2 are only slightly impacted (with insignificant increase of waiting time); however, ESI level 3 patients wait about 1 minute longer. Someone may say this is worth it because in return of one-minute tardiness we have 3 minutes reduction in total average waiting time. But this argument cannot be justified without considering the vulnerability of different patients. The results show that the reallocation of current resources in a well running ED will not increase benefits to all stakeholders uniformly.

The initial motivation of front loading the test-ordering process is to add value to the patients’ waiting time outside ED, and consequently, to have them spend less time in ED waiting for test results. Our results verify the effect of the “wait-with-value” strategy: all patients who need test(s)-ordering (ESI 1, 2, 3, 4) receive reduction in time spent in ED. We normalize the time in ED of interventions by baseline time to see their impact (shown in Figure 8). Both triage team and physician triage reduce ESI 1, 2, 3
cannot be justified without considering the vulnerability of different patients. The results show that the reallocation of current resources in a well running ED will not increase benefits to all stakeholders uniformly.

### Table 4.6

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ESI 2</td>
<td>83.373 [−0.314, 0.011]</td>
<td>[−10.646, −10.334]</td>
<td>[−7.804, −7.471]</td>
</tr>
<tr>
<td>ESI 3</td>
<td>65.706 [−0.398, -0.076]</td>
<td>[−11.690, −11.413]</td>
<td>[−8.382, −8.082]</td>
</tr>
<tr>
<td>ESI 4</td>
<td>44.126 [0.947, 1.479]</td>
<td>[−9.256, −8.784]</td>
<td>[−3.339, −3.992]</td>
</tr>
<tr>
<td>ESI 5</td>
<td>16.599 [0.544, 0.917]</td>
<td>[4.043, 4.537]</td>
<td>[3.131, 3.642]</td>
</tr>
<tr>
<td>Total</td>
<td>56.832 [0.350, 0.703]</td>
<td>[−7.775, −7.433]</td>
<td>[−7.258, −6.602]</td>
</tr>
</tbody>
</table>

**Fig. 4.4.** Time in ED in Regular Hours

and 4 patients’ time in ED about 10 – 20 %, but ESI 5 patients spend 20 – 30% more time in ED. Fast-track has little impact on the patients’ time in ED, but we observed a two-minute reduction for ESI 1 patients.

#### 4.2.2 Stress Test

In regular hours, triage team and physician triage perform almost equally well. However, in stress testing, physician triage method shows greater potential for handling an increased load in arrivals. According to Table 4.4, the performance of triage team degenerates while the physician triage performs consistently across the two scenarios. The improvement to ESI 1 and 2 patients brought by Triage Team is close to zero and ESI 5 patients wait even longer than baseline. This is due to the heavy duty undertaken by the triage team: they need to order tests for all patients passing through as well
Table 4.3. Time in ED (Regular Hours)

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>FT vs. Baseline</th>
<th>PT vs. Baseline</th>
<th>TT vs. Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESI 2</td>
<td>83.373</td>
<td>[-0.314, 0.011]</td>
<td>[-10.646, -10.334]</td>
<td>[-7.804, -7.471]</td>
</tr>
<tr>
<td>ESI 3</td>
<td>65.706</td>
<td>[-0.398, -0.076]</td>
<td>[-11.690, -11.413]</td>
<td>[-8.382, -8.082]</td>
</tr>
<tr>
<td>ESI 4</td>
<td>44.126</td>
<td>[0.947, 1.479]</td>
<td>[-9.256, -8.784]</td>
<td>[-3.339, -3.992]</td>
</tr>
<tr>
<td>ESI 5</td>
<td>16.599</td>
<td>[0.544, 0.917]</td>
<td>[4.043, 4.537]</td>
<td>[3.131, 3.642]</td>
</tr>
<tr>
<td>Total</td>
<td>56.832</td>
<td>[0.350, 0.703]</td>
<td>[-7.775, -7.433]</td>
<td>[-7.258, -6.602]</td>
</tr>
</tbody>
</table>

as treating some of them. When the arrival rate is high, triage team spends most of their capacity on treating patients so the benefit of front-loading is very small. We also note that in peak hours ESI 3 patients suffer a larger increase in waiting time (shown in Figure 4.5).

Table 4.4. Time to See Doctors (Stress Test)

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>FT vs. Baseline</th>
<th>PT vs. Baseline</th>
<th>TT vs. Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESI 1</td>
<td>4.214</td>
<td>[1.114, 1.414]</td>
<td>[-1.996, -1.805]</td>
<td>[-0.683, -0.457]</td>
</tr>
<tr>
<td>ESI 2</td>
<td>7.450</td>
<td>[1.272, 1.648]</td>
<td>[-2.747, -2.503]</td>
<td>[-0.673, -0.346]</td>
</tr>
<tr>
<td>ESI 5</td>
<td>58.843</td>
<td>[-42.891, -38.557]</td>
<td>[-31.552, -26.791]</td>
<td>[2.928, 9.780]</td>
</tr>
</tbody>
</table>

The patient’s time in ED does not change drastically between regular hours and peak hours. Yet interestingly, we find that during the stress test, physician triage and triage team improve the patients’ time in ED to a higher degree in comparison to the regular-hour case (see Figure 4.6). This is due to in peak hours, patients wait longer outside the ED, and hence less inside the ED.
When the arrival rate is high, triage team spends most of their capacity on treating patients so the benefit of frontloading is very small. We also note that in peak hours ESI 3 patients suffer a larger increase in waiting time (shown in Figure 10).

### Table 4.5. Time in ED (Stress Test)

<table>
<thead>
<tr>
<th>ESI</th>
<th>Time to See Doctor (TSD) unit: minutes</th>
<th>FT vs. Baseline</th>
<th>PT vs. Baseline</th>
<th>TT vs. Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>88.715</td>
<td>0.436, 1.045</td>
<td>-10.056, -9.539</td>
<td>-9.840, -9.280</td>
</tr>
<tr>
<td>3</td>
<td>71.629</td>
<td>-0.496, 0.121</td>
<td>-12.349, -11.805</td>
<td>-8.430, -7.801</td>
</tr>
<tr>
<td>5</td>
<td>21.591</td>
<td>0.559, 1.347</td>
<td>7.074, 8.073</td>
<td>7.717, 8.867</td>
</tr>
<tr>
<td>Total</td>
<td>63.888</td>
<td>2.364, 3.043</td>
<td>-7.557, -6.951</td>
<td>-7.478, -5.338</td>
</tr>
</tbody>
</table>
Figure 10. Mean waiting time comparison in steady-state test (TIE)

Table 68. Stress test (TIE)

<table>
<thead>
<tr>
<th>Time in ED (TIE)</th>
<th>Unit: minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
</tr>
<tr>
<td>Fasttrack</td>
<td></td>
</tr>
<tr>
<td>Physician Triage</td>
<td></td>
</tr>
<tr>
<td>Triage Team PDQ</td>
<td></td>
</tr>
</tbody>
</table>

ESI 1 | 87.254 [−3.795, −2.791], [−16.022, −15.004] | [−15.653, −14.613] |
ESI 2 | 88.715 [0.436, 1.045], [−10.056, −9.539] | [−9.840, −9.280] |
ESI 3 | 71.629 [−0.496, 0.121], [−12.349, −11.805] | [−8.430, −7.801] |
ESI 4 | 51.268 [3.331, 4.420], [−7.811, −6.749] | [−7.292, −6.658] |
ESI 5 | 21.591 [0.559, 1.347], [7.074, 8.073] | [7.717, 8.867] |

Total | 63.888 [2.364, 3.043], [−7.557, −6.951] | [−7.478, −5.338] |

The patient’s time in ED does not change drastically between regular hours and peak hours. Yet interestingly, we find that during the stress test, physician triage and triage team improve the patient’s time in ED to a higher degree in comparison to the steady state case (see Figure 11).

This is because during peak hours, patients wait longer outside the ED, and hence less inside the ED.

4.2.3 Result of Waiting-at-Risk

One natural question comes after comparison of mean statistics: is average waiting time an accurate measure of ED performance? To answer this question, again four patient flow interventions are compared under two scenarios by the measurement proposed in chapter 3, Waiting-at-Risk (WaR), to see the population most adversely affected. The mean and Waiting-at-Risk statistics are depicted in one figure for each intervention.

In the baseline case, the WaR increases drastically as the patient’s ESI level decreases. For patients with lowest ESI level, 5% of them may suffer a waiting time greater than 114 minutes. This happens even ED is operating in regular hours (86% utilization).

Generally speaking, for the baseline case, on average the WaR is 2–4 times more of the mean waiting time. The trend is that the lower ESI level, the larger WaR-to-mean.
Fig. 4.7. WaR Plot of Baseline Model

ratio. The harmful impact of static patient priority, starvation, is visualized. On the other hand, one thing even worse is that in contrast to prolonged mean waiting time, the impact of starvation is more serious on the WaR.

Previously, the Fast-Track strategy was shown to be merely a trade-off on average waiting time between patients of different ESI levels. But the plot of WaR tell us a another story: in terms of WaR, fast track is able to reduce waiting time for patients of almost every ESI level. Especially for fast track patients (ESI 4, 5), the reduction is huge (Comparing to baseline case, 100 mins for ESI 5, 30 mins for ESI 4).

But of course, the low level patients left out of fast track (ESI level 3) suffers at price of WAR increase of 9 minutes. Overall, fast track is necessary for ED even when it is running in regular hours.

As for the Physician Triage and Triage Team method, their WaR performance differs much greater than mean performance. The WaR of all ESI level of patients is low, especially for more urgent patients (ESI 1,2,3), their WAR are the lowest among four interventions.
Fig. 4.8. WaR Plot of FT Model

Fig. 4.9. WaR Plot of PT Model
On the contrary, adding more functionality to triage team could also imposes more uncertainty in the processing time. The WaR of Triage Team method clearly reflects this point. Even with the philosophy of ‘see and treat’, low ESI level (4 and 5) patients suffer a large variance in waiting time.

Now directing attention to the peak hour schedule (95% utilization), the baseline model shows a larger impact of ‘starvation’ on low ESI level patients (4 and 5). Physician Triage and Triage Team almost have the same shape of WaR, except for ESI level 5 patients. The WaR of ESI level 5 patients doubles in Triage Team, comparing to that of Physician Triage.

4.2.4 Test with Empirical Arrival Rate

We have already showed that the mean total waiting time is a biased performance indicator; thus, we expand our comparisons to the use of Vulnerability-Adjusted Waiting Time as our measure. 100 runs were completed at every test point for each model. The
Fig. 4.11. Peak Hour WaRs
patients were grouped with respect to arrival time and cumulative mean VAWT were calculated for each group.

\[ \text{Cumulative VAWT} = \sum_{ESI\text{level}=1}^{5} \text{Vulnerability} \times \text{Waiting Time of ESI level} \quad (4.1) \]

The results are shown in the Figure 4.12 and Table (4.2.4). The shaded regions indicate the 95% confidence intervals.

**Fig. 4.12.** Cumulative VAWT for Non-stationary Arrival

With the inclusion of patients' vulnerability, we observe that the baseline model and the studied interventions have different performances, and the performances under different scenarios are not consistent. This information is helpful when scheduling ED physicians and nursing forces. For instance, fast track achieves better performance in comparison to baseline when the arrival rate is increasing; however, SOF index decreases
slower when using fast track after the rush hour. Triage team also suffers the same property when arrival rate is decreasing. It seems that they have a “smoothing” effect on the patients’ waiting time. On the other hand, physician triage shows consistent improvement when facing both increasing and decreasing arrival rate. If we ignore the confidence interval, comparing with other models, physician triage also postpones the peak SOF from 12:00 to 14:00.

### 4.3 Comparing Dynamic Priority Scheduling Rules

Dynamic Priority rules are not exclusive to the patient flow interventions. They can be combined with the baseline, Physician Triage or Triage Team methods; therefore they are only tested in the baseline case for two different patient arrival rates. The purpose of experiments on dynamic priority is to discover the impact on reduction of starvation and therefore waiting time for low ESI level patients.

<table>
<thead>
<tr>
<th>Time</th>
<th>Baseline SOF</th>
<th>Baseline HW</th>
<th>Fast Track SOF</th>
<th>Fast Track HW</th>
<th>Physician Triage SOF</th>
<th>Physician Triage HW</th>
<th>Triage Team SOF</th>
<th>Triage Team HW</th>
</tr>
</thead>
<tbody>
<tr>
<td>0:00-2:00</td>
<td>0.933</td>
<td>0.112</td>
<td>1.074</td>
<td>0.110</td>
<td>0.944</td>
<td>0.129</td>
<td>0.817</td>
<td>0.111</td>
</tr>
<tr>
<td>2:00-4:00</td>
<td>0.834</td>
<td>0.125</td>
<td>0.820</td>
<td>0.133</td>
<td>0.945</td>
<td>0.128</td>
<td>0.729</td>
<td>0.117</td>
</tr>
<tr>
<td>4:00-6:00</td>
<td>0.968</td>
<td>0.110</td>
<td>0.963</td>
<td>0.124</td>
<td>0.962</td>
<td>0.123</td>
<td>0.866</td>
<td>0.108</td>
</tr>
<tr>
<td>6:00-8:00</td>
<td>2.087</td>
<td>0.667</td>
<td>2.461</td>
<td>0.924</td>
<td>1.583</td>
<td>0.229</td>
<td>2.083</td>
<td>0.629</td>
</tr>
<tr>
<td>8:00-10:00</td>
<td>23.532</td>
<td>2.976</td>
<td>22.275</td>
<td>3.421</td>
<td>12.011</td>
<td>1.897</td>
<td>16.224</td>
<td>2.034</td>
</tr>
<tr>
<td>10:00-12:00</td>
<td>34.223</td>
<td>2.350</td>
<td>30.903</td>
<td>3.180</td>
<td>19.308</td>
<td>2.430</td>
<td>22.406</td>
<td>2.004</td>
</tr>
<tr>
<td>12:00-14:00</td>
<td>31.329</td>
<td>2.619</td>
<td>30.263</td>
<td>3.027</td>
<td>21.368</td>
<td>3.757</td>
<td>21.073</td>
<td>1.831</td>
</tr>
<tr>
<td>14:00-16:00</td>
<td>26.852</td>
<td>2.425</td>
<td>25.614</td>
<td>2.453</td>
<td>19.827</td>
<td>2.743</td>
<td>20.399</td>
<td>2.083</td>
</tr>
<tr>
<td>16:00-18:00</td>
<td>22.213</td>
<td>2.140</td>
<td>21.216</td>
<td>2.617</td>
<td>17.598</td>
<td>2.506</td>
<td>18.447</td>
<td>2.129</td>
</tr>
<tr>
<td>18:00-20:00</td>
<td>16.404</td>
<td>2.243</td>
<td>15.269</td>
<td>2.217</td>
<td>11.570</td>
<td>2.068</td>
<td>15.805</td>
<td>2.135</td>
</tr>
<tr>
<td>20:00-22:00</td>
<td>8.924</td>
<td>1.566</td>
<td>10.040</td>
<td>1.960</td>
<td>5.576</td>
<td>1.417</td>
<td>10.063</td>
<td>1.882</td>
</tr>
<tr>
<td>22:00-24:00</td>
<td>4.371</td>
<td>1.142</td>
<td>4.794</td>
<td>1.512</td>
<td>2.644</td>
<td>0.981</td>
<td>5.503</td>
<td>1.637</td>
</tr>
</tbody>
</table>
The mean waiting time shows that with static priority, the waiting time is highest for ESI level 5 patients, but lowest for ESI level 1, 2 and 3 patients. The Cumulative-Waiting-Time rule seems to perform worst in three rules: the improvement in waiting time for low level patients is not much but at a cost of delaying urgent patients for too long. On the other hand, the due date strategy behind the Critical-Patients-First rule seems to generate a better trade-off.

![Comparison of Scheduling Policies on Mean](image.png)

Fig. 4.13. Comparison of Scheduling Policies on Mean

Generally speaking, from static priority to CPF, the mean waiting time is decreasing slower in pace with the increasing ESI level – Slopes in the figure are becoming more gentler. The WaR waiting time shows the trend even clearer. An unwanted result of the trade-off is also more obvious as CPF generates too much WaR waiting time for urgent patients.

In order to evaluate the trade-off between patients’ waiting time, Vulnerability-Adjusted-Waiting-Time (VAWT) is needed.
Fig. 4.14. Comparison of Scheduling Policies on WaR

A look at the VWAT shows, however, after adjusting the weight of waiting time by patient's vulnerability, static priority outperforms the dynamic priority scheduling on both mean and WaR (95% quantile) of the statistic. The mean result of cumulative VAWT is consistent with the work of Federgruen & Groenevelt (1988) on the optimal control policy for non-preemptive priority queues. He stated that suppose cost $C_k$ for one unit wait of class $k$ with an objective of minimizing $\sum_k C_k \lambda_k E(W_k)$, the optimal rule is to give the highest priority to the largest $C_k$. However, what is not expected is that, the WaR VAWT is also consistent with this rule. Further analysis or experiment is necessary to unveil the reasoning behind.

Table 4.7. Cumulative VAWT of Scheduling Policies in Regular Hours

<table>
<thead>
<tr>
<th>Measure</th>
<th>Static</th>
<th>CWT</th>
<th>CPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean VAWT</td>
<td>3.7</td>
<td>3.9</td>
<td>4.6</td>
</tr>
<tr>
<td>WaR VAWT</td>
<td>9.4</td>
<td>10.8</td>
<td>15.8</td>
</tr>
</tbody>
</table>
Chapter 5

Conclusions and Future Work

In this thesis, three popular patient flow interventions (fast track, physician triage and triage team) in addition to a basic emergency department model (baseline). A completely controlled environment was built using DES simulation for experimental purposes. We test the baseline model and interventions using different arrival scenarios. Under the steady-state scenario, physician triage and triage team performed equally well. The benefit of processing tests concurrent to waiting was verified. In the stress test, the performance of triage team degenerated because most time was spent on treatment for urgent patients. We showed that triage team intervention substantially reduced the waiting time for its targeted patients (ESI 4, 5); however, at a cost of a longer waiting time for patients with slightly higher priorities (e.g., ESI 3).

In order to justify the trade-offs in waiting times of patients at different ESI levels, we proposed a one-dimensional measure — Vulnerability-Adjusted Waiting Time to indicate the distinct vulnerability of patients. It was measured using recommended waiting time limits. Using this measure, we compared the performance of interventions under an empirically distributed non-stationary arrival process. The results showed that physician triage and triage team can increase performance; however, we also observed the inconsistent performances: some interventions had better performance when facing
an increasing rate of patients’ arrival but worse performance after peak hours. This fact is informative for making ED staff scheduling decisions.

We also noticed the adverse effect of ‘starvation’ in non-preemptive static priority queue. To resolve this problem, we proposed an algorithm to dynamically assign patients’ priorities at triage. The determination of the priority takes the waiting time and patient’s vulnerability to waiting into consideration. Various types of scheduling policies can be adopted here from Computer Science and Queueing literature to achieve different objectives.

At last, to overcome the lack of expressive power in mean statistics (Average LoS, Average WT, etc), the notion of Waiting-at-Risk is borrowed from the field of Finance to measure the 95% quantile of populations’ waiting time. The experiment shows the ‘chains of priority’ is the key source of large WaR. Fast track method shows a good potential in reducing WaR by breaking the priority chain into two short ones.
Reference


