University of Cincinnati

Date: 5/19/2016

I, Elham Torabi, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Business Administration.

It is entitled:
Capacity and Flow Management in Healthcare Delivery Systems with Multi-priority Patients

Student’s name: Elham Torabi

This work and its defense approved by:

Committee chair: Craig Froehle, Ph.D.

Committee member: Michael Magazine, Ph.D.

Committee member: Uday Rao, Ph.D.

Committee member: Denise White, Ph.D.
Capacity and Flow Management in Healthcare Systems with Multi-priority Patients

A dissertation submitted to the Graduate School of
the University of Cincinnati
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the Department of Operations, Business Analytics and Information Systems
of the Lindner College of Business

by

Elham Torabi

Bachelor of Science, Industrial Engineering, Isfahan University of Technology
Master of Science, Industrial Engineering, University of Houston

May 2016

Committee:
Craig M. Froehle, PhD (Chair)
Michael Magazine, PhD
Uday Rao, PhD
Denise White, PhD
Abstract

In healthcare services, the goal is to provide timely and high-quality care. However, given the high costs of healthcare, resource shortages, and increasing demand, efficiency is also crucial. High variability, together with cost pressures, make matching capacity and demand challenging for healthcare systems. Despite the need for being maximally efficient, health centers need to be responsive to patients as well, especially if there are patients who need emergent (immediate) care. Examples are operating rooms, with both elective (low-priority) and emergency (high-priority) patients, and emergency departments in which patients are prioritized into five classes based on medical urgency and resource requirements. The focus is to match supply and demand to ensure smooth patient flow in the system. This requires careful capacity and flow decisions, which involves balancing the tradeoff between efficiency and responsiveness in the presence of multiple priority classes of patients. Thus, we study two major operational decisions that arise when serving multi-priority streams of patients: 1) resource allocation, and 2) flow allocation. These decisions are interrelated and have a direct effect on system performance measured by many different metrics, such as patient wait time, patient flow time, and throughput. We start from a two-priority system motivated by operating room systems with emergency and non-emergency patients (Chapter 2). Then we explore the more complicated case of multi-priority systems motivated by emergency departments with several priority classes of patients (Chapters 3 and 4). For measuring the performance of different policies, we use patient waiting time as a proxy for system responsiveness. Using simulation and optimization methods, we identify capacity and flow allocation policies that minimize wait time, thus maximizing system responsiveness. We capitalize on statistical methods and data-mining techniques to help inform the operational and theoretical models, including those based on queueing theory, optimization, and simulation. The results indicate that in systems with multiple priority classes of patients,
both aggregating (pooling) and disaggregating capacity can be useful and helpful, depending on the patient populations involved and the system’s larger operational objectives. Our investigations highlight the importance of flow allocation decisions and the impact they can have on system responsiveness.
Acknowledgements

I am heartily grateful to my family, and all my friends in the OBAIS department, as I could not have finished this journey without their support. Not all doctoral students are as lucky as I have been; my deepest gratitude goes to my advisor, mentor, English teacher, and my dearest friend Professor Craig Froehle for his unconditional support during these years. Craig, no words can help me thank you enough for your kindness, patience and all that I learned from you.

I am much obliged to my wonderful committee members, Professors Uday Rao and Mike Magazine as well as Dr. Denise White for their technical guidance and mental support. Mike, I will be always in your debt for believing in me and giving me the chance to be here. Uday, thank you for teaching me to criticize my own work and to push myself without giving up.

Special thanks to Professors David Rogers, David Kelton and Jeffery Camm for answering my questions every now and then; also to Ms. Tricia Burger for being there for me when I needed a motherly hug. I would like to acknowledge my friends and mentors in the Department of Emergency Medicine especially Dr. Chris Lindsell and Dr. Chris Miller. Thank you for giving me the opportunity to work with you.

Most importantly, I would like to thank my great comrade since college, my beloved husband. Masoud, thank you for believing in me, being there every moment and inspiring me with your hard work and perseverance. Your love and kindness has been the warming light in my most difficult times, I love you.

This dissertation is dedicated to my family. To my mother, for I would not be here today without her sacrifices, and to my father, who taught me that being a decent human is much harder and more important than getting any academic degrees. Also, to my dearest sister who has paved the road for me all my life, and my brother, who is the delight of my life. And lastly to my husband, who is my life.
# Table of Contents

List of Figures .................................................................................................................. ix  
List of Tables ................................................................................................................... xi  
1. Introduction................................................................................................................ 1  
2. Optimal capacity allocation and flow control in healthcare systems with dual-priority patients .................................................................................................................. 5  
   Abstract ...................................................................................................................... 5  
   2.1. Introduction ....................................................................................................... 6  
   2.2. Background and literature review ..................................................................... 8  
      2.2.1. Capacity allocation .................................................................................... 8  
      2.2.2. Flow control ............................................................................................. 10  
      2.2.3. This paper ................................................................................................. 12  
   2.3. Problem description and performance evaluation ............................................ 13  
      2.3.1. Capacity allocation .................................................................................... 13  
      2.3.2. Flow control ............................................................................................. 14  
      2.3.3. Performance evaluation using queueing approximations .......................... 15  
   2.4. Mathematical model and Solution method ......................................................... 17  
      2.4.1. Mathematical model ................................................................................... 17  
      2.4.2. Proposed solution method .......................................................................... 18  
   2.5. Results and discussion ....................................................................................... 19  
      2.5.1. Optimal capacity allocation and flow control ........................................... 19  
      2.5.2. Sensitivity analysis ..................................................................................... 21  
   2.6. Extensions .......................................................................................................... 22  
      2.6.1. General arrival and service rates ................................................................. 23  
      2.6.2. Multi-priority patients .............................................................................. 23  
      2.6.3. Nonlinear combination of wait times and other objectives ........................ 24  
   2.7. Conclusions ........................................................................................................ 25  
3. Rethinking Patient Triage for Faster, More Cost-Effective Care in the Emergency Department ............................................................................................................. 27  
   Abstract ...................................................................................................................... 27  
   3.1. Introduction ....................................................................................................... 28  
   3.2. Background and Literature Review .................................................................... 28
Bibliography ........................................................................................................................................ 82
Appendices......................................................................................................................................... 94
  Appendix 2-A ..................................................................................................................................... 94
  Appendix 2-B ..................................................................................................................................... 98
  Appendix 3 ......................................................................................................................................... 100
  Appendix 4 ......................................................................................................................................... 101
List of Figures

1.1. Interrelationship among capacity and flow allocation decisions and system performance……………………………………………………………………………………………………… 3
2.1. The reciprocation among room configuration, waiting time, and flow allocation…… 11
2.2. Different room allocation policies……………………………………………………………………………………………………………………………………………………………………………….. 14
2.3. Schematic view of the queueing system…………………………………………………………………………………………………………………………………………………………………………………………… 15
2.4. Plot of the wait time objective as a function of arrivals to the flexible capacity…… 20
2.5. Wait time as a function of n_H and n_L for Example 1………………………………… 20
2.6. Optimal policy for λ_H = 0.16 and N = 20, as α and λ_L values change……………… 22
3.1. Breaking the trade-off between efficiency and quality in service operations…….. 29
3.2. ESI triage algorithm and typical flow allocation in EDs with a fast track………….. 32
3.3. Step-by-step methodology…………………………………………………………………………………………………………………………………………………………………………………………… 36
3.4. Partitioning tree for ESI3 patients presented with cough…………………………….. 38
3.5. Distribution of LOS for ESI3 patients with chief complaint cough…………………… 38
3.6. Schematic comparison of current and suggested patient routing…………………… 40
3.7. Distribution of Hospital X ED patients by ESI level…………………………………… 42
3.8. Distribution of number of ED diagnostics tests for patients in different ESI levels……………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………
4.3. Average number of arrivals per hour of the day in the example ED

4.4. Schematic view of the simulation logic

4.5. Scatter diagram of patient wait time based on care area

4.6. Average waiting time per patient category under different policies

4.7. Distance from the best performing policy for $1 < \alpha < 10$

4.8. Best-performing policies compared to the status quo

4.A-1. Distribution of waiting times in the fast track under dynamic policies

4.A-2. Distribution of waiting times in the main ED under dynamic policies

4.A-3. Distribution of average patients’ waiting time under static policies
List of Tables

2.1. Notations for the queuing model............................................................... 16
2.2. Problem parameters for Ex. 1................................................................. 19
3.1. Notations for the queuing model............................................................... 41
3.2. Waiting time approximations for each priority class................................. 42
3.3. Sample ESI31 subgroups found using partitioning..................................... 46
3.4. Input parameters estimated from Hospital X data..................................... 46
3.5. Mean patient wait time and system utilization under compared routing policies... 47
3.6. Sensitivity to system load: Mean overall patient wait times (minutes)............. 49
3.7. Input parameters under each LOS threshold value (T)............................... 50
3.8. Estimated waiting time (min.) under each LOS threshold value (T).............. 50
4.1. Service time distributions for discharged patients...................................... 64
4.2. Throughput, mean and 90th percentile of patient waiting by care area........... 69
4.3. Throughput and mean patient waiting by patient category........................... 69
4.4. Weighted waiting score based on care area, $W = \alpha W_{MED} + W_{FT}$............... 71
4.5. Average patient waiting score based on ESI category, $W = \sum_{i=1}^{5} \alpha_i W_i$........... 73
4.6. Patient wait time for SMLUtil-FT policy under different misclassification rate.... 74
3.A. Independent patient attributes available at triage used in the partitioning analysis... 100
1

Introduction

This dissertation is an effort to understand how complex service systems, such as those in healthcare, function and to identify managerial policies that produce optimal performance. In healthcare services, the goal is to provide timely and high-quality care (Horwitz and Bradley 2009). However, given the high costs of healthcare, resource shortages, and increasing demand, efficiency is also crucial (Graff et al. 2002, Lindsay et al. 2002). This makes healthcare operations an intriguing focus of service systems research. The industry poses complex problems from daily operations to strategic decisions and public health policy. Despite the expanding body of work in healthcare operations, there are many unresolved challenging problems in this area. High variability together with cost pressures make matching capacity and demand challenging for healthcare systems.

Despite the need for being maximally efficient, health centers need to be responsive to patients as well, especially if there are patients who need emergent (immediate) care. These systems try to improve responsiveness by properly allocating their existing capacity to the variable demand. Similar to many other problems, capacity and flow management in services has benefited from lessons learned in the manufacturing and supply chain literature. However, the key difference between services and manufacturing is the presence of customers’ input in the production process. The inherent variability in customer type, demand and service needs often disrupts operations and harms efficiency. Furthermore, because of the simultaneous production and consumption of the service product, it is almost impossible to fix a broken instance of
production. Thus, managing capacity and customer flow to ensure smooth operations is crucial, but complex, because there is an inherent tradeoff between responsiveness and efficiency.

In outpatient specialty clinics, when all patients are non-emergency, it is easier to match capacity and demand through conventional capacity planning and appointment scheduling practices. However, in environments like the emergency department or in operating rooms, the unpredictable arrival of emergency patients who need immediate care disrupts patient flow. Therefore, some health centers choose to dedicate part of their capacity to emergent cases to be more responsive to critically ill patients. With that dedication, the arrival of emergency patients will not disrupt the flow of other patients in the system. However, when resources are dedicated to different streams of patients, a patient might be waiting on a resource while there are other idle resources available (Karsten et al. 2015). In such cases, even “a little bit” of resource flexibility can make significant improvement in system responsiveness (Tsitsiklis and Xu 2012, Ferrand et al. 2014a).

In this dissertation, we closely examine healthcare systems that serve multi-priority patients. In multi-priority queues, the queueing rule is based on priority ranking and is typically “first come, first serve” within each priority class. Examples are operating rooms with both elective (low-priority) and emergency (high-priority) patients, and emergency departments in which patients are prioritized into multiple classes based on medical urgency and resource requirements. The focus is to match supply and demand to ensure smooth patient flow in the system. This requires careful capacity and flow decisions, which involves balancing the tradeoff between efficiency and responsiveness. Thus, we study two major operational decisions that arise when serving multi-priority streams of patients:
1- Resource allocation

- Should the resources be dedicated to each stream or pooled or a combination of these?

- If resources are not pooled, how many resources need to be dedicated to each class or shared among classes?

2- Flow allocation

- How should the patients be streamed into dedicated and shared resources?

These decisions are interrelated and have a direct effect on system performance measured by many different metrics, such as patient wait time, patient flow time, and throughput. The optimal patient flow allocation depends on how the resources are allocated (pooled, dedicated, or both and how many in each group). The optimal resource allocation depends on how the patients are streamed (Figure 1).

Flow Allocation  
\[ \rightarrow \]
Capacity Allocation  
\[ \rightarrow \]
System Performance

Figure 1.1. Interrelationship among capacity and flow allocation decisions and system performance.

We address both resource and flow allocation in the presence of multi-priority patients in queues with non-preemptive resume. We start from a two-priority system motivated by operating room systems with emergency and non-emergency patients (Chapter 2). Then we explore the more complicated case of multi-priority systems motivated by emergency departments with
multiple priority classes of patients (Chapters 3 and 4). For measuring the performance of different policies, we use patient waiting time functions as a proxy for system responsiveness. Using simulation and optimization methods, we identify capacity and flow allocation policies that minimize wait time, thus maximizing system responsiveness. We capitalize on statistical methods and data-mining techniques to help inform the operational and theoretical models, including those based on queueing theory, optimization, and simulation.
Optimal capacity allocation and flow control in healthcare systems with dual-priority patients

Abstract

There is a complex trade-off between efficiency and responsiveness in systems serving customers in different priority classes. Different resource allocation strategies have been adopted including full resource dedication (to specific classes) and full resource flexibility (where all resources shared). More recently, it has been shown that partial flexibility with some resources dedicated to a specific class and some shared among all classes can help balance this complicated trade-off. To identify the best resource allocation strategy, we introduce a two-stage optimization model involving flow control and M/M/n/non-preemptive queueing approximations. The model is motivated by healthcare systems with high- and low-priority patients, corresponding to emergency and non-emergency patients. The goal is to find how many dedicated and flexible (shared) resources, such as exam rooms or operating rooms, are needed and how the patients should be streamed into these resources so that patient waiting time is minimized. We show that the first-stage problem of finding the best flow allocation given any room configuration is jointly convex in flow variables. Thus, the optimal flow allocation can be found efficiently. The second stage problem of determining the allocation of resources is non-convex but can be solved using an iterative search on different room configurations. The computational results show that total flexibility is optimal for a wide range of problems. We identify conditions under which partial flexibility is optimal.
2.1. Introduction

High variability, together with cost pressures, make matching capacity and demand challenging for service systems. These systems try to improve responsiveness by properly allocating their existing capacity to the variable demand. Maximizing responsiveness is even more complicated when there are multiple priority classes of customers. Systems with multi-priority customers are common especially in the context of healthcare delivery. In addition to the emergency departments (ED), many healthcare delivery systems, such as operating rooms and diagnostic tests facilities, serve both emergency and non-emergency patients. These systems need to be efficient to reduce costs, but also to be responsive to high-acuity emergencies. Yet, there is a trade-off between efficiency and responsiveness (Frei 2006). Having more resources will result in better responsiveness, but less efficiency. On the other hand, as queueing theory demonstrates, when efficiency and utilization become high, responsiveness can degrade quickly. This tradeoff has been discussed in a wide range of operations management contexts from supply chain and logistics to call centers, theme parks and healthcare delivery.

Total resource flexibility (resource sharing) is often assumed to be the best approach in maximizing both responsiveness and efficiency in the presence of variability. However, there is evidence that resource sharing may not be the answer when demand is heterogeneous (Saghaﬁan et al. 2015). For instance, in systems with multi-priority patients, resource sharing may result in excess waiting times for lower priority patients (Ferrand et al. 2014). On the other hand, dedicating some of the available resources to different priority classes can increase responsiveness to them; but being more responsive to one class often necessitates being less responsive to others. Moreover, under total dedication, situations might occur in which customers are waiting while there is idle capacity in the system that they are not allowed to access because it is dedicated to another class of customers.
Inspired by Jordan and Graves (1995), a few studies suggest that partial flexibility which includes a combination of both dedicated and shared resources can break this efficiency-responsiveness tradeoff (Ferrand et al. 2012, Laker et al. 2015, Persson and Persson 2010). Yet, the question to be answered is: how can we find the optimal combination of shared and/or dedicated resources that maximizes responsiveness in systems with multiple priority classes of customers?

An interesting sub-problem that arises with partial flexibility is finding the optimal flow allocation. In other words, when there are both dedicated and flexible resources available to a class of patients, how should the arrivals be streamed to the available resources? The answer to this question (i.e., the flow allocation policy), directly affects the corresponding optimal capacity configuration. On the other hand, the optimal flow allocation decision depends on the number and proportion of dedicated and flexible resources. Therefore, there is an interrelationship among the flow control decisions and capacity allocation decisions that determines system performance.

In this paper, we find optimal capacity allocation and flow control decisions that maximize system responsiveness using total existing capacity. The remainder of this paper is organized as follows. Section 2.2 gives an overview of the related literature. In Section 2.3, we describe the problem and present a performance evaluation mechanism. In Section 2.4, we discuss the implications of the core problem of finding the best room allocation configuration and develop an optimization framework to find it. We discuss our numerical experiments, results, and sensitivity analysis in Section 2.5. In Section 2.6, we discuss the possible extensions. Finally, we draw conclusions in Section 2.7.
2.2. Background and literature review

2.2.1. Capacity allocation

2.2.1.1 Alternative policies

Flexibility has a long tradition in operations, especially in manufacturing and supply chain domains. Seminal studies, such as Van Meighem (1998), Jordan and Graves (1995), and Gupta et al. (1992), discuss how resource flexibility can improve responsiveness in the presence of uncertain demand and also highlight the importance of quantifying the costs and benefits of flexibility. In the context of service operations, Ward et al. (2014) explain the necessity and role of operational flexibility in emergency departments (ED). However, there is a trade-off between efficiency and responsiveness; a more responsive system may not be the most efficient. This trade-off has been the center of attention in many studies (Ferrand et al. 2014b, Akcay and Xu 2010, Randall et al. 2003).

In order to come up with a balance in this trade-off, Jordan and Graves (1995) developed a measure for comparing different levels of resource flexibility and showed that while complete flexibility maximizes responsiveness, most of the benefits can be achieved by partial flexibility at a much lower cost (i.e., more efficiently). A few studies have sought to apply the idea of partial flexibility to the healthcare context. One interesting application of resource flexibility in the healthcare is in serving patients with multiple priority classes. Ferrand et al. (2014) considered this problem in a hospital’s operating room (OR) environment involving a mix of emergency and elective surgeries. They studied allocation scenarios comparing different combinations of dedicated versus shared (or flexible) resources and draw useful conclusions. More particularly, using discrete event simulation, they showed that partial flexibility can outperform total dedication and total flexibility in the dual-priority system of operating rooms. Laker et al. (2014) is another good example of applying partial flexibility to healthcare. They
introduce the idea of “flex track” in the ED. The flex track is a class of flexible ED beds and other resources shared between the main ED and the minor care area. Again, using simulation modeling and analysis, the benefits of such partially flexible policies is illustrated. Yet, to the extent of our knowledge, no study has shown conditions under which partial flexibility can result in better responsiveness when compared with total flexibility.

2.2.1.2. Analytical methods to capacity allocation

Other than discrete-event simulation models, several analytical methods have been used in capacity planning and scheduling under uncertainty for ORs (Cardoen et al. 2010). Very few studies, including Zhang et al. (2008), Tancrez et al. (2009) and Persson and Persson (2010) consider partially flexible operating rooms. Zhang et al. (2008) use mixed integer programing (MIP) to allocate OR time to different surgery types. The results of the MIP are then subjected to uncertainty in a simulation model. To the best of our knowledge, no study has used a simulation-based optimization approach to this problem. The reason might be the cumbersome computational effort required. Tancrez et al. (2009) developed a continuous Markov chain for a system with only four operating rooms and calculated performance measures such as average wait times and disruption rate of emergencies to be used for surgery scheduling purposes. They provide useful managerial insights on how uncertainty in demand disrupts operations in that system but do not consider finding the optimal room configuration. Another limitation is the assumption that there are always rooms dedicated to low-priority patients. A Markov decision process also faces computational challenges when the number of operating rooms is large.

Another way to model the problem at hand is queueing methods. Hu and Benjaafar (2009) consider a multi-class, multi-server queuing system and compare resource dedication against sharing. They use a partitioning approach based on service times to partition the available
server capacity and dedicate each partition to a customer class. Although they do not consider multi-priority customers, the idea of using partitioning is an interesting approach to capacity allocation. Although there are many studies on calculating performance metrics in different queueing systems, the problem of flow and resource allocation in multi-priority systems has not yet been discussed broadly in the queueing literature.

Multi-class networks of queues have been analyzed in the literature (Cochran and Roche 2009, Harchol-Balter 2005, Shanthikumar and Yao 1992, Bertsimas et al. 1994). The term “multi-class customers” refers to groups of customers that have different characteristics. For instance, the arrival and service rate in one class might be different from another. On the other hand, the term “multi-priority customers” refer to customers that have different priorities to the system and may or may not have different arrival or service rates. Most of the literature around multi-priority systems assume preemptive resume; exceptions include Jouini et al. (2013).

Waiting lines with priorities have been studied since the 1950s (Miller 1960). Exact calculations for mean flow time and mean wait time for different priority classes are known for single-server problems. However, for problems with multiple servers, it is difficult to compute performance measures explicitly (Kao and Wilson 1999, Buzacott & Shanthikumar 1993, Gail et al. 1988). Therefore, results from single-server calculations are often used to approximate wait times in multi-server problems (Ulusçu and Altiok 2013, Shore 1988, Buzen & Bondi 1983, Bondi & Buzen 1984). In this paper, we use wait time approximations presented by Buzacott & Shanthikumar (1993) to evaluate the performance of each resource allocation policy.

2.2.2. Flow control

So far, we discussed the problem of capacity allocation, i.e., how to allocate resources to multiple classes of patients. Next, we discuss flow allocation, i.e., what proportion of each
customer class is sent to the dedicated and to the shared resources? As illustrated in Figure 2.1, capacity and flow allocation problems are inter-related (Ferrand et al. 2014). Therefore, we consider the flow allocation literature as well.

![Flow Allocation](Image)

**Figure 2.1.** The reciprocation among room configuration, flow allocation and system performance

To come up with reasonable flow allocation policies, many studies have used Markov chains, dynamic programming, fluid networks, prioritization, or load balancing methods (Liu and Whitt 2011, Harrison and Zeevi 2004, Chen and Yao 1991, Stidham 1985, Federgruen and Groenevelt 1988, Saghafian et al. 2012). Some studies, such as Schaak and Larson (1985), suggest threshold policies for flow management. In these studies, all resources are pooled, but there are separate queues for different classes of customers. The servers do not start serving lower priority classes unless there are no higher priority customers waiting and the queue length (or number of idle servers) has reached a threshold value.

The above flow allocation literature compares and contrasts different state-dependent flow allocation policies using discrete-event simulation or Markov chain models. Even there, the investigator has to assume some rule for sending patients to different segments of the capacity, e.g., cyclic, least busy, largest remaining capacity, etc. For instance, Akcay and Xu (2010) propose a dynamic assignment of jobs to flexible resources.
2.2.3. This paper

In this study, we model the system as a non-preemptive, multi-priority, multi-class queueing system with multiple servers. To find optimal capacity allocation, we develop a mathematical model to maximize responsiveness. As queueing theory indicates, patients would wait less in a more responsive system. Therefore, we use average waiting time of patients as our measure for system responsiveness. The queue at the dedicated resources for each priority class follows a first-come, first-served (FCFS) discipline. At the shared or flexible resources, the dispatching is based on patient priority class (M/M/n/NPP where NPP is non-preemptive priority). We use available closed-form queueing approximations to evaluate waiting time from Buzacott & Shanthikumar (1993). Alternatively, there are iterative approaches to approximate waiting times (Sleptchenko 2003), but here we choose to use closed-form approximations to achieve computational efficiency (see Section 2.3).

In the healthcare context, patient priority is usually based on medical urgency, thus the priority classes are set and known a priori, and the dispatching rule is non-preemptive. Since we approximate the system as a steady-state queueing system, flow allocation decision cannot be based on system status. Therefore, standard flow allocation methods used in the literature are not readily applicable to our problem. Therefore, we propose an optimization approach to find the best flow allocation policy based on capacity configuration (See Section 2.4). This paper contributes to the existing literature in the following ways:

- We develop an analytical tool that enables decision-makers not only to compare different capacity allocation policies, but also to find directions for improvement and the optimal solution that minimizes an weighted expected waiting time objective.
• We use closed-form queueing approximations to quantify the effects of capacity allocation decisions on patient wait time for different patient classes. This makes the computation effort less cumbersome, resulting in fast and convenient calculations.

• We prove that the expected waiting time approximation for non-preemptive dual priority queues is convex in flow control variables. This enables us to find the optimal flow allocation quickly for any capacity allocation. To the best of our knowledge, there has been no such convexity result for non-preemptive, multi-priority queueing systems.

• We show that, although total flexibility is the optimal solution to a wide range of problem instances, there are conditions under which partial flexibility outperforms total flexibility.

The results provide useful insights to the problem and a better understanding of the behavior of the system performance metrics, such as patients’ wait time. This information can be helpful both for practitioners and researchers investigating the advantages of partial flexibility. In practice, the results can help making capacity allocation decisions for current system parameters or for future changes and expansion plans. It is important to note that, although we model a healthcare environment the results are applicable to similar problems in other application areas with multi-priority customers and multiple identical resources.

2.3. Problem description and performance evaluation

2.3.1. Capacity allocation

We start from considering a system of operating rooms with two priority classes of patients, high-priority emergency cases, with index H, and low-priority cases, with index L. Later in Section 6, we discuss implications of extending the problem to multi-priority systems
such as emergency departments with more than two priority classes. In general, there are many possible alternative resource allocation policies, each resulting in different operational performance. Figure 2.2 shows a simple example of an OR system with five operating rooms.

Figure 2.2. Different room allocation policies: a) Total flexibility, b) Partial flexibility, dedicating to high priority, c) Partial flexibility, dedicating to low priority, d) Partial flexibility, e) Total dedication. [F] = Flexible; [H] = Dedicated to high-priority; [L] = Dedicated to low-priority.

The capacity allocation policy can differ from total flexibility at one extreme, with all rooms shared between high- and low-priority patients, to total dedication at the other extreme, with a certain number of rooms dedicated to each priority class. Between the two extremes are partially flexible resource allocation policies.

2.3.2. Flow control

As mentioned in the introduction, the optimal room configuration depends on the flow allocation policy. Figure 2.3 shows a more detailed flow diagram of patients through the ORs system.
Here we are interested in taking advantage of closed-form queueing approximations which hold only in steady state; thus, state-dependent flow allocation policies cannot be used. Instead, several allocation policies can be considered such as:

a. Optimum seeking: Allocate flow such that total wait time is minimized.

b. Service-level based: Allocated flow such that a desired service-level is achieved.

c. Room proportion-based: Allocate flow based on the ratio of number of station j rooms over total number of rooms.

d. Bound-based: Allocate flow based on some exogenous upper and lower bounds.

It is important to note that one can come up with many different heuristic allocation policies. We argue that although setting a policy like policies (b), (c) and (d) makes the problem easier, such policies requires fine tuning and might not be generalizable to all systems.

2.3.3. Performance evaluation using queueing approximations

We assume exponential processing times and inter arrival times (M/M/nj/FCFS for dedicated and M/M/ nj/NPP for flexible resources). Later, in Section 6, we show that all calculations can be adjusted to accommodate general processing times. The arrival rate of patient class i at station j is defined as the proportion γij of patients i going to station j multiplied by the arrival rate λi of the corresponding patient class. The arrival rate to the flexible capacity will be
the sum of the arrivals of high- and low-priority patients: \( \lambda_F = \lambda_{LF} + \lambda_{HF} \). Because we assume Poisson arrivals, the arrival process to the flexible station is also Poisson. Table 2.1 is a summary of queueing notation and waiting time formula.

Table 2.1. Notations for the queuing model

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>index for patient class, ( i = H, L )</td>
</tr>
<tr>
<td>J</td>
<td>index for stations, ( j = H, F, L )</td>
</tr>
<tr>
<td>N</td>
<td>total number of resources/rooms (known)</td>
</tr>
<tr>
<td>( \lambda_i )</td>
<td>arrival rate of class ( i ) patients to the system (known)</td>
</tr>
<tr>
<td>( t_i )</td>
<td>service time for patient class ( i )</td>
</tr>
<tr>
<td>( \gamma_{ij} )</td>
<td>percentage of ( i ) patients going to station ( j )</td>
</tr>
<tr>
<td>( \lambda_{ij} )</td>
<td>arrival rate of patient class ( i ) at station ( j ) (( \lambda_{ij} = \gamma_{ij} \lambda_i ))</td>
</tr>
<tr>
<td>( n_j )</td>
<td>number of resources at station ( j )</td>
</tr>
<tr>
<td>( \rho_j )</td>
<td>utilization rate for station ( j ) ( \rho_j = \frac{\sum_i t_i \lambda_{ij}}{n_j} )</td>
</tr>
<tr>
<td>( \rho_{ij} )</td>
<td>Traffic intensity of patient class ( i ) in station ( j ) ( \rho_{ij} = \frac{t_i \lambda_{ij}}{n_j} )</td>
</tr>
</tbody>
</table>

The waiting time approximation for the dedicated stations’ queue is

\[
E(W_{ij}) \approx \frac{t_i \rho_{ij}}{n_j} \left( \frac{1}{\rho_j} \right)^{\frac{1}{2(n_j+1)-1}}
\]  

(1)

The waiting time approximation for the flexible station queue is

\[
E(W_{HF}) \approx \frac{1 - \rho_F}{1 - \rho_{HF}} W_0
\]  

(2)

\[
E(W_{LF}) \approx \frac{1 - \rho_F}{(1 - \rho_{HF})(1 - \rho_{HF} - \rho_{LF})} W_0
\]  

(3)

where \( \rho_F = \frac{t_F \lambda_F}{n_F} \), \( t_F = \frac{\lambda_{HF} + \lambda_{LF}}{\lambda_F} \) and \( W_0 \approx \frac{t_F \rho_F^{\frac{1}{2(n_F+1)-1}}}{n_F} \).
2.4. Mathematical model and Solution method

As mentioned in Section 2.2, most of the related studies have used discrete-event simulation or Markov decision models to compare different policies while incorporating variability. Here, to solve the problem of resource allocation in systems with multi-priority customers, we develop an optimum-seeking approach based on closed-form queueing approximations presented in Section 2.4.1. Our approach provides insights while being applicable to a wide range of scenarios. This approach has several benefits. First, it is based on reliable and well-established theory. Second, while our example comes from the healthcare industry, this approach is useful for any queueing system with multi-priority customers.

Since queueing formulas are highly nonlinear, solving this problem is computationally hard. Therefore, in Section 2.4.2 we decompose this problem into a two-stage optimization model that can be solved more efficiently.

2.4.1. Mathematical model

Our goal is to find optimal room configuration (i.e., $n^*_H$ and $n^*_L$ values) as well as optimal flow allocation (i.e., $\lambda^*_ij$ values) that minimizes average patient waiting.

Minimize $W = \alpha W_H + W_L$ \hfill (4)

Subject to

$\rho_j < 1$ \quad for $j = L, F, H$ \hfill (5)

$\frac{\lambda_i l_i}{n_l + n_F} < 1$ \quad for $i = L, H$ \hfill (6)

$\sum_j n_j = N$ \quad for $j = L, F, H$ \hfill (7)

$n_j > 0 \ and \ integer \quad for \ j = L, F, H$ \hfill (8)

$0 \leq \gamma_{ij} \leq 1$ \quad for $i = H, L and j = L, F, H$ \hfill (9)

$\gamma_{iF} + \gamma_{ij} = 1$ \quad for $i, j = H, L$ \hfill (10)
The objective function (Equation 4) is a linear combination of wait times considering a higher weight, i.e., \( \alpha > 1 \), for higher-priority patients, where \( W_H = \lambda_{HF}W_{HF} + \lambda_{HH}W_{HH} \) and \( W_L = \lambda_{LF}W_{LF} + \lambda_{LL}W_{LL} \). Constraints Set 5 imposes upper and lower bounds on flow variables in each resource station. Other than the flow control constraints, we need to ensure that enough rooms are available for each patient class so that the system can handle the load of each class; otherwise, the problem is infeasible (Constraint Set 6). The total number of rooms is fixed and limited and the \( n_i \)'s are non-negative integers (Constraint Set 7 and 8). Constraint Set 9 keep the allocation proportion \( \gamma_{ij} \) is a positive value between 0 and 1. Finally all patients in each class have to use either the dedicated or the shared resources (Constraint Set 10). Note that the formulas and calculations provided in this section all assume dual-priority M/M/c/NPP queues. The extension to multi-priority M/G/c/NPP is presented in Section 6.

2.4.2. Proposed solution method

To achieve computational efficiency, we decompose the problem into two stages: 1) for each room configuration, find the proper flow allocation policy and calculate the total wait time; 2) find the optimal room configuration among all possible scenarios. For implementing this approach, the objective function in Equation 4 is replaced by the two stage objective function:

\[
\text{Min}_{n_H, n_L} \text{ Min}_{\lambda_{HF}, \lambda_{LF}} W = \alpha W(H) + \beta W(L))
\]

Based on the theorem below, the Stage 1 problem can be solved to optimality using conventional search algorithms such as the golden section method. The proof of this theorem is presented in the Appendix.

Theorem: Given any room configuration, \((n_H, n_L, n_F)\), the problem of finding the optimal flow allocation policy that minimizes average waiting times described in Equation 11, is jointly convex in flow allocation parameters \((\gamma_{HF}, \gamma_{LF})\). Thus, it can be efficiently solved to optimality.
The problem in Stage 2 is non-convex and hard to solve in general. However, the total number of rooms is fixed and only a relatively small number of room configurations are feasible. The traffic intensity of each class of patients poses constraints on the minimum number of rooms needed by each priority class, eliminating configurations that might lead to utilizations above 1. Thus, the Stage 2 problem can be solved by complete enumeration without considerable computational effort.

2.5. Results and discussion

2.5.1. Optimal capacity allocation and flow control

An illustrative instance (Example 1) was developed to reflect the practical environment that motivated this research with 20 ORs and two classes of patients, high- and low-priority. The problem parameters for Example 1 are shown in Table 2.2.

Table 2.2. Problem parameters for Example 1

<table>
<thead>
<tr>
<th>Parameter</th>
<th>High Priority</th>
<th>Low priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arrival rate (patients/min)</td>
<td>0.05</td>
<td>0.1</td>
</tr>
<tr>
<td>Service time (min)</td>
<td>Expo(120)</td>
<td>Expo(100)</td>
</tr>
<tr>
<td>Respective weight</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Total number of rooms</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

**Stage 1:** Due to the convexity property of the Stage 1 problem, the optimal flow allocation was identified efficiently. Figure 2.4 shows the case with $n_H=4$, $n_L=8$, and $n_F=8$ for illustrating this property. This figure suggests that patient wait time is not only jointly convex, but also smooth in the arrival parameters. Thus, even a good heuristic can perform reasonably well in finding an effective flow allocation ($\gamma_{HF}$, $\gamma_{LF}$) that results in a near-optimal total wait time for a given room configuration.

**Stage 2:** In the second stage, we solve the flow allocation problem for all possible combinations of rooms to find the optimal wait time, $W^*(n_H, n_L)$, for each room configuration,
(n_H, n_L). Figure 2.5 shows the waiting time response surface as a function of number of dedicated rooms for our problem instance. Any point on this surface represents (n_H, n_L, W*(n_H, n_L)) given the corresponding optimal Stage 1 flow values for the (n_H, n_L) configuration. The response surface is not smooth as was expected because of the highly nonlinear queueing formulae. For this problem instance, total flexibility is the optimal resource allocation.

Figure 2.4. Plot of the wait time objective as a function of arrivals to the flexible capacity in a system with n_H = 4, n_L = 8, and n_F = 8

Figure 2.5. Wait time as a function of n_H and n_L for Ex. 1
2.5.2. Sensitivity analysis

Using the optimization model described in Section 4, the optimal room configuration and flow allocation can be calculated for any system given the total number of rooms, and the arrival and service rates for the two patient classes. The optimal solution depends on these problem parameters which define the traffic intensity of each class of patients (i.e., $\rho_i = \frac{t_i\lambda_i}{N}$), as well as on the relative cost of patient waiting, $\alpha$. For Ex. 1 (the problem instance described in Section 2.5.1) and a wide range of its variants, total flexibility is the best allocation policy. However, it is interesting to investigate under what conditions the optimal policy may change.

In general, the queueing approximation formulations are complicated and quantitative analysis of the behavior of the solution surface is difficult. Therefore, we analyze the behavior of the waiting time objective in some special cases. We show that, partially flexible policies can outperform total flexibility when the traffic intensity of emergency patients is much bigger than that of non-emergency patients (Appendix B). In such conditions, it is optimal to dedicate some rooms to emergency patients and leave the rest shared between the two patient classes. In this condition, dedicating rooms to high-priority patients, reduces the level of interruption for the low-priority stream resulting in less waiting time for them.

To clarify, consider another problem instance, Ex. 2, which has the same total number of rooms and service times as Ex. 1, but the arrival rate of patients is changed to $\lambda_H = 0.16$ and $\lambda_L = 0.002$ patients per minute. Note that by changing the arrival rates we are in effect changing the traffic intensity of each patient class. For this problem instance the room configuration that minimizes total wait time is partial flexibility with $n_H^* = 19$, $n_L^* = 0$, so $n_F^* = 1$. The optimal flow allocation is $\lambda_{HH}^* = 0.97$, $\lambda_{HF}^* = 0.03$, $\lambda_{LL}^* = 0$, and $\lambda_{LF}^* = 1$. 
Other than traffic intensity, the weighting factor, \( \alpha \), can be a critical parameter. Figure 2.6 shows how the optimal policy changes in Ex. 2 for different \( \alpha \) and \( \lambda_L \) values when \( \lambda_H = 0.16 \) and \( N = 20 \) (restricting \( \lambda_L \) to be \( \leq 0.007 \)). This policy map can be generated for different combinations of patient arrivals.

![Figure 2.6. Optimal policy for \( \lambda_H = 0.16 \) and \( N = 20 \), as \( \alpha \) and \( \lambda_L \) values change.](image)

As discussed before and analytically shown in Appendix B, for cases similar to Example 2, dedicating rooms to emergency patients reduces wait time for low-priority patients. As shown in Figure 2.6, the optimality of the partial flexibility policy depends on the weighting factor \( \alpha \). At any combination of patient’s arrival rate, when \( \alpha \) increases, the improvement in low-priority waiting becomes less and less important to a point that it is no longer optimal to save low-priority patients from interruptions. As a result, total flexibility becomes optimal.

## 2.6. Extensions

For illustration purposes, in sections 3 to 5 we defined and solved the problem under assumptions such as Markovian arrival and service rates, two priority classes, and a linear
objective function. Here we discuss selected extensions and complications resulting from relaxing these assumptions.

2.6.1. General arrival and service rates

The queuing approximations shown in Section 3 can be easily generalized to G/G/c:NPP queues with multi-class patients. The approximations for the dedicated resources are trivial. Assume we have $l$ different patient classes then instead of the expected wait time formula shown in Equations 2 and 3; we can use the more generic approximation for expected wait time of patient class $l$ in a shared resource (Buzacott and Shanthikumar 1993):

$$E(W^l)_{G/G/c:NPP} \approx \frac{1 - \rho}{(1 - \rho)(1 - \rho_1 - \rho_{l-1})} \left( \frac{E(W)_{M/M/c:FCFS}}{E(W)_{M/M/1:FCFS}} \right) E(W)_{G/G/1:FCFS}$$ (12)

Using this approximation formula, handling general service rates is straightforward and the proposed solution method can be used for M/G/c:NPP queues with multi-class patients. However, in the case of general inter-arrival times, we can no longer take advantage of the properties of Markovian arrivals in patient streaming. For instance, there is no guarantee that the total arrival rate to the flexible capacity is no longer the sum of the high- and low-priority streams. As a result, in the case of general arrivals, the proposed method cannot be readily used and needs careful adjustments.

2.6.2. Multi-priority patients

A natural extension of this work is to generalize it to environments such as emergency departments with multi-priority patients. In this case, there are two possible approaches: a) since the dual-priority system shown in Section 4 is just a special case of multi-priority queues, we can extend the problem to multi-priority using the approximations explained earlier, or b) lump the
multiple classes of priority into two classes and use the work presented in this paper. Although approach (a) seems to be more theoretically sound, in practice, the convexity results that were proved for the case of dual-priority systems may not hold. Therefore, the solution method becomes computationally cumbersome. In fact, it may be more practical to utilize approach (b). For instance, in an emergency department we can consider patients with Emergency Severity Index (ESI) 4 and 5 as low priority patients who need minor care and ESI 2 and 3 patients as high priority patients. This is a very common assumption in practice because many emergency departments have “minor care” or “fast track” areas used to treat ESI 4 and 5 patients (Laker et al. 2014). In any environment where approach (b) is possible to set up (arrivals can be categorized as low- and high-priority), the methodology presented in our paper can be applied.

2.6.3. Nonlinear combination of wait times and other objectives

One of the most commonly used objective functions in multi-objective optimization problems is linear combinations in the form of a weighted sum of the individual objectives. The weights indicate the relative importance of one objective compared to others. In our problem, the weight can be cost of waiting or a utility function reflecting how patient’s health condition deteriorates overtime which are both hard to find. While linear combinations are convenient to work with, they may not reflect reality. The use of convex cost functions is common, for example Bipso (2013) used a convex cost function for finding optimal schedules in a single server queueing system. Here, as long as the objective function of interest is jointly convex in the arrival rates of high- and low-priority patients, the two-stage optimization model presented here can be solved efficiently.
2.7. Conclusions

Resource flexibility can help balance the tradeoff between efficiency and responsiveness in multi-priority service systems, such as operating rooms with elective and emergency surgeries. The question of how the operating rooms should be allocated, as dedicated or shared between the two priority classes, is difficult to answer. To answer this question, we propose a two-stage optimization approach to minimize the expected patient wait time using closed-form queuing approximations for multi-priority, multi-server systems with non-preemptive resume. The first stage involves addressing the flow allocation problem for a given room configuration and the second stage involves finding the best room configuration with the corresponding flow allocation. We show that the flow allocation problem is convex and can be solved to optimality. Once that problem is solved, the optimal room configuration can be found with complete enumeration. Our results show that although total flexibility is the optimal solution to a wide range of problems, partial flexibility can perform better under certain conditions. Thus, we closely investigate the sensitivity of the optimal solution to changes in problem parameters such as traffic intensity and relative cost of patient waiting. Using quantitative approach and numerical experiments, we show that when the traffic intensity of emergency patients is much higher than non-emergency patients, partial flexibility can outperform total flexibility.

The problem at hand is very complex. Therefore, for illustrative purposes, we made simplifying assumptions regarding patient arrival and service rate distributions and the number of patient classes. Relaxes these assumptions opens up an interesting line of investigation. For instance, an interesting extension of this research is to generalize it from dual-priority systems of queues to multi-priority systems. In the operating room environment, we assume only two classes of patients. However, in other environments, such as the emergency departments, there
are potentially multiple priority classes of patients/customers. For such environments, our approach may be extended to accommodate queueing networks with >2 priority classes.
Rethinking Patient Triage for Faster, More Cost-Effective Care in the Emergency Department

Abstract

Emergency Departments (ED) are under increasing pressure to be more efficient and to provide higher-quality care. One metric of care quality is patient satisfaction, and the timeliness of care is important both for optimal medical outcomes in the ED as well as maximizing patient satisfaction. Balancing cost-effectiveness and quality is especially challenging in the face of highly variable demand. In EDs triage algorithms direct patient routing, which affects capacity utilization and patient waiting. In EDs where a fast track has been implemented to ensure that lower-acuity patients do not face excessively long waiting times, decisions regarding which patients get routed to the fast track can exacerbate utilization and patient waiting problems. To address this problem, we propose an enhanced triage method. By examining an ED’s patient population, partitioning may be used to identify subgroups of moderate-acuity patients who may be routed to the lower-acuity fast track instead of the higher-acuity main ED. Using data from a large, urban teaching hospital, we first employ a partitioning method to identify these patient subgroups. Then, by developing queueing models, we are able to estimate our proposed triage policy’s effects on ED operational performance. We find that redirecting even a sixth of the moderate-acuity, medically appropriate patients from the main ED to the fast track can reduce patient waiting substantially for all higher- and moderate-acuity patients presenting to the ED. We examine the policy’s sensitivity to other scenarios and draw conclusions for both research and practice.
3.1. Introduction

EDs are under increasing pressure to be more efficient and to provide higher-quality care. One metric of care quality is patient satisfaction, and the timeliness of care is important both for optimal medical outcomes as well as maximizing patient satisfaction. Balancing cost-effectiveness and quality is especially challenging in the face of highly variable demand. In EDs, triage algorithms direct patient routing, which affects capacity utilization and patient waiting. In EDs where a fast track has been implemented to ensure that lower-acuity patients do not face excessively long waiting times, decisions regarding which patients get routed to the fast track can exacerbate utilization and patient waiting problems. To address this problem, we propose an enhanced triage method. By examining an ED’s patient population, partitioning may be used to identify subgroups of moderate-acuity patients who may be routed to the lower-acuity fast track instead of the higher-acuity main ED. Using data from a large, urban teaching hospital, we employ a partitioning method to identify these patient subgroups. Then, by developing queueing models, we are able to estimate our proposed policy’s effects on ED operational performance.

The remainder of this paper is as follows. In Section 2, we briefly introduce the related literature and how our work contributes to it. In Section 3, we present details on the data used and the queueing formulations. Section 4 includes numerical results. We assess the sensitivity of the results in Section 5. Finally, we draw conclusions and discuss limitations, extensions, and managerial implications in Section 6.

3.2. Background and Literature Review

3.2.1. Demand variability in services

Capacity and flow management in the emergency department (ED) is complicated primarily due to high variability in demand (e.g., different diseases, demographics, medical histories, and medical urgencies, among others). This inherent variability often frustrates the
achievement of the ultimate goal of the system, which is to provide high-quality care in a timely, cost-effective manner. One of the primary challenges in managing services in general, not just healthcare, is dealing with demand variability and uncertainty due to the trade-off between service quality and service efficiency. Being more responsive often requires having slack resources, which can be costly. Figure 3.1 shows a framework introduced by Frei (2006) for strategies in meeting highly variable demand in services.

Figure 3.1. Breaking the trade-off between efficiency and quality in service operations.  
*Adapted from Frei (2006)

The two extreme approaches are “classic reduction” (of variability) and “classic accommodation” (of variability). In classic reduction, the strategy is to reduce variability in demand as much as possible. An example of classic reduction is the use of appointment systems, which reduces the variability in arrival time of patients at the expense of being less accommodating to walk-in patients. In classic accommodation, high resource investments are made to ensure that all customers with variable needs are served in a timely manner. Traditionally, hospitals are in the zone of classic accommodation, which achieves high responsiveness and high quality of service, but at a high cost. Combining responsiveness with low cost is ideal, but achieving that has proven to be difficult.
Two strategies are suggested for low-cost-yet-responsive service (Frie 2006). One is low-cost accommodation, such as using self-service kiosks or email consultations. The second is uncompromised reduction, which is similar to what Bohmer (2005) referred to as separate-and-accommodate. In this approach, customers are separated into smaller, but more similar, groups and their specific group’s needs are accommodated through routinization. It is similar in concept to mass personalization (Kumar 2008), but relies on high-quality decision-making to identify commonalities within subsets of customers. Although separating demand and segmenting capacity reflect an anti-pooling strategy, this approach can be more efficient because within-group variability is less than overall variability.

3.2.2. The pros and cons of capacity pooling

Queueing theory makes it clear that pooling of homogeneous servers offers advantages in terms of system performance (Jack & Powers 2004). However, there are several studies in the literature that suggest queue-pooling may not be the best strategy (Song et al. 2015, Mandelbaum and Reiman 1998, Gilbert and Weng 1998, and Van Dijk and Van Der Sluis 2008, 2009). Under certain conditions, such as when the customers or jobs are heterogeneous or when there are incentives involved, it can be more operationally efficient to maintain separate queues.

For example, in services with multi-priority customers, such as the Emergency Department, pooling all resources can result in excessive wait times for low-priority customers. In such situations, dedicating some of the resources to low-priority customers seems to be effective at reducing their wait time. These customers often have a shorter service time as well. Studies such as Sanchez et al. (2006) show that implementation of a fast track can reduce wait time for low-acuity patients significantly, but would not eliminate the congestion in main ED. In fact, this improvement is achieved at the cost of taking away capacity from higher priority
customers, modestly increasing their waits. To solve this problem, Laker et al. (2014) proposed capacity sharing between the fast track and main ED using a “flex track.” They used discrete-event simulation to show that this partially flexible capacity-allocation policy improves patient flow in the ED. Thus, in the context of healthcare, where the cost of waiting may not be equal across different classes of patients, it makes sense to use a priority assignment and segment the scarce resources accordingly.

3.2.3. Prioritization and triage systems

Finding the optimal priority classification of customers who use a shared resource has been studied in many contexts (e.g., airport runways) in studies at least as far back as Oliver and Pestalozzi (1965). In most studies, the priority classification in the traffic systems is based on arrival time, service time, or a combination of those. Examples are the “$c\mu$ rule” and the “shortest-average-processing-time rule” (Xia et al. 2000).

In healthcare, the priority classification is based neither on patient arrival nor on patient service time, but rather primarily on medical urgency. Although medical urgency must be the primary concern, patient priority classification and streaming should be done in order to improve, if possible, overall performance of the system in terms of system throughput and patient wait time (Pekoz 2002).

Upon arrival to the ED, patients are classified into different priority classes through a triage process. One of the common triage systems used in the United States is the Emergency Severity Index (ESI), initially developed in 1999. It stratifies patients into five priority levels based on their acuity and resource needs (Gilboy et al. 2011). The ESI algorithm assigns the highest priority, level 1, to patients who need immediate, life-saving interventions. ESI level 2 is
assigned to patients who are either in high-risk situations or in severe pain/distress. Otherwise, the categorization is done based on resource requirements (see Figure 3.2).

Given the dual pressures on EDs to reduce costs while improving service quality, many have turned to partitioning their capacity (both beds and human resources) into two distinct areas: the main ED, and the “fast track” (a.k.a. “minor care”). The fast track, usually staffed with lower-cost providers, is a group of resources dedicated to lower-acuity patients. These patients often also have shorter service times, spending only a fraction as long in the ED as higher acuity patients do. This is an instance of Frei’s *uncompromised reduction*, or Bohmer’s *separate-and-accommodate* strategy, in which customers are grouped such that there is less variability within each group, resulting in more efficient flow management.

![Figure 3.2. ESI triage algorithm and typical flow allocation in EDs with a fast track](image)

As shown in Figure 3.2, in EDs with a fast track, patients categorized as ESI levels 1, 2, and 3 are routed to the main ED beds, where they receive more targeted service, sometimes by
more-skilled care providers. ESI level 4 and 5 patients, who need lower-intensity care and fewer resources, are directed to the fast track, which is often staffed with lower-cost providers like residents, nurse practitioners, and physician assistants. The use of a fast track has been shown to be effective at reducing wait times for low-acuity patients and improving patient satisfaction (Garcia et al. 1995). Although dedicating part of the capacity to low-acuity patients can lead to more cost-effective use of human resources and may reduce wait time for ESI level 4 and 5 patients, it can make the ED less responsive to higher-acuity patients by shifting capacity away from the main ED (assuming a fixed budget and/or space for the entire emergency department), illustrating the downside of capacity segmentation.

However, observations of large, urban teaching hospitals’ EDs tends to reveal excessive wait times in the main ED area while fast track areas go underutilized, with most hospitals that use them shutting them down completely several hours each day. This observation implies that, despite the effort to improve responsiveness by segmenting capacity and demand, there is still a mismatch between supply and demand in both areas of the ED.

This suboptimal flow and capacity balance may be at least partly due to the triage algorithm used for patient classification. Widely used ED triage systems do not attempt to consider operational consequences of triage decisions in the ED. Therefore, relying on current triage frameworks can contribute to suboptimal capacity management and patient routing in the ED, potentially resulting in more congested emergency departments. Allocation of physical capacity, such as exam rooms, is often difficult to change after it is implemented. So, given the physical capacity allocated to each area, policies regarding demand segmentation and patient flow allocation merit additional study.
Saghafian et al. (2012) show that streaming of the ESI2 and ESI3 patients based on their expected disposition (‘discharge’ versus ‘admit’) can improve responsiveness of the ED significantly. Upon arrival, the triage nurse has to predict whether the patient is going to be admitted or discharged and classifies the patient into one of two groups with dedicated resources in the main ED area. Saghafian et al. (2015), propose an alternative triage system that takes into account complexity as well as urgency. In that work, the level of complexity is set by the estimated number of interactions required between the patient and care givers. If a patient needs more than one interaction, his case is considered “complex” (otherwise, “simple”), and that approach yields modest performance improvement.

There still seems to be opportunity for improvement, however. Although the ESI triage system has an operational component to it (i.e., consideration of care resource requirements), it does not consider the mix of the patient population. Therefore, relying only on the ESI triage algorithm is likely to result in sub-optimal patient streaming if different areas of the ED are not identically utilized.

3.2.4. Proposed new approach

Based on the above observations and existing literature, we propose that the ED triage/routing policy itself can be modified to both reduce wait times for higher-acuity patients who need to be treated in the main ED and help address underutilization in the fast track. An analysis of historical data on ED visits shows that approximately half of the patients at a large, urban academic ED are categorized as ESI level 3. Given the problems of ED crowding, we ask this research question: *Would redirecting specific subgroups of ESI level 3 patients to the fast track improve patient flow in the ED?*
To answer this question, we propose a new routing policy that further stratifies $ESI3$ patients based on their expected length of stay (LOS), which is a common metric in ED operations (Welch et al 2011). Under this new policy, subgroups of $ESI3$ patients whose LOS is expected to be less than a threshold value (e.g., 120 minutes) may be routed to the fast track instead of to a bed in the main ED. The proposed subgroups are identified using partitioning methods based on demographics, vital signs, and other patient attributes evident at the time of triage. Those $ESI3$ patients who can be directed to the fast track are denoted as $ESI3^f$ patients, whereas $ESI3$ patients who need to receive care in the main ED are denoted as $ESI3^h$ patients.

To assess the performance of this new policy, we use queueing theory for multi-server, multi-priority systems to model patient flow in the ED. The use of queueing theory to model ED operations is common (Cochran and Roche 2009, Green et al. 2006, Panayiotopoulos and Vassilacopoulos 1984). We use insights and results from studies such as Davis (1965), Iravani and Balcioglu (2008), Pekoz (2002) and Buzacott and Shanthikumar (1993) to formulate patient flow time and wait time under our proposed routing policy and compare it to current practice.

3.3. Methods

Streaming these $ESI3$ subgroups to the fast track will affect arrival and service rates in both areas of the ED, which are estimated first using partitioning methods on an actual patient data set obtained from a large, urban, academic medical center’s ED. We then develop queueing models to compare the performance of the suggested policy to the ED’s current approach in practice, using patient wait time and system utilization as performance metrics. We examine the sensitivity of the results to system load and the LOS threshold value used for classifying the $ESI3$ patients. Figure 3.3 summarizes the steps in the methodology we use to examine this problem.
3.3.1. Recursive partitioning

First, we identify $ESI^3_l$ and $ESI^3_h$ subgroups in order to estimate their service rates and arrival rates for the queueing analysis discussed in the next section. We base these subgroups on information available at the time of triage because, in practice, this tends to be when the patient-routing decision is made. The triage decision is made upon patient arrival by a triage nurse based on the patient’s chief complaint, vital signs, demographics, and a brief medical history.

To identify the subgroups, we use Recursive Partitioning Analysis (RPA), a multivariable statistical method that seeks to correctly classify a data set into partitions based on independent variables or factors. The partitioning is done such that the within-partition distance is minimized. This method is widely used in medical decision-making (Zhang and Singer 1999, Scott et al. 2012) and as a data-mining technique because it is suitable for exploring relationships when a good a priori model is not available. It can also handle large problems effectively and provide easily interpretable results (Hawkins 2009).

The independent variables or factors can be either continuous or categorical. If a factor is continuous, the partitions are created by identifying a cutting value; the sample is then divided into values above and below this cutting value. If the factor is categorical, the sample is divided into two groups of levels. The response variable can also be either continuous or categorical (nominal or ordinal). If the response variable is continuous, then the platform fits means. If it is categorical, then the fitted value is a probability. In either case, the split is chosen to maximize

Figure 3.3. Step-by-step methodology.

1. Identify $ESI^3_l$ & $ESI^3_h$ patients using recursive partitioning on past data
2. Develop queueing models
3. Estimate arrival & service rates using past data and compare routing policies
the difference in the response variables between the two branches of the split. Node splitting was based on the log-worth statistic, or $-\log_{10}(p\text{-value})$. The adjusted p-value accounts for the number of different ways splits can occur. This adjusted p-value is fairer than either the unadjusted p-value, which favors variables with many levels, or the Bonferroni p-value, which favors variables with small numbers of levels.

In this study, the response variable of interest is mean patients’ length-of-stay (LOS), with LOS defined as the time duration starting when a patient is first seen by a physician and ending at the patient’s departure from the ED. Independent variables are shown in the Appendix. Since patients’ chief complaints are a major source of the variability in LOS, we first classify patients according to their chief complaint and then partition them into subgroups using the other variables (demographics, vital signs, and list of diagnostic tests needed). Starting from the entire initial population of ESI3 patients presenting with each chief complaint, we continue splitting until we find one or more partitions having a mean LOS less than or equal to a set threshold.

Figure 3.4 shows an example partition tree for ESI3 patients presenting with “cough” as the chief complaint. The partitioning results show that patients who need neither labs nor respiratory tests have a mean LOS of 114.4 minutes. With a 120-minute threshold value, as long as this group of patients can receive appropriate care in the fast track, they can be directed there for care instead of to a bed in the main ED. Consistent with prior observations (White et al. 2011), these subgroups tend to have both lower mean lengths-of-stay as well as lower variability in their LOS and a lower coefficient of variation (Figure 3.5).
Figure 3.4. Partitioning tree for ESI3 patients presented with cough. Source: JMP v12 (SAS)

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>LogWorth Difference</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Rows</td>
<td>722</td>
<td></td>
<td>200.45983</td>
<td>186.17468</td>
</tr>
<tr>
<td>Lab&lt;1</td>
<td>336</td>
<td></td>
<td>129.29762</td>
<td>77.664611</td>
</tr>
<tr>
<td>Lab&gt;1</td>
<td>386</td>
<td></td>
<td>262.40415</td>
<td>226.70591</td>
</tr>
<tr>
<td>Resp&lt;1</td>
<td>187</td>
<td></td>
<td>114.37968</td>
<td>56.303328</td>
</tr>
<tr>
<td>Resp&gt;1</td>
<td>149</td>
<td></td>
<td>148.02013</td>
<td>95.052366</td>
</tr>
</tbody>
</table>

Figure 3.5. Distribution of LOS for ESI3 patients with chief complaint cough.
The partitioning process is repeated for each chief complaint. Once all $ESI^3$-eligible partitions are identified, they are grouped together to form the overall $ESI^3$ subgroup; the remainder of the $ESI^3$ patients would form the $ESI^3$ subgroup. Once the subgroups are identified by the partitioning process, they are presented to medical professionals to ensure that routing these patients to the fast track is medically appropriate for them. All partitioning is performed using the JMP 12 analytics suite (SAS Inc).

3.3.2. Queueing models

In order to develop queueing representations of patient flows in the ED, we must first understand the care delivery process. Patients can arrive to the ED by air care, ambulance, or as walk-ins. Upon arrival, trauma patients (those needing immediate, life-saving care; often $ESI$ level 1) will be directly taken to a resuscitation unit. Other patients are triaged and wait in the waiting area. At this point, if an appropriate (with respect to the patient’s needs) bed is available in the ED, the patient is directed to that bed; otherwise, patients will be taken to a triage bed for a second, more detailed triage process. There, the triage nurse collects more information from the patient, checks his vitals, and may initiate laboratory or radiology requests. After the second triage, the patient waits in the waiting area until a bed is available for him. Once the patient is placed in a bed, the first ED provider (physician or mid-level provider, such as a nurse practitioner or physician’s assistant) will meet with and assess the patient. That care provider may order labs, tests, consults, and treatments. Additional care may be provided before and/or after those outside services. Once care is complete, the provider determines the patient’s disposition; he is either discharged home or admitted to an inpatient hospital ward. A subset of patients are placed in an observation unit for up to 24 hours for additional evaluation and treatment before being discharged. As mentioned, the length-of-stay metric begins when the first
ED provider starts to care for the patient and ends when the patient leaves the ED (i.e., is discharged, admitted, or sent to the observation unit).

The main ED and the fast track work as independent, priority-based queues with non-preemptive resume (NPP) and can be modeled as G/G/n/NPP. Under the current patient routing policy (Fig. 3.6a), ESI1 patients are sent to the resuscitation unit (separate from the main ED and not included in our analysis), ESI2 and ESI3 patients are directed to the main ED, and ESI4 and ESI5 patients are directed to the fast track.

In our proposed patient routing policy (Fig. 3.6b), ESI1 go to the resuscitation unit (again, not included in the analysis), ESI2 patients are directed to the main ED, and ESI4 and ESI5 are directed to the fast track. ESI3 patients are partitioned into two groups: ESI3^l and ESI3^h. All ESI3^h patients are sent to the main ED. ESI3^l patients, who are expected to spend shorter amounts of the time in the ED and may be cared for using fast track personnel and resources, will be sent to the fast track. We assume that sending qualified patients to the fast track will neither compromise nor enhance care quality and patient outcomes. We also assume that the primary streaming decision is final and patients will not be redirected from the fast track to main ED or vice versa after they are initially assigned to one of those routes.

Figure 3.6. Schematic comparison of the queueing inputs of current and suggested patient routing
In Figure 3.6a, the main ED can be considered a dual-priority system with $ESI2$ patients as high-priority and $ESI3$ patients as low-priority. In the fast track, there are two priority classes: high ($ESI4$ patients) and low ($ESI5$ patients). In Figure 3.6b, our proposed system, the main ED is a dual-priority system with $ESI2$ patients as high-priority and $ESI3^h$ patients as low-priority. In the fast track, however, there are now three priority classes: $ESI4$, $ESI5$, and $ESI3^l$.

Table 3.1. Notations for the queuing model

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i$</td>
<td>index for patient class, $i = H, L$</td>
</tr>
<tr>
<td>$j$</td>
<td>index for stations, $j = M, FT$</td>
</tr>
<tr>
<td>$N$</td>
<td>total number of rooms (known)</td>
</tr>
<tr>
<td>$\lambda_i$</td>
<td>arrival rate of class $i$ patients to the system (known), $i = H, L$</td>
</tr>
<tr>
<td>$n_j$</td>
<td>number of resources at station $j$ (known), $j = M, FT$</td>
</tr>
<tr>
<td>$\rho_j$</td>
<td>utilization rate for station $j$, $j = M, FT$</td>
</tr>
<tr>
<td>$\rho_{ij}$</td>
<td>Traffic intensity of patient class $i$ in station $j$</td>
</tr>
<tr>
<td>$W_{ij}$</td>
<td>wait time for patient class $i$ in station $j$ queue, $i = H, L$, $j = M, FT$</td>
</tr>
<tr>
<td>$t_i$</td>
<td>service time for patient class $i$ (known), $i = H, L$</td>
</tr>
<tr>
<td>$P_j$</td>
<td>Set of patient classes that are routed to station $j$</td>
</tr>
</tbody>
</table>

Using the Buzen and Bondi (1983) approximation, wait time of patients in G/G/n queues with two and three priority classes can be calculated conveniently (Whitt 1993). Let $\rho_{ij} = \frac{\lambda_i t_i}{n_j}$ be the traffic intensity caused on station $j$ by patient class $i$; $t_{e_j} = \frac{\sum_{i \in P_j} \lambda_i t_i}{\sum_i \lambda_i}$, the effective service time at station $j$; and $W_{0j}$ the waiting time approximation in G/G/n$_j$/FCFS in station $j$. Patient waiting times can then be calculated using the approximation formulae in Table 3.2.
Table 3.2. Waiting time approximations for each priority class.

<table>
<thead>
<tr>
<th>Patient Class</th>
<th>Waiting time approximation</th>
<th>Status quo</th>
<th>Proposed routing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status quo</td>
<td></td>
<td>$\frac{1 - \rho_M}{1 - \rho_2} W_{0,M}$</td>
<td>$\frac{1 - \rho_M}{1 - \rho_2} W_{0,M}$</td>
</tr>
<tr>
<td>$ESI_2$</td>
<td></td>
<td>$\frac{1 - \rho_M}{(1 - \rho_2)(1 - \rho_2 - \rho_3)} W_{0,M}$</td>
<td>$\frac{1 - \rho_M}{(1 - \rho_2)(1 - \rho_2 - \rho_3 L)} W_{0,M}$</td>
</tr>
<tr>
<td>$ESI_{3h}$</td>
<td></td>
<td>$\frac{1 - \rho_M}{(1 - \rho_2)(1 - \rho_2 - \rho_3)} W_{0,M}$</td>
<td>$\frac{1 - \rho_M}{(1 - \rho_2)(1 - \rho_2 - \rho_3 H)} W_{0,M}$</td>
</tr>
<tr>
<td>$ESI_{3l}$</td>
<td></td>
<td>$\frac{1 - \rho_M}{(1 - \rho_3)} W_{0,FT}$</td>
<td>$\frac{1 - \rho_M}{(1 - \rho_3)} W_{0,FT}$</td>
</tr>
<tr>
<td>$ESI_{4}$</td>
<td></td>
<td>$\frac{1 - \rho_M}{(1 - \rho_4)} W_{0,FT}$</td>
<td>$\frac{1 - \rho_M}{(1 - \rho_4)(1 - \rho_3 L - \rho_4)} W_{0,FT}$</td>
</tr>
<tr>
<td>$ESI_{5}$</td>
<td></td>
<td>$\frac{1 - \rho_M}{(1 - \rho_5)(1 - \rho_4 - \rho_5)} W_{0,FT}$</td>
<td>$\frac{1 - \rho_M}{(1 - \rho_5)(1 - \rho_3 L - \rho_4 - \rho_5)} W_{0,FT}$</td>
</tr>
</tbody>
</table>

3.3.3. Empirical dataset and analysis

To empirically test our analytical results and proposed patient routing policy, we provide a numerical analysis using 115,194 ED patient visit records collected over a period of 18 months from an emergency department at a large, urban, academic medical center (Hospital X). Of these patients, 48% were classified as ESI level 3. The example emergency department has 22 rooms.

![Figure 3.7. Distribution of Hospital X ED patients by ESI level.](image)
in the main ED dedicated to *ESI*2 and *ESI*3 patients. There are 8 additional rooms in the fast track dedicated to *ESI*4 and *ESI*5 patients. There are also three trauma rooms available for critical care of *ESI*1 patients. Figure 3.7 shows the distribution of patients by acuity class. Figures 3.8 and 3.9 show the distribution of the number of diagnostics tests for patients and their LOS by ESI category. It can be observed from these distributions that there are many ESI3 patients who have shorter LOS and lower resource requirements. Thus, it may be more appropriate operationally to direct them to the fast track instead of the main ED area. The results of applying the partitioning and queueing analyses on this dataset are presented in Section 3.4.

Figure 3.8. Distribution of number of ED diagnostics tests (electrocardiogram or EKG, Labs, Respiratory) done for patients in different ESI levels
Figure 3.9. Distribution of doc-to-depart time for all patients in different ESI levels

3.4. Results

As explained in Section 3.3.1, recursive partitioning was used to identify $ESI^{3_l}$ patients. Figure 3.10 shows how the data were disaggregated to arrive at the final $ESI^{3_l}$ and $ESI^{3_h}$ subgroups. Out of the 115,194 ED visits, 50,862 were classified as $ESI^{3}$ patients. Of those, we only consider the discharged patients because admitted patients are typically not able to be treated in a fast track environment. So, discharged $ESI^{3}$ patients are far more likely to qualify (both medically and operationally) as $ESI^{3_l}$. We then identified the 50 most frequent chief complaints, each of which had at least 150 cases, or 100 cases per year on average, for further partitioning (for other chief complaints, small N’s made it impossible to create useful partitions).
Figure 3.10. Breakdown of ED visits at Hospital X.

For exploratory purposes, consultation with the ED director at Hospital X suggested that starting with a 120-minute threshold value for Doc-to-depart time would be appropriate. Because of the degree of variation and asymmetry in the LOS distribution, considering mean LOS alone can be misleading. Therefore, we used two criteria for our recursive partitioning: $ESI_{3}^{l}$ patients would be those who had both a mean LOS $\leq 120$ minutes and a $90^{th}$ percentile LOS $\leq 180$ minutes. Several $ESI_{3}^{l}$ partitions are presented in Table 3.3. In sensitivity analyses, we also considered threshold levels of 150 and 180 minutes for mean LOS (see Section 3.5).
Table 3.3. Sample $ESI^3$ subgroups found using partitioning

<table>
<thead>
<tr>
<th>Chief complaint</th>
<th>Description of patient characteristics</th>
<th>Patient LOS (min.)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Dental Pain</td>
<td>TEMP&lt;98.8 &amp; SpO2&gt;=98</td>
<td>57.2</td>
<td>54.7</td>
</tr>
<tr>
<td>Rash</td>
<td>BP2&gt;=65 &amp; PULSE&lt;74</td>
<td>62.0</td>
<td>51.0</td>
</tr>
<tr>
<td>Medication refill</td>
<td>PULSE&lt;118 &amp; AGE&lt;60</td>
<td>88.8</td>
<td>68.7</td>
</tr>
<tr>
<td>Laceration</td>
<td>arrival means(Self) &amp; RESP&lt;22 &amp; AGE&lt;32 &amp; 96&lt;=SpO2&lt;100</td>
<td>77.1</td>
<td>40.2</td>
</tr>
<tr>
<td>Knee Pain</td>
<td>PULSE&lt;107 &amp; 98&lt;=SpO2&lt;99 &amp; RESP&lt;22 &amp; 75 &lt;= BP2&lt;102 &amp; Gender(M)</td>
<td>64.0</td>
<td>30.4</td>
</tr>
<tr>
<td>Sore throat</td>
<td>BP2&gt;=72 &amp; RESP&lt;21 &amp; PULSE&lt;120 &amp; Gender(F) &amp; BP1&lt;149 &amp; SpO2&lt;98</td>
<td>79.5</td>
<td>48.0</td>
</tr>
<tr>
<td>Shoulder Pain</td>
<td>SpO2&gt;=95 &amp; arrival-means(Self) &amp; BP2&lt;83</td>
<td>98.5</td>
<td>76.8</td>
</tr>
<tr>
<td>Shortness of Breath</td>
<td>AGE&lt;26 &amp; PULSE&lt;86 &amp; BP1&lt;131</td>
<td>111.8</td>
<td>58.9</td>
</tr>
</tbody>
</table>

After identifying $ESI^3$ patients, who comprise ~10% of $ESI3$ discharged patients, we estimate the LOS (service time) and arrival rate for each priority class. Using the queueing approximation formula presented in Table 3, waiting time approximations are calculated. Table 3.4 shows the empirical model parameters estimated from the retrospective data from Hospital X. Table 3.5 summarizes the queueing calculation results for this problem using parameters representative of Hospital X’s physical configuration (22 main ED beds and 8 fast track beds).

Table 3.4. Input parameters estimated from Hospital X data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Statistic</th>
<th>$ESI^2$</th>
<th>All</th>
<th>$ESI^3$</th>
<th>$ESI^3$</th>
<th>$ESI^4$</th>
<th>$ESI^5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$ESI^3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>mean</td>
<td>348.9</td>
<td>242.2</td>
<td>120.2</td>
<td>266.7</td>
<td>118.8</td>
</tr>
<tr>
<td>Service time</td>
<td>stdev</td>
<td>244.5</td>
<td>190.3</td>
<td>81.1</td>
<td>196.4</td>
<td>138.1</td>
<td>81.7</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>0.70</td>
<td>0.79</td>
<td>0.67</td>
<td>0.74</td>
<td>1.16</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>76.2</td>
<td>30.5</td>
<td>179.1</td>
<td>35.6</td>
<td>49.7</td>
<td>222.8</td>
</tr>
<tr>
<td>Interarrival time</td>
<td>stdev</td>
<td>99.6</td>
<td>41.3</td>
<td>220.2</td>
<td>49.6</td>
<td>72.8</td>
<td>306.8</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>1.31</td>
<td>1.35</td>
<td>1.23</td>
<td>1.39</td>
<td>1.46</td>
<td>1.38</td>
</tr>
</tbody>
</table>
Table 3.5: Mean patient wait time and system utilization under compared routing policies

<table>
<thead>
<tr>
<th></th>
<th>Wait time (min.)</th>
<th>Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Current</td>
<td>Proposed</td>
</tr>
<tr>
<td><strong>ESI2</strong></td>
<td>2.1</td>
<td>1.4</td>
</tr>
<tr>
<td><strong>ESI3</strong></td>
<td>8.7</td>
<td>4.8</td>
</tr>
<tr>
<td><strong>ESI3</strong></td>
<td>8.7</td>
<td>3.6</td>
</tr>
<tr>
<td><strong>ESI4</strong></td>
<td>2.9</td>
<td>9.2</td>
</tr>
<tr>
<td><strong>ESI5</strong></td>
<td>6.0</td>
<td>17.4</td>
</tr>
<tr>
<td><strong>Main ED</strong></td>
<td>7.3</td>
<td>3.9</td>
</tr>
<tr>
<td><strong>Fast Track</strong></td>
<td>3.4</td>
<td>9.2</td>
</tr>
</tbody>
</table>

As shown in Table 3.5, the new routing policy will result in shorter wait times for all 

*ESI2* and *ESI3* patients throughout the ED. The wait time reduction in the main ED is due to the reduced work load from shifting the *ESI3* patients to the fast track. The wait time for *ESI3* patients in the fast track is also shorter because they now have the highest priority among all the patients in the fast track. These improvements in responsiveness to ED’s higher-acuity patients (*ESI2* and *ESI3*) come at the expense of reduced responsiveness to lower-acuity patients (*ESI4* and *ESI5*). However, since the risk associated with forcing higher-acuity patients to wait is substantially greater, the benefits may very well outweigh the loss (whether they actually do or not would need to be a judgment call made by hospital administration and ED management).

Furthermore, it is shown that wait time is a major factor in percentage of patients who leave without being seen (LWBS) especially for *ESI3* patients (Batt and Terwiesch 2015). According to Lucas et al. (2013), *ESI3* patients have the highest LWBS ratio compared to patients in other acuity levels. Thus, given the high volume of *ESI3* patients, reducing their waiting time improves LWBS ratio more significantly.

In addition to reduction in waiting time and LWBS ratios for higher-acuity patients, the utilization of the fact track increases under the new routing policy, creating a better balance between capacity and demand. Under the new routing policy, more patients will be treated using
lower-cost resources. Thus, it is expected that implementing this anti-pooling policy would improve system performance by reducing the overall costs of waiting and providing care.

3.5. Sensitivity Analyses

Systems with highly variable arrival rates and/or service times can be highly sensitive to utilization levels. In order to examine the sensitivity of our results to system utilization, we consider three different cases: a) the base case with normal load; b) a crowded scenario (10% higher utilization); and c) an overcrowded scenario (20% higher utilization). The results are summarized in Table 3.6 and graphically demonstrated in Figure 3.11.

As expected, when system utilization increases, the waiting times of all groups of patients increase. However, in all cases, the rerouting policy results in significant improvement in average waiting times of ESI2 and ESI3 patients. In the overcrowded case, the average waiting time of ESI5 patients reaches 40 minutes. Thus, at times that the ED is overcrowded, it might not be a good policy to send all ESI3/ patients to the fast track.

Another parameter that may impact the results is the threshold LOS value used to identify ESI3/ patients because it directly influences the number of patients who may be eligible to be served in the fast track. Initially, based on expert opinion and fast track operational goals, we assumed a 120-minute threshold value (T) for LOS. Here, we examine the effects of using 150- and 180-minute thresholds instead. The resulting inter-arrival and service times (queueing model input parameters) are presented in Table 3.7. The waiting time results are summarized in Table 3.8. By increasing the LOS threshold, the number of patients sent to fast track increases, along with their average LOS. These two factors result in longer patient wait times in the fast track while the average wait time in the main ED decreases only slightly.
Table 3.6. Sensitivity to system load: Mean overall patient wait times (minutes)

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th></th>
<th></th>
<th>Crowded</th>
<th></th>
<th></th>
<th>Overcrowded</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Baseline Utilization)</td>
<td>(Utilization +10%)</td>
<td>(Utilization +20%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESI2</td>
<td>2.1</td>
<td>1.4</td>
<td>4.2</td>
<td>3.0</td>
<td>8.0</td>
<td>6.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESI3_h</td>
<td>8.7</td>
<td>4.8</td>
<td>17.8</td>
<td>10.3</td>
<td>33.5</td>
<td>20.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESI3_l</td>
<td>8.7</td>
<td>3.6</td>
<td>17.8</td>
<td>5.6</td>
<td>33.5</td>
<td>8.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESI4</td>
<td>2.9</td>
<td>9.2</td>
<td>5.0</td>
<td>14.4</td>
<td>8.2</td>
<td>21.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESI5</td>
<td>6.0</td>
<td>17.4</td>
<td>10.6</td>
<td>27.1</td>
<td>17.1</td>
<td>40.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main ED</td>
<td>7.3</td>
<td>3.9</td>
<td>14.8</td>
<td>8.4</td>
<td>28.0</td>
<td>16.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fast Track</td>
<td>3.4</td>
<td>9.2</td>
<td>6.0</td>
<td>14.4</td>
<td>9.8</td>
<td>21.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.11. Visual comparison of patients waiting time as system utilization increases
Table 3.7. Input parameters under each LOS threshold value (T)

<table>
<thead>
<tr>
<th>Service time</th>
<th>120 min</th>
<th>150 min</th>
<th>180 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length-of-Stay Threshold Value (T)</td>
<td>$ESI^3_l$</td>
<td>$ESI^3_h$</td>
<td>$ESI^3_l$</td>
</tr>
<tr>
<td>N</td>
<td>4,296</td>
<td>21,396</td>
<td>5,894</td>
</tr>
<tr>
<td>Mean</td>
<td>120.22</td>
<td>266.67</td>
<td>133.66</td>
</tr>
<tr>
<td>S.D.</td>
<td>81.12</td>
<td>196.38</td>
<td>90.48</td>
</tr>
<tr>
<td>CV</td>
<td>0.67</td>
<td>0.73</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Interarrival time

<table>
<thead>
<tr>
<th>Current Policy</th>
<th>120 min</th>
<th>150 min</th>
<th>180 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ESI^2$</td>
<td>2.1</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>$ESI^3_h$</td>
<td>8.7</td>
<td>4.8</td>
<td>4.7</td>
</tr>
<tr>
<td>$ESI^3_l$</td>
<td>8.7</td>
<td>3.6</td>
<td>4.6</td>
</tr>
<tr>
<td>$ESI^4$</td>
<td>2.9</td>
<td>9.2</td>
<td>13.3</td>
</tr>
<tr>
<td>$ESI^5$</td>
<td>6.0</td>
<td>17.4</td>
<td>24.4</td>
</tr>
<tr>
<td>Main ED</td>
<td>7.3</td>
<td>3.9</td>
<td>3.8</td>
</tr>
<tr>
<td>Fast Track</td>
<td>3.4</td>
<td>9.2</td>
<td>12.8</td>
</tr>
</tbody>
</table>

Note: Mean and S.D. are shown in minutes.

Table 3.8. Estimated waiting time (min.) under each LOS threshold value (T)

<table>
<thead>
<tr>
<th>Current Policy</th>
<th>T=120 min</th>
<th>T=150 min</th>
<th>T=180 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ESI^2$</td>
<td>2.1</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>$ESI^3_h$</td>
<td>8.7</td>
<td>4.8</td>
<td>4.7</td>
</tr>
<tr>
<td>$ESI^3_l$</td>
<td>8.7</td>
<td>3.6</td>
<td>4.6</td>
</tr>
<tr>
<td>$ESI^4$</td>
<td>2.9</td>
<td>9.2</td>
<td>13.3</td>
</tr>
<tr>
<td>$ESI^5$</td>
<td>6.0</td>
<td>17.4</td>
<td>24.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Proposed Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>T=120 min</td>
</tr>
<tr>
<td>T=150 min</td>
</tr>
<tr>
<td>T=180 min</td>
</tr>
</tbody>
</table>

3.6. Conclusions

3.6.1. Discussion

EDs are under increasing pressure to provide more and better care with fewer resources. How an ED’s capacity is allocated to different physical spaces (e.g., beds) and levels of care (e.g., a trauma bay vs. fast track), and how patients are routed within the ED to these care areas can have dramatic influence on the efficiency of the ED’s operations. In an attempt to improve
the ability of EDs to provide care to patients more quickly, we examined a proposed new triage policy that splits ESI3 patients into two subgroups based on their expected length-of-stay and care requirements. We then analytically examined that policy using queuing models and evaluated its potential efficacy using data from a large, urban teaching hospital’s ED.

In this study, we found additional evidence to support earlier findings that pooling queues can, under certain conditions, lead to poor service being provided to certain groups of customers. In the case of systems with multi-priority customers and heterogenous service requirements, like EDs, it can be more desirable to put similar customers into smaller groups and accommodate their needs accordingly. Through data partitioning, this retrospective study of ED visits data identified subgroups of short-LOS ESI3 patients who do not necessarily need the medically intense resources available in the main ED area. Using queuing models, we found that these ESI3 patients can be treated safely and more quickly if routed to the fast track instead. Overall, while lower-acuity ESI4 and ESI5 patients waited slightly longer, wait times for the higher-acuity ESI2 and ESI3 patients (a majority of all patients receiving care in the ED) were reduced significantly. Additionally, utilization was reduced in the main ED (enhancing its responsiveness) while it was increased in the fast track (better capitalizing on the lower-cost resources used there). Considered together, this policy seems to have the potential for increasing service quality (through reduced patient waiting) while improving the ED’s cost-effectiveness.

In order to assess our policy’s generalizability, we examined the sensitivity of our results to both ED busyness and the length-of-stay threshold value. Evaluation of crowded and overcrowded (10% and 20% higher utilization than baseline, respectively) scenarios showed no change to the performance of the proposed policy. Gains in terms of reduced waiting for higher-acuity patients were sustained.
The results of the sensitivity analysis on the LOS threshold value were more interesting. When increasing the LOS threshold value above the baseline 120 minutes, waiting for $ESI2$ patients decreased (as before) and waiting for $ESI4$ and $ESI5$ patients increased (as before), but the benefits to the $ESI3$ patients (in terms of waiting time) were attenuated. Under the current policy, $ESI3$ patients wait an average of 8.7 minutes. Under the proposed policy with a LOS threshold value of 120 minutes, $ESI3^h$ patients wait 4.8 minutes and $ESI3^l$ patients wait 3.6 minutes on average; the $ESI3^l$ patients benefit more. When $T=150$, the population of $ESI3^l$ patients increases (from 16.7% of all $ESI3$ at $T=120$ to 22.9% at $T=150$) and those values climb to 4.7 and 4.6 minutes, respectively. While still below the current policy’s baseline values, the improvements are more evenly distributed between the two subgroups. At $T=180$ minutes, the $ESI3^l$ population grows even larger (35.8% of all $ESI3$), with waiting times now unevenly apportioned in the opposite direction from $T=120$ minutes: $ESI3^h$ patients wait an average of 2.5 minutes while $ESI3^l$ patients wait 11.3 minutes. This situation forces the new $ESI3^l$ subgroup to wait more than if they had just been routed to the main ED under the current policy. Clearly, as the $ESI3^l$ patient population grows as a proportion of all $ESI3$ patients, we begin to unduly stress the fast track, resulting in inappropriately high utilization there, and “starving” the main ED. But, again, if the hospital is willing to let lower-acuity patients wait longer in order to ensure that higher-acuity patients wait less, this disproportionate allocation of the benefits of our proposed triage policy may, in fact, be seen as desirable.

3.6.2. Limitations and Extensions

While this research makes substantial contributions to our theoretical and practical understanding of emergency department operations, it has a few limitations, some of which represent opportunities for additional research. While our queueing analysis provides
straightforward calculations for estimating the benefits of the proposed routing policy, these calculations do not consider system status. A more robust policy for patient routing should consider the current (if not also near-future) state of the ED, in terms of utilization, patient demand, shift changes, etc. As our sensitivity analysis shows, the best routing decision depends on the status of the system. If the system is overcrowded directing all $ESI^3$ patients to the fast track can have adverse results on patient satisfaction and number of patients reneging. Therefore, an interesting extension of this study is to find out whether or not $ESI^3$ patients should be directed to fast track given current system load and expected future arrivals.

A second limitation is that the variables we considered in the partitioning analysis are based on the patient and his condition. Such an analysis is only useful as long as both (a) the ED’s patient population remains similar to the sample used to develop the partitions, and (b) care practices do not significantly change the expected length-of-stay for these patients. Since both of those elements are subject to evolving over time, the partitioning exercise should be redone on a regular basis (e.g., every 3 years) in order to ensure that no significant subgroups are being missed while selected subgroups remain appropriate for routing to the fast track. A potential extension of this work would be to design a heuristic by which $ESI^3$ patients are identified by their attributes instead of membership in a predetermined, static subgroup.

3.6.3. Implications for Practice

Clearly, EDs are being required to develop new ways of serving more patients more quickly without the benefit of added resources. While the problem and policy proposed here was motivated by the general state of EDs in the United States, it benefits from having been analyzed using data from a large, urban, level I trauma center ED. To that end, implementation could significantly benefit EDs. By breaking up the dominant $ESI^3$ triage category and shifting a
portion of that population into the fast track, the ED can provide more timely care overall while
doing so with lower-cost resources. The identification of triage-ready subgroups means that
triage nurses could use a relatively simple checklist (ideally integrated into the registration
information system) to identify those ESI3 patients who should be routed to the fast track instead
of a main ED bed. Most of the subgroups are identifiable by chief complaint with a small number
(typically less than four) qualifiers, such as normal blood pressure and temperature and age <65.

Another implication for EDs is the general take-away that both aggregating (pooling) and
disaggregating capacity can be useful and helpful, depending on the patient populations involved
and the ED’s larger operational objectives. Where a patient spends his time, the capacity of each
process step relative to patient demand, and the value provided by each step should all factor into
our consideration of how capacity is allocated and how patients are routed within the ED.
Improving emergency department responsiveness through dynamic patient routing

Abstract

The triage decision in the Emergency Department potentially determines the patient’s pathway through the system, affecting both resource usage and patient flow. For optimal operational performance, triage decisions should be based not only on patient medical needs, but also on the operational consequences of assigning a patient to a certain level of care. In EDs with a fast track, it has been shown that redirecting a subset of shorter-stay, moderate-acuity patients to the fast track can reduce the waiting time of higher-acuity patients by reducing the work load in the main ED area. However, implementation of this policy merits closer study because both volume (i.e., how many patients to redirect) and timing (i.e., when to redirect) are potentially dependent upon the status of the system. An improper decision can increase fast track wait time dramatically, harming patient satisfaction and increasing the number of patients who leave without being seen. We ask the question: Which routing policy for short-stay, moderate-acuity patients most improves ED responsiveness? Using discrete-event simulation, we compare various time-dependent and status-dependent routing policies. The results show that system-status-dependent policies outperform static and time-dependent policies. We examine the sensitivity of our results to the cost of waiting for patients in different acuity levels, to patient misclassification at the time of triage, and to the possibility of changes in the patient’s clinical path.
4.1. Introduction

Emergency Departments (ED) need to be more responsive to cope with the growing demand and expanding roles (Morganti et al. 2013). The Institute of Medicine called ED crowding a “national epidemic” in 2006. ED crowding occurs when the demand for services exceeds the system capacity to provide care, leading to unreasonable waiting times (McCarthy et al. 2009). According to the National Hospital Ambulatory Medical Care Survey 2011, only 27% of 136.6 million patients who visited EDs were seen within 15 minutes of ED arrival (Center for Disease Control and Prevention).

High variability in demand is one of the most important factors that make matching capacity and demand challenging in the ED. EDs traditionally use triage systems to group patients with similar needs and medical urgency together. The Emergency Severity Index (ESI) is a common triage system used in many EDs in the US. ESI stratifies patients to five groups based on their medical urgency and resource requirements. ESI level 1 patients, who have the highest priority, are those who need immediate life-saving intervention. ESI level 2 are highly-acute patients who are either in high-risk situations or in severe pain/distress. ESI level 3 patients are moderate-acuity patients whose situation is not risky but their treatment requires multiple different resources. Finally, patients in ESI 4 and 5 levels are considered low-acuity as their condition is not life threatening and their treatment is not resource intensive.

The use of triage systems allows EDs to manage their capacity according to the needs of different groups of patients. For instance, most EDs have dedicated highly equipped rooms for trauma patients. Similarly, some EDs have a minor care or fast track area dedicated to low-acuity patients. Therefore, among all the ED front-end operations, the triage decision has an important role in matching capacity and demand. It has been shown that considering the case complexity
and expected dispositions (i.e., admit vs. discharge) at the time of triage can lead to better patient streaming in the ED and lower patient wait times (Saghafian et al. 2012 and 2015). The triage decision potentially determines the patient’s pathway through the system affecting both resource usage and patient flow. For optimal operational performance, triage decisions should be based not only on the patient’s medical needs but also on the operational consequences of assigning a patient to a certain level of care.

In EDs with a fast track, routing lower-acuity patients to the dedicated fast track service line improves responsiveness to them, but negatively affects the higher-acuity patients by taking away capacity from them. It has been shown that redirecting a subset of shorter-stay, moderate-acuity patients to the fast track can improve waiting time of higher-acuity patients by reducing the work load in the main ED area (Torabi et al. 2016). However, implementation of this policy warrants closer study because both the volume (i.e., how many patients to redirect) and the timing (i.e., when to redirect) are potentially dependent upon the status of the system. An improper decision can increase the fast track wait time dramatically, potentially harming patient satisfaction and increasing the number of patients who leave without being seen (LWBS).

There are several policies we could use to route these moderate-acuity patients based on the operational status of the ED (detailed in the Section 4.3.2). In this study we ask the question: Which routing policy for short-stay, moderate-acuity patients most improves ED responsiveness? In answering this question, we also consider implications in evaluating the cost of waiting for patients in different acuity levels, the problem of misclassification at the time of triage, and the possibility of changes in clinical pathway after the patient is seen by the physician.

The remainder of this paper is organized as follows. In Section 4.2, we provide background information and a brief review of related literature. In Section 4.3, we introduce our
methods and data. We present results and discuss our findings in Section 4.4. Finally, we conclude our discussion in Section 4.5.

4.2. Background and literature review

4.2.1. Optimal control of arrivals

EDs can be modeled as multi-class, multi-server, priority queueing systems. Several studies have analyzed multi-server queues with multiple priority classes (Harchol-Balter et al. 2005, Sleptchenko et al. 2005). An interesting, yet difficult question that arises in multi-class queueing systems is the optimal control of arrivals (Huang et al. 2015, Li and Neely 2012, Bretthauer 2000, Helm and Waldmann 1984, Stidham 1978). This problem has been widely studied in job shops and call centers with the primary question being which job/customer should be served next. Several static job/customer prioritization rules, such as the “cμ rule,” (Mandelbaum and Stolyar 2004) have been explored with the goal of identifying policies that optimize the performance metric of interest (e.g., total average lead time, average caller wait time, and such).

When there are two or more parallel queues, the problem becomes more complicated; the additional question to be answered is to which queue the customer should be assigned (Glazebrook and Nino-Mora 2000). The goal is to optimize the overall performance of the system as measured by average waiting for all customers, make span, or other performance metrics. This patient-routing decision is sometimes made with additional considerations, such as distributing the workload fairly among servers (Armony and Ward 2010).

In systems with high variability in arrivals, such as EDs, static policies can be easy to implement, but can also have significant and negative operational consequences (Armony et al. 2015). Therefore, time-dependent and state-dependent analysis of queueing networks has been
studied and various system-status-dependent policies have been examined (Towsley 1980, Cruz and Smith 2007). Examples are joining the shorter queue (Turner 2000) and joining the queue with shortest expected delay (Adan and Vessels 1995, Hlynka et al. 1994). Queue length threshold policies have also been studied (Schaack and Larson 1986, Feng et al. 2001, Bell and Williams 2001). These dynamic policies were shown to be optimal under certain conditions but there is no guarantee of optimality when problem assumptions change (Whitt 1985).

4.2.2. Patient triage and streaming in the ED

Timeliness of care can have a strong effect on both patient satisfaction and patient health outcomes (Derlet and Richards 2000, Lindsay et al. 2002). Thus, EDs have been making considerable effort to increase responsiveness by improving their front-end operations. A list of implemented strategies, such as bedside registration and immediate bedding, and their operational outcomes are presented in Wiler et al. (2010). Many EDs have used strategies such as Physician/Practitioner at Triage, Advanced Triage Protocols, and Rapid Medical Assessment to achieve higher accuracy and initiate patient evaluation tasks like ordering images or labs (Traub 2015). However, they do not make fundamental changes to the level of care to which the patient is assigned.

In the operations management literature, there are studies that propose alternative streaming policies that can be used to complement the traditional triage systems (Saghafian et al. 2015). These streaming policies suggest that grouping patients based on their predicted length of stay (LOS) at the time of triage would improve patient flow in the ED. These findings also suggest that grouping similar customers together such that within group variability is minimized is an efficient approach in dealing with highly variable demand (Bohmer 2005).
Saghafian et al. (2012) classify patients based on their predicted disposition to admit versus discharge groups. Patients in each group are then prioritized based on their medical urgency. Assuming that patients with more complex cases are likely to have longer LOS, Saghafian et al. (2014) propose a complexity-augmented triage policy. They note the complexity of a patient’s case at the time of triage is a better indicator of LOS than predicted ED disposition. In both studies, using simulation modeling and analytical methods, they show that combining traditional triage with LOS-based patient streaming policies can improve patient flow in the ED significantly. It is also noted that EDs with higher variability in LOS among the identified groups of patients, benefit most from these streaming policies.

Torabi et al. (2016) study patient streaming in EDs with a fast track. Their proposed streaming is based both on the Emergency Severity Index (ESI) and patients’ LOS (Fig. 4.1). ESI is a five-level triage algorithm assigning the value 1 to highest-acuity patients and value of 5 to lowest-acuity patients (Gilboy et al. 2011). In EDs with a fast track, the lower-priority patients (i.e., ESI4s and ESI5s) are sent to the fast track, whereas the higher priority patients (i.e., ESI2s and ESI3s) are sent to the main ED area. Using data-mining methods, they further stratify moderate-acuity patients (ESI3) based on their LOS to identify sub-groups with shorter LOS and lower risk (denoted by ESI3L). Remainder of ESI3 patients are designated as ESI3H. Using multi-class multi-server non-preemptive priority queueing approximations, they show that streaming these patients to the fast track can reduce waiting time in the main ED and increase fast track utilization. Their sensitivity analysis shows that system responsiveness is highly sensitive to the policy as system workload goes up. Thus, the decision whether or not a moderate-acuity, short-stay patient should be sent to the fast track depends on system status.
4.2.3. This study

For non-trauma patients being cared for in an ED with a fast track, there are two separate non-preemptive priority queues feeding two sets of dedicated servers (i.e., the main ED and the fast track). As shown in Figure 4.1, ESI4 patients and ESI5 patients are treated in the fast track and ESI2 and ESI3<sup>H</sup> patients are treated in the main ED. However, ESI3<sup>L</sup> patients could be served either in the fast track, where they have the highest priority (but receive a lower level of care), or in the main ED, where they have the lowest priority (but receive a higher level of care). So, the optimal arrival control problem narrows down to deciding where to direct these moderate-acuity patients such that system responsiveness is maximized (Figure 4.2). Failure in finding the proper policy can lead to long waiting times and high abandonment rates, which are both important metrics in managing ED operations (Lucas et al. 2013).
The arrival process is non-stationary with multi-priority, multi-class patients arriving at a non-preemptive, priority-queueing system with multiple non-identical servers. To the best of our knowledge, there is not any study addressing the arrival-control problem in such queueing systems. We use simulation modeling and analysis to find the dynamic routing policy that better improves system responsiveness. In the context of this study, misclassifying a long-stay patient as a short-stay one can lead to congestion in the fast track. Therefore, we also examine the sensitivity of our results to various misclassification rates.

4.3. Methods

4.3.1. Performance measurement

The primary goal of this study is to improve ED responsiveness. We focus on two metrics: 1) patient waiting time, often referred to in practice as door-to-bed time; and 2) patient abandonment, measured by ratio of patients who leave without being seen (LWBS). It has been shown that LWBS rates highly depend on patient wait times for moderate- and low- acuity patients (Batt and Terwisch 2015), and LWBS rates have begun to be used for modifying reimbursement rates in some settings.
Considering waiting times, the question is what waiting time function should be used. We define two main waiting time functions as linear weighted averages calculated as

\[ W = \sum_{i=1}^{5} \alpha_i W_i \]

where \( W_i \) is the average waiting of patients and \( \alpha_i \) is the coefficient corresponding to cost of waiting for patient category \( i \). Due to the difference in the volume of patients in each care area and each patient category, \( W_i \) is calculated as a volume-weighted average.

4.3.2. Alternative policies

We consider a collection of seven static and dynamic policies for patient streaming.

Policies 1 and 2 are static policies:

1. \( \text{All MED} \): all \( ESI3^L \) patients are sent to the main ED (MED).
2. \( \text{All FT} \): all \( ESI3^L \) patients are sent to the fast track (FT).

Policies 3 through 6 are status-dependent heuristic policies:

3. \( \text{All MED-Jockey} \): all \( ESI3^L \) patients are sent to main ED but if a FT bed becomes available they can quit the main ED queue and get into service in FT.
4. \( \text{Shorter Q} \): is the popular “join the shorter queue” rule.
5. \( \text{SML Util-FT} \): join the server that has smaller utilization, ties to be sent to FT
6. \( \text{SML Util-MED} \): join the server that has smaller utilization, ties to be sent to main ED

Policy 7 is the optimal time-dependent policy:

7. \( \text{Time-dependent} \): identified by a simulation-based optimization approach explained in Section 4.4, \( X = [1,1,1,0,0,0,0,1,1,0,1,1] \) is the optimal time-dependent policy. Starting at 12 midnight, for each two-hour time period if \( x_i \) is 1,
all $ESI^3$ patients arriving at that time period will be routed to FT; otherwise they are routed to the main ED.

4.3.3. Simulation model

No known, tractable analytical model can fully include all the details of a system as complex as a modern ED. Therefore, we develop a discrete-event simulation model to compare the performance of the different system-status-dependent and time-dependent routing policies described in Section 4.3.2.

We simulated a medium-sized ED with five beds in the fast track and 15 beds in the main ED. To establish reasonable estimates of key parameters, such as service times and inter-arrival times of patients in different $ESI$ categories, we use data from 67,399 patient visits to an urban, teaching hospital ED. The service time distributions are summarized in Table 4.1. As shown in Figure 4.3, the arrival of patients in the ED is non-stationary. To model non-stationary arrivals in the simulation model, we use the thinning method described in Lewis and Shedler (1979). The schematic view of the simulation logic is presented in Figure 4.4. The number of replications was set to 1200 replications for each policy, achieving a 95% half-width for waiting time outputs ($\leq 5\%$ of the mean). To insure a fair comparison of policies, we use common random numbers in generating random variates in all different scenarios.

Table 4.1. Service time distributions for discharged patients

<table>
<thead>
<tr>
<th>Patient Class</th>
<th>N</th>
<th>Service time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ESI2$</td>
<td>11,072</td>
<td>Lognormal (322, 250)</td>
</tr>
<tr>
<td>$ESI3^H$</td>
<td>21,396</td>
<td>Lognormal (266, 181)</td>
</tr>
<tr>
<td>$ESI3^L$</td>
<td>4,296</td>
<td>Lognormal (123, 88.6)</td>
</tr>
<tr>
<td>$ESI4$</td>
<td>25,277</td>
<td>Lognormal (111, 88.5)</td>
</tr>
<tr>
<td>$ESI5$</td>
<td>5,358</td>
<td>Lognormal (77.4, 63.1)</td>
</tr>
</tbody>
</table>
Figure 4.3. Average number of arrivals per hour of the day in the example ED
Figure 4.4. Schematic view of the simulation logic. Thin refers to the thinning process.
4.3.4. Simulation-based optimization

Incorporating the status-dependent policies in the simulation model is straightforward. However, modeling time-dependent policies poses a greater challenge. In a time-dependent policy, the routing decision for each patient is based on the time of day. Time of day can be defined as periods of time with a fixed duration, e.g., two-hour time periods. Therefore, a time-dependent policy can be characterized as a binary vector $X$ of size $n$ (number of time periods in 24 hours). If $x_t = 1$ at the time of $ESI^{L}$ patient arrival, the patient is to be sent to the fast track; otherwise, the patient is directed to main ED. Instead of comparing all $2^n$ possible time-dependent policies with the system-status-dependent policies, we first use a simulation-based optimization approach to find the optimal time-dependent policy that minimizes average weighted waiting.

The optimization problem is $O(2^n)$, making solving for large values of $n$ difficult. However, in the context of this problem, and for the policy to be practical in the ED, we consider 2-hour time periods resulting in $2^{12}$ possible solutions. Complete enumeration of the 4096 cases takes roughly one hour in OptQuest for Arena (Rockwell Automation) on a typical desktop PC. However, the heuristic search algorithm can find the optimal solution within the first 200 simulations.

4.4. Results and discussion

4.4.1. Comparison of policies

Table 4.2 shows throughput, mean and 90th percentile of patient waiting time under each policy based on care area. Distribution of average waiting times under each policy is presented in the appendix. The results imply that throughput is not significantly different under different
policies. However, the waiting time varies because the number of $ESI3^L$ patients who are served in FT varies from one policy to another, affecting waiting time of patients in both areas.

Considering individual areas, static policies result in shortest waiting times in one area at the cost of much longer waiting time in the other area. The All FT policy results in the shortest achievable waiting time in the main ED with mean of 20 minutes with 90% of patients wait less than 50 minutes. This is at the cost of a 30-minute increase in average waiting time in the FT with more than 40% of patients waiting more than 60 minutes.

As shown in Figure 4.5, All MED, SML Util-MED, All MED-Jockey, SML Util-FT, and All FT policies create a frontier. If the cost of waiting for patients in both areas is assumed to be equal, All MED-Jockey and Util-Main ED seem to result in shorter total average waiting times (Table 4.2). Statistical comparisons for average waiting times implies that All FT and SML Util-FT, All MED, and SML Util-MED result in shortest waiting times in one area, but longest waiting time in the other. Again, All MED-Jockey seems to perform well in both areas.

![Figure 4.5. Scatter diagram of patient wait time based on care area](image_url)

Figure 4.5. Scatter diagram of patient wait time based on care area
Table 4.2. Throughput, mean and 90th percentile of patient waiting by care area

<table>
<thead>
<tr>
<th>Policy</th>
<th>Main ED</th>
<th></th>
<th>Fast Track</th>
<th></th>
<th>No. of ESI3* served</th>
<th></th>
<th>Total avg. waiting time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average No. Served</td>
<td>Waiting time (min)</td>
<td>Average No. served</td>
<td>Waiting time (min)</td>
<td>No. of ESI3* served</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>90th percentile</td>
<td>Mean</td>
<td>90th percentile</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All MED</td>
<td>111 (0.5)</td>
<td>31.0 (1.6)</td>
<td>68.9</td>
<td>69 (0.5)</td>
<td>20.6 (1.1)*</td>
<td>49.1</td>
<td>16</td>
</tr>
<tr>
<td>All FT</td>
<td>96 (0.5)</td>
<td>20.4 (1.2)*</td>
<td>49.9</td>
<td>84 (0.5)</td>
<td>50.3 (2.1)</td>
<td>96.8</td>
<td>0</td>
</tr>
<tr>
<td>All MED-Jockey</td>
<td>107 (0.5)</td>
<td>22.1 (1.1)†</td>
<td>50.6</td>
<td>72 (0.5)</td>
<td>27.9 (1.5)†</td>
<td>64.7</td>
<td>12</td>
</tr>
<tr>
<td>Shorter Q</td>
<td>100 (0.5)</td>
<td>23.2 (1.3)†</td>
<td>54.5</td>
<td>80 (0.4)</td>
<td>31.3 (1.3)</td>
<td>64.6</td>
<td>4</td>
</tr>
<tr>
<td>Util-FT</td>
<td>101 (0.5)</td>
<td>21.2 (1.2)*†</td>
<td>50.7</td>
<td>78 (0.5)</td>
<td>32.1 (1.6)</td>
<td>71.4</td>
<td>6</td>
</tr>
<tr>
<td>Util-MED</td>
<td>104 (0.5)</td>
<td>26.5 (1.5)</td>
<td>60.9</td>
<td>76 (0.4)</td>
<td>23.0 (1.6)*†</td>
<td>52.1</td>
<td>9</td>
</tr>
<tr>
<td>Time-dependent</td>
<td>104 (0.5)</td>
<td>25.3 (1.4)</td>
<td>59.4</td>
<td>76 (0.5)</td>
<td>28.2 (1.4)†</td>
<td>62.5</td>
<td>8</td>
</tr>
</tbody>
</table>

*first best (statistically undifferentiated)  †second best (statistically undifferentiated)

Table 4.3. Throughput and mean patient waiting by patient category.

<table>
<thead>
<tr>
<th>Policy</th>
<th>ESI2</th>
<th>ESI3*</th>
<th>ESI3*</th>
<th>ESI4</th>
<th>ESI5</th>
<th>Total avg. waiting time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Avg. wait</td>
<td>N</td>
<td>Avg. wait</td>
<td>N</td>
<td>Avg. wait</td>
</tr>
<tr>
<td>All MED</td>
<td>33</td>
<td>9.0 (0.4)</td>
<td>62</td>
<td>36.1 (2.1)</td>
<td>16.1</td>
<td>39.6 (2.4)</td>
</tr>
<tr>
<td>All FT</td>
<td>33</td>
<td>7.6 (0.4)*</td>
<td>62</td>
<td>24.3 (1.7)*</td>
<td>16.5</td>
<td>10.9 (0.4)</td>
</tr>
<tr>
<td>All MED-Jockey</td>
<td>33</td>
<td>8.8 (0.4)*†</td>
<td>62</td>
<td>29.3 (1.7)*†</td>
<td>16.5</td>
<td>11.2 (0.5)</td>
</tr>
<tr>
<td>Shorter queue</td>
<td>33</td>
<td>8.3 (0.4)*†</td>
<td>62</td>
<td>27.9 (1.7)*†</td>
<td>16.6</td>
<td>9.4 (0.6)*†</td>
</tr>
<tr>
<td>Util-FT</td>
<td>33</td>
<td>8.2 (0.4)*</td>
<td>62</td>
<td>26.7 (1.7)*†</td>
<td>16.6</td>
<td>4.1 (0.3)*</td>
</tr>
<tr>
<td>Util-MED</td>
<td>33</td>
<td>8.5 (0.4)*†</td>
<td>62</td>
<td>31.3 (1.9)</td>
<td>16.5</td>
<td>19.3 (1.5)</td>
</tr>
<tr>
<td>Time-dependent</td>
<td>33</td>
<td>8.4 (0.4)</td>
<td>62</td>
<td>31.0 (1.9)</td>
<td>16.5</td>
<td>16.4 (0.9)</td>
</tr>
</tbody>
</table>

*first best (statistically undifferentiated)  †second best (statistically undifferentiated)
Table 4.3 summarizes throughput, mean and maximum waiting time of patients by ESI category. In general, waiting time of ESI2 patients is not largely affected by the routing policy as they have the highest acuity. Also, the relative volume of ESI3L patients is not big enough to create major blockage in the main ED. On the other hand, routing policy impacts waiting time of other patients significantly. The reduction in average wait time can be up to 12 minutes for ESI3H and up to 30 minutes for ESI3L patients. For ESI4 and ESI5 patients, routing policy can increase average waiting time up to 31 minutes and 80 minutes, respectively. The results imply that All MED-Jockey and SML Util-FT result in shorter total average waiting (Figure 4.6).

Figure 4.6. Average waiting time per patient category under different policies (primary axis) and total volume-adjusted average waiting (secondary axis)
4.4.2. Sensitivity analyses

4.4.2.1. Patients’ relative cost of waiting

As discussed in Section 4.3.2, we consider $\alpha$ as the relative cost of waiting. Determining the right value for this parameter is difficult and there is no existing literature quantifying the cost of patient waiting. Therefore, we examined different values of $\alpha$ to find out how it affects choice of dynamic policy. Table 4.4 shows the weighted waiting scores under different policies as $\alpha$ changes from 1 (no difference in waiting cost) to 20 (waiting time of main ED patients is 20 times more costly than FT patients).

Table 4.4. Weighted waiting score based on care area, $W = \alpha W_{MED} + W_{FT}$

<table>
<thead>
<tr>
<th>Policy</th>
<th>$\alpha$:</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>No. served in FT</th>
</tr>
</thead>
<tbody>
<tr>
<td>All MED</td>
<td>27.0</td>
<td>46.1</td>
<td>103.5</td>
<td>199.1</td>
<td>390.2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>All FT</td>
<td>34.4</td>
<td>45.2</td>
<td>77.9</td>
<td>132.3</td>
<td>241.1</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>All MED-Jockey</td>
<td>24.4</td>
<td>37.6</td>
<td>77.3</td>
<td>143.3</td>
<td>275.4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Shorter queue</td>
<td>26.8</td>
<td>39.7</td>
<td>78.4</td>
<td>142.8</td>
<td>271.7</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>SML Util-FT</td>
<td>25.9</td>
<td>37.9</td>
<td>73.8</td>
<td>133.6</td>
<td>253.2</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>SML Util-MED</td>
<td>25.0</td>
<td>40.3</td>
<td>86.3</td>
<td>162.8</td>
<td>315.9</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Time-dependent</td>
<td>26.5</td>
<td>41.1</td>
<td>85.0</td>
<td>158.1</td>
<td>304.3</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.7 shows the distance of each policy from the best performing policy for different values of $\alpha$. As shown in Figure 8, for small values of $\alpha$ (i.e., $1 \leq \alpha \leq 2.21$), All MED-Jockey is the best performing policy. For $2.21 \leq \alpha \leq 8.77$, SMLUtil-FT and for $\alpha > 8.77$, All FT are the best policies. In other words, as the relative cost of waiting for higher-acuity patients increases, it is more advantageous to reduce the workload on the main ED as much as possible by sending more $ESI3^L$ patients to the fast track.
Figure 4.7. Distance from the best performing policy for $1 < \alpha < 10$

Figure 4.8. Best-performing policies compared to the status quo
Our second waiting time function is based on patient category. Starting from equal weights for all patients, we tested the sensitivity of results to $\alpha_i$. Results show that sending the patients to the server with smallest utilization is always the best (Table 4.5). When the relative cost of waiting is uniform ties should be routed to the main ED to avoid extended waiting times for lower-acuity patients. Yet, as the lower-acuity waiting becomes less important, more patients can be routed to the FT. Therefore, ties should be routed to the FT.

<table>
<thead>
<tr>
<th>Policy</th>
<th>not weighted</th>
<th>weighted by urgency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1,1,1,1,*</td>
<td>2,1,1,1,1,1,1,1,1,1</td>
</tr>
<tr>
<td>All MED</td>
<td>25.6</td>
<td>27.3</td>
</tr>
<tr>
<td>All FT</td>
<td>33.0</td>
<td>34.4</td>
</tr>
<tr>
<td>All MED-Jockey</td>
<td>23.9</td>
<td>25.5</td>
</tr>
<tr>
<td>Shorter queue</td>
<td>25.6</td>
<td>27.2</td>
</tr>
<tr>
<td>SML Util-FT</td>
<td>24.8</td>
<td>26.3</td>
</tr>
<tr>
<td>SML Util-MED</td>
<td>23.8</td>
<td>25.4</td>
</tr>
<tr>
<td>Time-dependent</td>
<td>25.4</td>
<td>26.9</td>
</tr>
</tbody>
</table>

*weights for $ESI_2$, $ESI_3^{H}$, $ESI_3^{L}$, $ESI_4$ and $ESI_5$ respectively.

4.4.2.2. Misclassification

Aside from the reliability of the ESI triage system, ED operations are also affected by misclassifications at the time of triage. Since the triage decision is based on limited information, there is a possibility of misclassification. There are also discrepancies among nurses because the triage decision can be subjective (Fields et al. 2013).

Argon and Ziya (2009) found that in classifying customers based on imperfect information, it is better to give higher priority to customers who provide stronger signal, i.e., customers for whom the probability of belonging to a certain class is high. Saghafan et al. (2015) consider misclassification by using “error-impacted” rates in their Markov chain model.
Van Der Zee and Theil (1961) study a dual-priority system in which the priority assignment is subject to misclassification. They show that, if the classifier puts the cases he/she is unsure of in a mixed group, the waiting cost to the system would be less than if they misclassify them. Then the customers in the mixed group are assigned to one of the two priority groups based on the relative cost of misclassification. The cost of misclassification depends on the arrival of customers in the mixed group and the classification error rate.

Under triaging an ESI3\textsuperscript{H} patient as an ESI3\textsuperscript{L} will lead to have a patient in FT with relatively longer LOS. Prolonged LOS in FT might also occur if patient clinical pathway changes after the consultation or after the results of diagnostic tests come in. These situations are instances of misclassifying a long-stay patient as a short-stay one. Thus, it is important to examine the sensitivity of waiting time results to misclassification. Table 4.6 shows average waiting time of patients under different misclassification rates for SML Util-FT (the leading dynamic policy in Section 4.4.2.1).

<table>
<thead>
<tr>
<th>Category</th>
<th>Misclassification rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>ESI2</td>
<td>8.2 (0.4)</td>
</tr>
<tr>
<td>ESI3\textsuperscript{H}</td>
<td>26.7 (1.7)</td>
</tr>
<tr>
<td>ESI3\textsuperscript{L}</td>
<td>4.1 (0.3)</td>
</tr>
<tr>
<td>ESI4</td>
<td>27.7 (1.4)</td>
</tr>
<tr>
<td>ESI5</td>
<td>75.6 (4.8)</td>
</tr>
<tr>
<td>Main ED</td>
<td>21.2 (1.2)</td>
</tr>
<tr>
<td>Fast track</td>
<td>32.1 (1.6)</td>
</tr>
</tbody>
</table>

Note: Values shown are mean (95% half width) patient wait time, in minutes.

The sensitivity analysis results show that patient triage misclassification up to 10% does not result in statistically significant difference from a system assumed to be making no misclassification errors. However, even with 20% misclassification, the change in waiting time
is, while statistically significant, small in magnitude and unlikely to be managerially meaningful. These results suggest that the SML Util-FT policy seems robust to misclassification.

4.5. Conclusion

In this paper, we examine the problem of patient streaming in EDs with a fast track. High variability in arrivals and service requirements lead to mismatch between capacity and demand often resulting in crowding in the main ED area. To create a better balance between capacity and demand, patient streaming should be based on status of the system. There are different dynamic policies that could be used. We compared a variety of static, system-status-dependent, and time-dependent policies using different weighted waiting performance metrics.

The results show that there is a trade-off between waiting time of higher- versus lower-priority patients. The policy that is best for one group can be the worst for the other. In general, when costs of patient waiting is uniform, the best policy is to direct \( ESI_3^L \) to the main ED queue but if a room becomes available in the fast track while they are waiting, redirect them to the fast track. As the relative cost of waiting for higher-acuity patients go up, it is better to redirect more \( ESI_3^L \) patients to the fast track. In this case, our results indicate that the best routing policy is to send \( ESI_3^L \) patients to the area with smaller utilization and route them to fast track if there is a tie. If the relative cost of waiting for higher-acuity patients is very large, the static policy of sending all \( ESI_3^L \) patients to the fast track would be best. The sensitivity results indicate that this policy is quite robust to misclassification. The policy is simple and easy to implement with no additional costs to the ED.

Any ED with an imbalance between utilization of the fast track and the main ED can take more advantage from routing \( ESI_3^L \) patients to the fast track. EDs might be different in mix of population and capacity allocation. Our simulation model is based on data from an urban,
teaching hospital, thus it might not be representative of all EDs. Despite this limitation, this study quantifies the advantage and necessity of considering operational consequences of triage decision about moderate-acuity patients. Moreover, the insights from this study shed light on dynamic customer streaming in on other multi-priority systems.
Conclusion

Healthcare delivery is a challenging area of operations that needs to satisfy competing goals of the system from minimizing costs and maximizing throughput to maintaining quality of care and patient and provider satisfaction. In healthcare delivery, resources can be costly and scarce and the penalty for not satisfying demand can be tragic. Thus, numerous studies in the healthcare operations management literature try to optimize capacity allocation and flow management practices to ensure that clinical resources are optimally aligned to meet patient demand. However, the high uncertainty in demand, such as patient arrivals, disease and needs, makes the problem even more difficult. This dissertation is a close investigation of capacity allocation and flow management is healthcare systems with multi-priority patients. It brings together numerous methods such as data mining, discrete-event simulation, and optimization, to find optimal static and dynamic solutions to the problem.

In Chapter 2, we start from a system of operating rooms with both emergency, high-priority surgeries and elective, low-priority surgeries. We compare different capacity allocation policies such as total flexibility, partial flexibility and dedication and find that although total flexibility is the optimal solution in many cases, when there is a large difference in the traffic intensity imposed by different groups of patients, it might be optimal to dedicate some resources to higher-priority patients. To find optimal room and flow allocation policy we introduce a two-stage optimization model involving queueing approximations for M/M/n non-preemptive, multi-priority systems with optimal flow control. The goal is to find how many dedicated and flexible rooms are needed and how the patients should be streamed into the resources so that total
weighted patient wait time is minimized. We show that the first-stage problem of finding the best flow allocation given any room configuration is jointly convex in flow variables, so that the optimal flow allocation can be found efficiently. The second stage involves an iterative search on different room configurations. Extensions to general service times and multi-priority patients are explored. Sensitivity analysis is also performed to examine how the optimal solution changes as a function of patient arrivals and the relative cost of waiting.

In healthcare systems such as Emergency Departments, the variability in demand is so high that demand and capacity segmentation is the only way to be able to reduce variability. Demand segmentation is grouping similar patients together such that within group variability is smaller. Thus, the question to be answered is how the segmentation should be done and how the different groups of patients should be assigned to existing capacity. To answer these questions, in Chapters 3 and 4 we consider the problem of patient streaming in the emergency department given the existing resource allocation. Given the cost pressures on emergency departments, many have turned to partitioning their capacity (both beds and human resources) into two distinct areas: (a) the main ED and (b) minor care (or fast track). Patients categorized as ESI levels 1, 2, and 3 are routed to the main ED beds where they receive more targeted service by typically more skilled care providers. ESI 4 and 5 patients, who need lower intensity care and fewer resources, are redirected to the fast track, which is often staffed with lower-cost providers, such as nurse practitioners or physician assistants. Although dedicating part of the capacity to low-acuity patients can lead to more cost-effective use of human resources and may reduce patient length-of-stay (LOS) for ESI 4 and 5 patients, it can make the ED less responsive to higher-acuity patients by sequestering capacity away from the main ED (illustrating the downside of capacity segmentation).
Analyzing historical data on ED visits shows that approximately half of all patients are categorized as ESI level 3. Given the problems of ED crowding, we ask this research question: *Would redirecting specific sub-groups of ESI level 3 patients to the fast track improve patient flow in the ED?*

The proposed sub-groups are identified based on demographics, vital signs, and other patient attributes at the time of triage by the triage nurse. Applying partitioning methods using the information available at triage, we identify ESI level 3 patients who could be sent to the fast track. The operational criteria for sending patients to the fast track is to have an average length of stay (LOS) which is less than a threshold value e.g., $<120$ minutes. Streaming these sub-groups to the fast track will affect arrival and service rates in both areas of the ED. Thus, we first estimate the effects of this new policy on arrival and service rates in both areas of the ED, then develop queueing models to compare the performance of the suggested policy to the current approach in terms of patient LOS and waiting time. We find that redirecting even a sixth of the moderate-acuity, medically appropriate patients from the main ED to the fast track can reduce patient waiting substantially for all higher- and moderate-acuity patients presenting to the ED. We examine the policy’s sensitivity to other scenarios and draw conclusions for both research and practice.

In Chapter 3 we show that, in EDs with a fast track, redirecting a subset of shorter-stay, moderate-acuity patients to the fast track can reduce the waiting time of higher-acuity patients by reducing the work load in the main ED area. However, implementation of this policy merits close study because both volume (i.e., how many patients to redirect) and timing (i.e., when to redirect) are important and dependent upon the status of the system. An improper decision can increase fast track wait time dramatically, harming patient satisfaction and increasing the number
of patients who leave without being seen. Thus in Chapter 4 we ask the question: *Which routing policy for short-stay, moderate-acuity patients most improves ED responsiveness?* Using discrete-event simulation, we compare various time-dependent and status-dependent routing policies. The results show that system-status-dependent policies outperform static and time-dependent policies. We examine the sensitivity of our results to the cost of waiting for patients in different acuity levels, to patient misclassification at the time of triage, and the possibility of changes in the patient’s clinical path.

Considering the results and insights form the above studies, in systems with multiple priority classes of patients, there are several factors which needs to be closely considered to best match capacity and demand:

1. population mix, i.e., number of priority groups and the volume of patients in each group
2. the relative traffic intensity of high- versus low-priority patients
3. the relative cost of waiting for high- versus low-priority patients
4. capacity mix, i.e., number of dedicated and shared resources
5. fluctuations in demand throughout the day

Our results show that both aggregating (pooling) and disaggregating capacity can be useful and helpful, depending on the systems larger operational objectives. It is important to note that the operational objective in one segment of capacity serving a specific group of patients may be different from that of other segments. For example, EDs need to be maximally responsive to high-acuity patients, served in the main ED area, but maximally efficient when treating low-acuity patients served in minor care.

Most demand management strategies, such as optimal panel sizing and appointment scheduling, that are used in many outpatient clinics may not be readily applicable to
environments such as EDs. Therefore, there is a need to find effective ways to manage variable demand. Our investigations highlight the importance of flow allocation decisions and the impact they can have on system responsiveness. In other words, having flexibility in streaming demand can result in better balance in capacity and demand. Flexible streaming can be achieved through careful patient classification. As we show in Chapters 3 and 4, further stratification of ESI3 patients enabled us to identify a group of patients who can be served in either segments of capacity. This flexibility enables the system to react to under- and over-utilization of capacity segments in real time. Since real time change in streaming policy may be easier and less costly than changing capacity allocation, the flexible streaming policy is a valuable recurs that can be used as an instant recovery strategy for systems with high demand variability and segmented capacity.

In conclusion, healthcare delivery systems, especially those with multi-priority patients, need to consider capacity allocation and patient streaming problems together to best improve patient flow. In this regard, healthcare delivery systems can gain valuable insights from common practices in airlines and hotel industries, such as yield management that has been long used to earn maximal profit by managing perishable capacity. Nonetheless, the investigator needs to consider special characteristics of healthcare systems when adopting solution approaches. These characteristics include but are not limited to presence of knowledge workers, high variability in customer needs and expectations, and possible tragic costs of not satisfying demand.
Bibliography


https://iom.nationalacademies.org/Reports/2006/Hospital-Based-Emergency-Care-At-the-Breaking-Point.aspx


Lindsay, P., Schull, M., Bronskill, S., et al. (2002). The development of indicators to measure


Appendices

Appendix 2-A

**Theorem:** Given any room configuration, \((n_{HF}, n_L, n_F)\), the problem of finding the optimal flow allocation policy that minimizes average waiting times described as \(W = \alpha W_H + W_L\), is jointly convex in flow allocation parameters \((\gamma_{HF}, \gamma_{LF})\).

Proof: We prove this theorem through a series of claims.

Assume that arrival process is Poisson and service duration is exponential. Waiting time is for each patient class at each station is defined by Equations (1)-(3) in Section 3.3. Let \(W = \alpha W_H + W_L\) for \(\alpha \geq 0\), where

\[
W_H = \lambda_{HH} W_{HH} + \lambda_{HF} W_{HF} = \rho_{HH} \frac{n_H}{t_H} W_{HH} + \rho_{HF} \frac{n_F}{t_H} W_{HF}
\]

\[
W_L = \lambda_{LL} W_{LL} + \lambda_{LF} W_{LF} = \rho_{LL} \frac{n_L}{t_L} W_{LL} + \rho_{LF} \frac{n_F}{t_L} W_{LF}
\]

**Claim 1:** \(W_{HH}\) is increasing convex in \(\rho_{HH}\) and \(\rho_{HF}\), \(W_{LL}\) is increasing convex in \(\rho_{LL}\) and \(\rho_{LF}\),

and \(W_0\) is increasing convex in \(\rho_F = \rho_{HF} + \rho_{LF}\).

Proof: Number of studies such as Shanthikumar and Yao (1991) and Lee and Cohen (1983) have proved the convexity of waiting time in utilization in multi-server queueing systems with a single class of customers. Also, from definition:

\[
\rho_{HH} = \frac{\lambda_{HH} t_H}{n_H} = \lambda_H (1 - \gamma_{HF}) \frac{t_H}{n_H} = (\lambda_H - \lambda_{HF}) \frac{t_H}{n_H} = \frac{\lambda_H t_H}{n_H} - \rho_{HF} = \rho_{0H} - \rho_{HF}.
\]

Similarly, \(\rho_{LL} = \frac{\lambda_{LL} t_L}{n_L} - \rho_{LF} = \rho_{0L} - \rho_{LF}\), where \(\rho_{0L}\) and \(\rho_{0H}\) are constant. Since \(\rho_{HH}\) is linear in \(\rho_{HF}\), and \(\rho_{LL}\) is linear in \(\rho_{LF}\), \(W_{HH}\) and \(W_{LL}\) are convex in \(\rho_{HF}\) and \(\rho_{LF}\) respectively.

**Claim 2:** \(W_{HF}\) and \(W_{LF}\) are increasing convex in \((\rho_{HF}, \rho_F)\) when \(\rho_F \geq \frac{1}{2}\).
Proof: Both $W_{HF}$ and $W_{LF}$ are of the form $\frac{1}{1-\rho_{HF}} \tilde{w}$ for $\tilde{w}$ convex in $\rho_F$.

$$\tilde{w}_{LF} = \frac{c\rho^a_F}{1-\rho_F} \text{ and } \tilde{w}_{HF} = c\rho^a_F .$$

$$f(\rho_{HF}, \rho_F) = \frac{1}{1-\rho_{HF}} \tilde{w}$$

$$f_H = \frac{\partial f}{\partial \rho_{HF}} = \frac{1}{(1-\rho_{HF})^2} \tilde{w} + \frac{1}{1-\rho_{HF}} \tilde{w}_H$$

$$f_{HH} = \frac{2}{(1-\rho_{HF})^3} \tilde{w} + \frac{2}{(1-\rho_{HF})^2} \tilde{w}_H + \frac{1}{1-\rho_{HF}} \tilde{w}_{HH}$$

$$f_{HF} = \frac{1}{(1-\rho_{HF})^2} \tilde{w}_F + \frac{1}{(1-\rho_{HF})} \tilde{w}_{HF}$$

$$f_{FF} = \frac{1}{1-\rho_F} \tilde{w}_{FF}$$

By design, $\tilde{w}_H = 0$. Thus $\det \left( \text{Hess} (f) \right) = f_{HH}f_{FF} - f_{HF}^2 = \frac{2}{(1-\rho_{HF})^4} \tilde{w}\tilde{w}_{FF} - \frac{1}{(1-\rho_{HF})^4} \tilde{w}_F^2$. So,

$$\det \left( \text{Hess} (f) \right) = \frac{1}{(1-\rho_{HF})^4} (2\tilde{w}\tilde{w}_{FF} - \tilde{w}_F^2).$$

Case 1. $\tilde{w} = c\rho^a_F$

$$\tilde{w}_F = ca\rho^a_F \text{ and } \tilde{w}_{FF} = ca(a-1)\rho^{a-2}_F \text{. After simplifying } 2\tilde{w}\tilde{w}_{FF} - \tilde{w}_F^2 = c^2a(a-1)\rho^{2a-2}_F \geq 0. \text{ Thus } W_{HF} \text{ is convex in } (\rho_{HF}, \rho_F).$$

Case 2. $\tilde{w} = \frac{c\rho^a_F}{1-\rho_F}$

$$\tilde{w}_F = \frac{ca\rho^{a-1}_F}{1-\rho_F} + \frac{c\rho^a_F}{(1-\rho_F)^2}$$

$$\tilde{w}_{FF} = \frac{ca(a-1)\rho^{a-2}_F}{1-\rho_F} + \frac{2ca\rho^{a-1}_F}{(1-\rho_F)^2} + \frac{2c\rho^a_F}{(1-\rho_F)^3}$$
\[
2\tilde{w}_{FF} - \tilde{w}_F^2 = \frac{c^2}{(1-\rho_F)^4} \left\{ a^2 \rho_F^2(1-\rho_F)^2 + 2a \rho_F^2(1-\rho_F) + \rho_F^2 \right\}
\]

\[
2\tilde{w}_{FF} - \tilde{w}_F^2 \equiv A \rho_F^{2}\left( \rho_F^2 - (1-\rho_F)^2 \right)
\]

\[
2\tilde{w}_{FF} - \tilde{w}_F^2 \geq 0 \text{ if and only if } \rho_F \geq 1-\rho_F \text{ or } \rho_F \geq \frac{1}{2}.
\]

Therefore, we have the convexity result in Claim 2 proven only when the utilization in flexible station is at least 50%. We can argue that in situations when utilization is small (<50%), patient wait times are relatively low as well, making the need for optimization low.

**Claim 3:** \( \rho_{hh} \frac{n_{hh}}{t_h} W_{hh} \) and \( \rho_{ll} \frac{n_l}{t_l} W_{ll} \) are convex in \( \rho_{HF} \) and \( \rho_{LF} \), respectively.

**Proof:** \( W_{hh} \) is convex in \( \rho_{hh} \), therefore \( \rho_{hh} \frac{n_{hh}}{t_h} W_{hh} \) is convex in \( \rho_{hh} \) and thereby in \( \rho_{HF} \).

Similarly, \( \rho_{ll} \frac{n_l}{t_l} W_{ll} \) is convex in \( \rho_{ll} \) and \( \rho_{LF} \).

**Note:** \( g(x) = xf(x) \Rightarrow g'(x) = xf'(x) + f(x) \Rightarrow g''(x) = xf''(x) + 2f(x) > 0 \)

**Claim 4:** \( \rho_{HF} \frac{n_{HF}}{t_h} W_{HF} \) is convex in \( (\rho_{HF}, \rho_F) \) if \( \rho_{HF} \leq \frac{2}{3} \). Similarly \( \rho_{LF} \frac{n_{LF}}{t_l} W_{LF} \) is convex in \( (\rho_{HF}, \rho_F) \).

\[
f(\rho_{HF}, \rho_F) = \left( \frac{\rho_{HF}}{1-\rho_{HF}} \right) \tilde{w}
\]

\[
\tilde{w} = g(\rho_F) = t_F \rho_F^{\frac{1}{2}(n_F + 1)}
\]

\[
\frac{\partial f}{\partial \rho_{HF}} = f_h = \left[ \frac{\rho_{HF}}{(1-\rho_{HF})^2} + \frac{1}{1-\rho_{HF}} \right] \tilde{w} + \frac{\rho_{HF}}{1-\rho_{HF}} \tilde{w}_h
\]

\[
\frac{\rho_{HF}}{1-\rho_{HF}} \tilde{w}_h = 0
\]
\[ \frac{\partial^2 f}{\partial \rho_{HF}^2} = f_{HH} = \left[ \frac{2\rho_{HF}}{(1-\rho_{HF})^3} + \frac{2}{(1-\rho_{HF})} + \frac{1}{(1-\rho_{HF})^2} \right] \tilde{w} \geq 0 \]
\[ \frac{\partial^2 f}{\partial \rho_{HF} \rho_F} = f_{HF} = \left[ \frac{\rho_{HF}}{(1-\rho_{HF})^2} + \frac{1}{(1-\rho_{HF})} \right] \tilde{w}_F \]
\[ \frac{\partial^2 f}{\partial \rho_F^2} = f_{FF} = \left( \frac{\rho_{HF}}{1-\rho_{HF}} \right) \tilde{w}_{FF} \]

\[ \text{det Hessian}(f) = f_{HH} \cdot f_{FF} - f_{HF}^2 = \left[ 2\rho_{HF}^2 + (1-\rho_{HF})(2-2\rho_{HF} + 1) \right] \tilde{w}_{FF} - \left[ \rho_{HF}^2 + (1-\rho_{HF})(1+\rho_{HF}) \right] \tilde{w}_F^2 \]
\[ \geq \left[ \rho_{HF}^2 + (1-\rho_{HF})(1+\rho_{HF}) \right] (\tilde{w}_{FF} - \tilde{w}_F^2) \]

\[ \text{det Hessian}(f) = f_{HH} \cdot f_{FF} - f_{HF}^2 \geq 0 \text{ only if } \rho_{HF} \leq \frac{2}{3}. \] So \( \rho_{HF} \frac{n_F}{t_H} W_{HF} \) is convex if \( \rho_{HF} \leq \frac{2}{3} \).

Similarly for \( \rho_{LF} \frac{n_F}{t_L} W_{LF} \) we can work with \( g(\rho_{LF}, \rho_F) = \left( \frac{\rho_{LF}}{1-\rho_F + \rho_{LF}} \right) \tilde{w} \) or
\[ g(\rho_{HF}, \rho_F) = \left( \frac{\rho_F - \rho_{HF}}{1-\rho_{HF}} \right) \tilde{w} = \left( 1 - \frac{1-\rho_F}{1-\rho_{HF}} \right) \tilde{w} \]

**Claim 5:** \( W = \alpha W_H + W_L \) is convex in \( (\rho_{HF}, \rho_F) \) under the conditions of Claims 2 and 4

**Proof:** \( W \) is a linear function of \( W_H \) and \( W_L \). Claims 1-4 indicate that \( W_H \) and \( W_L \) are both convex in \( (\rho_{HF}, \rho_F) \). Therefore, \( W \) is also convex in \( (\rho_{HF}, \rho_F) \). Since \( (\rho_{HF}, \rho_F) \) are linear in flow parameters \( \gamma_{HF} \) and \( \gamma_{LF} \), \( W \) is convex in \( (\gamma_{HF}, \gamma_{LF}) \) under conditions specified in Claims 2 and 4.
Appendix 2-B

Claim: When traffic intensity of emergency high-priority patients, u, is strictly larger than traffic intensity of non-emergency low-priority patients, partial flexibility by dedicating to high-priority patients can reduce waiting time for low-priority patients better than total flexibility.

Let \( u = \frac{t_{H}^{\lambda_{H}}}{N} \) and \( v = \frac{t_{L}^{\lambda_{L}}}{N} \) be the traffic intensity of high- and low-priority patients respectively. We compare the total flexibility policy (Case 1) with partially flexible policy (Case 2) in which some rooms is dedicated to high-priority patients. For ease of calculation we consider two asymptotic situations: a) when \( u \) is strictly less than \( v \), and b) when \( u \) is strictly greater than \( v \).

Case 1: Total flexibility. \( n_{F} = N \) and \( n_{H} = n_{L} = 0 \)

\[
W_{H} = W_{HF} \approx \frac{(u + v)^{N}}{N(1-u)} \\
W_{L} = W_{LF} \approx \frac{(u + v)^{N}}{N(1-u)(1-u-v)}
\]

Case 2: Partial flexibility, dedicating to H. \( n_{F} = N - n_{H} \) and \( n_{L} = 0 \)

\[
\begin{align*}
\rho_{HF} &= \frac{\gamma \lambda_{H} t_{H}}{N - n_{H}} = \frac{\gamma Nu}{N - n_{H}} \\
\rho_{HH} &= \frac{(1-\gamma)\lambda_{H} t_{H}}{n_{H}} = \frac{(1-\gamma)Nu}{n_{H}}
\end{align*}
\]

\[
\hat{W}_{HF} = \frac{(u' + v)^{N-n_{H}}}{(N - n_{H})(1-u')} \quad , \quad \hat{W}_{HH} = \frac{u'^{n_{H}}}{n_{H}(1-u'')}
\]

\[
\hat{W}_{LF} = \frac{(u' + v)^{N-n_{H}}}{(N - n_{H})(1-u')(1-u' - v)}
\]

Let \( \Delta W_{L} = \hat{W}_{LF}^{1} - \hat{W}_{LF}^{2} \) be the difference between low-priority waiting in the two cases above. Then \( \Delta W_{L} > 0 \) implies that partial flexibility results in smaller waiting time for low-priority patients. We analyze \( \Delta W_{L} \) under the two asymptotic conditions.
Condition a) $u \ll v$

$$\Delta \tilde{W}_{LF} = \frac{v^N}{N(1-v)} - \frac{v^{N-n_H}}{(N-n_H)(1-v)} = \frac{(N-n_H)v^N - Nv^{N-n_H}}{N(N-n_H)(1-v)} < 0$$

$$N - n_H - Nv^{-n_H} < 0 \quad \text{because} \quad N - n_H < N \quad \& \quad \frac{N}{v^{-n_H}} > N$$

Thus under this condition, dedicating to high-priority patients increases waiting time for low-priority patients.

Condition b) $u \gg v$

$$\Delta \tilde{W}_{LF} = \frac{u^N}{N(1-u)^2} - \frac{u^{N-n_H}}{(N-n_H)(1-u)^2}$$

If we replace $u'$ with $\frac{yNu}{(N-n_H)}$, then

$$\Delta \tilde{W}_{LF} = \frac{u^N}{N(1-u)^2} - \frac{(yNu)^{N-n_H}}{(N-n_H)^{N-n_H+1}(1-u)^2}$$

$$\Delta \tilde{W}_{LF} > 0 \text{ if } \frac{u^N(N-n_H)^{N-n_H+1}(1-u)^2 - N(1-u)^2(yNu)^{N-n_H}}{N(1-u)^2} > 0$$

Thus $\Delta \tilde{W}_{LF} > 0$ if

$$\frac{(1-u)^2}{u^{N-n_H}} > \frac{N(1-u)^2}{u^{N-n_H}+1}$$
## Appendix 3

**Table 3.A.** Independent patient attributes available at triage used in the partitioning analysis.

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>levels</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systolic Blood Pressure</td>
<td>70-240</td>
<td>Arterial blood pressure during heart contraction</td>
</tr>
<tr>
<td>Diastolic Blood Pressure</td>
<td>40-120</td>
<td>Arterial blood pressure between heartbeats</td>
</tr>
<tr>
<td>Pulse Rate</td>
<td>40-180</td>
<td>Heartbeats per minute</td>
</tr>
<tr>
<td>Body Temperature</td>
<td>96-104</td>
<td>Core body temperature</td>
</tr>
<tr>
<td>Respiratory Rate</td>
<td>8-36</td>
<td>Number of breaths per minute</td>
</tr>
<tr>
<td>Oxygen Saturation (SpO2)</td>
<td>88-100%</td>
<td>Arterial oxygen saturation in the blood</td>
</tr>
<tr>
<td>Means of Arrival</td>
<td>Aircare/ EMS/ Walk-in</td>
<td>Patient means of arrival to the ED</td>
</tr>
<tr>
<td>Laboratory test</td>
<td>0/1</td>
<td>1 if patient needs lab tests, 0 otherwise</td>
</tr>
<tr>
<td>EKG (electrocardiogram)</td>
<td>0/1</td>
<td>1 if patient needs an EKG, 0 otherwise</td>
</tr>
<tr>
<td>Radiology</td>
<td>0/1</td>
<td>1 if patient needs Radiology imaging, 0 otherwise</td>
</tr>
<tr>
<td>Age</td>
<td>1-90</td>
<td>In years</td>
</tr>
<tr>
<td>Gender</td>
<td>Male/Female</td>
<td></td>
</tr>
</tbody>
</table>
Appendix 4

Figures 4.A-1 and 4.A-2 show patients average waiting time distributions under different dynamic policies in the fast track and the main ED respectively. Each data point is the corresponding average waiting time in one replication of the simulation model under the given policy. Figure 4.A-3 shows the distribution of average patients’ waiting time under static policies.
Figure 4.A-1. Distribution of waiting times in the fast track under dynamic policies: (a) All to MED with jockeying, (b) shorter queue (c) Smaller utilization- ties to FT, (d) Smaller utilization- ties to MED, and (e) time-dependent.
Figure 4.A-2. Distribution of waiting times in the main ED under dynamic policies: (a) All to main ED with jockeying, (b) shorter queue (c) Smaller utilization- ties to FT, (d) Smaller utilization- ties to main ED, and (e) time-dependent.
Figure 4.A-3. Distribution of average patients’ waiting time under static policies in: a) the main ED, and b) the fast track