Optimal Decision Making for Healthcare Operations: Models and Implementation

by

Liangyuan Na

B.A., University of California, Berkeley (2018)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Author . . . .

Sloan School of Management
May 4, 2023

Certified by . . . .

Dimitris Bertsimas
Boeing Leaders for Global Operations Professor of Management
Associate Dean for Business Analytics
Thesis Supervisor

Accepted by . . . .

Georgia Perakis
William F. Pounds Professor of Management Science
Co-Director, Operations Research Center

Optimal Decision Making for Healthcare Operations: Models and Implementation

by

Liangyuan Na

Submitted to the Sloan School of Management
on May 4, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

Healthcare systems face financial and operational challenges to provide high-quality patient care. Advances in machine learning and scalable optimization have led to opportunities for utilizing analytics to improve healthcare operations. However, only a limited number of models have been extensively deployed in practice due to the complex nature and strict regulations of healthcare systems. This thesis aims to develop and deploy practical analytics models that support strategic, tactical, and operational decision making in healthcare systems.

A large part of the thesis involves close collaborations with Hartford HealthCare (HHC), the largest hospital network in Connecticut, spanning seven hospitals with $5 billion annual revenue. In Chapter 2, we optimize nurse staffing at the Emergency Department (ED) of Hartford Hospital. We develop a two-phase methodology: (a) a robust optimization model to allocate aggregate staffing levels, followed by (b) mixed integer optimization models to schedule each nurse. Then in Chapter 3, we develop machine learning models to predict eight patient operational outcomes related to discharge, mortality, and intensive care for all inpatients at seven hospitals. We build an online daily pipeline from data extraction to prediction-driven decision support.

More importantly, we implement our models into a two-module end-to-end software deployed in large-scale production, supporting daily decision making of over 400 users of doctors, nurses, and managers across seven hospitals at HHC. The nurse scheduling module provides a labor-free process from input collection to schedule output, improving patient coverage and nurse satisfaction with reduced cost. The patient outcome prediction module is deeply integrated into medical providers' daily workflow, identifying timely discharges and patient exacerbation. HHC reports better staff workflow and patient care, together with substantial benefits in reduced length of stay and increased financial margins.

In the final part of the thesis, we collaborate with a mobile health (mHealth) application, Hearsteps, designed to reduce sedentary behavior and promote physical activity in individuals with hypertension. In Chapter 4, we develop innovative batch off-policy learning methods to optimize the app's digital intervention by sending anti-sedentary messages to users. Our interpretable decision tree-based policy improves treatment effects and guides future clinical trials.

Thesis Supervisor: Dimitris Bertsimas
Title: Boeing Leaders for Global Operations Professor of Management
Associate Dean for Business Analytics

Acknowledgments

First and foremost, I express my deepest gratitude to my advisor, Dimitris Bertsimas, for his phenomenal mentorship and influence on me over the past five years. I could not have asked for a better advisor, who truly believes in me, motivates me, and guides me to be a better researcher and person. Dimitris inspires me to genuinely help people and make a positive impact in the world, especially in healthcare. He strikes me with his boundless energy, pioneering ideas, and remarkable courage. Beyond research, he taught me communication, leadership, endurance, and patience. I learned how to collaborate with different stakeholders, make changes to the world, push back occasionally despite my personality, and never give up. I am touched by his deep care for his students and for me. I treasure our memorable times together: (more than) weekly meetings, 8 am Zoom calls with hospitals, road trips to Connecticut, half-day-long classes, celebrations at INFORMS, and discussions about life, among many others. Thank you, Dimitris, for being my role model, caring mentor, close friend, and teammate on my side - I will always be grateful to you.

I would like to thank my thesis committee members, Jónas Jónasson and Nikos Trichakis, for serving on my committee and providing support and guidance to my work. I thank Jónas for teaching me different perspectives on healthcare delivery and providing me with an action-learning experience in the Healthcare Lab class, which prepared me to work on healthcare operations. Furthermore, I greatly appreciate our extensive discussions of my work which he provided excellent suggestions. I thank Nikos for providing valuable feedback on my work over the years of my PhD, since he chaired my General Examination committee. I have been deeply inspired by his work on robust optimization and healthcare operations, and his insights have been instrumental in my research. I also want to thank Alexandre Jacquillat for serving on my General Examination committee.

I am extremely fortunate to have collaborated with incredible professors, who mentored me significantly and helped me grow as a researcher. I cannot thank Jean Pauphilet enough for all his help, which contributed significantly to my thesis and provided invaluable advice for my PhD. I deeply admire his kindness, his extensive knowledge across optimization, statistical learning, and hospital applications, as well as his outstanding research capabilities and creativity. Our Monday meetings have been pivotal moments for me, resolving any research questions and bottlenecks with his consultation and fading all my concerns away with his optimism and encouragement. I would like to thank Bartolomeo Stellato for closely working with me on robust optimization and patiently mentoring me during the earlier times of my PhD. He gave me invaluable advice on proofs, papers, and presentations, and introduced to me programming practices, parallel computing, and perfect TikZ figures. These fundamental skills have been beneficial for my later research. I am grateful to Susan Murphy for our exciting collaboration on the mobile health work of my thesis, and for introducing me to batch off-policy learning. I am highly motivated by her energy and passion for digital health. I also thank our collaborator Predrag Klasnja who provided excellent insights from the clinical trials perspective. I thank my undergraduate research advisors from Berkeley, Anil Aswani and Mariana Olvera-Cravioto, for introducing me to research in the first place.

A significant milestone of my PhD journey is the extensive and close collaboration with Hartford HealthCare (HHC) on the Holistic Hospital Optimization (H2O) initiative. It was the most rewarding experience for me to deploy my models in multiple hospitals at large scales and see the impact of improving daily healthcare operations and patient outcomes. This unique partnership and achievement would not have been possible without the exceptional team from HHC. Barry Stein, Chief Clinical Innovation Officer of HHC, gave us excellent vision, kind support, generous resources, and strong championship, which played an important role in the successful implementation. Daniel Kombert, Associate Vice President of Medical Affairs, led the HHC team for the patient outcome prediction project, provided excellent feedback to model development and evaluation from the medical perspective, and coordinated the gradual rollout of our tool. Melissa Boisjoli-Langlois and Andrew Castiglione, our earliest and closest physician champions, showed me how patient cases are analyzed in progression rounds and how the predictions can be improved from the medical perspective.

I had the pleasure to interact with many doctor users of the tool, especially Pooja Hebbal and Maram Khalifa, together with Melissa, Andrew, Dan, Barry, and many more recent users, whose feedback was critical in advancing the model improvement and implementation.

The nurse staffing project would not have been possible without the amazing nurse managers, Patricia Veronneau and Audrey Silver. We spent numerous meetings talking about the complexities and constraints of scheduling, and I am genuinely touched by their great energy in the extremely high-pace environment, tireless efforts in improving the nurse experience, and persistence in overcoming obstacles in the project. The implementation also received excellent directions from Cheryl Ficara, Senior Vice President of Operations at HHC, as well as much assistance from other nurses, especially Nicole Vogt, Yolanda Johnson, and Katherine Martinez-Taveras. Regarding other important components of H2O, I would like to acknowledge the support of the IT department, especially Frank Damiano and Qun Yu, for their help in data extraction, as well as the Care Logistics Center, and in particular Chris Biernat, Katherine Battle, and Elizabeth Ciotti, for showing me how every single patient move is made. I thank them all for their hard work, great partnership, weekly meetings, passion for new models, and determination in overcoming challenges together. I learned that healthcare people are one of the kindest in the world and I am deeply inspired by their true care for patients and their persistence in improving healthcare. The valuable opportunity to visit HHC multiple times was eye-opening and inspiring for me to see how the hospital operates in my eyes and was memorable to meet and bond with the best healthcare collaborators I could have asked for. Furthermore, the software implementation would not have been made possible without Ali Haddad-Sisakht from Dynamic Ideas LLC, whose tremendous work and accommodation to any user request are greatly appreciated. I also thank Louis Raison for his hard work in staffing software automation and the development team for the interface front-end development.

Another thing I love about MIT is working with the brightest and kindest fellow students. I feel extremely grateful to have Kimberly Villalobos Carballo, who is not only an incredible teammate but also a close friend. I am fortunate to get to know her as an exceptional teacher, researcher, and person, whom I admire for her kind heart, sharp mind, fast actions, hard work, great empathy, open honesty, and much more. She greatly influences me with her positive energy, outgoing personality, curiosity, creativity, and determination. We shared many moments together: the ups and downs, laughsters and tears, research meetups and debugging sessions, catchups and conversations, music and dances. Thank you, Kim, for always being there to offer me the warmest help and support, and I am so excited about all the amazing things you will achieve next. I treasure the magic to share many things with Yu Ma - born on the same day, from Berkeley to the ORC, we continued to work on projects and TA classes together in the research group, and of course, celebrate our joint birthdays. The Holistic Artificial Intelligence in Medicine (HAIM) initiative brought me many other awesome collaborators and friends: Especially, I thank Cynthia Zeng for our deep talks at girls' nights, Leonard Boussioux for his inspiring enthusiasm, Michael Li for his support and apartment parties, Holly Wiberg for guiding and encouraging my first TA experience, Luis Soenksen for bringing new ideas and expertise, and Ignacio Fuentes for his support and resources from MIT Jameel Clinic.

I am also incredibly lucky to have many wonderful friends and mentors at the Operations Research Center (ORC). Yuchen Wang was instrumental in helping me navigate the ORC since day one, offering valuable support and guidance for me to learn and grow. I also benefited greatly from the wisdom and advice of other senior students like Arthur Delarue, Jack Dunn, Patricio Foncea, Jing Lu, Sebastien Martin, Agni Orfanoudaki, Ted Papalexopoulos, Deeksha Sinha, Li wang, Andy Zheng, Daisy Zhuo, as well as the great support of people from my cohort. I thank many friends for making my time memorable in the ORC: Amine Bennouna, Moise Blanchard, Wenyu Chen, Qinyi Chen, Raluca-Ioana Cobzaru, Shuvomoy Das Gupta, Vassilis Digalakis, Andreea Georgescu, Xiaoyue Gong, Adam Kim, Driss Lahlou Kitane, Jason Liang, Manuel Pelaez, Rainy Niu, Yuan Shi, Fransisca Susan, Leann Thayaparan, Zikai Xiong, Evan Yao, Ghali Zerhouni, Sabrina Zhai, Emily Zhang, Renbo Zhao, Jiayu Zhao, and many others. I thank the Executive MBA students I taught and MBAn students I mentored, who made me experience the joy of teaching. I am grateful to Georgia Perakis and Patrick Jaillet for their dedication to keeping the ORC a vibrant place for students, and for checking in and supporting me along the way. I thank Laura Rose and Andrew Carvalho for their help with all our questions.

I would like to thank my friends and family who have always been by my side. I thank my friends who keep me company in Boston, folks who play music with me, my friends from college and from China. I want to give special thanks to Sebastian Palacios for the great support throughout my PhD and to Anais Miller, Jaewon Saw, and Hilary Wang for their support from the west coast. I am extremely grateful to my aunt and uncle, Ning Liang and Chang Yu, who introduced me to the possibility of coming to the US in the first place, convinced me of pursuing a PhD, and gave me invaluable advice over the years. I thank their family with my dear cousins, Sheeline Yu and Sheerea Yu, for always welcoming me to a warm home in the US. Last but not the least, I express my wholehearted gratitude to my family in China. In particular, I am indebted to my grandparents, Shuren Huang and Weizhuang Liang, who played such an important role in raising me and are always truly proud of me. I owe everything to my parents, Shuang Liang and Yan Na, who made immense sacrifices for me to study in the US and love me under no conditions. This thesis is dedicated to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We re-imagine a patient's journey in healthcare systems for the near future:

Encountering a medical emergency, the patient calls the ambulance which takes her to an emergency department (ED) of a hospital network. The ED is well-prepared for the new arrival with additional nurses scheduled by automatic staffing software, which detected a recent increase in ED volume ahead of time. Skillfully trained nurses quickly triage the patient into a treatment room, where nurses provide high-quality care for the patient with sufficient resources. Meanwhile, the hospital database updates the patient's medical record and sends an anticipated hospitalization request. A real-time capacity prediction and bed allocation system admits the patient into an available intensive care unit (ICU) inside the hospital.

During hospitalization, a machine learning tool guides the caring team on various decisions. As the patient stabilizes, the tool predicts the likelihood of leaving the ICU and updates the bed management center with an expected ICU capacity in the upcoming days. The care provider team monitors her mortality risk predictions daily, receiving a red alert for an increase in the risk score. After several days of improvement, the tool sends a green alert predicting a discharge to a skilled nursing facility in the next 48 hours. The case management team begins contacting nurse facilities, obtains signatures for authentication, and prepares for a timely discharge, until the attending physician orders an informed discharge.

After several weeks at the facility, the patient is discharged home with continued care provided by digital health services. She wears an activity tracker that monitors her heart rate and steps taken every minute, and the data gets passed to a mobile health application. After several hours of lack of movement, she receives a phone notification suggesting a one-minute exercise. Calculated with a maximized expected treatment effect by the built-in reinforcement learning algorithm, the message is sent just in time to reduce the patient's sedentary time and remain healthy.

The pathway toward this vision of smarter healthcare requires a joint effort by operations researchers and healthcare workers in two aspects. First, models: leveraging big data and artificial intelligence to develop methodologies to tackle operational bottlenecks. Second, implementation: putting the innovations in deployment to transform healthcare practices.

1.1 Decision Making in Healthcare Operations

We study three problems in different healthcare organizations and develop methodologies from optimization, machine learning, and policy learning to support their decision making processes. The implications benefit a wide range of stakeholders in healthcare systems, including doctors, nurses, human resources, case managers, bed managers, mobile health app designers, and most of all, patients.

1.1.1 Nurse Staffing Optimization

The ED is essential in treating urgent patients and managing a large source of patient arrivals into the hospital. One challenge is to schedule the "right" number of nurses, i.e., large enough to accommodate future patient demand but not too large to save limited staffing resources and nursing hours. Another challenge is to assign a desirable schedule to each nurse for their higher satisfaction. To improve patients' and nurses' experiences, optimizing nurse staffing and scheduling is necessary for the ED. Improved staffing levels that better match patient demand can improve cost-effectiveness while meeting patient need with fewer staff. Individual nurse schedules that better satisfy each nurse's preferences can increase nurse satisfaction. Additionally, increased nurse training can strengthen nurse capacities and resources. In Chapter 2, we partner with an ED to develop and implement such optimization models in practice, to create a positive impact on nurses and patients as well as the overall performances of the ED and the hospital (Bertsimas et al. 2023a).

1.1.2 Patient Outcome Prediction

Access to accurate predictions of patients' outcomes can enhance the medical staff's decision-making, which ultimately benefits all stakeholders in the hospital. Anticipating short-term discharges informs about bed availability and can facilitate resource utilization while identifying discharge barriers provides clinical guidance to personalize the delivery of care. Furthermore, detecting patients with high mortality or ICU risk (or changes thereof) can alert the medical team and call their attention to those who need it the most. More broadly, discharge is a complex process involving the coordination of different stakeholders and resources, which could be anticipated given accurate predictions on discharge disposition early in a patient's stay. There is a collection of models for such inpatient flow predictions in the literature. We refer to Awad et al. (2017a) for a comprehensive survey on mortality and length of stay (LOS) predictions, and to Mees et al. (2016) for discharge destination predictions. Accordingly, in Chapter 3, we develop machine learning models that predict various patient outcomes, including imminent discharge, mortality risk, discharge disposition, and ICU risk, and deploy the models in a large hospital network (Na et al. 2023).

1.1.3 Mobile Health Intervention

To promote healthy behaviors, many mobile health applications provide message-based interventions, such as tips, motivational messages, or suggestions for healthy activities. Ideally, the intervention policies should be carefully designed so that users obtain the benefits without being overwhelmed by overly frequent messages. As part of the HeartSteps (Liao et al. 2018, Klasnja et al. 2019) physical-activity intervention, users receive messages intended to disrupt sedentary behavior. HeartSteps uses an algorithm to uniformly spread out the daily message budget over time but does not attempt to maximize treatment effects. This limitation motivates constructing a policy to optimize the message delivery decisions for more effective treatments. Moreover, the learned policy needs to be interpretable to enable behavioral scientists to examine it and to inform future theorizing. Thus in Chapter 4, we address this problem by learning an effective and interpretable policy that reduces sedentary behavior (Bertsimas et al. 2022).

1.2 Deployment in a Large Hospital Network

Apart from tackling these problems with novel methodologies, the thesis also highlights successful implementations in a large-scale hospital network. Specifically, the work presented in Chapters 2 and 3 is deployed in Hartford HealthCare (HHC), the largest hospital system in Connecticut. The projects are part of a broader initiative named Holistic Hospital Optimization (H2O), where we improve hospital operations holistically with analytics tools. In this section, we provide an overview of HHC and discuss the main challenges encountered during the collaboration. Moreover, we summarize key elements that contributed to the successful implementation, sharing insights and inspirations for future efforts of developing and deploying other decision-making support tools in healthcare organizations.

1.2.1 Hartford HealthCare

Hartford HealthCare is the largest and most comprehensive healthcare network in Connecticut. With 36,000 employees, HHC operates in over 400 locations, including seven acute care hospitals, several behavioral health, physical therapy, and rehabilitation facilities, a multi-specialty physician group, a clinical care organization, services of regional home care and senior care, and a mobile neighborhood health program. In 2021, the system covered 104,696 transitions from inpatient care (i.e., movement from HHC to another healthcare setting), 555,358 patient days of hospitalization, and 406,949 emergency department visits, generating an operating revenue of $5 billion.

The unified network provides a high standard of care with enhanced access, affordability, and equity in crucial specialties and institutes. HHC contains seven diverse hospitals, ranging from Hartford Hospital (HH), one of the largest teaching hospitals in New England (867 licensed beds) to smaller (150 beds) community hospitals.

We further present in more detail the seven hospitals of the network. The main hospital of HHC, Hartford Hospital, is one of the largest teaching hospitals in New England, in collaboration with the University of Connecticut School of Medicine. HH is a tertiary hospital, recognized to be high performing in a variety of procedures, conditions, and specialties. Backus Hospital (BH) is an acute-care community teaching hospital and a trauma center in the east region, with an outstanding specialty in stroke. Charlotte Hungerford Hospital (CH) is a general acute care community hospital serving a healthcare resource in the northwest region. The Hospital of Central Connecticut (HOCC) is a central region community teaching hospital providing comprehensive inpatient and outpatient services in a variety of specialties and participating in residency programs with the University of Connecticut School of Medicine. MidState Medical Center (MMC) is another community hospital serving the central region with a wide collection of services. St. Vincent's Medical Center (SV) is a tertiary community teaching hospital and mission-driven Catholic hospital that provides care with special attention to the most vulnerable and poor patients. Windham Hospital (WH) is a community hospital in the east region and does not provide an ICU level of care. Descriptive information and statistics about the seven hospitals are provided in Table 1.1.

1.2.2 Challenges in Healthcare Analytics

Developing and implementing new analytics models in the healthcare field poses numerous challenges. Based on our experience of deploying H2O decision-making support tools in Hartford HealthCare, we summarize various challenges encountered during different phases of the projects.

Table 1.1: Descriptive Statistics of the Seven Hospitals in HHC in 2021.

Hospital	Beds	Units	Services	Highest Level	Patient Days	Operating Revenue
BH	233	13	38	Critical Care	52,328	$449.9 million
CH	122	11	18	Intensive Care	27,912	$175 million
HH	867	47	48	Intensive Care	261,954	$2 billion
HOCC	446	20	34	Critical Care	76,325	$552.8 million
MMC	156	12	31	Intensive Care	39,972	$385.4 million
SV	520	26	43	Intensive Care	85,322	$466.5 million
WH	130	5	14	Step Down	11,545	$127.5 million

Regulation hurdles in model development. The healthcare industry places the highest priority on patient safety and outcomes, leading to a highly conservative and rigorous environment. Hospital institutions, in particular, are subject to high risks at stake and stringent regulations. Starting analytics projects in hospitals requires the initiatives to be documented, reported, and evaluated in detail. Obtaining patient data, especially transferring electronic medical records out of hospitals' fireworks, is extremely challenging, subject to data privacy concerns and Institutional Review Board protocols. In addition, some hospitals have inadequate historical data recording systems, unorganized data structures, and limited capabilities to extract data frequently from their IT systems. These bottlenecks can cause significant delays, up to months or even years to build optimization and machine learning models. These hurdles remain present throughout the development phase, for example, when requesting additional data and project expansion to improve and extend models.

Conservatism in result screening. After developing the models, conveying results to healthcare stakeholders is another critical phase. The medical staff needs to build understanding, trust, and approval of the resulting models. To prevent adverse effects on patients from using the models, doctors and nurses seek confident model interpretations that align with their medical knowledge, as well as high standards for accurate results and low error tolerance. Hospital leadership relies on rigorous evidence of significant margins of improvement to consider incorporating the results, especially given the potential changes to medical staff's workflows. Some stakeholders have disagreements due to reasons such as unwillingness to change schedules, restrictions by contract, hesitations to adopt new machine intelligence, and reluctance to overcome decades of conventions in place. Pushing back such resistance to achieve consensus is necessary for large hospital institutions to move forward. Overall, the complications of involving humans in the decision making loops can be larger than technical complexities by an order of magnitude, leading to a slow process of implementing new tools at a larger scale.

Complexity of decision-making process integration. After obtaining consensus on implementing the changes, adopting new tools in medical decision-making processes is deemed the most challenging phase. As healthcare workers need a highly efficient workflow, developing automated user-friendly end-to-end software implementation is critical. Injecting new software with existing commercial software in use poses considerable challenges involving third parties and might not be feasible, in which case healthcare workers need additional steps to use the new tool. The integration of human-machine decision-making requires further investigation. As machines may not capture all clinical and operational factors to make complex human decisions, medical users need to be able to abide by or override the model's decisions when in disagreement. Moreover, some cases require medical decision-makers to run models on demand to incorporate dynamic information changes, which needs training in new technical skills. Addressing these challenges requires close collaboration between researchers and healthcare providers to upgrade the system.

1.2.3 Key Factors of Successful Implementation

From our journey of overcoming these challenges and implementing advanced analytics solutions on a large scale, we identify three key components that contributed to our success.

Close collaboration with Hartford HealthCare. Over the past decade, HHC and MIT have been closely collaborating to improve decision-making in various parts of the healthcare system, conducting over 20 analytics-based projects together. For example, the HHC-MIT partnership developed machine learning (ML) models to quantify different clinical risks such as stroke risk (Orfanoudaki et al. 2020), COVID-19 mortality risk (Bertsimas et al. 2020), neutropenic risk (Wiberg et al. 2021), and postoperative outcomes (Orfanoudaki et al. 2022). The success and positive impact of these projects in different parts of HHC built the necessary trust to envision larger network-wide implementations of analytics tools under the H2O initiative. We interact closely with different stakeholders, ranging from hospital leadership, doctors, nurses, information technology (IT), case management teams, via weekly meetings and visit trips. From iterations of improving models based on feedback from our clinical collaborators, we create a sense of co-ownership together which is essential for successful implementation.

End-to-end software integration. For an initiative of this scale, we also partnered with Dynamic Ideas LLC, a data consultancy company with deep expertise in the implementation of operations research and analytics tools. Together we developed software interfaces to share model outputs with medical users in a convenient, interpretable, and streamlined way. The software features model explanations facilitating transparency and trust in models and interactive adjustment supporting editing and overriding the outputs. To further facilitate the integration of the H2O tools, we provided tutorials and lectures about machine learning to doctors at HHC and reviewed human-machine cooperation regularly to improve the new decision-making process. This end-to-end software integration coupled with multi-party collaboration significantly accelerated the adoption of the tool and its continuous improvement.

Staggered progression over time. Our success in two large-scale implementations at HHC also lies in the way we organized the progression of the projects over time. Instead of implementing all models for all the HHC hospitals simultaneously, we adopted a gradual approach and started with HH, the main hospital of the network. We first tested the patient outcome prediction tool with four physician champions who are lead hospitalists of five medical units at HH. After a positive evaluation of the impact of this pilot use, HHC decided to extend our work to other hospitals in two batches, three more hospitals at a time, until fully deployed in all hospitals. Compared with the prediction support tool, the nurse staffing implementation directly making decisions on 200+ nurses' schedules and changing their lives was even more challenging effort. We pushed the incorporation at HH gradually, starting from offline strategic adoption and transitioning into online operational use. Through numerous back-and-forth iterations, the ED first adjusted hiring strategies to accommodate our staffing level recommendations and slowly accommodated changes in nurses' schedules. Eventually, the software automated online scheduling, where nurses enter preferences as needed and nurse managers generate schedules dynamically. After the staggered progression, HHC plans to extend the innovation to all nurse staffing in seven hospitals in the future.

1.3 Overview of Main Contributions

We summarize our solutions to address the aforementioned problems and challenges in correspondence to the outline of the thesis. Our key contributions are two-fold, spanning from models with demonstrated benefits through experiments to implementation with impacts created in practice.

1.3.1 Models Pushing State-of-the-art

Chapter 2. We present an integrated approach for optimizing nurse staffing at the Emergency Department (ED) of Hartford Hospital, leveraging a combination of data, optimization, and software. We develop and implement two-phase optimization models to schedule nurses for 12-hour shifts across different positions over each 6-week staffing cycle. In the first phase, we develop a robust optimization model to allocate staffing levels for uncertain patient demand. Given the optimized levels along with nurse preferences, we then develop a pair of mixed integer problems to generate individual schedules including work, trainee, and preceptor shifts for each nurse. Experimental results demonstrate that our proposed approach leads to less costly (5–8%) staffing with more coverage of patient care (8–25%) and higher nurse satisfaction (5%). Compared with the previous schedule, nurses can work fewer shifts on weekends (17%), holidays (14%), and overtime (85%) as well as be assigned to more diverse positions (3.6) and more daily training opportunities (0.95).

Chapter 3. In collaboration with Hartford HealthCare, we develop machine learning models that predict short-term and long-term outcomes for all inpatients across their seven hospitals. In particular, we predict the probability of patients being discharged and that of patients being transferred from/to an intensive care unit in the next 24/48 hours, as well as the probabilities of mortality and other discharge dispositions. All models achieve high out-of-sample accuracy (in the 75.7%–92.5% range) and are well-calibrated.

Chapter 4. We design an effective and interpretable policy to optimize digital intervention delivery for more effective treatments in a mobile health application. We make decisions for HeartSteps physical-activity intervention to send users messages intended to disrupt sedentary behavior. We propose Optimal Policy Trees + (OPT+), an innovative batch off-policy learning method, that combines personalized threshold learning and an extension of Optimal Policy Trees under a budget-constrained setting. We implement and test the method using data collected in HeartSteps V2/V3 clinical trials. Computational results demonstrate a significant reduction in sedentary behavior with a lower delivery budget. OPT+ produces a highly interpretable and stable output decision tree thus enabling theoretical insights to guide future research.

1.3.2 Implementation Impacting Real-world Systems

Chapter 2. We implement our framework into an automated end-to-end scheduling optimization software, which is deployed for use at Hartford Hospital since March 2023. The software collects preferences from 200 ED nurses, enables managers to generate optimized schedules based on decision preferences, and provides guided dynamic adjustments until exporting final daily schedules. This transformative implementation streamlines a labor-free staffing process and delivers robustly sufficient, cost-effective, and desirable schedules. Ultimately, it benefits nurses, patients, hospital leadership, and other stakeholders at Hartford Hospital, demonstrating the potential of revolutionizing nurse staffing at healthcare institutions.

Chapter 3. We implement an automated pipeline that extracts new data every morning and dynamically updates our predictions, as well as user-friendly software and a color-coded alert system to communicate these patient-level predictions (alongside explanations) to the clinical teams. We have been gradually deploying the tool, and training medical staff at Hartford HealthCare since May 2022. As of January 2023, over 200 doctors, nurses, and case managers across seven hospitals use it every day in their patient review process. In addition to increased accuracy in discharge date prediction, we observe a significant reduction in average length-of-stay following the tool's adoption and anticipate substantial financial benefits for the system.

Chapter 2

Optimization Automates Nurse Scheduling at Hartford Hospital Emergency Department

2.1 Introduction

The Emergency Department (ED) is a crucial component of hospitals, providing urgent care to patients in need. However, ED overcrowding can lead to increased waiting times, compromise care, and result in adverse outcomes for patients, such as dissatisfaction and increased mortality rates (Bernstein et al. 2009). Additionally, ED congestion can affect patient flows into the hospital, further impacting patient care (Elder et al. 2015).

Nurses play a vital role in providing care to patients, especially ED patients in critical conditions. However, healthcare systems face a worsening nurse shortage and resource limitation, which hinders their ability to meet nurse-patient-ratios and thus leads to higher patient mortality, higher nurse burnout, and dissatisfaction (Aiken et al. 2002). To patients, higher hours of care by registered nurses, lower nurse workload, and higher nurse skillset are associated with reduced mortality rate and shorter hospital stays, among various patient outcome improvements (Needleman et al. 2002, Aiken et al. 2014, Twigg et al. 2019). To nurses, overtime hours during understaffing periods could result in employee attrition, and dissatisfaction with schedule flexibility is also related to more intention to leave (Leineweber et al. 2016). To the whole system, shortage of nurse staffing also negatively impacts throughput metrics in the ED such as increased ED stay and leaves without being seen (Ramsey et al. 2018).

Our objective is to develop nurse staffing optimization models that address the aforementioned challenges in ED operations. More importantly, we aim to implement the models to bring tangible benefits to the ED. The latter is not achievable without close collaboration with an ED in practice. In the next section, we introduce our partner ED and Hartford Hospital and describe their nurse staffing problem.

2.1.1 ED Nurse Staffing Problem at Hartford Hospital

As introduced in Section 1.2.1, Hartford Hospital is HHC's flagship facility, a teaching hospital and a tertiary care center with 867 beds. With over 160 years of experience, the hospital serves over 40,000 discharges and 100,000 ED visits annually.

The ED at HH has around 200 nurses, who work in different positions of two categories: management positions covering logistics as well as "pods" treating patients. The set of management positions includes Chief Nurse Leadership (CNL), first nurse, resource, triage, and Front End Provider (FEP). There are various pods to treat patients from different categories, including the main pods (consisting of blue, green, and orange pods and their hallways with additional capacity, for the majority of patients), red pod (for resuscitation/emergent patients), purple pod (for behavioral patients), iTrack (for less urgent patients), and Emergency Department Observation Unit (EDOU for observation patients). During the early surge of COVID-19, HH opened up a new trailer pod at the ED to treat COVID-19 patients and removed hallway pods that could have exposed patients to infection. Nurses are classified into different tiers based on their years of experience and qualifications for these positions. The nurses work on one of three shift types (7 am–7 pm, 11 am–11 pm, and 7 pm–7 am). Every 6 weeks, a staffing schedule assigns nurses to shifts at these positions throughout the period. Additionally, the leadership can arrange training for nurses for positions with higher eligibility requirements, requiring the scheduling of both trainees and preceptors. We use the term shift for both the 12-hour time slot as well as assignment of a nurse to a slot.

Currently, HH has a fixed staffing level for the number of nurses working in each position for every shift, which remains the same every day. Throughout the day, the levels are also kept the same for most positions except for a few positions like triage and iTrack. However, patient demand in pods varies over time, and patterns are seen within the week and within the day. We elaborate in the Appendix such inconsistency between fluctuating demand levels (see Figure A-1 in Appendix Section A.1, Figure A-3a and Figure A-4a in Section A.2) and fixed staffing levels (Figure A-4b in Section A.2). This contrast can lead to overstaffing at some times, increasing hospital staffing costs, and understaffing at other times, compromising patient care quality. Furthermore, COVID-19 has further complicated the situation by disrupting patient demand patterns and increasing employee attrition.

The current nurse scheduling process is as follows. Staffing cycles rotate every 6 weeks. Before each staffing cycle, nurses enter their preferences and availability. Schedules and ED nurse managers manually generate a schedule that aims to balance staffing levels and preferences between nurses. In addition, ED leadership also manually attempts to add training sessions to the schedule when possible. After the schedule is announced, nurses can ask for amendments to better satisfy individual preferences and increase fairness among nurses. In turn, managers and schedulers need to constantly re-compute schedules and accommodate feasible requests. To announce the schedule, scheduling managers take an additional step to manually convert the schedule of approximately 3600 shifts (on average, three weekly shifts per nurse for 200 nurses over 6 weeks) into a “team sheet” in a specified format. The entire scheduling process takes managers and schedulers many hours and days of manual work and can be prone to errors.

We improve the ED nurse scheduling process with an integrated approach consisting of a two-phase optimization methodology and a software implementation. First, we use a data-driven approach that allocates limited staffing resources cost-effectively and staffs sufficient nurses to cover target nurse-to-patient ratios for uncertain patient demand. Second, we generate an optimized individual-level schedule that improves nurse satisfaction by matching individual requests and preferences and expanding training opportunities. Finally, we implement an automated scheduling software that relieves manual labors and operations burdens for nurse managers and schedulers. Overall, our decision-support tool leverages a combination of data, optimization, and software to achieve a holistic improvement for ED staffing.

2.1.2 Related Work

The operations research literature contains numerous models aimed at improving ED operations. We summarize several main categories of methodologies used and refer readers to Saghafian et al. (2015) for a comprehensive review. Simulation-based approaches involve system simulation of the ED environment, which can be used to evaluate alternative scenarios or combined with optimization to allocate resources (Chen and Wang 2016), staff (doctors, lab technicians, and nurses) (Ahmed and Alkhamis 2009), and particularly nurses (Draeger 1992). Queueing theory methods typically assume a Poisson process with an arrival rate, sometimes considered as uncertain (Maman 2009). In a multiclass queueing system, Chan et al. (2021) examine beneficial yet challenging dynamic shift assignments for ED nurses. Hu et al. (2021) assume a doubly stochastic Poisson process and integrate with ED demand forecast (Hu et al. 2023) to derive a two-stage nurse staffing policy for both base and surge levels. Mathematical optimization is widely used to formulate personnel scheduling as formal optimization problems with constraints and objectives (Brucker et al. 2011). Such nurse staffing models can be solved using optimization solvers (Svirsko et al. 2019), iteratively by each objective (Rerkjirattikal et al. 2022), or using heuristic algorithms (Hamid et al. 2020). Additionally, goal programming models are developed to achieve a set of assumed target goals for ED nurse staffing (Ang et al. 2018, Mohammadian et al. 2019). Among the different approaches, we deem mathematical optimization as the most suitable to our problem, as it can learn from our data without any distribution assumptions, incorporate specific constraints to generate realistic schedules in practice and optimize for multiple objectives without having pre-set target values.

Mathematical optimization methods can account for uncertainty sources of uncertainty in staffing problems on the demand side or supply side. Van Hulst et al. (2017) use robust optimization (Ben-Tal et al. 2009, Bertsimas and den Hertog 2022) to generate shifts for workforce planning against adversarial workload predictions, while Lim and Mobasher (2011) apply relative robust optimization method (Kouvelis and Yu 2013) to model a common nurse scheduling problem with nurse-patient-ratio objectives subject to patient workload variability. In addition to demand uncertainty, changes in nurse availability due to sickness or unexpected events can make the schedule infeasible, which is another bottleneck for model implementation in practice. Clark et al. (2015) review the challenges of rescheduling and call for methods to support rescheduling. Wickert et al. (2019) study strategies to reconstruct schedules for multi-skilled nurses using an integer optimization problem with some relaxations. In this work, we address both uncertain components in our two-phase models. To protect against demand uncertainty, we develop a robust optimization model that characterizes the target nurse-patient-ratios with an uncertainty set and solves for robust optimal solutions that satisfy constraints with the worst-case objective value under all realizations of uncertain demand in the set. Furthermore, we implement a software interface that allows nurses to update availabilities and enables ED managers to re-run the integer optimization problem with updated information and to edit the schedule dynamically.

The most related work to ours is Ang et al. (2018), who develop a mixed-integer sequential goal programming model that optimizes for multiple objectives, including nurse–patient-ratios, shift preferences, and nurse rest days in the ED. They integrate the model into an online decision support system to facilitate implementation. This paper provoked discussion and critics, see, e.g., Park et al. (2022) that calls for “no more unimplementable nurse workforce planning”. Despite the abundance of operations research work in the past decade on nurse workforce scheduling, there is still a significant gap between research and practice, with supposedly only 30% of nurse scheduling models from research being implemented and even fewer remaining in use (Kellogg and Walczak 2007). Another obstacle to implementation is the lack of optimization knowledge and understanding in nurses (Park et al. 2022). To bridge this gap and advance nursing science, we work closely with nurse managers and iterate on optimization models that can be realistically implemented. We also develop a user-friendly end-to-end software interface for nurses to use and train nurse leadership to run the scheduling optimization on their own. Our model-software integration, jointly with trust built with medical collaborators results in the successful deployment at Hartford HealthCare.

2.1.3 Main Contributions

We summarize our contributions as follows. First, we develop a two-phase optimization approach to optimize nurse staffing in the ED:

1. We leverage robust optimization and historical demand data to optimize aggregate nurse staffing levels.

2. Given aggregate staffing levels, we develop mixed integer optimization problems to generate an individual-level schedule that prioritizes individual nurse preferences and training opportunities.

Next, on computational experiments, we demonstrate the benefit of our optimization approach and its flexibility to adjust under post-COVID-19 demand fluctuations. The schedule cuts costs by 5–8% during overstaffing periods (early COVID-19) and reduces insufficient nurse-patient-ratio coverage by 8–25% during understaffing periods (more recently). In addition, we conduct various experiments that compare different modeling variants, illustrate metric trade-offs controlled by parameters, and provide strategic insights for ED leadership. The optimized individual-level schedule improves individual nurse satisfaction by 5% as well as introduces position diversity, training opportunities, and fairness. Finally, we implement the models into automatic end-to-end scheduling software. Near 200 ED nurses enter their availability and preferences into the interface, and nurse managers run the models on their own to generate, edit, and announce schedules via simple clicks. In addition to the benefit of optimization, the software revolutionizes ED nurse scheduling with minimal manual burdens.

2.2 ED Aggregate Staffing Optimization

In the first phase, we develop a robust optimization model to allocate staffing levels based on the historical patterns of uncertain demand.

2.2.1 Data

Indices and Sets Notation. We first define the indices and sets used in the model. For the rest of this thesis, we use $[J]$ for $J \in \mathbb{Z}_+$ to denote the set $\{1, 2, \dots, J\}$. We use the following list of indices with respective cardinalities:

- Nurses are assigned to work at a position j: CNL, first nurse, resource, triage, FEP, blue, green, orange, hallway, red, purple pod, iTrack, and EDOU ($J = 13$).

- Some of the pods, blue, green, orange pods, and hallway, have the same functionalities of treatment and thus belong to the same pod type as the main pods. Pod type n: main pods, red pod, purple pod, iTrack, and EDOU ($N = 5$).

- As nurses can float to work between red pod, main pods, and iTrack during each shift, we define a pod floating group as a set of pods where nurses can float with each other. Pod floating group m: main pods + red pod + iTrack, purple pod, EDOU ($M = 3$).

- Nurses work on one of the shift types. Shift i: 7 am–7 pm, 7 pm–7 am, 11 am–11 pm ($I = 3$).

- Based on nurses' years of experience and training qualifications, they are categorized into a nurse tier q ($Q = 9$).

- Based on years of experience, the hospital classifies nurses into two groups that restrict the frequency of weekend shifts. Nurse weekend group g: working every other weekend, every third weekend ($G = 2$).

- The optimization model schedules for an entire staffing cycle (6-week period) with some week-over-week regularity. We discretize time over the cycle into each week indexed by w ($W = 6$), each day d ($D = 42$), and each 1-hour time period t ($T = 1,008$). To account for schedule patterns from week to week, we also introduce hour s ($S = 168$) and day e of each week ($E = 7$).

We define the following discrete subsets:

- J_n, $J_m \subseteq [J]$: set of pods j of type n and of floating group m, respectively.

- $Dn \subseteq [D]$: set of weekend days of the cycle.

- $De \subseteq [E]$: set of days of the first week of the cycle.

- D_w, $Dn_w \subseteq [D]$: set of days and set of weekend days in week w, respectively.

Demand Data Collection. Hartford HealthCare's IT system transfers records of ED patients to our data repository on Amazon Web Services using a secure file transfer protocol server. The data for each patient contains details such as Emergency Severity Index (acuity level), accommodation, service, ICD10 code, and discharge disposition (e.g., admitted to ICU, surgery, interventional radiology, or discharge). Based on ED rules, we map each patient from a combination of the aforementioned information into an appropriate pod and a target nurse-patient-ratio for the patient (1:1 or 1:2 for severe patients in the red pod, 1:7 for behavioral patients in the purple pod, 1:12 for low acuity level patients in iTrack, and 1:5 for remaining patients in main pods and EDOU). To obtain the demand for nurses, we calculate the time range for which each patient is treated in each pod with the corresponding target nurse-patient-ratio. We then aggregate the data by patients to compute the historical demand h^{hist}_{mt}, representing the number of nurses needed to meet the target nurse-patient-ratio to treat the patients in pod floating group m during hour t. We automatically receive the data files every day and aggregate them into demand data every week.

Modeling Uncertain Demand. Every 6 weeks, we follow the timeline shown in Figure 2-1 to model the uncertain demand for each staffing cycle spanning from week w^{start} to week $w^{\text{start}}+5$. The uncertainty is $\tilde{h}_{msw}$: the number of nurses needed to meet the target nurse-to-patient ratio for patients in pod floating group m during hour s of week w in the next staffing cycle. We use historical demand data h^{hist}_{msw} to construct an uncertainty set $\mathcal{U}$ and assume $\tilde{h}_{msw} \in \mathcal{U}$. The schedule is typically planned and announced to nurses approximately three weeks ahead of the beginning of the upcoming staffing cycle. To ensure adequate preparation time, we construct the uncertainty set four weeks prior to week w^{start} using historical demand data up to the end of week $w^{\text{start}}-5$. We first utilize the demand of the most recent 6 weeks of the period, week $w^{\text{start}}-10$ to $w^{\text{start}}-5$, as the nominal demand $\bar{h}_{msw}$. To account for fluctuations, we estimate demand changes between 6-week periods that are 10 weeks apart. To do this, we compute a list of absolute differences between pairs of historical demand data starting from week $w^{\text{start}}-20$:

$$\hat{\varepsilon}_{msw} = |h^{\text{hist}}_{msw} - h^{\text{hist}}_{msw-9}|, \quad \forall m \in [M], s \in [S], w \in \{w^{\text{start}}-11, \dots, w^{\text{start}}-5\}.$$

Finally, we construct an uncertainty set that bounds deviations of uncertainty $\tilde{h}_{msw}$ from nominal demand $\bar{h}_{msw}$ parameterized by $\hat{\varepsilon}_{msw}$:

$$\mathcal{U} = \left\{ \tilde{h}_{msw} \;\middle|\; \begin{array}{ll} \tilde{h}_{msw} \geq 0, & \forall m \in [M], s \in [S], w \in [W] \\ |\bar{h}_{msw} - \tilde{h}_{msw}| \leq \epsilon_{1ms}, & \forall m \in [M], s \in [S], w \in [W] \\ |\sum\limits_{s \in [S]} \bar{h}_{msw} - \sum\limits_{s \in [S]} \tilde{h}_{msw}| \leq \epsilon_{2m}, & \forall m \in [M], w \in [W] \\ |\sum\limits_{m \in [M], s \in [S]} \bar{h}_{msw} - \sum\limits_{m \in [M], s \in [S]} \tilde{h}_{msw}| \leq \epsilon_3, & \forall w \in [W]. \end{array} \right\},$$

where we estimate pairwise inter-10-week demand fluctuations ϵ_{1ms} for each m, s as the 80^{th} percentile of $\hat{\varepsilon}_{msw}, \forall w \in \{w^{\text{start}}-11, \dots, w^{\text{start}}-5\}$, and scale ϵ_{1ms} accordingly to obtain ϵ_{2m} and ϵ_3. From Table 2 in Bertsimas et al. (2021a), the uncertainty set

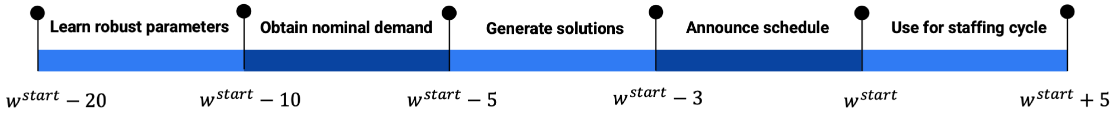


Figure 2-1: Timeline of Uncertain Demand Modeling for Every Staffing Cycle.

can obtain a probabilistic guarantee on constraint violation if ϵ_{2m} roughly scales with $\sqrt{S}\epsilon_{1ms}$ and ϵ_3 roughly scales with $\sqrt{MS}\epsilon_{1ms}$. Thus we use:

$$\epsilon_{2m} = \beta_\epsilon/\sqrt{S}\sum_{s\in[S]}\epsilon_{1ms}, \quad \forall m \in [M], \quad \epsilon_3 = \beta_\epsilon/\sqrt{MS}\sum_{m\in[M],s\in[S]}\epsilon_{1ms}$$

with an adjustable discount factor $\beta_\epsilon \in (0, 1]$.

Other Data and parameters. Other than the demand data, we collect the following data from ED leadership:

- Z_q: total number of nurses of type q.
- Z_q^g: total number of nurses of type q and weekend group g.
- z_{jie}^{curr}: number of nurses working in pod j during shift i on day e of the week in the current schedule.
- $\underline{n}_j, \overline{n}_j$: minimum and maximum number of nurses required at unit j, respectively.
- $\overline{Tn_q}$: maximum total number of shifts each nurse of type q can work in a week.
- N_U: set of tuples (q, j) such that nurse of type q cannot work in unit j due to qualifications and years of experience.
- E_{qt}: number of nurses of type q available to work on period t.

We also introduce σ_{idt} (or equivalently σ_{idsw}) to indicate whether shift i on day d contains time period t.

2.2.2 Optimization Model Formulation

Decision Variables. In the next 6-week cycle, on each day of the week, we decide the number of nurses of each tier to be staffed for each position during each shift. To do so, we introduce decision variables:

- $z_{qjidg} \in \mathbb{Z}^+$: number of nurses of tier q from the weekend group g working at position j during shift i on day d.

Objective. The objective of the model is to minimize the total number of nurse shifts scheduled while penalizing shortage for demand and changes from the current schedule, with parameters to control trade-offs between the three terms.

To achieve this, we use two auxiliary variables to track the second and third terms:

- npr: the weighted sum of insufficiency in the worst case w.r.t. the demand uncertainty set, where insufficiency is defined as the number of nurse shifts missing to satisfy the target nurse-to-patient ratios to treat the patients.

- Δz_{jie}: the absolute difference in the number of scheduled nurses working at position j during shift i on day e of the week from the current schedule.

We use the notation $f(x)_+$ to denote the positive part of the function $f(x)$, i.e., $f(x)_+ = \max\{f(x), 0\}$. We impose constraints on the auxiliary variables:

$$npr \geq \sum_{m\in[M],s\in[S],w\in[W]} \omega_w(\tilde{h}_{msw} - \sum_{q\in[Q],j\in J_m,i\in[I],d\in[D],g\in[G]} z_{qjidg}\sigma_{idsw})_+, \quad \forall \tilde{h} \in \mathcal{U}, \tag{2.1}$$

$$\Delta z_{jie} \geq |z_{jie}^{\text{curr}} - \sum_{q\in[Q],d\in De,g\in[G]} z_{qjidg}|, \quad \forall j \in [J], i \in [I], e \in [E]. \tag{2.2}$$

Constraint (2.1) is a robust constraint with parameter ω_w increasing with w to penalize more on later weeks to capture the most recent demand trend.

Our objective function is to minimize a weighted combination of three terms:

$$\begin{aligned}
\min \quad & \sum_{q\in[Q],j\in[J],i\in[I],d\in[D],g\in[G]} z_{qjidg} && \text{(Minimize number of scheduled work shifts)} \\
+ \quad & \mu_1 \cdot npr && \text{(Penalize when staffed below target ratios)} \\
+ \quad & \mu_2 \sum_{j\in[J],i\in[I],e\in[E]} \Delta z_{jie} && \text{(Penalize number of changed work shifts),}
\end{aligned}$$

with parameters $\mu_1, \mu_2 \geq 0$ to control the trade-off between the objective terms.

Constraints. The schedule must adhere to a range of staffing constraints to be feasible. These include:

- Each position j has a minimum and maximum number of nurses to staff during time t:

  $$\underline{n}_j \leq \sum_{q\in[Q],i\in[I],d\in[D],g\in[G]} z_{qjidg} \cdot \sigma_{idt} \leq \overline{n}_j, \quad \forall j \in [J], t \in [T].$$

  Staffing levels are fixed at certain positions (such as CNL and first nurse) for logistical reasons (with $\underline{n}_j = \overline{n}_j$ to be the same as current levels), while other positions at different pods can be adjusted to demand.

- Certain units are not eligible based on nurse tiers:

  $$z_{qjidg} = 0, \quad \forall (q, j) \in N_U, i \in [I], d \in [D], g \in [G].$$

- Staffing levels among the three main pods are allocated in proportion to their capacities:

$$\begin{aligned}
&\sum_{q\in[Q],i\in[I],d\in[D],g\in[G]} z_{q4idg} \leq \sum_{q\in[Q],i\in[I],d\in[D],g\in[G]} z_{q3idg} \leq \sum_{q\in[Q],i\in[I],d\in[D],g\in[G]} z_{q5idg},\\
&\sum_{q\in[Q],i\in[I],d\in[D],g\in[G]} z_{qjid5} - \sum_{q\in[Q],i\in[I],d\in[D],g\in[G]} z_{qjid3} \leq 1,\\
&\sum_{q\in[Q],i\in[I],d\in[D],g\in[G]} z_{qjid5} - \sum_{q\in[Q],i\in[I],d\in[D],g\in[G]} z_{qjid4} \leq 1.
\end{aligned}$$

- Staffing levels are kept to be consistent from week to week for a stable schedule:

$$\begin{aligned}
\sum_{q\in[Q],g\in[G]} z_{q,j,i,k+E(w-1),g} = \sum_{q\in[Q],g\in[G]} z_{q,j,i,k+E(W-1),g},\\
\forall k \in [E], w \in [W-1], i \in [I], j \in [J].
\end{aligned} \tag{2.3}$$

ED leadership indicates a preference to consider only the constraints above when deciding on staffing levels, based on demand only. However, they have the option to also consider nurse availability and supply information in this stage of scheduling by including additional constraints in the model:

- Assignments are capped by the number of available nurses of each tier during each period:

$$\sum_{j\in[J],i\in[I],d\in[D],g\in[G]} z_{qjidg} \cdot \sigma_{idt} \leq E_{qt}, \quad \forall q \in [Q], t \in [T].$$

- Total number of weekly working hours is bounded for each nurse:

$$\sum_{d\in D_w,j\in[J],i\in[I],g\in[G]} z_{qjidg} \leq \overline{Tn}_q Z_q, \quad \forall q \in [Q], w \in [W].$$

- Nurses working every other weekend can have at most two weekend shifts every two consecutive weeks (allowing one shift on Saturday and another one on

Sunday on the working weekend):

$$\sum_{d\in Dn_w \cup Dn_{w+1}, j\in[J], i\in[I], g\in[G]} z_{qjidg} \leq 2Z_q^1, \quad \forall q \in [Q], w \in [W-1].$$

- Nurses working every third weekend can have at most two weekend shifts every three consecutive weeks:

$$\sum_{d\in Dn_w \cup Dn_{w+1} \cup Dn_{w+2}, j\in[J], i\in[I], g\in[G]} z_{qjidg} \leq 2Z_q^2, \quad \forall q \in [Q], w \in [W-2].$$

2.2.3 Solving the Robust Optimization Model

We describe our method and implementation of solving the robust optimization problem outlined in Section 2.2.2. Due to the fact that both the decision variables and uncertainties are integers, duality theory is not applicable and thus a closed-form robust counterpart cannot be obtained. In light of the discrete and finite nature of the uncertainty set, a cutting plane method is employed to identify a subset of the most restrictive constraints and to approximate the worst-case objective value among the uncertainty set. We solve the model with Algorithm 1, in which the master problem is the optimization problem defined in Section 2.2.2 with $\mathcal{U}$ being replaced by a subset $\mathcal{U}^\kappa$ in constraint (2.1).

Algorithm 1 Cutting plane algorithm

```
Given the nominal demand $h^{\text{hist}} : \mathcal{U}^0 \leftarrow \{h^{\text{hist}}\}$, $\kappa \leftarrow 1$, and tolerance $\eta$, initiate the master problem
repeat
    Solve the master problem and obtain a solution $z^\kappa$
    $npr_\kappa(\tilde{h}, z^\kappa) \leftarrow \sum_{m\in[M],s\in[S],w\in[W]} \omega_w(\tilde{h}_{msw} - \sum_{q\in[Q],j\in J_m,i\in[I],d\in[D],g\in[G]} z^\kappa_{qjidg}\sigma_{idsw})_+$
    $h^\kappa \leftarrow \arg\max_{\tilde{h}\in\mathcal{U}} npr_\kappa(\tilde{h}, z^\kappa)$    ▷ Maximize insufficiency w.r.t. demand uncertainty set
    $\mathcal{U}^\kappa \leftarrow \{h^{\text{hist}}, h^1, \ldots, h^\kappa\}$
    $\kappa \leftarrow \kappa + 1$
until $(npr^\kappa - npr^{\kappa-1})/npr^{\kappa-1} < \eta$ or $\kappa = \kappa^{\max}$    ▷ Reach violation gap or maximum iterations
```

All models in this work are implemented using the JuMP package (Dunning et al. 2017) in Julia programming language (Bezanson et al. 2017) and are solved by the Gurobi solver (Gurobi Optimization, LLC 2023). The master problem is converted into a mixed integer linear problem, where we linearize constraints (2.1) with positive parts and constraints (2.2) with absolute values by including additional auxiliary variables and constraints. We set a stopping criterion to solve the master problem with either a tolerance of 0.01 optimality gap for the mixed integer optimization problem or a 20-minute time limit. For the cutting planes implementation, we utilize solver callbacks with added lazy constraints in JuMP to accelerate the model compilation and solving process. We set the tolerance of violation gap for cutting planes $\eta = 0.1$ and the maximum number of iterations of cutting planes $\kappa^{\max} = 20$. To linearize the objective of maximizing the summation of positive parts for each subset $\mathcal{U}^{\kappa}$, we introduce big M constraints with $M = 100$.

Parameter Tuning. In preparation for each upcoming staffing cycle starting on week w^{start}, we fine-tune a combination of parameters, $\beta_\epsilon, \mu_1, \mu_2, \omega_w$, using historical data. To validate our choices, we examine the performance of different combinations of parameters over a validation period of $w^{\text{start}} - 10$ to $w^{\text{start}} - 5$. By comparing the values of the objective terms, we determine the optimal combination of parameters.

Deterministic Model. For comparative purposes, we also consider a deterministic model where we only use the nominal demand instead of incorporating the uncertainty set. In this case, we solve a single master problem using the optimization model defined in Section 2.2.2, but with $\mathcal{U}$ replaced by the set of historical demand only $\{h^{\text{hist}}\}$ in constraint (2.1), which is equivalent to setting $\epsilon_1 = \epsilon_2 = \epsilon_3 = 0$ in $\mathcal{U}$.

2.2.4 Quantifying Strategic Decisions

Our model not only optimizes staffing levels for each cycle but also provides a way for nurse leadership to quantify strategic decisions. By designing different model variations, we can evaluate different options and guide the leadership in making decisions such as: Should they maintain the same staffing levels each day or week of the cycle? And should they consider incorporating new shift types in addition to the three existing ones? To guide these strategic decisions, we use our model to compare different variations and quantify the operational costs associated with each. This approach allows us to evaluate the trade-off between different objectives and impose additional constraints as needed.

Stable staffing levels. Currently, the staffing levels remain constant every day. Motivated by the different demand patterns by day of the week, we allow the staffing levels to vary by day of the week while keeping the same levels among the 6 weeks with a weekly stability constraint (2.3). To evaluate the benefits of this variation, we compare it to two other options:

1. Daily stability: Staffing levels can be fixed from day to day, by replacing (2.3) with the following constraint:

$$\sum_{q\in[Q],g\in[G]} z_{qjidg} = \sum_{q\in[Q],g\in[G]} z_{qjiDg}, \quad \forall d \in [D-1], i \in [I], j \in [J]. \tag{2.4}$$

   This option is easier for the leadership to manage but might result in unnecessary overstaffing and understaffing on some days.

2. No stability: Staffing levels can be allowed to vary every day, by not including constraint (2.3) and (2.4). This option has additional flexibility to fit demand patterns but could be potentially overfitting the data.

The operational costs and values of the three options can be compared to determine the best approach.

Change shift designs. ED leadership also considers changing or adding shift types. Such changes would be disruptive and require some nurses to change their lifestyles to accommodate the new working hours but they may lead to more cost-effective staffing. With many potential shift type options available, it is essential to determine which ones could result in the most improvement. We support this by introducing binary variables $y_i \in \{0,1\}$ that indicate whether a shift type i is generated. We have the set of existing shift types I_{exist} and consider a feasible set of potential shift types $I_{\text{feasible}} \supseteq I_{\text{exist}}$ with σ_{idt} for each $i \in I_{\text{feasible}}$. We add the following constraints to the model:

- A shift type is generated if and only if used in any shifts:

$$\sum_{q\in[Q],j\in[J],d\in[D],g\in[G]} z_{qjidg} \leq M_y \cdot y_i, \quad \forall i \in [I]$$

  with a constant M_y.

- The number of shifts types is bounded with a parameter $Y_{\max}$:

$$\sum_{i\in I_{\text{feasible}}} y_i \leq Y_{\max}.$$

- If considering only new shift types without changing the existing shift types, then all current shift types are kept:

$$y_i = 1, \quad \forall i \in I_{exist}.$$

By solving different variations and comparing their objective values, we can identify the most valuable shift type candidates.

2.3 ED Individual Scheduling Optimization

For the second phase, we develop another optimization model to schedule individual nurses based on the recommended aggregate schedule.

2.3.1 Data

Indices and Sets. We introduce additional indices and sets besides those defined in Section 2.2.1:

- Individual nurse ℓ ($L \sim 200$ representing the total number of nurses).
- Pattern u ($U = 3$) out of patterns to have shifts on and off:
    1. All shifts in a row (e.g., working on Monday, Tuesday, and Wednesday consecutively),
    2. Every other day (e.g., Monday on, Tuesday off, Wednesday on, ...),
    3. Two on two off (e.g., Monday and Tuesday on, Wednesday and Thursday off, ...).
- Criterion o ($O = 9$) to measure each nurse's dissatisfaction, such as match or mismatch in shift assignment with nurses' preference on dates and patterns.

and sets:

- L_q: set of nurses of type q.
- J^{slack}: set of positions j where a shortage or surplus of at most one is allowed.

Data Input. One of the inputs of the second phase model is the output aggregate levels z^{agg}_{jid} from the first phase, aggregating the number of nurses to schedule for position j during shift i on day d from the solution $z^{\star}_{qjidg}$:

$$z^{\text{agg}}_{jid} = \sum_{q \in [Q], g \in [G]} z^{\star}_{qjidg}, \quad \forall j \in [J], i \in [I], d \in [D].$$

While the assignment per nurse type q and weekend group g in the first phase is included to ensure a feasible assignment, it is subject to change during the second phase. In addition to the aggregate levels, the model takes into account a range of individual preferences and availabilities for each nurse:

- $I^{\text{pref}}_{\ell i}, I^{\text{unpref}}_{\ell i}, I^{\text{feas}}_{\ell i}$: whether nurse ℓ prefers (typically their current shift type), does not prefer, and has the feasibility to work at shift type i, respectively.
- $F_{\ell w}$: maximum number of shifts for nurse ℓ on week w, obtained by subtracting Paid Time Off (PTO) and education time of each week from the total number of weekly shifts.

- $k_{\ell w}$: whether nurse ℓ is assigned to work on the weekend of week w.

- P^{pref}_{ld}, P^{unpref}_{ld}, P^{off}_{ld}, P^{avail}_{ld}: whether nurse ℓ prefers, does not prefer, is off, and is available to work on day d, respectively.

- $A^{\text{pref}}_{\ell u}$, $A^{\text{unpref}}_{\ell u}$: whether nurse ℓ prefers to have, and not to have the work pattern u, respectively.

- $r^{\text{feas}}_{\ell j}, p^{\text{feas}}_{\ell j}$: whether nurse ℓ is eligible to be a trainee, and a preceptor (training the trainees), respectively, at position j.

2.3.2 Integer Optimization Model on Shift Scheduling

To generate an individual schedule that prioritizes various staff preferences as well as training opportunities given these inputs, we develop a mixed integer linear problem that comprises the following components.

Decision Variables. The core decision variables track nurse assignments:

- $b_{\ell jid} \in \{0, 1\}$: whether nurse ℓ works at position j during shift i on day d.

- $r_{\ell jid} \in \{0, 1\}$: whether nurse ℓ is assigned as trainee at position j during shift i on day d.

- $s_{\ell i} \in \{0, 1\}$: whether nurse ℓ is assigned to shift type i.

To optimize computational efficiency, we implement all variables as sparse matrices and tensors under conditions as follows. For instance, binary variables $b_{\ell jid}$ are only needed if the values are allowed to be one, and thus are only defined for indices ℓ, j, i, d such that nurse ℓ is eligible to work at position j and is available during shift i on day d. Such sparse indexing naturally incorporates some eligibility, availability, and feasibility constraints.

Objective. Our primary objective is to maximize nurse satisfaction across various metrics, such as individual preference on dates, times, and shift patterns, diversity in pod assignments, and reduction in weekend, holiday, and overtime shifts. Meanwhile, we penalize shortages and surpluses in aggregate staffing levels while rewarding scheduled training shifts.

We introduce auxiliary variables to track different terms of the objective:

- $c_{\ell o} \in \mathbb{R}$: computing nurse ℓ's dissatisfaction penalty score on criterion o.
- $f^+_{\ell w} \in \{0, 1\}$: whether nurse ℓ works an overtime shift on week w.
- $k^+_{\ell w} \in \{0, 1\}$: whether nurse ℓ works on a weekend of week w where she is supposed to be off that weekend.
- $v_{\ell j} \in \{0, 1\}$: whether nurse ℓ has at least one shift at position j during the staffing cycle.
- $a_{\ell u d} \in \{0, 1\}$: whether nurse ℓ starts work pattern u on day d.
- $z^-_{jid}, z^+_{jid} \in [0, 1]$: whether there is one shortage or surplus at position j during shift i on day d, respectively.
- $f^-_\ell \in \mathbb{R}^+$: number of unassigned shifts for nurse ℓ.

We impose a range of constraints on the auxiliary variables:

- Number of each nurse's weekly shifts cannot exceed the maximum (allowing at most one overtime shift per week if $f^+_{\ell w} = 1$):

$$\sum_{d \in D_w, j \in [J], i \in [I]} (b_{\ell jid} + r_{\ell jid}) \leq F w_{\ell w} + f^+_{\ell w}, \quad \forall \ell \in [L], w \in [W].$$

- Nurse can only work (at most two shifts) on the weekend if assigned the weekend (allowing shifts to schedule on an unassigned weekend if $k^+_{\ell w} = 1$):

$$\sum_{d \in Dn_w, j \in [J], i \in [I]} (b_{\ell jid} + r_{\ell jid}) \leq 2(k_{\ell w} + k^+_{\ell w}), \quad \forall \ell \in [L], w \in [W].$$

- Tracks whether nurse ℓ has at least one shift at position j:

$$v_{\ell j} \leq \sum_{d \in [D], i \in [I]} b_{\ell j i d}, \quad \forall \ell \in [L], j \in [J].$$

- Tracks work patterns:

  - Nurse cannot work more than three shifts in a row:

$$a_{\ell,1,d} + a_{\ell,1,d+1} \leq 1, \quad \forall \ell \in [L], d \in [D-3].$$

  - All shifts in a row pattern:

$$\begin{aligned} 3a_{\ell 1 d} \quad & \leq \sum_{j \in [J], i \in [I]} (b_{\ell,j,i,d} + b_{\ell,j,i,d+1} + b_{\ell,j,i,d+2}) \\ & + \sum_{j \in [J], i \in [I]} (r_{\ell,j,i,d} + r_{\ell,j,i,d+1} + r_{\ell,j,i,d+2}), \quad \forall \ell \in [L], d \in [D-2], \\ 3a_{\ell 1 d} \quad & \geq \sum_{j \in [J], i \in [I]} (b_{\ell,j,i,d} + b_{\ell,j,i,d+1} + b_{\ell,j,i,d+2}) \\ & + \sum_{j \in [J], i \in [I]} (r_{\ell,j,i,d} + r_{\ell,j,i,d+1} + r_{\ell,j,i,d+2}) - 2, \quad \forall \ell \in [L], d \in [D-2]. \end{aligned}$$

  - Every other day pattern:

$$\begin{aligned} 3a_{\ell 2 d} \quad & \leq \sum_{j \in [J], i \in [I]} (b_{\ell,j,i,d} + 1 - b_{\ell,j,i,d+1} + b_{\ell,j,i,d+2}) \\ & + \sum_{j \in [J], i \in [I]} (r_{\ell,j,i,d} + 1 - r_{\ell,j,i,d+1} + r_{\ell,j,i,d+2}), \quad \forall \ell \in [L], d \in [D-2], \\ 3a_{\ell 2 d} \quad & \geq \sum_{j \in [J], i \in [I]} (b_{\ell,j,i,d} - 1 - b_{\ell,j,i,d+1} + b_{\ell,j,i,d+2}) \\ & + \sum_{j \in [J], i \in [I]} (r_{\ell,j,i,d} - 1 - r_{\ell,j,i,d+1} + r_{\ell,j,i,d+2}), \quad \forall \ell \in [L], d \in [D-2]. \end{aligned}$$

- Two on two off pattern:

$$\begin{aligned}
4a_{\ell 3d} &\le \sum_{j\in[J],i\in[I]} (b_{\ell,j,i,d} + b_{\ell,j,i,d+1} + 2 - b_{\ell,j,i,d+2} - b_{\ell,j,i,d+3}) + \sum_{j\in[J],i\in[I]} (r_{\ell,j,i,d} \\
&\qquad + r_{\ell,j,i,d+1} + 2 - r_{\ell,j,i,d+2} - r_{\ell,j,i,d+3}), \quad \forall \ell \in [L], d \in [D-3], \\
4a_{\ell 3d} &\ge \sum_{j\in[J],i\in[I]} (b_{\ell,j,i,d} + b_{\ell,j,i,d+1} - 1 - b_{\ell,j,i,d+2} - b_{\ell,j,i,d+3}) + \sum_{j\in[J],i\in[I]} (r_{\ell,j,i,d} \\
&\qquad + r_{\ell,j,i,d+1} - 1 - r_{\ell,j,i,d+2} - r_{\ell,j,i,d+3}), \quad \forall \ell \in [L], d \in [D-3].
\end{aligned}$$

- Computes individual nurse penalty score for each $l \in [L]$ counting the number of:

  - Shifts on dates supposed to be off: $c_{\ell 1} = \sum_{d\in[D],j\in[J],i\in[I]} P^{\text{off}}_{ld}(b_{\ell jid} + r_{\ell jid})$.
  - Unassigned weekends turned on: $c_{\ell 2} = \sum_{w\in[W]} k^{+}_{\ell w}$.
  - Overtime shifts: $c_{\ell 3} = \sum_{w\in[W]} f^{+}_{\ell w}$.
  - Unpreferred shift types: $c_{l4} = \sum_{d\in[D],j\in[J],i\in[I]} I^{\text{unpref}}_{\ell i} b_{\ell jid}$.
  - Unpreferred dates: $c_{l5} = \sum_{d\in[D],j\in[J],i\in[I]} P^{\text{unpref}}_{ld}(b_{\ell jid} + r_{\ell jid})$.
  - Preferred dates: $c_{l6} = \sum_{d\in[D],j\in[J],i\in[I]} P^{\text{pref}}_{ld}(b_{\ell jid} + r_{\ell jid})$.
  - Different positions assigned: $c_{l7} = \sum_{j\in[J]} v_{\ell j}$.
  - Unpreferred patterns: $c_{l8} = \sum_{d\in[D-2],u\in[2]} A^{\text{unpref}}_{\ell u} a_{\ell ud} + \sum_{d\in[D-3]} A^{\text{unpref}}_{\ell 3} a_{\ell 3d}$.
  - Preferred patterns: $c_{l9} = \sum_{d\in[D-2],u\in[2]} A^{\text{pref}}_{\ell u} a_{\ell ud} + \sum_{d\in[D-3]} A^{\text{pref}}_{\ell 3} a_{\ell 3d}$.

- The individual nurse schedule matches aggregate staffing levels with shortages and surpluses:

$$z^{\text{agg}}_{jid} = z^{-}_{jid} - z^{+}_{jid} + \sum_{\ell\in[L]} b_{\ell jid}, \quad \forall j \in J^{\text{slack}}, i \in [I], d \in [D].$$

- Some positions do not allow shortages or surpluses:

$$z^-_{jid} = z^+_{jid} = 0, \quad j \in J \setminus J^{\text{slack}}, i \in [I], d \in [D].$$

- The number of unassigned shifts for each nurse is the number of maximum shifts minus assigned shifts:

$$f^-_\ell \geq \sum_{w \in [W]} F_{\ell w} - \sum_{j \in [J], i \in [I], d \in [D]} (b_{\ell jid} + r_{\ell jid}).$$

Our objective function is to minimize a weighted combination of four terms:

$$\begin{aligned}
\min \quad & \mu_3 \sum_{\ell \in [L], o \in [O]} \lambda_o c_{\ell o} && \text{(Minimize total weighted dissatisfaction score)} \\
+ \quad & \mu_4 \sum_{j \in [J], i \in [I], d \in [D]} w^{\text{shortage}}_j z^-_{jid} && \text{(Penalize shortage to aggregate staffing levels)} \\
+ \quad & \mu_5 (\sum_{j \in [J], i \in [I], d \in [D]} z^+_{jid} + \sum_{\ell \in [L]} f^-_l) && \text{(Penalize unassigned nurse shifts and surpluses)} \\
- \quad & \mu_6 \sum_{\ell \in [L], j \in [J], i \in [I], d \in [D]} r_{\ell jid} && \text{(Reward total training shifts assigned)}
\end{aligned}$$

with parameters:

- $\lambda_1, \ldots \lambda_5, \lambda_8 > 0, \lambda_6, \lambda_7, \lambda_9 < 0$ as weights for each penalty score metric,
- $\mu_3, \mu_4, \mu_5, \mu_6 > 0$ to control the trade-off between the objective terms,
- $w^{\text{shortage}}_j \geq 0$ as penalization weights for shortage at position j.

Constraints. Besides the constraints associated with the auxiliary variables for the objective, the schedule is enforced to satisfy the following constraints:

- Nurses are limited to working at most one location for each available shift, and cannot work on unavailable dates:

$$\sum_{j \in [J]} (b_{\ell jid} + r_{\ell jid}) \leq P^{\text{avail}}_{ld}, \quad \forall \ell \in [L], i \in [I], d \in [D].$$

- Each nurse is assigned at most one shift type over the 6-week period:

$$\sum_{i\in[I]} s_{\ell i} \leq 1, \quad \forall \ell \in [L].$$

- Each nurse can be only assigned a feasible shift type:

$$s_{\ell i} \leq I^{\text{feas}}_{\ell i}, \quad \forall \ell \in [L], i \in [I].$$

- Each nurse can only work on her assigned shift type (with $M_s = 2D$):

$$\sum_{d\in[D],j\in[J]} (b_{\ell jid} + s_{\ell jid}) \leq M_s \cdot s_{\ell i}, \quad \forall \ell \in [L], i \in [I].$$

- Nurses of each level can only work at their eligible positions:

$$b_{\ell jid} = 0, \quad \forall q \in [Q], (q, j) \in N_U, \ell \in L_q, i \in [I], d \in [D].$$

- For EDOU, at least one EDOU resource nurse needs to be present at 7-7 shifts:

$$\sum_{\ell\in L_{\text{EDOU resource}}} b_{\ell,\text{EDOU},i,d} \geq 1, \quad \forall i \in \{1, 3\}, d \in [D].$$

- Each nurse can only train at positions she is eligible to be a trainee:

$$r_{\ell jid} \leq r^{\text{feas}}_{\ell j}, \quad \forall \ell \in [L], j \in [J], i \in [I], d \in [D].$$

- There can be at most three training shifts for each training over the period:

$$\sum_{i\in[I],d\in[D]} r_{\ell jid} \leq 3, \quad \forall \ell \in [L], j \in [J].$$

- The number of trainees assigned cannot exceed the number of eligible preceptors working at each position during each shift:

$$\sum_{\ell \in [L]} r_{\ell jid} \leq \sum_{\ell \in [L]} b_{\ell jid}\, p_{\ell j}^{\text{feas}}, \quad j \in [J], i \in [I], d \in [D].$$

- To ensure fairness among nurses, each nurse's penalty score cannot exceed a bound $c^{\max}$:

$$\sum_{o \in [O]} c_{\ell o} \leq c^{\max}, \quad \ell \in [L]. \tag{2.5}$$

2.3.3 Integer Optimization Model on Preceptor Scheduling

After obtaining the optimal solution for scheduled work shifts $b^{\star}_{\ell jid}$ and trainee shifts $r^{\star}_{\ell jid}$ from solving the model defined in Section 2.3.2, we develop and solve another mixed integer linear optimization model to schedule preceptor shifts to train the corresponding trainees. The overall goal is to distribute trainees more evenly among eligible preceptors. To achieve this, we introduce decision variables:

- $p_{\ell jid} \in \{0, 1\}$: whether nurse ℓ serves as a preceptor at position j during shift i on day d.

- $p^{\max} \in \mathbb{R}$: an auxiliary variable to track the maximum number of preceptor shifts assigned among all nurses.

We set the objective function to:

$$\min \quad p^{\max} \qquad \text{(maximum number of preceptor shifts assigned to each nurse).}$$

The model is subject to several constraints:

- $p^{\max}$ is at least the number of preceptor shifts for each nurse:

$$p^{\max} \geq \sum_{j \in [J], i \in [I], d \in [D]} p_{\ell jid}, \quad \forall \ell \in [L].$$

- Each nurse can only be a preceptor if she is eligible to train the trainees at each position:

$$p_{\ell jid} \leq p^{\text{feas}}_{\ell j}, \quad \forall \ell \in [L], j \in [J], i \in [I], d \in [D].$$

- Each preceptor shift is in tandem with the nurse's work shift at the same position:

$$p_{\ell jid} \leq b^{\star}_{\ell jid}, \quad \forall \ell \in [L], j \in [J], i \in [I], d \in [D].$$

- Each trainee shift is covered by a preceptor. Specifically, the number of preceptors working is at least the number of trainees assigned at each position during each shift:

$$\sum_{\ell \in [L]} r^{\star}_{\ell jid} \leq \sum_{\ell \in [L]} p_{\ell jid}, \quad \forall j \in [J], i \in [I], d \in [D].$$

2.3.4 Flexible Options for Nurse Leadership

We provide various options for ED nurse leadership to generate alternative schedules based on their preferences. These options include:

- The option to forbid shift type change from each nurse's pre-assigned shift type by setting $s_{\ell i} = I^{\text{pref}}_{\ell i}, \quad \forall \ell \in [L], i \in [I]$.

- The ability to exclude trainee scheduling by setting $r_{\ell jid} = 0, \quad \forall \ell \in [L], j \in [J], i \in [I], d \in [D]$.

- The option to decompose the model into ED and EDOU and solve each of the two partitions separately.

- Control over the priority importance of different terms in the objective function by adjusting μ's. Currently, the order of importance is shortage, satisfaction, unassigned/surplus, and training. If training is not intended to be scheduled, it can be excluded by setting its weight to zero.

- Adjustment of the relative priority of coverage between positions using w^{shortage}_{j}. Currently, positions that require more senior nurse tiers (such as CNL and triage) have higher penalization weight for shortage.

- Fine-tuning of the relative importance of satisfaction score metrics using λ's. To reduce model solving time, the preferred pattern and/or unpreferred pattern term(s) can be turned off by assigning a weight of zero if desired.

- Inclusion of fairness among nurses as a hard constraint in Equation (2.5), with $c^{\max}$ controlling the degree of worst-case penalty score, or as a soft constraint by adding the term $c^{\max}$ in the objective function with a corresponding weight factor. Alternatively, fairness can be excluded altogether, which could also reduce solving time.

2.4 Results

We present experimental results of various components of our approach in different staffing periods. As shown in Figure A-1 in Appendix section A.1, the ED volume has fluctuated over the recent years. We first developed the first phase aggregate model in late 2020, when demand significantly dropped due to the beginning of COVID-19. In Section 2.4.1, we compared model variations and parameter trade-offs, illustrated schedule and demand patterns, and demonstrated the benefit of optimization in this overstaffing period. Then in 2021, we iterated with the nurse leadership to adjust the aggregate model and to develop the second phase individual model. In Section 2.4.2, we present the progression of the recommended schedules as well as the benefits of optimization for individual nurses. More recently in 2022, the ED patient demand started bouncing back to the pre-COVID-19 period, and understaffing exceeded more than ever. In Section 2.4.3, we demonstrate our calibrated model's benefit in reducing insufficiency during this understaffing period. Moreover, we developed software for automatic nurse staffing and investigated the results of resolving the models.

2.4.1 Reducing Staffing Costs during Low Demand Periods

We first define the quality metrics and a variety of approaches for evaluation. We conduct experiments using data for staffing cycles from October 26 to December 20

2020 to compare approaches and demonstrate metric trade-offs. Lastly, we illustrate recommended schedule changes and schedule patterns corresponding to demand patterns.

Quality measures. We use two metrics to evaluate staffing. (a) *Cost*: We obtain the total number of nurses assigned to work shifts, which is not subject to uncertainty. (b) *Insufficiency*: We first compute the fractional number of nurses missing to satisfy the target nurse-to-patient ratios during each hour of the upcoming staffing cycle, which we sum over all hours of the 6 weeks and then divide by the number of days and the number of hours per shift (12). We call this metric average daily insufficiency evaluated out of sample, subject to uncertain demand during the future 6 weeks. For example, a current average daily insufficiency of 0.17 reported in Table 2.1 means that we need to add on average of 0.17 12-hour shifts per day to the exactly needed time and spot to satisfy target ratios. We aim to reduce both cost and insufficiency.

Different approaches for evaluation. We consider the following 6 approaches that generate schedules to compare their performances:

1. *Current schedule*: The baseline is the current staffing levels that the ED uses every day.

2. *Oracle approach*: We design the oracle approach to understand what would be the best we could have done if we had perfect demand information. For each 6-week period, we apply the optimization model on the demand observed in these 6 weeks and test the schedule during the same period. Since it is not possible to know perfect demand information ahead, this approach is not realistic to implement in practice but rather illustrates the best possible improvement as a benchmark.

3. *Non-robust optimization*: In the remaining approaches, we only use information known previously to make prospective decisions for the future to demonstrate the feasible improvement achievable in practice. We first apply the optimization

model on the previous 6-week period's data to generate the schedule, and then apply the schedule on the following new 6-week period to evaluate the performance. In the non-robust scenario, we solve a deterministic model to optimize the schedule on the data in the previous period 6-week period.

4. *Robust optimization (a)*: As demand keeps changing over time, the deterministic approach might not be robust against uncertain demand deviations between the previous and new 6-week periods. We thus use the robust model described in Section 2.2.2 to protect against possible insufficiency due to such demand uncertainty.

5. *Robust optimization (b)*: As mentioned in Section 2.2.4, ED leadership is interested in quantifying the benefits of strategic changes in the shift type design. This variation considers changing one of the shift types to another 12-hour period.

6. *Robust optimization (c)*: The last variation considers adding a new shift type spanning another 12-hour period on top of the three existing ones.

Comparison results among approaches. We test the above approaches on data from three overlapping 6-week periods (beginning one week apart) from October 26 - December 6 to November 9 - December 20 in 2020. Table 2.1 shows the daily staffing cost and insufficiency as well as relative reductions from the current status among the three tested periods using the 6 different approaches. Currently, each day has 56 nurse shifts and an average daily insufficiency of 0.17 nurse-shifts. On average, rearranging current shifts optimally given oracle information can reduce the daily cost to 49.81 (11.05% reduction) and daily insufficiency to 0.02 (86.96% reduction). This suggests significant room for improvement from the current status quo in both quality measures. By rearranging the current shifts prospectively, the non-robust schedule reduces cost by 9.95% on average but increases insufficiency by 22.55%. We generate robust schedules with more protection against demand deviations with three variations. With the current shift types, robust schedule (a) can on average reduce

Table 2.1: Results of Average Outcomes (± Standard Feviation) from 6 Approaches.

Schedule approach	Shift types	Daily cost	Cost reduction [%]	Daily insufficiency	Insufficiency reduction [%]
Current	7-7, 11a-11p	56	-	0.17 (± 0.06)	-
Oracle	7-7, 11a-11p	49.81 (± 0.46)	11.05 (± 0.82)	0.02 (± 0.00)	86.96 (± 0.33)
Deterministic	7-7, 11a-11p	50.42 (± 0.38)	9.95 (± 0.67)	0.21 (± 0.04)	-22.55 (± 79.57)
Robust (a)	7-7, 11a-11p	51.33 (± 0.50)	8.33 (± 0.90)	0.12 (± 0.03)	27.70 (± 81.31)
Robust (b)	7-7, 12p-12a	50.57 (± 0.49)	9.69 (± 0.88)	0.15 (± 0.07)	13.18 (± 72.64)
Robust (c)	7-7, 11a-11p, 2p-2a	50.57 (± 0.38)	9.69 (± 0.67)	0.15 (± 0.07)	11.45 (± 74.95)

cost to 51.33 (8.33% reduction) and reduce insufficiency to 0.12 (27.70% reduction). Changing or adding 12-hour shift types, robust schedules (b) and (c) can further reduce average cost to 50.57 (9.69% reduction) with average insufficiency of 0.15 (11.45-13.18% reduction). The results show that adding or changing shift types could reduce staffing cost by an additional 1.36% while increasing daily insufficienty with 0.03. After reviewing with the nurse leadership, we decide to recommend the robust schedule (a) for its ability to reduce both cost and insufficiency, as well as its feasibility to be implemented without disruptive changes in shift types. Meanwhile, the potential benefit of adding an afternoon shift uncovers this possibility to be later implemented in 2023.

Trade-off between cost and insufficiency. There exists a trade-off between cost and insufficiency, as staffing more nurses facilitates more sufficiency with increased cost, and vice versa. We illustrate the flexibility of the optimization model to control such trade-offs by varying parameters μ_1 and β_ϵ in Figure 2-2, where scatter points from each approach represent the daily average insufficiency and cost of schedules generated with different parameters. This earlier version of the model did not include the objective term with μ_2 or the parameter ω_w. The current schedule has a daily average cost of 56 and insufficiency of 0.17. Retrospectively, given perfect information, schedules optimized with different parameters range from having 50.43 cost and 0.01 insufficiency to having 48.57 cost and 0.06 insufficiency. Prospectively with the data-driven approach, the non-robust schedules lead to a reduction of cost to 49.43 - 50.90 with the trade-off of higher insufficiency between 0.26 and 0.20. The robust schedules

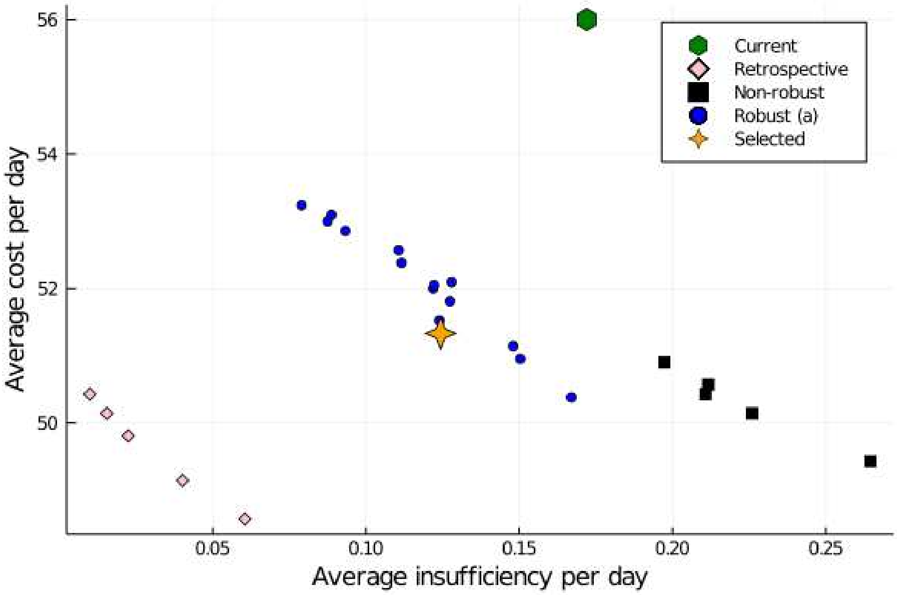


Figure 2-2: Cost-insufficiency Trade-offs.

Table 2.2: Recommended Schedule Change (Shift 1: 7a-7p, 2: 11a-11p, 3: 7p-7a).

Pod	Main			Red			Purple		
Shift	1	2	3	1	2	3	1	2	3
Mon	-2	+4	-2	-1	0	-1	-1	0	0
Tue	-2	+3	0	-1	0	-1	0	0	0
Wed	-2	+3	0	-1	+1	-1	0	0	-1
Thu	-1	+3	-2	-1	0	-1	-1	0	0
Fri	-1	+2	-1	-1	+1	-1	0	0	-1
Sat	-4	+1	-4	-1	+1	-1	-1	0	-1
Sun	-3	+1	-3	-1	0	-1	-1	0	-1

(a) incorporate uncertain demand deviations to trade off some cost reduction for more sufficiency, ranging from giving 50.38 cost and 0.17 insufficiency to 53.24 cost and 0.08 insufficiency. One of them is selected for schedule illustration next.

Illustration of schedule change. We illustrate the robust schedule (a) for November 2 - December 13 with selected parameter values. Compared with the current schedule, Table 2.2 reports the number of additional nurses compared with the status quo for each shift on each day of the week at main, red, and purple pods (e.g. +2 denotes adding 2 more shifts and -1 denotes reducing 1 shift). Staffing levels at other positions (CNL, first nurse, triage, and iTrack) are not changed per request from ED

Table 2.3: Recommended Level Change at Iteration 1.

Pod	Blue			Green			Orange			Purple		
Shift	7a	11a	7p	7a	11a	7p	7a	11a	7p	7a	11a	7p
Sun	-1	0	-1	-1	0	-1	-2	0	-2	-1	0	-1
Mon	-1	+1	-1	0	0	0	-1	+1	-1	0	0	0
Tue	0	0	0	0	0	0	-1	+1	-1	0	0	0
Wed	0	0	0	0	0	0	0	0	0	0	0	0
Thur	-1	+1	-1	0	0	0	-1	+1	-1	0	0	0
Fri	0	0	0	0	0	0	-1	0	0	0	0	0
Sat	-1	+1	-1	-1	0	0	-2	+1	-2	-1	0	0

leadership. We recommend having more 11 am–11 pm shifts (especially on weekdays) and fewer 7-7 shifts (especially on weekends). Changes in each shift vary from 0–4 for the main pods and 0–1 for the red pod or purple pod. This recommendation reduces average daily cost by 7.40% (from 56 to 51.86) and insufficiency by 47.87% (from 0.18 to 0.10) during this period. We further show our better match of staffing levels with demand patterns and provide the full schedule by the optimization approach in the Appendix section A.2.

2.4.2 Iterating Towards Implementable Schedules

Adjusting aggregate schedule with ED leadership feedback. After presenting the positive results to Hartford Hospital in the previous section, we gained support from the hospital executives. In 2021, we worked closely with the nurse leadership and had multiple iterations to make the schedules more practical for implementation. We made adjustments to the model's input parameters to better align with the needs of the nursing staff. For instance, we fixed the red pod staffing to three all the time and introduced the possibility for nurses to float from the red pod to the main pods when demand in the red pod is low. After adjusting the parameters, we obtained the first iteration of output aggregate level changes (numbers of nurses increased or decreased from the current schedule), as shown in Table 2.3, for the staffing cycle from May 30 - July 10, 2021.

The results suggested the need for more 11 am–11 pm shifts, fewer 7 am–7 pm, 7 pm–7 am, and weekend shifts, which would reduce the total number of shifts by 8%

Table 2.4: Recommended Level Change at Iteration 2.

Pod	Blue			Green			Orange			Purple		
Shift	7a	11a	7p	7a	11a	7p	7a	11a	7p	7a	11a	7p
Sun	-1	0	-1	-1	0	-1	-2	0	-2	-1	0	-1
Mon	0	0	0	0	0	0	0	0	0	0	0	0
Tue	0	0	0	0	0	0	-1	+1	-1	0	0	0
Wed	0	0	0	0	0	0	0	0	0	0	0	0
Thur	0	0	0	0	0	0	0	0	0	0	0	0
Fri	0	0	0	0	0	0	-1	0	0	0	0	0
Sat	-1	+1	-1	-1	0	-1	-1	0	-1	-1	0	0

and better match demand patterns for this overstaffing period. However, the nursing leadership was reluctant to reduce staffing levels significantly, citing various concerns. First, since the ED staffing levels were kept in the current way for decades, it was challenging for human decision-makers to break the tradition and adopt innovations. To lessen this concern, we added the penalty term for the number of changes from the current schedule with weight μ_2, which represents the willingness to change. Second, due to the high risk of ED patients, the nurses are among the most conservative people to reduce staffing, despite their high cost and employee dissatisfaction. We addressed this problem by tuning the trade-off parameters μ_1 and β_ϵ and providing options for them to adjust the parameters based on their safety level. Third, being aware of the larger fluctuations in ED demands than usual due to COVID-19, the leadership raised concerns about how well the previous weeks' data can capture trends for the future 6-week period. Thanks to this concern, we added the parameter ω_w to increase penalization weights on more recent weeks' data to capture more recent trends. The output schedule changes after these adjustments, shown in Table 2.4, had a reduced magnitude of changes, especially on weekdays, which made nurse leadership more comfortable in adopting our recommendation. Overall, these adjustments helped build closer relationships and trust between the research team and the nursing team, resulting in more practical and implementable schedules.

Benefits of individual schedule optimization. With an implementable aggregate schedule, we solve the individual scheduling model based on collected nurse

Table 2.5: Comparison of Alternative Registered Nurse (RN) Schedules.

Variation	Default	Less overtime	Shift change	EDOU float
Overtime shifts	72	52	8	48
Training shifts	72	60	40	60
Weekends turned on	11	11	6	11
Holidays turned on	2	2	2	2
RN with shift type changed	0	0	4	0
EDOU RN relocated to ED	0	0	0	2

preferences to make practical assignments. We consider a roster of 110 ED nurses (excluding per diem nurses) and 26 EDOU nurses. We solve and compare four schedules with alternative settings and parameters in Table 2.5. Compared with the solution with default parameters, a variation with increased weights on overtime shifts reduces overtime with a trade-off of less training. On top of the variation, allowing changes in nurses' shift types significantly reduces overtime shifts by changing four nurses' shift types. Alternatively, assigning two EDOU nurses to float to work at ED can bring several more overtime reductions. After reviewing with ED leadership, we decided to use the schedule with shift changes due to its best overall metrics.

We then show a metric comparison between the selected schedule and the current one in Table 2.6. The optimization reduces the number of daily shifts by 5%, with a higher reduction on weekends (17%) and holidays (14%). Additionally, the optimization leads to a remarkable 85% reduction in overtime shifts. Such savings in staffing cost and nurse workload at the ED is achieved while maintaining sufficient staffing: less than 0.1% of patient demand exceeds target patient-to-nurse ratios. Furthermore, the optimized schedule expands training shift opportunities by nearly one shift per day on average. By assigning each nurse to over four different positions on average per staffing cycle, the optimized schedule introduces desirable diversity, whereas the current schedule assigns most nurses to the same position throughout the staffing cycle. The optimized schedule reduces the individual nurses' dissatisfaction penalty score by 5%, while also ensuring fairness among their scores. Overall, our optimization significantly improves operational efficiency and nurse satisfaction in the ED.

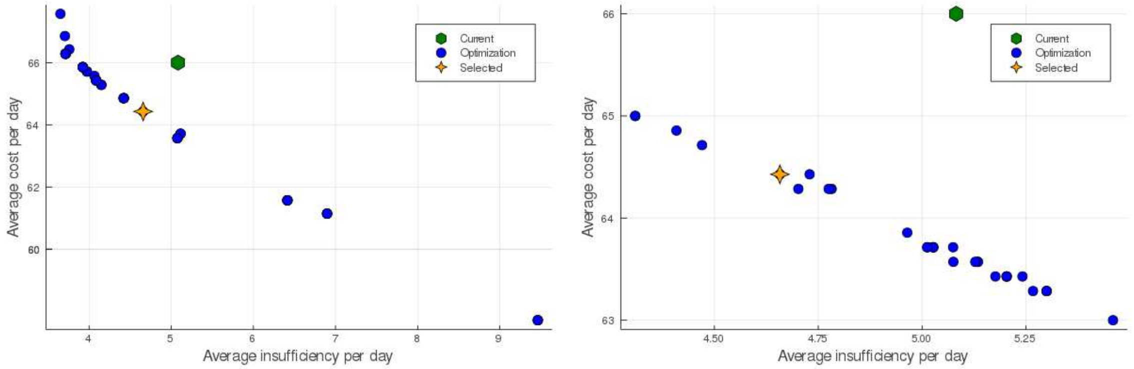
Table 2.6: Comparison of Current vs Optimized Schedule.

Metric	Current	Recommended	Reduction
Work shifts per day	48.3	45.7	5%
Per weekend day	48.0	40.0	17%
Per holiday	48.5	41.5	14%
Overtime per day	1.3	0.2	85%
Insufficiency per day	<0.1	<0.1	-
Training shifts per day	0	0.95	-
Different positions per nurse	1	4.6	-
Weighted dissatisfaction score	101	96	5%

2.4.3 Reducing Insufficiency during High Demand Periods

During the second half of 2022, the demand in the ED kept increasing and reached even higher than in the pre-COVID-19 period, resulting in severe nurse understaffing. In correspondence to the demand change, the ED changed the structure of their positions and increased the number of nurses. We apply the same model developed from the previous period into this new period while adjusting the ED positions and tiers input and re-tuning model parameters. We demonstrate the ability of our model to shift the primary benefit from reducing overstaffing under low demand into reducing insufficiency under high demand. To further facilitate automating the scheduling in adjustment to demand variation, we investigate the benefit of resolving the aggregate model and adjusting the schedule every week during the 6-week staffing cycle.

Parameter Tuning Illustration. We re-tune the parameters of the updated model, which incorporated additional parameters μ_2 and ω_w, to illustrate the objective trade-offs. We first tune μ_1 and β_ϵ, the two main parameters that control the trade-off between cost and insufficiency. In Figure 2-3a, we show the different combinations where μ_1 drawn from the list {0, 0.2, 0.3, 0.4, 0.6, 0.8, 1} and β_ϵ from {0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6} with fixed μ_2 and ω_w. From the plot, different parameter combinations lead to a wide range of metric trade-offs. Several parameters result in higher insufficiency than the current schedule to save cost, which is not desirable for the ED as patient care is the priority. Several other parameters increase staffing to



(a) Varying Parameters μ_1 and β_ϵ.

(b) Varying Parameters μ_2 and ω_w.

Figure 2-3: Objective Trade-offs during Staffing Cycle December 4, 2022 - January 14, 2023.

reduce insufficiency, which provides strategic insights on how hiring more nurses and allocating them with optimization can reduce understaffing more significantly. For schedule implementation with the current nurses, we select $\mu_1 = 0.2$ and $\beta_\epsilon = 0.3$ that reduce both cost and insufficiency simultaneously.

With μ_1 and β_ϵ fixed, we then fine-tune the other parameters. We vary μ_2 to regularize the number of changes from the current schedule and ω_w to adjust penalization weights based on the recency of training data. We use an auxiliary parameter $\omega \in [0, 1]$ and let $\omega_1, \ldots, \omega_6$ to be uniformly distributed between $1 - \omega$ and $1 + \omega$. In Figure 2-3b, we vary μ_2 from {0, 0,1, 0.2, 0.3, 0.4, 0.5} and vary ω from {0, 0.1, 0.2, 0.25, 0.3, 0.4, 0.5}. We finally select $\mu_2 = 0.2$ and $\omega = 0.3$ together with $\mu_1 = 0.2$ and $\beta_\epsilon = 0.3$, which reduces insufficiency 8% (from 5.08 to 4.66 shifts per day) with 2.38% less staffing (from 66 to 64.43 daily shifts).

Adjustment from resolving aggregate model weekly. We consider another variation to solve the aggregate model every week. We conduct experiments on three adjacent 6-week staffing cycles between July 31 and December 3, 2022. We compare three alternatives:

1. Current schedule.

2. The model is solved every 6 weeks and generates a schedule for each 6-week

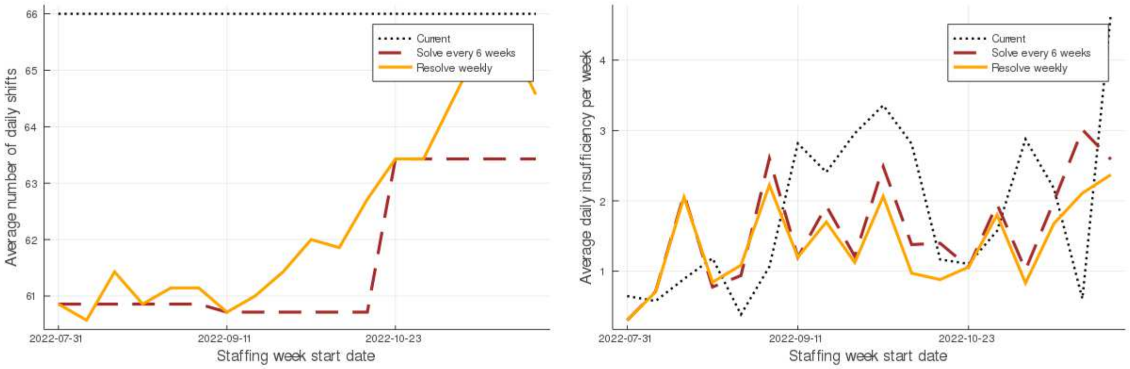
Table 2.7: Average Daily Results and Change (δ) from Current Schedules on Three Staffing Cycles during July 31 - December 3, 2022.

Schedule approach	Cost [δ]	Insufficiency [δ]	δz from current	δz by week
Current	66	1.85	0	0
Solve every 6 weeks	61.67 [-6.57%]	1.59 [-13.86%]	4.71 [7.14%]	0
Resolve weekly	62.37 [-5.51%]	1.39 [-24.83%]	4.26 [6.46%]	2.08 [3.33%]

staffing cycle. This requires solving the model three times for the three staffing cycles.

3. For each staffing cycle, the model is resolved every week using the training data one week later and update the schedule for the remaining weeks of the cycle. This involves solving the model 18 times for the 18-week period.

For each week, we compute the average number of daily shifts and average daily insufficiency during that week's schedule, and plot the three variations during the time period in Figure 2-4. Resolving the model weekly changes in staffing levels from week to week and leads to improved insufficiency in most weeks. We summarize the average daily metrics among the 18 weeks for the three alternatives in Table 2.7. On average, resolving weekly changes the staffing levels by 2.08 shifts from week to week. This would require the nurse leadership to ask several nurses to request shift changes from their pre-arranged schedule. Solving the model every 6 weeks (resp. every week) on average changes the staffing levels from the current schedule by 4.71 (resp. 4.26) shifts. In return, resolving weekly can reduce the shortage of nurse shifts to meet target nurse-patient-ratios by 24.83% from the current schedule, which is larger than 13.86% reduction by solving the model every 6 weeks. The benefit of reducing insufficiency is particularly useful for this period when the ED is extremely understaffed and is achieved with 5.51% (resp. 6.57%) less staffing cost from solving the model every 6 weeks (resp. every week).



(a) Daily Number of Shifts.

(b) Daily Insufficiency.

Figure 2-4: Change from Resolving Weekly during Three Staffing Cycles.

2.5 Implementation

In collaboration with a data consultancy company and a development team, we implement the models into end-to-end software for ED nursing to optimize scheduling. The process is summarized in Figure 2-5, which outlines the four phases of scheduling: preparation, input entrance, solution generation, and schedule output. The flowchart depicts the interactions between the software (blue rounded rectangle), ED nurse managers (red rectangle), and individual nurses (green oval) at Hartford Hospital

Our software revolutionizes the ED nurse scheduling process as follows. The implementation begins with nurse managers setting up general configurations and nurse information in the software, which then enables individual nurses to log into their accounts and enter general preferences. This phase is completed prior to the first use and can be updated whenever necessary, whereas the remaining three phases repeat for every 6-week staffing cycle. Approximately five weeks before the start of each cycle, the software collects nurses' and managers' requests specific to the upcoming cycle and then passes them as inputs to the optimization model. Four weeks before the cycle, the software automatically solves the first-stage aggregate model, allowing managers to edit the output aggregate levels and run the program to generate schedules with parameters of their choice. After managers edit, compare, and select the schedules, the software outputs reports based on templates for managers to announce to nurses three weeks ahead of the staffing cycle. If nurses request to swap or change

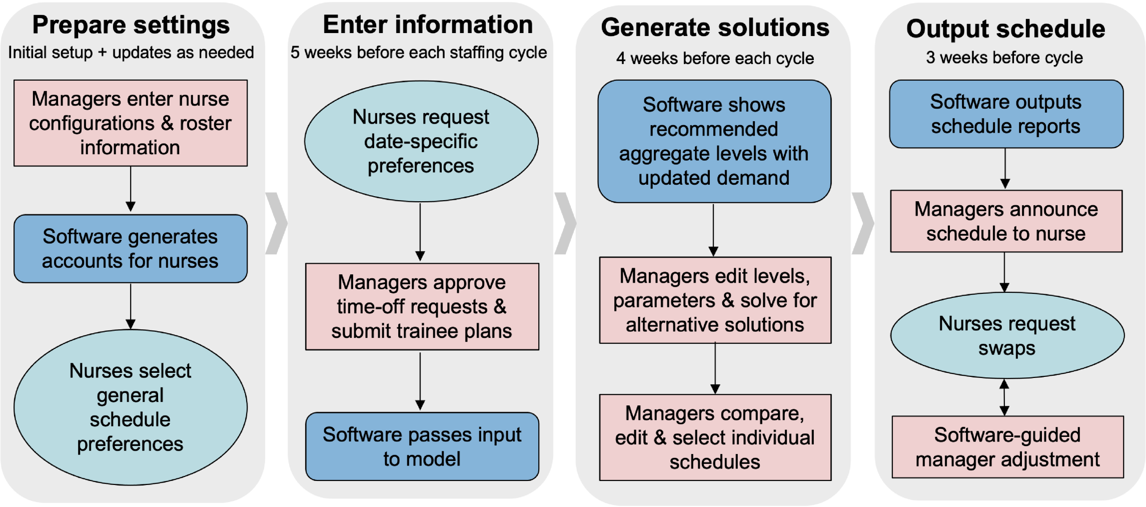

Prepare settings
Initial setup + updates as needed
Managers enter nurse configurations & roster information
Software generates accounts for nurses
Nurses select general schedule preferences
Enter information
5 weeks before each staffing cycle
Nurses request date-specific preferences
Managers approve time-off requests & submit trainee plans
Software passes input to model
Generate solutions
4 weeks before each cycle
Software shows recommended aggregate levels with updated demand
Managers edit levels, parameters & solve for alternative solutions
Managers compare, edit & select individual schedules
Output schedule
3 weeks before cycle
Software outputs schedule reports
Managers announce schedule to nurse
Nurses request swaps
Software-guided manager adjustment

Figure 2-5: Process of Integrating Decision-support Software into ED Nurse Staffing.

shifts from the announced schedule, the software guides managers to accept or reject the changes based on staffing shortages and surpluses at each shift. The end-to-end implementation was partially used in March 2023 and fully deployed since April 2023.

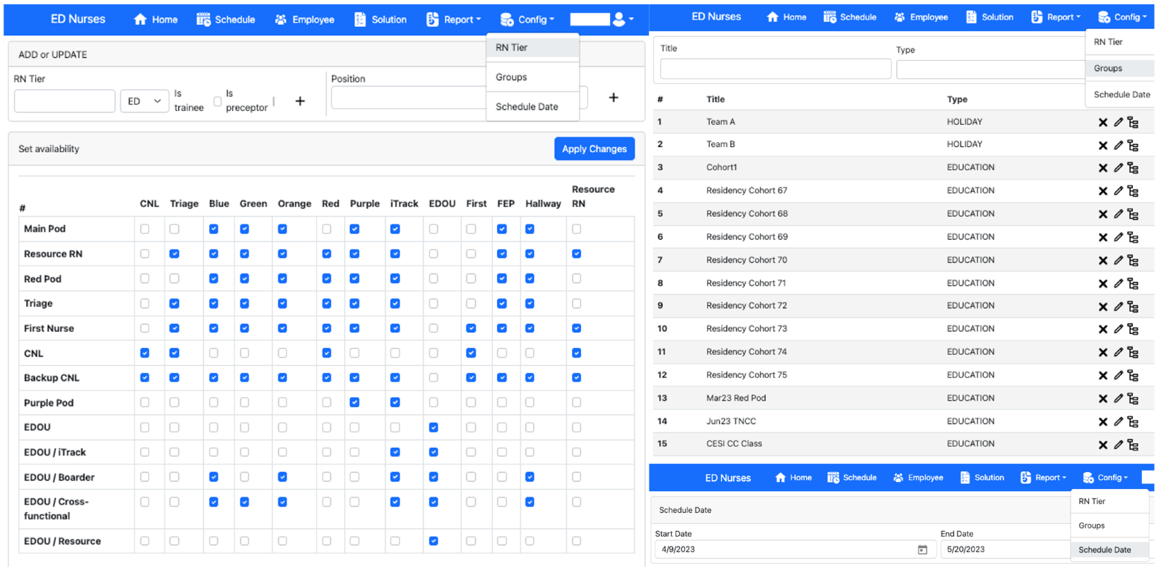
2.5.1 Software Illustration

In this section, we demonstrate various components of the software from input collection to output generation.

Collecting input. Figure 2-6 shows the two sections to gather input from nurse managers and individual nurses. The Config tab, as shown in Figure 2-6a, enables nurse managers to set general staffing inputs: the RN Tier page specifies eligibilities for nurses of each tier to work at each position; the Groups page assigns nurses into different cohorts with corresponding unavailability dates for each cohort due to holiday off or education; the Schedule Date page sets the start and end date of the next staffing cycle to schedule for. These configurations set up the structures for the Employee tab shown in Figure 2-6b, which collects information and preferences from individual nurses. Managers (ED leadership, head nurses, and scheduling assistants) can modify the nurse roster by adding, removing employees, or importing an Excel file. They can also edit each nurse's account settings and basic information with a click. Once the forms are released to all nurses, they can log into their accounts to enter additional shift availability, shift pattern preferences, dates they prefer to work on and request paid or unpaid time off. After nurses enter their preferences, managers can approve or deny their requested days off and finalize all nurse inputs for the software to pass to the optimization models.

Generating solutions. After collecting nurse input, nurse managers can use the software's output sections, as demonstrated in Figure 2-7, to generate and output schedules. In the Solutions Tab, managers can set parameters, solve the model, and obtain a solution with a specified name, as illustrated in Figure 2-7a. They can review and edit the recommended aggregate levels in the table, and select parameters such as whether or not to allow nurses to change shift types, to include training schedules, and to incorporate fairness. The software offers additional flexibility to solve for a subset of nurses or tiers, such as solving for ED only or EDOU nurse only. The managers can experiment with different parameters and generate alternative solutions to compare, as shown in Figure 2-7b. In the schedule table, each nurse is represented by a row, containing basic information and assigned shifts for each position (in different colors) on specific dates. Options to filter or search for specific nurses, as well as control which columns to display, are also available.

Adjusting schedules. Due to dynamic changes in staffing needs and nurse availabilities, the support for schedule adjustments is deemed essential. As seen at the bottom of Figure 2-7b, aggregate staffing levels from the individual schedule above (first number) are presented in comparison to target levels (second number) for each shift or 4-hour block at any subset of positions selected. For example, on Thursday, April 13, the red pod staffs 4 nurses during blocks 7a-11a, 11a-3p, and 3p-7p, which has a surplus of one nurse during the first block (highlighted in red) and is the same as the target aggregate level during the last two blocks. On Friday and Saturday, the red pod has a shortage of one nurse during 11a-3p and 3p-7p (highlighted in orange). Every week, the software also resolves the model with updated demand and



ADD or UPDATE

RN Tier: ED ☐ Is trainee ☐ Is preceptor +

Position: +

Set availability — Apply Changes

#	CNL	Triage	Blue	Green	Orange	Red	Purple	iTrack	EDOU	First	FEP	Hallway	Resource RN
Main Pod	☐	☐	☑	☑	☑	☐	☑	☑	☐	☐	☑	☑	☐
Resource RN	☐	☑	☑	☑	☑	☑	☑	☑	☐	☐	☑	☑	☑
Red Pod	☐	☐	☑	☑	☑	☑	☑	☑	☐	☐	☑	☑	☐
Triage	☐	☑	☑	☑	☑	☑	☑	☑	☐	☐	☑	☑	☐
First Nurse	☐	☑	☑	☑	☑	☑	☑	☑	☐	☑	☑	☑	☑
CNL	☑	☑	☐	☐	☐	☑	☐	☐	☐	☑	☐	☐	☑
Backup CNL	☑	☑	☑	☑	☑	☑	☑	☑	☐	☑	☑	☑	☑
Purple Pod	☐	☐	☐	☐	☐	☐	☑	☑	☐	☐	☐	☐	☐
EDOU	☐	☐	☐	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐
EDOU / iTrack	☐	☐	☐	☐	☐	☐	☐	☑	☑	☐	☐	☐	☐
EDOU / Boarder	☐	☐	☑	☐	☑	☐	☐	☑	☑	☐	☐	☑	☐
EDOU / Cross-functional	☐	☐	☑	☑	☑	☐	☐	☑	☑	☐	☐	☑	☐
EDOU / Resource	☐	☐	☐	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐

Title — Type

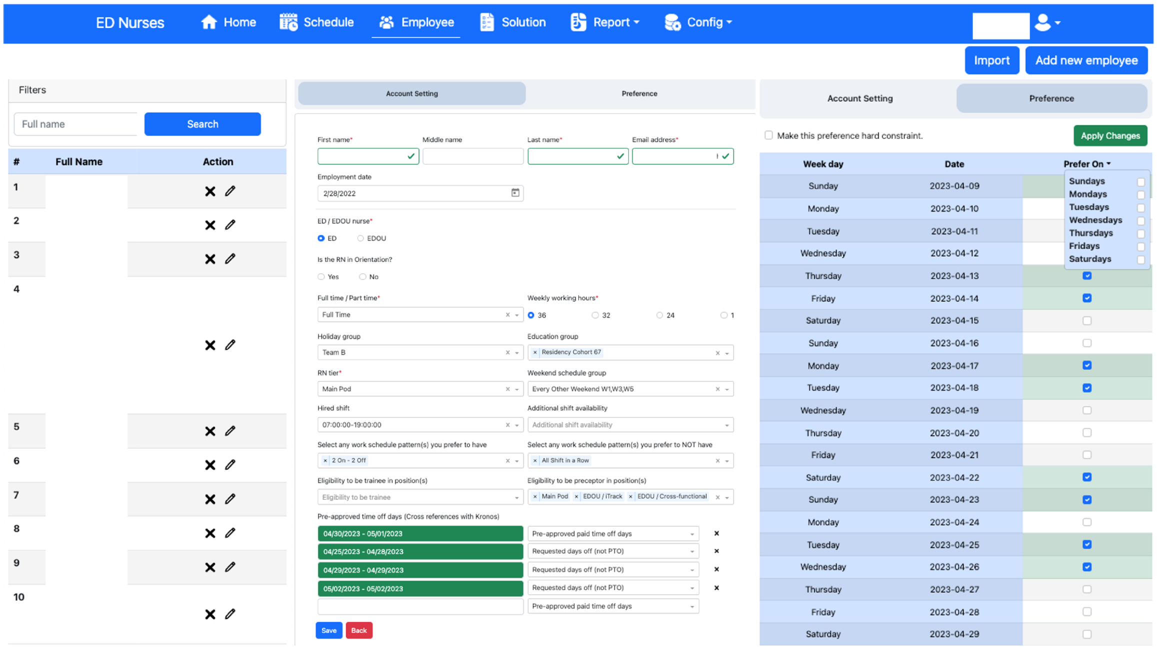
#	Title	Type
1	Team A	HOLIDAY
2	Team B	HOLIDAY
3	Cohort1	EDUCATION
4	Residency Cohort 67	EDUCATION
5	Residency Cohort 68	EDUCATION
6	Residency Cohort 69	EDUCATION
7	Residency Cohort 70	EDUCATION
8	Residency Cohort 71	EDUCATION
9	Residency Cohort 72	EDUCATION
10	Residency Cohort 73	EDUCATION
11	Residency Cohort 74	EDUCATION
12	Residency Cohort 75	EDUCATION
13	Mar23 Red Pod	EDUCATION
14	Jun23 TNCC	EDUCATION
15	CESI CC Class	EDUCATION

Schedule Date

Start Date: 4/9/2023

End Date: 5/20/2023

(a) Configuration Settings.



#	Full Name	Action
1		✕ ✎
2		✕ ✎
3		✕ ✎
4		✕ ✎
5		✕ ✎
6		✕ ✎
7		✕ ✎
8		✕ ✎
9		✕ ✎
10		✕ ✎

First name* · Middle name · Last name* · Email address*

Employment date: 2/28/2022

ED / EDOU nurse*: ● ED ○ EDOU

Is the RN in Orientation? ○ Yes ○ No

Full time / Part time*: Full Time

Weekly working hours*: ● 36 ○ 32 ○ 24 ○ 1

Holiday group: Team B

Education group: Residency Cohort 67

RN tier*: Main Pod

Weekend schedule group: Every Other Weekend W1,W3,W5

Hired shift: 07:00:00-19:00:00

Additional shift availability: Additional shift availability

Select any work schedule pattern(s) you prefer to have: 2 On - 2 Off

Select any work schedule pattern(s) you prefer to NOT have: All Shift in a Row

Eligibility to be trainee in position(s): Eligibility to be trainee

Eligibility to be preceptor in position(s): Main Pod, EDOU / iTrack, EDOU / Cross-functional

Pre-approved time off days (Cross references with Kronos)

04/30/2023 - 05/01/2023	Pre-approved paid time off days	✕
04/25/2023 - 04/28/2023	Requested days off (not PTO)	✕
04/29/2023 - 04/29/2023	Requested days off (not PTO)	✕
05/02/2023 - 05/02/2023	Requested days off (not PTO)	✕
	Pre-approved paid time off days	

☐ Make this preference hard constraint. — Apply Changes

Prefer On: Sundays ☐ Mondays ☐ Tuesdays ☐ Wednesdays ☐ Thursdays ☐ Fridays ☐ Saturdays ☐

Week day	Date	Prefer On
Sunday	2023-04-09	
Monday	2023-04-10	
Tuesday	2023-04-11	
Wednesday	2023-04-12	
Thursday	2023-04-13	☑
Friday	2023-04-14	☑
Saturday	2023-04-15	☐
Sunday	2023-04-16	☐
Monday	2023-04-17	☑
Tuesday	2023-04-18	☑
Wednesday	2023-04-19	☐
Thursday	2023-04-20	☐
Friday	2023-04-21	☐
Saturday	2023-04-22	☑
Sunday	2023-04-23	☑
Monday	2023-04-24	☐
Tuesday	2023-04-25	☑
Wednesday	2023-04-26	☑
Thursday	2023-04-27	☐
Friday	2023-04-28	☐
Saturday	2023-04-29	☐

(b) Employee Information and Preferences.

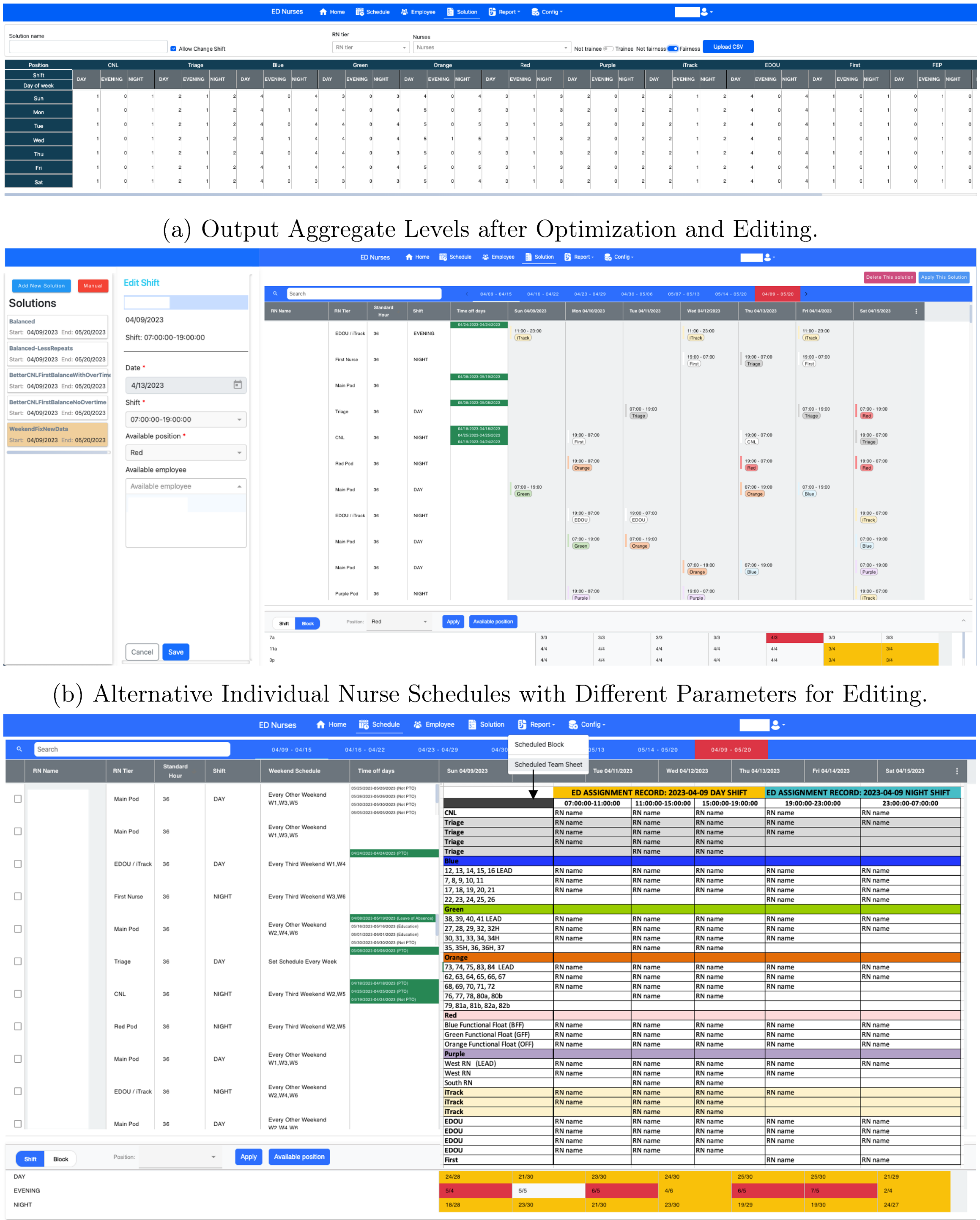
Figure 2-6: Software Sections for Nurse Input Collection.

alerts managers of any expected changes. This component advises managers where shortages or surpluses exist across the pods and blocks after the schedule is built, suggesting adjustments to be made. Also demonstrated in Figure 2-7b, managers can execute any schedule changes by editing shifts, dates, positions, or nurses with simple clicks. The combination of these functionalities supports a variety of schedule adjustments: It suggests how each shortage can be filled if specific nurses can work overtime to cover the spot, advising managers to know which nurses to ask and staff at which shift to reduce the shortage. If nurses request shift swaps due to availability or preference change, the aggregate summary encourages managers to approve those requests to swap from a surplus block into a shortage block. If two nurses request to swap with each other, the list of available positions and employees while editing shifts automatically checks the feasibility of the switch and enables managers to approve it if possible. To cover additional shortages, managers can add shifts on per diem nurses (who do not work full-time but pick up shifts on demand anytime) to fill in the shortage shifts.

Announcing Outputs. Finally, ED managers can output the finalized schedule as shown in Figure 2-7c. The Schedule tab shows the determined schedule, where nurse managers have a holistic view and each nurse can view their own schedule. In addition, the Report tab enables managers to download and print reports of the schedule based on the ED template, including a summary of staffing for each cycle and a daily team sheet for each date. The screenshot includes an example team sheet on April 9, 2023, which assigns nurse names at each position during each time block for both day and night shifts of the day. Overall, the software provides decision support as well as aligns with ED operations.

2.5.2 Financial Benefit Estimation

The schedule optimization with its implementation translates into substantial financial benefits projected by Hartford Hospital. The reduction of nursing hours from Table 2.6 and Table 2.8 converts into estimated cost savings in Table 2.9. By saving



Solution name

☑ Allow Change Shift

RN tier: RN tier

Nurses: Nurses

Position	CNL			Triage			Blue			Green			Orange			Red			Purple			iTrack			EDOU			First			FEP		
Shift Day of week	DAY	EVENING	NIGHT	DAY	EVENING	NIGHT	DAY	EVENING	NIGHT	DAY	EVENING	NIGHT	DAY	EVENING	NIGHT	DAY	EVENING	NIGHT	DAY	EVENING	NIGHT	DAY	EVENING	NIGHT	DAY	EVENING	NIGHT	DAY	EVENING	NIGHT	DAY	EVENING	NIGHT
Sun	1	0	1	2	1	2	4	0	4	3	0	3	4	0	4	3	1	3	2	0	2	2	1	2	4	0	4	1	0	1	0	1	0
Mon	1	0	1	2	1	2	4	1	4	4	0	4	5	0	5	3	1	3	2	0	2	2	1	2	4	0	4	1	0	1	0	1	0
Tue	1	0	1	2	1	2	4	0	4	4	0	4	5	0	5	3	1	3	2	0	2	2	1	2	4	0	4	1	0	1	0	1	0
Wed	1	0	1	2	1	2	4	1	4	4	0	4	5	1	5	3	1	3	2	0	2	2	1	2	4	0	4	1	0	1	0	1	0
Thu	1	0	1	2	1	2	4	0	4	4	0	3	5	0	5	3	1	3	2	0	2	2	1	2	4	0	4	1	0	1	0	1	0
Fri	1	0	1	2	1	2	4	1	4	4	0	4	5	0	5	3	1	3	2	0	2	2	1	2	4	0	4	1	0	1	0	1	0
Sat	1	0	1	2	1	2	4	0	3	3	0	3	5	0	4	3	1	3	2	0	2	2	1	2	4	0	4	1	0	1	0	1	0

(a) Output Aggregate Levels after Optimization and Editing.

Solutions

Balanced
Start: 04/09/2023 End: 05/20/2023

Balanced-LessRepeats
Start: 04/09/2023 End: 05/20/2023

BetterCNLFirstBalanceWithOverTime
Start: 04/09/2023 End: 05/20/2023

BetterCNLFirstBalanceNoOvertime
Start: 04/09/2023 End: 05/20/2023

WeekendFixNewData
Start: 04/09/2023 End: 05/20/2023

Edit Shift

04/09/2023

Shift: 07:00:00-19:00:00

Date *: 4/13/2023

Shift *: 07:00:00-19:00:00

Available position *: Red

Available employee: Available employee

Search | 04/09 - 04/15 | 04/16 - 04/22 | 04/23 - 04/29 | 04/30 - 05/06 | 05/07 - 05/13 | 05/14 - 05/20 | 04/09 - 05/20

RN Name	RN Tier	Standard Hour	Shift	Time off days	Sun 04/09/2023	Mon 04/10/2023	Tue 04/11/2023	Wed 04/12/2023	Thu 04/13/2023	Fri 04/14/2023	Sat 04/15/2023
	EDOU / iTrack	36	EVENING	04/24/2023-04/24/2023	11:00 - 23:00 iTrack			11:00 - 23:00 iTrack		11:00 - 23:00 iTrack	
	First Nurse	36	NIGHT					19:00 - 07:00 First	19:00 - 07:00 Triage	19:00 - 07:00 First	
	Main Pod	36		04/08/2023-05/19/2023							
	Triage	36	DAY	05/08/2023-05/08/2023			07:00 - 19:00 Triage			07:00 - 19:00 Triage	07:00 - 19:00 Red
	CNL	36	NIGHT	04/18/2023-04/18/2023 04/25/2023-04/25/2023 04/19/2023-04/24/2023		19:00 - 07:00 First			19:00 - 07:00 CNL		19:00 - 07:00 Triage
	Red Pod	36	NIGHT			19:00 - 07:00 Orange			19:00 - 07:00 Red		19:00 - 07:00 Red
	Main Pod	36	DAY		07:00 - 19:00 Green				07:00 - 19:00 Orange	07:00 - 19:00 Blue	
	EDOU / iTrack	36	NIGHT			19:00 - 07:00 EDOU	19:00 - 07:00 EDOU				19:00 - 07:00 iTrack
	Main Pod	36	DAY			07:00 - 19:00 Green	07:00 - 19:00 Orange				07:00 - 19:00 Blue
	Main Pod	36	DAY					07:00 - 19:00 Orange	07:00 - 19:00 Blue		07:00 - 19:00 Purple
	Purple Pod	36	NIGHT			19:00 - 07:00 Purple		19:00 - 07:00 Purple			19:00 - 07:00 iTrack

Position: Red

7a	3/3	3/3	3/3	3/3	4/3	3/3	3/3
11a	4/4	4/4	4/4	4/4	4/4	3/4	3/4
3p	4/4	4/4	4/4	4/4	4/4	3/4	3/4

(b) Alternative Individual Nurse Schedules with Different Parameters for Editing.

Search | 04/09 - 04/15 | 04/16 - 04/22 | 04/23 - 04/29 | 04/30 | 05/13 | 05/14 - 05/20 | 04/09 - 05/20

	RN Name	RN Tier	Standard Hour	Shift	Weekend Schedule	Time off days
□		Main Pod	36	DAY	Every Other Weekend W1,W3,W5	05/25/2023-05/26/2023 (Not PTO) 05/26/2023-05/26/2023 (Not PTO) 05/30/2023-05/30/2023 (Not PTO) 06/05/2023-06/05/2023 (Not PTO)
□		Main Pod	36		Every Other Weekend W1,W3,W5	
□		EDOU / iTrack	36	DAY	Every Third Weekend W1,W4	04/24/2023-04/24/2023 (PTO)
□		First Nurse	36	NIGHT	Every Third Weekend W3,W6	
□		Main Pod	36		Every Other Weekend W2,W4,W6	04/08/2023-05/19/2023 (Leave of Absence) 05/16/2023-05/16/2023 (Education) 06/01/2023-06/01/2023 (Education) 05/30/2023-05/30/2023 (Not PTO)
□		Triage	36	DAY	Set Schedule Every Week	05/08/2023-05/08/2023 (PTO)
□		CNL	36	NIGHT	Every Third Weekend W2,W5	04/18/2023-04/18/2023 (PTO) 04/25/2023-04/25/2023 (PTO) 04/19/2023-04/24/2023 (Not PTO)
□		Red Pod	36	NIGHT	Every Third Weekend W2,W5	
□		Main Pod	36	DAY	Every Other Weekend W1,W3,W5	
□		EDOU / iTrack	36	NIGHT	Every Other Weekend W2,W4,W6	
□		Main Pod	36	DAY	Every Other Weekend W2,W4,W6	

Sun 04/09/2023 | Tue 04/11/2023 | Wed 04/12/2023 | Thu 04/13/2023 | Fri 04/14/2023 | Sat 04/15/2023

	ED ASSIGNMENT RECORD: 2023-04-09 DAY SHIFT			ED ASSIGNMENT RECORD: 2023-04-09 NIGHT SHIFT	
	07:00:00-11:00:00	11:00:00-15:00:00	15:00:00-19:00:00	19:00:00-23:00:00	23:00:00-07:00:00
CNL	RN name	RN name	RN name	RN name	RN name
Triage	RN name	RN name	RN name	RN name	RN name
Triage	RN name	RN name	RN name	RN name	
Triage	RN name	RN name	RN name		
Triage		RN name	RN name		
Blue					
12, 13, 14, 15, 16 LEAD	RN name	RN name	RN name	RN name	RN name
7, 8, 9, 10, 11	RN name	RN name	RN name	RN name	RN name
17, 18, 19, 20, 21	RN name	RN name	RN name	RN name	RN name
22, 23, 24, 25, 26				RN name	RN name
Green					
38, 39, 40, 41 LEAD	RN name	RN name	RN name	RN name	RN name
27, 28, 29, 32, 32H	RN name	RN name	RN name	RN name	RN name
30, 31, 33, 34, 34H	RN name	RN name	RN name	RN name	
35, 35H, 36, 36H, 37		RN name	RN name		
Orange					
73, 74, 75, 83, 84 LEAD	RN name	RN name	RN name	RN name	RN name
62, 63, 64, 65, 66, 67	RN name	RN name	RN name	RN name	RN name
68, 69, 70, 71, 72	RN name	RN name	RN name	RN name	
76, 77, 78, 80a, 80b		RN name	RN name		
79, 81a, 81b, 82a, 82b					
Red					
Blue Functional Float (BFF)	RN name	RN name	RN name	RN name	RN name
Green Functional Float (GFF)	RN name	RN name	RN name	RN name	RN name
Orange Functional Float (OFF)	RN name	RN name	RN name	RN name	RN name
Purple					
West RN (LEAD)	RN name	RN name	RN name	RN name	RN name
West RN	RN name	RN name	RN name	RN name	
South RN		RN name	RN name		
iTrack	RN name	RN name	RN name	RN name	
iTrack	RN name	RN name	RN name		
iTrack		RN name	RN name		
EDOU	RN name	RN name	RN name	RN name	RN name
EDOU	RN name	RN name	RN name	RN name	RN name
EDOU	RN name	RN name	RN name	RN name	RN name
EDOU	RN name	RN name	RN name		
First				RN name	RN name

Position:

DAY	24/28	21/30	23/30	24/30	25/30	25/30	21/29
EVENING	5/4	5/5	6/5	4/6	6/5	7/5	2/4
NIGHT	18/28	23/30	21/30	23/30	19/29	19/30	24/27

(c) Output Final Schedule.

Figure 2-7: Software Sections for Generating Schedules.

Table 2.8: Estimated Time Spent by ED Leaders on Manual Scheduling.

Responsibility component	Responsible staff	Total time
Schedule build/balancing	Scheduler	12 hours
Schedule approval	Scheduler/Manager	1 hour
Management of shift switches	Scheduler/Manager/Assistants	4 hours weekly
Orientation scheduling	Educators/Managers/Scheduler	3 hours
Daily team sheet building	Manager/Assistants	8 hours weekly
Total hours per 6-week staffing cycle		88 hours

Table 2.9: Projected Savings of Nursing Hours and Cost for HH ED.

Component	Daily hours	Annual hours	Hourly cost	Annual cost
Base shifts	31.2	11,388.00	$50	$569,400.00
Overtime (extra)	13.2	4,818.00	$25	$120,450.00
Manual scheduling	2.1	764.76	$50	$38,238.10
Total				$728,088.10

2.6 shifts (12-hour long) daily, ED will save an average of 31.2 nursing hours per day, which annualizes into 11,388 hours per year. An average salary of an entry-level registered nurse with an average fringe of 24% is approximately $50 per hour at Hartford Hospital. With overtime salary being 50% higher than the base salary, the reduction of 1.1 overtime shifts per day cuts the cost with an additional $25 per hour. Combined with the 88-hour manual work per 6 weeks saved from Table 2.8, the hospital financial department projects that the optimization can save $728,088.10 annually in nursing costs.

2.5.3 Overcoming Challenges in Deployment

Deploying the optimization tool at Hartford Hospital automated and revolutionized ED nurse staffing. Such an initiative of changing the lives of 200 nurses, nurse schedulers, and managers carried numerous challenges. The deployment of the tool involved transitioning from an offline to an online scheduling process. In this section, we discuss how we overcame challenges, progressed with the deployment, and achieved practical impact.

Offline strategic adoption. The development started with offline iterations at the beginning of 2021. Researchers generated schedules offline for ED leadership to review, and then incorporated feedback and constraints into the models. As described in Section 2.4.2, researchers adapted the optimization models and parameters to overcome the nurse leadership's conservatism in reducing staffing levels. With refined models, the team communicated with hospital leadership to adopt our optimized schedules. The model suggested fewer 7-7 shifts and more 11-11 shifts in response to demand patterns. However, as each nurse was pre-assigned to a fixed shift type upon being hired, it was difficult to swap shifts without disrupting their lives. After strategic discussions, hospital leadership decided to hire more 11 am–11 pm shift nurses as opposed to 7-7 shift nurses to accommodate the recommendation. During this overstaffing period, the optimized schedule needed fewer nurse shifts than the hired nurse resources as shown in Sections 2.4.1 and 2.4.2. This presented another challenge as the hospital was still required to utilize all nursing hours as per the contract. After evaluating several options, the ED decided to turn the saved work shifts into additional training shifts, which count as working hours but do not serve patient demand.

Motivation from Offline to Online. Following the successful strategic adoptions at the ED, our next goal was to implement the generated individual nurse schedule for the ED to use. However, experimentation for the next staffing cycle showed that offline iterations between researchers solving models and managers providing feedback were not feasible to be adopted. Managers needed to propose changes to the schedule frequently for reasons such as nurse dissatisfaction, availability changes, and manager preferences. The frequent back-and-forth communications with the research teams caused delays and were not sustainable. On the other hand, manually reshuffling the machine-generated schedule was infeasible for nurse managers without optimization training. To address these issues, we decided to develop software for automation that would allow nurse managers to regenerate and edit schedules online.

Online implementation. In 2022, we began software development for staffing automation. The research team collaborated with consultancy and development teams and integrated the optimization models into the software. Given imported input data, users were able to solve the models and obtain the schedule under the Solution and Schedule tabs. However, the upcoming main challenge was to connect the software with the hospital to collect input from nurses and pass output for the ED to use. One complication arose as the entire nurse staffing at Hartford HealthCare relied on commercial software that links shift assignments to the nurse payroll system. Despite the team's attempts to integrate with the commercial software throughout 2022, the complexity of involving a third-party organization was beyond our control. After careful discussions between the team and executives at Hartford HealthCare, a decision was made to build a standalone software instead. In early 2023, we built the information collection component of the software. After frequent weekly meetings and iterations between researchers, the development team, nurse managers, and schedulers, the interface evolved to cover all ED nurse staffing functionalities. We further improved the schedule generation process, enabling dynamic schedule changes by nurse leadership to adapt to quick information changes. The ED executed more model recommendations after the long collaboration, such as swapping shift types for some nurses. The fine-tuned schedule announced on April 30, 2023 has no overtime shifts, satisfies 92% preferred date assignments, only 1 undesirable shift pattern among all nurses across 6 weeks, and assigns nurses with desirable diversity to average 2.43 different positions per week. Following the pilot ED deployment at Hartford Hospital, the Hartford HealthCare executive team aims to extend the models and software to cover all nurse staffing in 7 hospitals of the network in the future.

2.6 Conclusions

We develop models and implementation to optimize nurse staffing at Emergency Department at Hartford Hospital. Our methodology consists of two phases to optimize each 6-week staffing cycle. First, we learn an uncertainty set from patient demand data and develop a robust optimization model solved by the cutting planes method to allocate aggregate staffing levels and quantify strategic decisions. Next, we develop a pair of two mixed integer optimization models to generate individual nurse schedules for work, trainee, and preceptor shifts. Experimental results demonstrate the versatile benefit of the first-phase aggregate model: cutting staffing costs by 5–8% in low-demand periods versus reducing insufficiency by 8–25% during high-demand periods. In addition, we analyze how the outcomes change with different model variations and parameter trade-offs as well as illustrate schedule iterations and demand patterns. The second-phase individual schedule optimization brings significant benefits to ED nurses, reducing 17% weekend, 14% holiday, and 85% overtime shifts while increasing 5% satisfaction score, 3.6 more diverse positions, and 0.95 training per day. The models are implemented into end-to-end software that supports scheduling from staffing preparation, input collection, and solution generation, to schedule output. After numerous challenges were overcome, the software was deployed starting March 2023 at Hartford Hospital, automating and transforming the nurse scheduling at the Emergency Department. The implementation relieved manual scheduling burdens and is projected to save $728,088.10 in annual nursing costs. brought more cost-effective, sufficient, and desirable staffing into practice, benefiting various stakeholders in the hospital system.

Chapter 3

Patient Outcome Predictions Improve Hospital Operations at Hartford HealthCare

3.1 Introduction

The collection of patient-level information from Electronic Medical Records (EMRs) combined with advances in machine learning methodology creates opportunities to enhance hospital operations and clinical decision making, especially for inpatients, who constitute a major part of hospital activities and revenues.

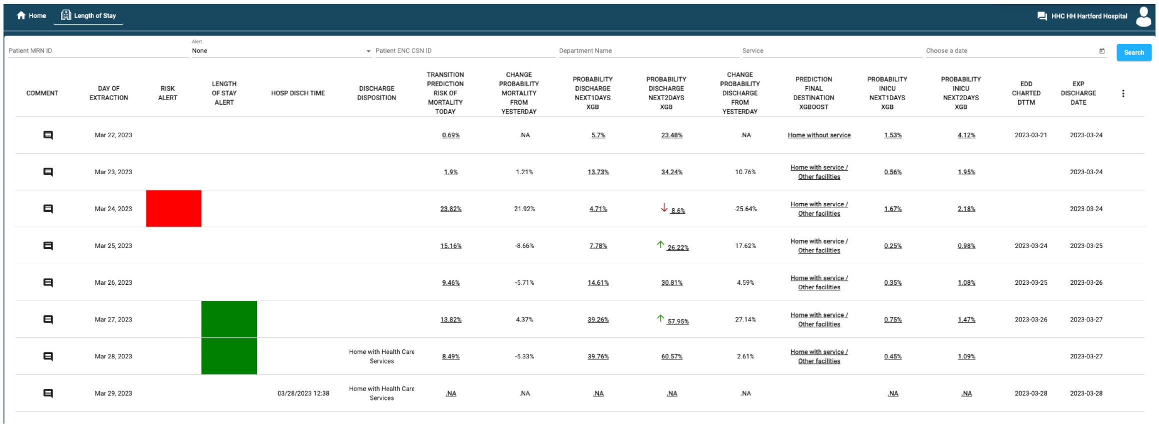
There is a recent, rich, and growing academic literature on machine learning models trained to predict inpatient outcomes. However, only a limited number of these models are deployed in practice and make an impact. As introduced in Section 1.2.1 and Section 1.2.3, we collaborate with Hartford HealthCare and Dynamic Ideas LLC and succeed in developing and implementing models in practice that improves hospital operations. Since 2020, our joint team of healthcare providers, academics, and data consultants, have been working together to develop predictive models on eight operational metrics for seven different hospitals (covering over 2,400 inpatient beds), to deploy these models in the entire hospital network, and to measure their impact on the operational performance of the hospitals.

3.1.1 Summary of Contributions

First, we build an end-to-end ML pipeline, from data extraction and processing to a user-friendly software interface, and apply it to the seven HHC hospitals. Using comprehensive data from patient EMRs, including patient status, clinical measurements, and laboratory results, we train ML models (XGBoost) to daily predict eight inpatient outcomes: mortality risk, probability of discharge in the next 24 and 48 hours, discharge disposition, and intensive care unit (ICU) risk in the next 24 and 48 hours. Our models achieve state-of-the-art accuracy (75.7%–92.5% area under the receiver operating curve, i.e., AUC across all tasks and all hospitals) and are well calibrated. Moreover, we demonstrate that integrating our discharge predictions into physicians' decision making process can identify more discharge opportunities with higher accuracy and lower readmission risk.

Second, through multiple iterations with clinicians, we develop a software tool for doctors, nurses, and case managers, to integrate these predictions into their daily workflow. In addition to presenting raw risk scores, the tool provides patient-level explanations for each of the predictions as well as a color-coded alert system to help them quickly identify at-risk patients (red alert) and potential imminent discharges (green alert). The features of our software help increase trust, clinical adoption, and easy integration into the patient progression rounds.

Finally, we have been gradually deploying our solution in the hospitals, training physicians and nurses, and measuring its operational impact. As of January 2023, our tool is benefiting over 200 users and supporting daily operational decisions at the seven hospitals. We observe a significant reduction in the average length of stay (LOS) by 0.67 days per patient from our solution and project annual revenue uplift of $35.89 million dollars from our deployment in the HHC system.



Patient MRN ID

Alert: None

Patient ENC CSN ID

Department Name

Service

Choose a date

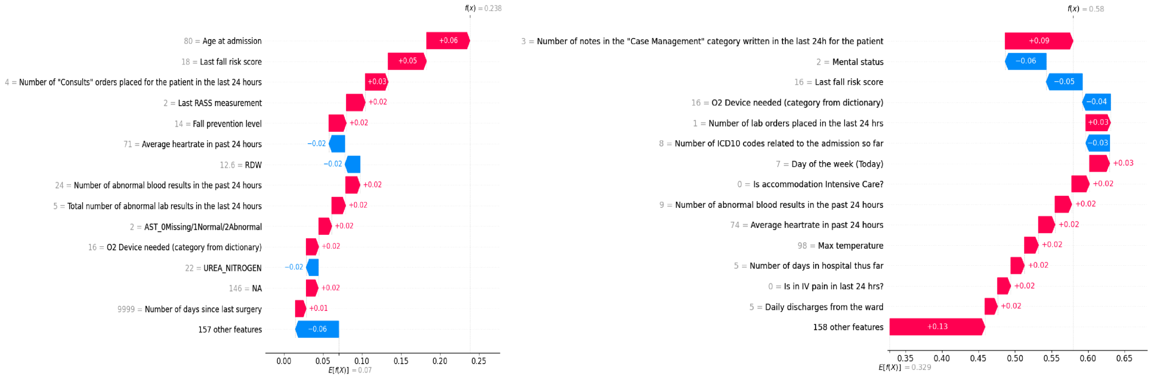
COMMENT	DAY OF EXTRACTION	RISK ALERT	LENGTH OF STAY ALERT	HOSP DISCH TIME	DISCHARGE DISPOSITION	TRANSITION PREDICTION RISK OF MORTALITY TODAY	CHANGE PROBABILITY MORTALITY FROM YESTERDAY	PROBABILITY DISCHARGE NEXT1DAYS XGB	PROBABILITY DISCHARGE NEXT2DAYS XGB	CHANGE PROBABILITY DISCHARGE FROM YESTERDAY	PREDICTION FINAL DESTINATION XGBOOST	PROBABILITY INICU NEXT1DAYS XGB	PROBABILITY INICU NEXT2DAYS XGB	EDD CHARTED DTTM	EXP DISCHARGE DATE
💬	Mar 22, 2023					0.69%	.NA	5.7%	23.48%	.NA	Home without service	1.53%	4.12%	2023-03-21	2023-03-24
💬	Mar 23, 2023					1.9%	1.21%	13.73%	34.24%	10.76%	Home with service / Other facilities	0.56%	1.95%		2023-03-24
💬	Mar 24, 2023	🟥				23.82%	21.92%	4.71%	↓ 8.6%	-25.64%	Home with service / Other facilities	1.67%	2.18%		2023-03-24
💬	Mar 25, 2023					15.16%	-8.66%	7.78%	↑ 26.22%	17.62%	Home with service / Other facilities	0.25%	0.98%	2023-03-24	2023-03-25
💬	Mar 26, 2023					9.46%	-5.71%	14.61%	30.81%	4.59%	Home with service / Other facilities	0.35%	1.08%	2023-03-25	2023-03-26
💬	Mar 27, 2023		🟩			13.82%	4.37%	39.26%	↑ 57.95%	27.14%	Home with service / Other facilities	0.75%	1.47%	2023-03-26	2023-03-27
💬	Mar 28, 2023		🟩		Home with Health Care Services	8.49%	-5.33%	39.76%	60.57%	2.61%	Home with service / Other facilities	0.45%	1.09%		2023-03-27
💬	Mar 29, 2023			03/28/2023 12:38	Home with Health Care Services	.NA	.NA	.NA	.NA	.NA		.NA	.NA	2023-03-28	2023-03-28

Figure 3-1: Trajectory of Predictions for an Example Patient.

3.1.2 Our Tool in Action

We now illustrate how our tool works on one patient trajectory. Figure 3-1 is a screenshot of our software solution for a particular patient. Each row corresponds to a day. The patient is admitted on March 22, 2023. On the first day of the stay, the patient is assigned low probabilities of mortality and discharge. Two days after admission, the condition of the patient deteriorates: discharge probabilities decrease and mortality risk increases, peaking at 23.82% on March 24. The system detects an exacerbation of the patient's condition and raises a red alert, which calls the attention of the caring team. Over the following days, the mortality risks decrease towards patient recovery, and the probabilities of discharge gradually increase until discharge. The system correctly delivers green alerts on the last two days of the stay. In particular, the probability of discharge in the next 48 hours (resp. 24 hours) started exceeding 0.4 (resp. 0.25) on the day before (resp. the day of) discharge. On March 28, the patient is discharged to Home with Health Care Services, as correctly predicted by the discharge disposition prediction model (from the final destination column).

By clicking on each predicted value, the clinician can also access a waterfall plot to understand the patient-specific factors that explain such a prediction. Figure 3-2 displays these plots for the mortality prediction on March 24 (Figure 3-2a) and the 48-hour discharge prediction on March 27 (Figure 3-2b) for our patient. For a given



(a) Mortality Risk on March 24.

(b) 48-hr Discharge Probability on March 27.

Figure 3-2: Prediction Explanation Plots for the Patient Example.

risk score f, starting from the baseline risk in the population, $\mathbb{E}[f(X)]$ (here, the variable X denotes all patient-level information used to make predictions), at the bottom of the plot, we add up the contributions of each variable to finally reach the patient-specific estimate, $f(x)$. Figure 3-2a, for example, explains why the patient's mortality risk score on March 24 is 0.238, while the average prediction among all patients is 0.07. The main factors explaining a higher-than-average prediction are the fact that the patient has a high age (+0.06), a high fall risk score assessment (+0.05), multiple consultation orders placed in the last 24 hours, and agitated status indicated by a Richmond Agitation Sedation Scale (RASS) measurement equal to 2 (+0.02), among others. Meanwhile, these aspects are partially counterbalanced by several other variables that decrease the predicted mortality risk; for instance, the average heart rate in the past 24 hours and red cell distribution width (RDW).

3.2 Problem Statement and Related Literature

We focus our attention on the following patient outcomes: length of stay, mortality risk, discharge disposition, and ICU risk.

3.2.1 Length of Stay

Models in the literature predict a variety of LOS related outcomes, such as daily discharge volume (Zhu et al. 2015), next 24-hour discharge (Safavi et al. 2019, Bertsimas et al. 2021b), long LOS (Bardak and Tan 2021, Bertsimas et al. 2021b), and remaining LOS (Wang et al. 2022). Since our priority is prompt discharge identification, the medical team at HHC suggests predicting whether each patient will be discharged without expiration (i.e., without death) in the next 24 and 48 hours. We note that patient death or transition to hospice are not considered as discharges. Predicting the next 48h discharges is critical and particularly useful for HHC since it helps physicians identify and prioritize patients who are ready for discharge while giving case management teams enough time to accelerate discharge preparations, which ultimately reduces patient burdens and direct operating costs in healthcare systems.

3.2.2 Mortality Risk

Preventing death is one of the major responsibilities of hospitals. Physicians examine and evaluate patient daily charts with clinical knowledge and experience, but they encounter challenges in simultaneously processing hundreds of measurements and giving quick dynamic assessments for all patients. We build models to help predict each patient's mortality risk, defined as the probability of each patient expiring or going to hospice at the end of the hospital stay. The high accuracies of mortality risk prediction models are demonstrated in previous works such as Awad et al. (2017b), Rajkomar et al. (2018), Jin et al. (2018), Bardak and Tan (2021). With such aid, the medical staff can give prioritization and rapid treatment plans to high-risk patients, which could potentially prevent their death. Moreover, detection of increasing mortality risk over time alerts the care team of patients with worsening conditions, which can lead to more timely intervention and improve patient outcomes.

3.2.3 Discharge Disposition

It is also important to anticipate the patient's final destination after discharge. We divide the dispositions into three categories: home without service (i.e., discharged back home with self-care), expired or hospice (died at the hospital or transferred into hospice, where the latter is considered as near death), and home with service or other facilities (such as skilled nursing facility, rehab facility, long term acute care hospital). Differentiating discharge destinations early helps anticipate the need for post-discharge resources. In particular, the third discharge disposition requires additional case management efforts, such as contacting and obtaining approval from the care facility or ordering devices for service at home. Combined with discharge time prediction, it can help the case management team coordinate post-discharge services and prevent logistical delays.

3.2.4 ICU Risk

Uncertain intensive care demand and limited bed space make it particularly challenging to manage flows out of and into the ICUs, one of the most scarce and expensive resources in a hospital. Patients who need to enter an ICU and cannot do so promptly are often placed in secondary units, leading to higher readmission risk and extended LOS (Kim et al. 2016). On the other side, some patients are ready to leave the ICU but experience delays due to congestion in step-down units or delayed ICU discharge identification, which in turn congests the ICU, overflows other parts of the hospital (Long and Mathews 2018), and prolongs boarding from the emergency department (Mathews et al. 2018). As exacerbated during the COVID-19 pandemic, predictions of ICU admission and mortality can help manage these valuable units (Zhao et al. 2020, Covino et al. 2020, Subudhi et al. 2021). We build models to predict the probability of being in the ICU in the next 24 and 48 hours, respectively. Specifically, we predict the probability of entering (resp. leaving) the ICU for patients currently not in (resp. currently in) the ICU.

Our objective, however, is not to predict for the sake of prediction but, more importantly, to implement our innovations in hospitals for real-world impact. Under such context, we aim to develop practical ML systems and deploy them in the entire hospital network. In this regard, our approach is similar to that of Bertsimas et al. (2021b) who study a similar set of patient outcomes and integrate their predictions into hospital daily operations. In addition to using different ML models, their analysis, however, is single-center while we train, deploy, and evaluate the benefits of our models in different locations. Furthermore, the operational decisions and end-users are different. While they create a dashboard with unit-level predictions for the hospital flow management center, we work with doctors, nurses, and medical staff to leverage these patient-level predictions for the management of each patient individually.

3.3 Methods

In this section, we describe our methodological approach, from data collection and feature engineering to predictive modeling and its integration as a decision making support tool.

3.3.1 Data Collection

We first collect and build relevant data extracts for our various prediction tasks. Since HHC uses a third-party EMR solution and since this is their first ML project applied to all inpatients in the network, there was no pre-existing pipeline for us to use, and thus we build the data extracts and pipeline from scratch.

We start by replicating the pipeline of Bertsimas et al. (2021b) and using the same set of variables as those they identify as important, followed by weekly discussions with doctors, nurses, and the IT team at HHC to include other useful data sources and variables. Based on those discussions as well as data availability, we build 10 data extracts summarized in Table 3.1. The extracts provide information including demographics (e.g., age at admission), patient status (e.g., current service, oxygen device), clinical measurements (e.g., oxygen concentration, blood pressure), laboratory

Table 3.1: Summary of Data Extracts.

Extract	Extract description	Granularity	Example columns
1	Admission, discharge, transfer events	Event	Hospital, department destination
2	Admission, discharge, transfer orders	Order	Service, order type
3	Lab results with normal ranges	Patient day	Platelet count with normal range
4	Clinical measurements	Patient day	Blood pressure, respiratory rate
5	Preparation for discharge	Patient day	Discharge time, future surgery date
6	DNR status	Patient day	DNR
7	Time invariant patient information	Patient	Age, discharge disposition
8	Summary statistics of notes	Patient day	Diagnosis, number of notes written
9	Surgery	Surgery case	Procedure name, start time, end time
10	Summary statistics of orders	Patient day	Number of orders, pending labs

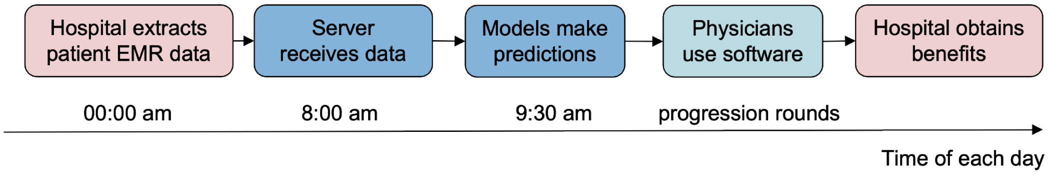


Figure 3-3: End-to-end Pipeline in Daily Production.

results (e.g., albumin, bilirubin), diagnoses, orders, procedures, notes, and others.

All the data is extracted from the EMRs directly, on HHC's IT system, password-protected, and then transferred to a dedicated database hosted on Amazon Web Service (AWS) machines by a Secure File Transfer Protocol server. Daily extracts are scheduled every day at midnight so that up-to-date data about current inpatients are received by 7:40 am every morning (see Figure 3-3). EMRs were gradually deployed in the network so data availability for training purposes varies by hospital. Our database starts on January 2016 for Hartford Hospital, January 2021 for St. Vincent Medical Center, and January 2020 for the remaining five hospitals.

3.3.2 Feature Curation

Since our pipeline and predictions are updated daily, we curate a feature space where each row represents each patient day. We create features from the following six groups.

1) Current conditions (e.g., department, whether in ICU);

2) Lab results (e.g., albumin, white blood cell count);

3) Clinical measurements (e.g., temperature, respiratory rate, heart rate);

4) Time series summary statistics of operational variables (e.g., days in ICU);

5) Patient information prior to current admission (e.g., age, previous admission);

6) Auxiliary operational variables which are not patient-specific (e.g., day of the week).

Creating this collection of features requires a large amount of data processing, such as imputing missing data, parsing string formats, and encoding categorical variables. A more comprehensive list of variables and details on data processing can be found in Section B.1 of the Appendix.

3.3.3 Machine Learning Modeling

Inclusion-Exclusion Criteria: We consider all inpatients during their hospitalization days in HHC. Emergency Department patients and outpatients who are not admitted as an inpatient are excluded. For each inpatient, we consider the range starting from their admission date until their discharge date. We further filter the data depending on the prediction target. We identify a set of special discharge dispositions: left against medical advice/AMA, still a patient, admitted as an inpatient, court/law Enforcement, ED dismiss/diverted Elsewhere. For mortality and discharge disposition predictions, we exclude patients whose discharge disposition is missing because the target is missing, and exclude patients who have a special discharge disposition to reduce noise in the target. For 24/48-hr discharge predictions, we exclude patients whose discharge disposition or discharge time is missing and exclude data points where the patient is discharged in the next 24/48 hours to one of the special dispositions. For 24/48-hr ICU predictions, we include only patients who still are in the hospital in the next 24/48 hours. Moreover, we include only patients who are currently in the ICU for leaving ICU predictions and include only patients currently not in the ICU for entering ICU predictions.

Split of Training, Validation, and Testing Sets: For each hospital, we sort the data by record date and split it into 50% training, 20% validation, and 30% testing sets. The train/validation/test set split is performed separately for each hospital and chronologically to reflect real-world implementation, where the models are trained on past data and utilized for future prediction. Both imputation and machine learning models are trained on the training set, tuned on the validation set, and evaluated on the testing set.

Prediction Models: We consider a variety of machine learning models to make predictions. Since patient predictions should be interpretable for medical staff, we started with interpretable ML models, including Optimal Classification Trees (OCT, Bertsimas and Dunn 2017) and sparse classification (Bertsimas et al. 2021c), implemented in the Interpretable AI software (Interpretable AI, LLC 2022). OCT is a state-of-the-art interpretable decision tree model that divides the feature space using simple and understandable rules. The resulting OCT trees were closely examined and discussed by the doctors, which uncovered important insights as well as established doctors' initial understanding and trust in the algorithm. As doctors got more receptive to using and understanding the models, we then shifted the objective to improve the performance of predictions. We consider other machine learning models including XGBoost (Chen and Guestrin 2016), LightGBM (Ke et al. 2017), and Tabnet (Arik and Pfister 2021). XGBoost is a gradient boosting method that ensembles a set of decision trees to make predictions in an additive fashion. It consistently outperformed the other five methods across all prediction tasks on preliminary experiments so we decided to use it for our final models. To obtain higher accuracy, we also considered ensembling up to 10 XGBoost models together or creating more sophisticated features from time series measurement using the `tsfresh` package (Christ et al. 2018), but decided that the marginal gain in accuracy did not justify the additional computational burden and lost of interpretability. We train a separate model for each prediction task (a multi-class classification model for discharge disposition and binary classification models for the other target variables) and for each hospital. We use the validation set to calibrate the hyper-parameters (depth of trees, learning rate, number of estimators, loss function, and L2 regularization rate).

Model Calibration: For interpretability purposes, it is important for classification models to be well calibrated, i.e., that the numerical scores (scaled between 0 and 1) returned by the models correspond to the probability of the event of interest to occur. For example, if the predicted probability of discharging a patient in the next 48 hours is 0.2, a doctor will likely expect interpret that this patient has a 20% chance of being discharged. Note that calibration and accuracy are different issues (for example, dividing all scores by a factor 2 impacts calibration but keeps AUC constant). Therefore, we ensure that all our models are well calibrated by chronologically splitting the testing set into two halves, using the isotonic regression method (Zadrozny and Elkan 2002) to calibrate the model on the first half, and then assessing the final calibration on the second half.

Computing Resources: Data processing, feature engineering, model training, and offline testing are conducted in Python 3.5.2 with a parallelization strategy on MIT Supercloud (Reuther et al. 2018) with 32GB RAM Intel Xeon-P8 CPU per instance. The online daily pipeline in production is run in Julia 1.5 and Python 3.8 on a cloud computing server via AWS with 64GB RAM and 8 CPUs.

3.3.4 Predictive Analytics for Decision Making

To turn these predictive analytics into a decision support tool that is sustainably used by practicionners, we complement the raw probability predictions with an alert system and visual explanations for each prediction.

Color-coded Alert System: It can be difficult for clinicians to quickly grasp the implications of a raw probability score and use it efficiently for decision making. For instance, it is not obvious how to interpret a 0.28 risk of mortality or a 0.57 probability of discharge. Moreover, medical staff are responsible for many patients, and they cannot read and process hundreds of probabilities including eight predictions for all their patients every day as part of their filled daily schedule. To overcome these challenges, we design an alert system to highlight the set of patients who are getting likely to be discharged and patients who are exacerbating with different colors. We send a green alert for patients that are ready for short term discharge, i.e., whenever their probability of 24-hr or 48 hr discharge is above certain thresholds. On the other hand, we send a red alert to warn for patients who have a high risk or are exacerbating, i.e., if the mortality risk or the increase of mortality risk from the previous day reaches certain thresholds.

Prediction Explanation Plot: As doctors require sufficient clinical reasoning to make major decisions like patient discharge, it is critical to provide them with some interpretation of model predictions. We compute Shapley values and SHapley Additive exPlanations (SHAP Štrumbelj and Kononenko 2014) on the XGBoost models to derive each feature's attribution to model predictions. We use SHAP summary plots to visualize the top features of each model and their overall effect of the predictions, which is useful to audit the models and assess their validity with clinicians. We also produce SHAP plots at the individual level on every patient's prediction on each day, to provide a visual explanation of each predicted probability.

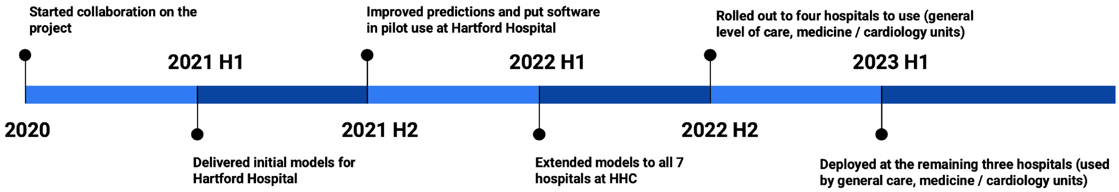

Started collaboration on the project
Improved predictions and put software in pilot use at Hartford Hospital
Rolled out to four hospitals to use (general level of care, medicine / cardiology units)
2021 H1
2022 H1
2023 H1
2020
2021 H2
2022 H2
Delivered initial models for Hartford Hospital
Extended models to all 7 hospitals at HHC
Deployed at the remaining three hospitals (used by general care, medicine / cardiology units)

Figure 3-4: The Collaboration Timeline.

3.4 Project Management Approach

A major factor in the success of the large-scale deployment of an advanced analytics solution like this one lies in the way we organized the progression of the project over time and scheduled its extension from one to several hospitals within HHC.

3.4.1 Project Timeline

We have been collaborating on this solution since 2020. Our first challenge was to interface with the third-party software HHC uses for managing all of their EMRs. In 2020, we worked on an automated pipeline to daily extract data from the EMR system to dedicated database. We then focused our attention to HH, the largest of HHC hospitals, to develop a first proof of concept for our predictive models (including only a limited number of operational outcomes) and our doctor-facing interface. With this prototype, we constituted in 2021 a group of physician champions to serve as beta testers, collected feedback from their utilization and understanding of the models (regarding model accuracy, missing predictive information, additional relevant outcomes to predict, and the user interface), and iteratively improved the models and the software tool. The tool was then rolled out for usage by the entire HH in 2022. Concurrently, we re-trained the models initially developed for HH to the other six hospitals at HHC and deployed them in production progressively between May 2022 and January 2023. The timeline of the project is sketched in Figure 3-4.

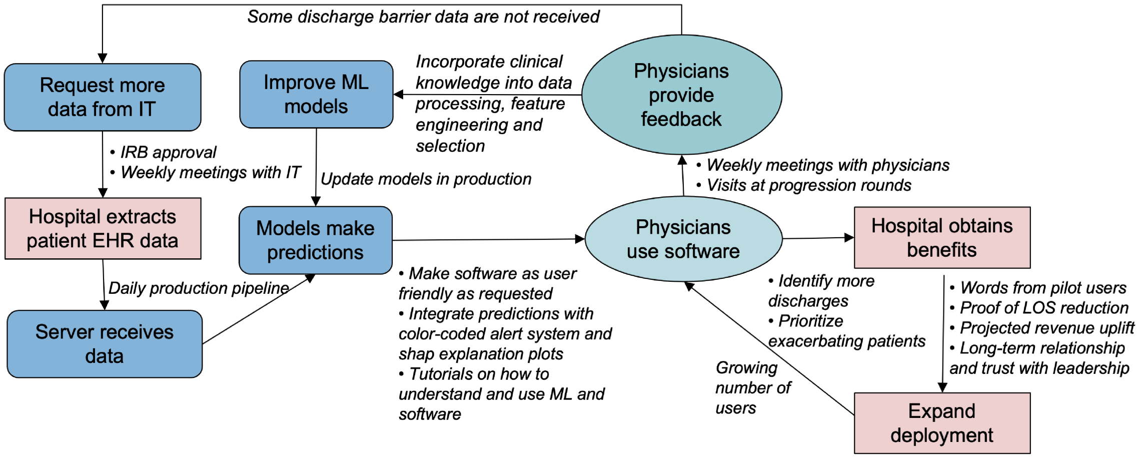

Some discharge barrier data are not received
Request more data from IT
Improve ML models
Incorporate clinical knowledge into data processing, feature engineering and selection
Physicians provide feedback
• IRB approval
• Weekly meetings with IT
Update models in production
• Weekly meetings with physicians
• Visits at progression rounds
Hospital extracts patient EHR data
Models make predictions
Physicians use software
Hospital obtains benefits
Daily production pipeline
• Make software as user friendly as requested
• Integrate predictions with color-coded alert system and shap explanation plots
• Tutorials on how to understand and use ML and software
• Identify more discharges
• Prioritize exacerbating patients
• Words from pilot users
• Proof of LOS reduction
• Projected revenue uplift
• Long-term relationship and trust with leadership
Server receives data
Growing number of users
Expand deployment

Figure 3-5: Implementation Feedback Loops.

3.4.2 Pilot Implementation at Hartford Hospital

Instead of developing our models for all the HHC hospitals simultaneously, we adopted a gradual approach and started with HH, the main hospital of the network. Since the second half of 2021, HHC tested the tool with four physician champions who are lead hospitalists of five medical units at HH, including two teaching units (BLISS 7 EAST and CONKLIN 4) and three non-teaching units (CENTER/NORTH 12 and CONKLIN 5). Every week, the clinical and analytics teams met to review technical issues on the deployment of the model in production as well as to incorporate feedback from the physician leaders.

Figure 3-5 represents the different feedback loops between the technical team (blue rounded rectangle), the hospitals (red rectangle), and specific physicians from HHC (green oval). Feedback from the phyisicians has been crucial to identify other sources of data to integrate into our model, as well as define the most relevant uses cases for the prediction and improve the functionalities of our software tool. On a weekly basis, physicians reviewed patients and their predictions and provided written comments for each prediction. For example, when patients had high predicted discharge probability but were not ready for discharge yet, physicians would note their current discharge barrier and the MIT team would propose solutions to account for it. Through this process, the MIT team incorporated physicians' clinical knowledge into the feature processing and engineering steps, e.g., by creating delta variables and time series summary statistics for some important features or by adjusting the missing value imputation based on medical knowledge. Clinical expertise was also needed to decide the depth and breadth of the data extracts. For example, we initially received only one measurement of the RASS score per day (the latest) but later decided to extract all daily measurements to capture deterioration/improvement in the patient's anxiety level.

Having a software interface to share the outputs of our models in a convenient, interpretable, and streamlined way with physicians significantly accelerated the adoption of the tool and its continuous improvement. The MIT team further facilitated the integration of the tool by providing tutorials and lectures about machine learning to doctors at HHC. We cannot emphasize enough how instrumental this pilot implementation with physician champions at HH has been for establishing a close relationship and trust between the analytics team (MIT-Dynamic Ideas) and the doctors and nurses at HHC, and for enabling the wider roll-out of the solution.

3.4.3 Scaling to Multiple Hospitals

Only once built the data extraction pipelines, the ML models, and ran through a couple of iterations with teams at HH, did we start replicating this development (including the data extraction part) to other hospitals. After an evaluation of the impact of our solution on the pilot conducted at HH (presented in 3.6), HHC decided to extend our work to three other hospitals—Hospital of Central Connecticut (HOCC), Backus Hospital (BH), and Charlotte Hospital (CH)—chosen for the diverse set of activities they cover and for their appetite for experimentation. Finally, we included the remaining three hospitals—Midstate Medical Center (MMC), St. Vincent's Medical Center (SV), and Windham Hospital (WH).

By extending our data extraction and processing to different institutions, we designed a fairly robust and generalizable feature processing pipeline (described in Section B.1 of the Appendix). After feature curation and processing, we chose to train separate ML models for each hospital for different reasons. First, as described in

Section 1.2.1, the seven HHC hospital are very diverse in size, service, levels of care, and patient populations. Second, developing and deploying models for a large hospital network requires a staggered roll-out. As opposed to developing a model for all hospitals at once, we developed and implemented them gradually to earn trust and support from the leadership. Third, since each hospital has its own encoder and imputer (see B.1 of the Appendix), we cannot apply one common model for all hospitals. In an early stage, we tried applying the model trained for HH directly to three other hospitals and, as expected, we obtained low performance due to different ways of encoding (e.g., department and service).

3.5 Results

We present statistics, evaluation, and analysis of the models. Unless specified otherwise, all results in this section are evaluated in the testing sets for each hospital.

3.5.1 ML Model Evaluation

Accuracy: The performance of the binary classification models is assessed by the AUC. For discharge disposition, multiclass AUCs are computed on each class against the rest. Table 3.2 presents out-of-sample AUCs of eight prediction tasks for seven hospitals. For hospitals BH, CH, HH, HOCC, MMC, and SV, AUCs range from 0.905–0.925 for mortality, 0.858–0.884 for discharge disposition, 0.812–0.848 for 24-hr discharge, 0.816–0.852 for 48-hr discharge, 0.850–0.872 for 24-hr entering ICU, 0.812–0.896 for 24-hr leaving ICU, 0.811–0.847 for 48-hr entering ICU, and 0.833–0.896 for 48-hr leaving ICU. Our models achieve state-of-the-art performances compared with the literature described in Section 3.2. Compared with the other six hospitals, WH has lower AUCs in mortality and discharge-related predictions, which are likely due to the smaller data size (see Table B.1 and Table B.2) and the higher complexity of differentiating among less critical patients who have a much higher proportion of being discharged in the next 48 hours than other hospitals (see Table B.4). A full summary statistics of the data at each hospital for each prediction task is reported in Section

Table 3.2: AUC Metrics for All Predictions in All Seven Hospitals.

Hospital Prediction	BH	CH	HH	HOCC	MMC	SV	WH
Mortality	0.915	0.902	0.919	0.905	0.925	0.925	0.888
Discharge Disposition	0.858	0.871	0.884	0.884	0.879	0.869	0.802
Discharge 24 hr	0.832	0.812	0.857	0.844	0.837	0.848	0.757
Discharge 48 hr	0.830	0.816	0.852	0.843	0.836	0.841	0.768
Enter ICU 24 hr	0.867	0.853	0.868	0.868	0.872	0.850	No ICU
Leave ICU 24 hr	0.896	0.830	0.871	0.883	0.887	0.812	
Enter ICU 48 hr	0.834	0.813	0.818	0.811	0.847	0.820	
Leave ICU 48 hr	0.896	0.848	0.865	0.880	0.876	0.833	

B.2.1 in the Appendix. We also report a breakdown of the AUCs by department at HH as an example in Table B.5 in Section B.2.3 Appendix.

Model Calibration: As discussed in Section 3.3, it is important for classification models to be well calibrated. Accordingly, we calibrate our models on the first half of the testing set and assess the calibration on the second half using calibration curves. Detailed assessment of the calibration of our models are presented in Section B.2.2 of the Appendix.

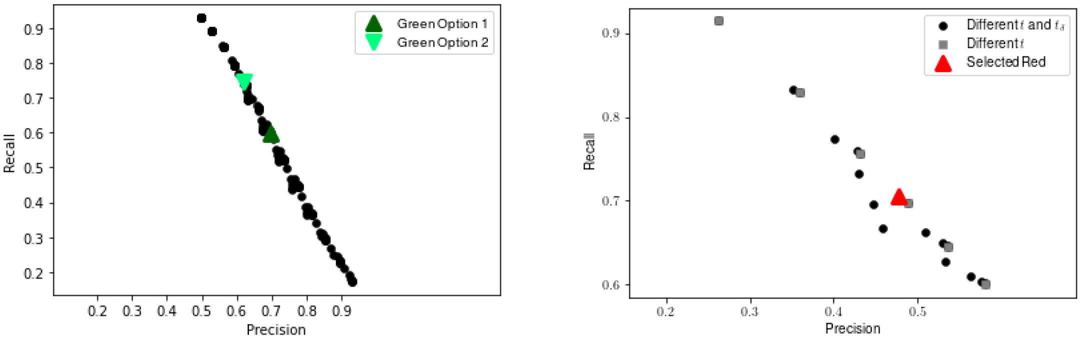
3.5.2 Alert System Assessment

We select the thresholds for the color-coded alert system depending on the resulting precision/recall trade-off they provide for identifying discharges (green alert) and high-risk patients (red alert), as illustrated in Figure 3-6 for HH.

For discharge prediction, our objective is to correctly mark (with a green alert) the patients that will be discharged in the next 48 hours. Precision represents the proportion of actual discharges among patients that are marked (also referred to as positive predicted value) and recall is the proportion of patients marked among all actual discharges (true positive rate). We raise an alert whenever the predicted probabilities of being discharged in the next 24 or in 48 hours exceed a threshold, t_{24} and t_{48} respectively. Figure 3-6a represents the precision and recall for different threshold values, $t_{24}, t_{48} \in [0, 1]$. We observe that the precision-recall relationship is almost linear. HHC wants to achieve a precision of around 0.7 for 48-hour discharge prediction, so we define the option one of the green alert as $t_{24} = t_{48} = 0.5$ (dark green upward triangle in the plot), yielding 0.698 precision and 0.598 recall. After three months of utilization, the medical team expressed the need to identify more potential discharges, at the expense of lower precision. Accordingly, we lowered the threshold for a green alert, defined as $t_{24} = 0.25$, $t_{48} = 0.4$ (light green downward triangle in the plot), which gives 0.621 precision and 0.746 recall.

For mortality prediction, our objective is to mark with a red alert on two types of patients: those who will likely expire (i.e., death or hospice), and those who have worsening conditions. Accordingly, we raise an alert whenever the predicted mortality probability exceeds a threshold t, or when its absolute change compared with the previous day exceeds t_δ. Figure 3-6b represents the precision/recall trade-off for $t \in [0.05, 0.3], t_\delta = 1$ (grey squares) and for $t \in [0.05, 0.3], t_\delta \in [0.05, 0.3]$ (black circles). We observe that incorporating a criterion based on variations in mortality scores ($t_\delta < 1$) provides more granularity in terms of precision/recall trade-off, although it can lead to strictly worst performance if not calibrated carefully. Also, it spotlights patients whose condition is deteriorating, which the medical team can find even more helpful than predicting an absolute probability of mortality. For example, doctors find the tool more helpful in identifying a young patient's mortality risk increased from low probabilities on previous days, which would call for the provider team's action of interference, compared with reporting an 80-year-old patient's consistently high mortality risk, which typically would have been already expected from their clinical assessment. Given that patient lives are at stake, doctors emphasize having a high recall and are less concerned about low precision (i.e., high false alarm rate). Therefore, we define red alerts as $t = 0.2$ and $t_\delta = 0.1$, which gives 0.477 precision and 0.705 recall.

We report the accuracy, precision, and recall for our previous green ($t_{24} = t_{48} = 0.5$), new green ($t_{24} = 0.25, t_{48} = 0.4$), and red ($t = 0.2, t_\delta = 0.1$) alerts at all seven hospitals in Table 3.3. Among the hospitals, previous green alerts have 0.696-0.782 accuracy, 0.645-0.698 precision, and 0.546-0.684 recall; new green alerts have



(a) Green Alerts for Discharge.

(b) Red Alert for Mortality.

Figure 3-6: Out-of-Sample Precision-Recall Curves under Different Thresholds for Colored Alert System at HH.

Table 3.3: Precision and Recall under Selected Thresholds for Alerts.

Alert	Hospital	BH	CH	HH	HOCC	MMC	SV	WH
Previous Green	Accuracy	0.757	0.739	0.782	0.771	0.759	0.78	0.696
	Precision	0.672	0.645	0.698	0.682	0.684	0.692	0.68
	Recall	0.653	0.679	0.598	0.628	0.684	0.546	0.646
New Green	Accuracy	0.734	0.714	0.767	0.751	0.739	0.768	0.687
	Precision	0.604	0.588	0.621	0.611	0.623	0.617	0.629
	Recall	0.786	0.8	0.746	0.764	0.796	0.701	0.78
Red	Accuracy	0.901	0.895	0.899	0.881	0.896	0.886	0.925
	Precision	0.55	0.574	0.477	0.492	0.528	0.505	0.53
	Recall	0.668	0.553	0.705	0.691	0.715	0.663	0.471

0.687-0.768 accuracy, 0.588-0.629 precision, and 0.701-0.8 recall; red alerts have 0.885-0.925 accuracy, 0.477-0.55 precision, and 0.471-0.715 recall. We further report a department-level alert evaluation at all departments of HH in Section B.2.3 of the Appendix.

3.5.3 Assistance to Medical Staff

To evaluate how our models can help the decision making for doctors, nurses, and case management teams, we compare their accuracy with that of physicians. Every day at progression rounds, a team of attending physicians, residents, and nurses reviews each patient's chart information, discusses their case, and enters (or updates) an Expected

Table 3.4: Comparison of AUC Metrics between Doctors' and Models' Discharge Predictions.

	48-hr Discharge AUC			24-hr Discharge AUC		
Hospital	Doctor	Model	Increment	Doctor	Model	Increment
BH	0.689	0.803	0.114	0.732	0.811	0.079
CH	0.647	0.786	0.139	0.696	0.789	0.093
HH	0.644	0.834	0.190	0.678	0.843	0.165
HOCC	0.603	0.817	0.214	0.642	0.824	0.182
MMC	0.582	0.809	0.227	0.606	0.816	0.210
SV	0.524	0.82	0.296	0.548	0.831	0.283
WH	0.683	0.746	0.063	0.718	0.742	0.024

Discharge Date (EDD) for each patient. To compare with our models' outputs, we consider EDD as a score (by converting it into a time to discharge) to rank patients based how soon they are likely to be discharged and we evaluate its AUC. In line with our outcome definition for the discharge prediction models, we compare the doctors' and models' discharge predictions on patients whose discharge dispositions are neither hospice nor expiration, and report their respective AUC in Table 3.4. We observe that our models achieve higher AUCs than the doctors for all hospitals, and that the improvement is higher for 48-hr than for 24-hr predictions. The improvement ranges between 0.063–0.296 for 48-hr discharge and ranges between 0.024–0.283 for 24-hr discharge.

We also investigate how doctors perform (in terms of precision and recall) at identifying 48-hour discharges compared with our green alert system. We report the precision and recall of doctors for each of the seven hospitals in Table 3.5. In comparison with the green alert (see, e.g., Table 3.3), doctors generally demonstrate lower precision but higher recall, thus predicting more discharges than the green alerts. Greater benefits can be obtained by combining the two predictions together. In particular, if we predict a discharge when both the doctor's EDD is less than 48 hours away and a green alert is raised ('Doctor `AND` Green' column in Table 3.5), we increase precision by 0.121–0.294 compared with doctors'. Furthermore, we can achieve higher recall by predicting a discharge whenever the doctor predicts discharge or a green alert is raised ('Doctor `OR` Green' column). By doing so, we complement

Table 3.5: Precision and Recall Improvement (Delta) of Doctors' 48-h Discharge Predictions by Referencing Green Alert.

Metric	Precision			Recall		
Hospital	Doctor	Doctor AND Green	Delta	Doctor	Doctor OR Green	Delta
BH	0.522	0.713	0.191	0.775	0.917	0.143
CH	0.510	0.671	0.161	0.792	0.926	0.134
HH	0.500	0.720	0.220	0.681	0.878	0.197
HOCC	0.478	0.691	0.213	0.656	0.882	0.226
MMC	0.510	0.705	0.195	0.605	0.892	0.287
SV	0.411	0.705	0.294	0.547	0.824	0.277
WH	0.596	0.718	0.121	0.806	0.906	0.100

doctors' input to identify potential discharge cases that doctors would otherwise have missed and increase the recall of doctors by 0.1–0.287 across seven hospitals. These results suggest that by referencing the green alert system, doctors, nurses, and the case management team can identify more discharges and do so more precisely, which can further reduce patients' LOS.

3.5.4 Representation of Readmission Risk

In addition to accurate, timely discharges, it is also important to ensure the safety of the discharge. According to Medicare, around 20% of hospital-discharged patients are readmitted within 30 days, resulting in billions of dollars in annual costs (Jencks et al. 2009). Early readmissions, which occur within 7 days post-discharge, are particularly closely linked to premature discharges, where a tradeoff exists between LOS and readmission (Koekkoek et al. 2011). As predicting post-discharge outcomes can be challenging due to the many complex factors involved, models predict 30-day and 7-day readmission with moderate discriminative ability (Zhou et al. 2016, Saleh et al. 2020). Although our models for patient outcomes up to and including discharge do not consider readmission or other post-discharge outcomes, we are interested in exploring their potential associations. Specifically, we examine the relationship between our discharge prediction tool and the risk of patient readmission within 7 days and 30 days, respectively. We evaluate our findings using data from January to April 2022, which represents an out-of-sample period in the testing sets for all 7 hospitals prior

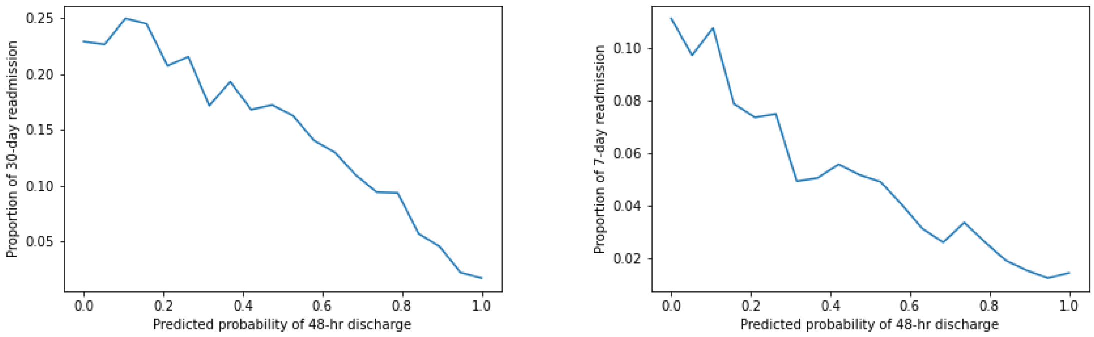


Figure 3-7: Readmission Risk vs 48-hr Discharge Probability.

to the models being put into production.

We begin by discretizing the probability of discharge within the next 48 hours into fixed buckets with 5% probability intervals. For each bucket, we compute the proportions of 30-day readmission and 7-day readmission among patients within that probability bucket and who get discharged within the next 48 hours. As shown in Figure 3-7, we observe a negative, almost linear, relationship between the two. Specifically, higher predicted discharge probabilities correspond to lower readmission risks for both 30-day and 7-day readmissions. This suggests that our model's confidence in a patient's likelihood of being discharged within 48 hours is also indicative of the safety of the discharge. Importantly, this relationship holds even though readmission outcomes are not included as inputs or outputs during the training of our discharge prediction models.

We also investigate the relationship between our green alert system and readmission risk. Specifically, we compare the proportions of patients who are subsequently readmitted within 30 days or 7 days, among those discharged with and without a green alert, respectively. As shown in Figure 3-8, patients who receive a green alert 48 hours before discharge have 3.07% lower 30-day readmission risk and 1.13% lower 7-day readmission risk compared to those discharged without a green alert. One-sided Welch's t-tests with different variance groups reject null hypotheses with p-value $< 10^{-8}$ for 30-day and p-value $< 10^{-3}$ for 7-day, concluding that patients discharged with green alerts have statistically significantly lower readmission rates than those

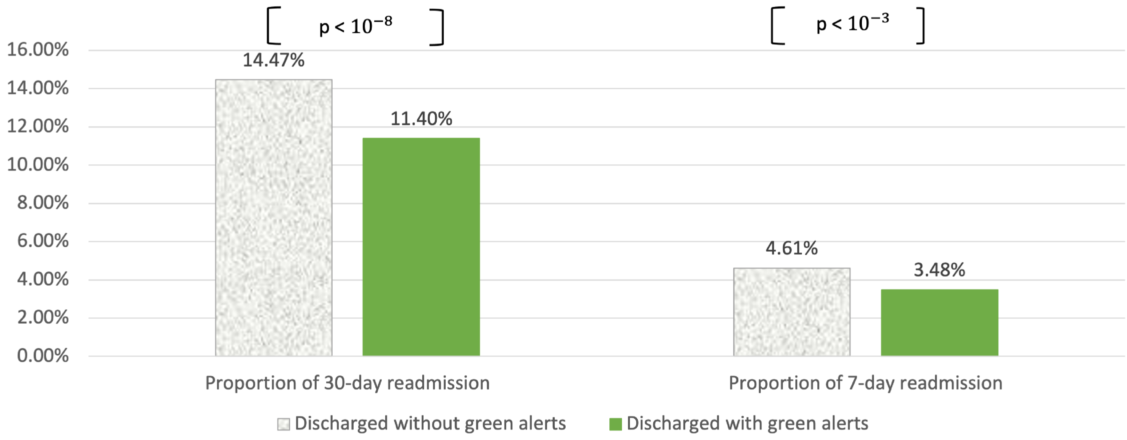


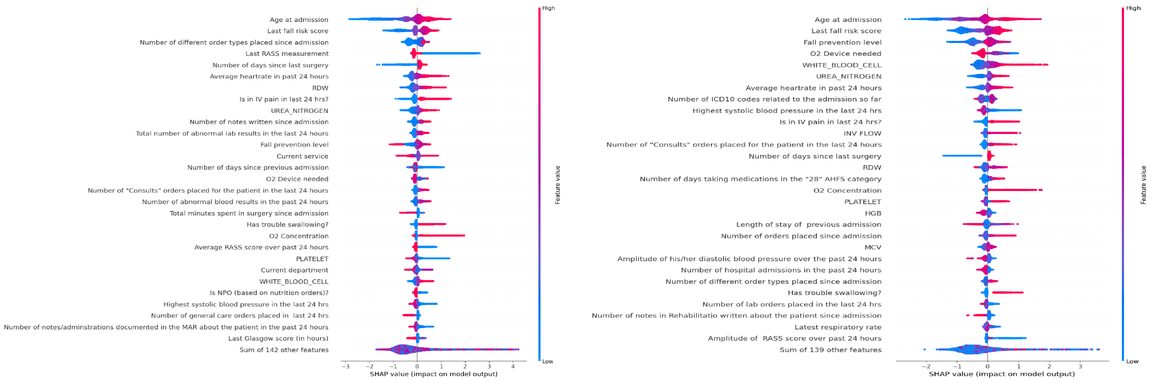
Figure 3-8: Readmission Risks for Patients Discharged without vs with Green Alerts.

discharged without green alerts. Odds ratios for 30-day and 7-day readmissions for patients without a green alert compared to those with are 1.32 and 1.34, respectively. This indicates that the odds of being readmitted within 30 days and 7 days are increased each by over 30% for patients who do not receive a green alert. These findings suggest an additional potential use of our green alert system to screen readmission risks and advise more confidence discharges.

3.5.5 Understanding Predictions

We show SHAP summary plots of example prediction models on sample hospitals to understand how the top 30 features of each model impact the prediction in the testing set. In the plots, positive SHAP values represent an increase in predicted probability whereas negative values represent a decrease; the colors ranging from red to blue represent high to low values of the features.

We analyze Figure 3-9 for mortality models. For some important variables, such as age, fall risk score, heart rate, RDW, intravenous (IV) pain, and white blood cells, higher values drive higher mortality risks. Other important variables include the number of order types since admission, the number of days since the last surgery, the oxygen (O2) device, fall prevention level, and more. While most of these variables align with doctors' experiences and intuitions, others differ from what doctors typically use to make their assessments. For example, according to our model, RDW plays



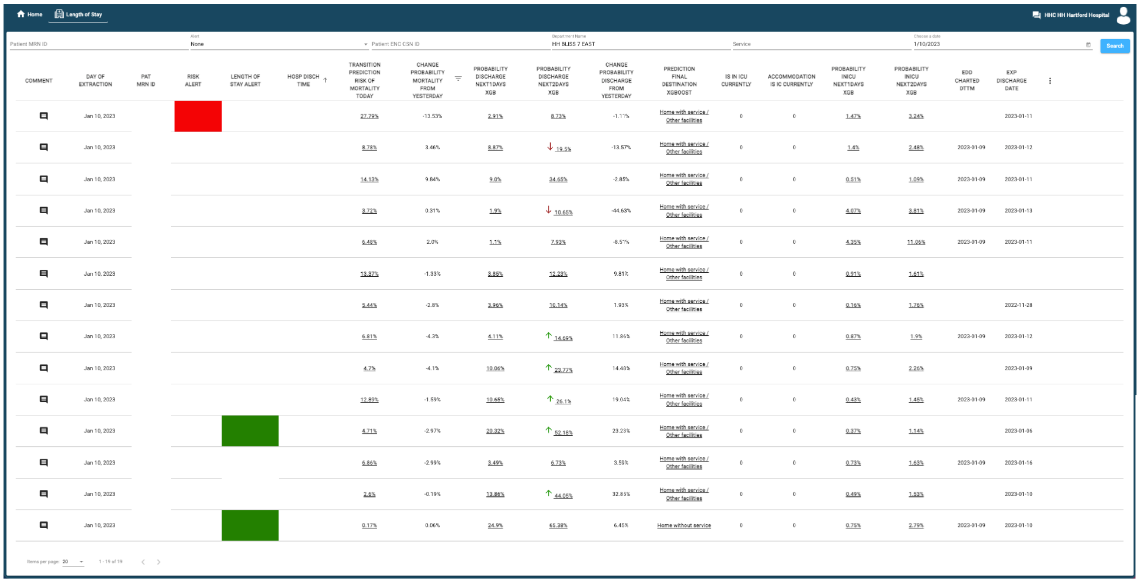
(a) HH Mortality.

(b) WH Mortality.

Figure 3-9: SHAP Summary Plots for Mortality Predictions.

a major role in mortality predictions; however, this feature was not one of the main variables doctors considered during their assessment. After some literature review, we learned that multiple papers found a similar significance of RDW in mortality predictions (Şenol et al. 2013, Wang et al. 2018, Soni and Gopalakrishnan 2021), and following careful examination doctors agreed to incorporate RDW in their mortality assessment. This anecdote illustrates how interpretable machine learning can help doctors learn and enhance their clinical evaluation.

A comparison of the HH model (Figure 3-9a) and the WH model (Figure 3-9b) shows that while the majority of the important variables are the same between the two hospitals, a small number of differences exists. For instance, the current department is a crucial factor in the HH model's predictions, but it does not rank among the top 30 features in the WH model. Reviews with doctors suggest this disparity might be due to the fact that HH has a broad range of departments catering to a diverse patient population, with varying levels of care and specialties, whereas WH has a small collection of departments that do not include critical or intensive care for a more homogeneous patient group. Further, we present additional SHAP plots on discharge and ICU predictions in Figure B-2 and analyze them in Section B.2.4 of the Appendix.



Patient MRN ID | Alert: None | Patient ENC CSN ID | Department Name: HH BLISS 7 EAST | Service | Choose a date: 1/10/2023 | Search

COMMENT	DAY OF EXTRACTION	PAT MRN ID	RISK ALERT	LENGTH OF STAY ALERT	HOSP DISCH TIME ↑	TRANSITION PREDICTION RISK OF MORTALITY TODAY	CHANGE PROBABILITY MORTALITY FROM YESTERDAY	PROBABILITY DISCHARGE NEXT1DAYS XGB	PROBABILITY DISCHARGE NEXT2DAYS XGB	CHANGE PROBABILITY DISCHARGE FROM YESTERDAY	PREDICTION FINAL DESTINATION XGBOOST	IS IN ICU CURRENTLY	ACCOMMODATION IS IC CURRENTLY	PROBABILITY INICU NEXT1DAYS XGB	PROBABILITY INICU NEXT2DAYS XGB	EDD CHARTED DTTM	EXP DISCHARGE DATE
🗨	Jan 10, 2023		🟥			27.79%	-13.53%	2.91%	8.73%	-1.11%	Home with service / Other facilities	0	0	1.47%	3.24%		2023-01-11
🗨	Jan 10, 2023					8.78%	3.46%	8.87%	↓ 19.5%	-13.57%	Home with service / Other facilities	0	0	1.4%	2.48%	2023-01-09	2023-01-12
🗨	Jan 10, 2023					14.13%	9.84%	9.0%	34.65%	-2.85%	Home with service / Other facilities	0	0	0.51%	1.09%	2023-01-09	2023-01-11
🗨	Jan 10, 2023					3.72%	0.31%	1.9%	↓ 10.65%	-44.63%	Home with service / Other facilities	0	0	4.07%	3.81%	2023-01-09	2023-01-13
🗨	Jan 10, 2023					6.48%	2.0%	1.1%	7.93%	-8.51%	Home with service / Other facilities	0	0	4.35%	11.06%	2023-01-09	2023-01-11
🗨	Jan 10, 2023					13.37%	-1.33%	3.85%	12.23%	9.81%	Home with service / Other facilities	0	0	0.91%	1.61%		
🗨	Jan 10, 2023					5.44%	-2.8%	3.96%	10.14%	1.93%	Home with service / Other facilities	0	0	0.16%	1.76%		2022-11-28
🗨	Jan 10, 2023					6.81%	-4.3%	4.11%	↑ 14.69%	11.86%	Home with service / Other facilities	0	0	0.87%	1.9%	2023-01-09	2023-01-12
🗨	Jan 10, 2023					4.7%	-4.1%	10.06%	↑ 23.77%	14.48%	Home with service / Other facilities	0	0	0.75%	2.26%		2023-01-09
🗨	Jan 10, 2023					12.89%	-1.59%	10.65%	↑ 26.1%	19.04%	Home with service / Other facilities	0	0	0.43%	1.45%	2023-01-09	2023-01-11
🗨	Jan 10, 2023			🟩		4.71%	-2.97%	20.32%	↑ 52.18%	23.23%	Home with service / Other facilities	0	0	0.37%	1.14%		2023-01-06
🗨	Jan 10, 2023					6.86%	-2.99%	3.49%	6.73%	3.59%	Home with service / Other facilities	0	0	0.73%	1.63%	2023-01-09	2023-01-16
🗨	Jan 10, 2023					2.6%	-0.19%	13.86%	↑ 44.05%	32.85%	Home with service / Other facilities	0	0	0.49%	1.53%		2023-01-10
🗨	Jan 10, 2023			🟩		0.17%	0.06%	24.9%	65.38%	6.45%	Home without service	0	0	0.75%	2.79%	2023-01-09	2023-01-10

Items per page: 20 | 1 - 19 of 19

Figure 3-10: Table of Predictions (with Patient IDs De-identified).

3.6 Implementation and Impact

As presented in Section 3.4, the solution was first piloted with physicians at HH. This pilot phase helped refine the user interface and conduct a first impact evaluation, before extending the deployment to the other hospitals.

3.6.1 Software Implementation

Each doctor or medical staff can log into their account with two-factor authentication, select their hospital, and enter the application. In the tool, patients can be filtered by a variety of features (e.g., department, date, alert, patient ID). For example, Figure 3-10 shows a list of patients and their associated predictions in the HH BLISS 7 East unit on January 10th, 2023. Each row corresponds to one patient on a given day. Each user can customize their interface and the subset of columns to display among the following five categories:

1. Basic information, e.g., current date, patient ID, current department, room, bed number, service, and level of care.

2. Predicted probabilities for the eight patient operational characteristics on the

current day.

3. The predicted probabilities from the previous day or the change in these probabilities. An arrow ↑ (resp. ↓) is prefixed if the 48-hour discharge probability increases (resp. decreases) by at least 0.1 compared to the previous day.

4. Color-coded (green and red) alerts.

5. Doctor's expected discharge date with charted time.

The software also supports printing and can be accessed from mobile devices. Upon clicking the comment icon, the user can provide feedback on the prediction for the team to review.

As discussed in Section 3.1, the clinician has also access to a SHAP waterfall plot for each patient prediction (upon clicking on the predicted value) to understand the patient-specific factors that explain the prediction. Figure 3-2 displays two examples of SHAP waterfall plots. For each feature on the vertical axis, we represent its contribution (either negative, in blue, or positive, in red) to the final prediction for this particular patient.

3.6.2 Progression of Hospital Deployment

We describe the progress of software use at HHC, from the pilot deployment with physician champions at Hartford Hospital, and following the initial benefits revealed, expanded scales of utilization in all hospitals at HHC.

Financial Benefits from Pilot Implementation at HH: Before rolling out the tool to the other hospitals, HHC's leadership and financial departments evaluated the potential benefits of the tool, based on the result of the pilot implementation at HH. They conduct a simple before-and-after analysis and compare the lengths of stay of patients in Q4 2020 (277 such patients) and Q4 2021 (351 patients), for those patients whose attending discharge physician is one of the four physician champions who had access to the tool in the second half of 2021. Compared with Q4 2020, they observe a reduction in average LOS by 0.35 days (from 5.84 to 5.49) in Q4 2021. Given HH's high utilization rate, HHC assumes that any additional bed made available thanks to a reduction in LOS will be occupied. Accordingly, they estimate (see details in Section B.3.2 of the Appendix) that the pilot implementation could have generated a $711,348.44 increase in annual contribution margin (CM), hence motivating further deployment.

However, we acknowledge that this initial evaluation has several limitations. The pilot implementation only involves a limited number of physicians/units/patients, in a single hospital, and over a limited period of time. The magnitude of the benefits is likely to be different once the tool is more deeply integrated into the workflows of all medical staff, and deployed to other units and hospitals. Moreover, in their before-after analysis, HHC compares LOS in 2020 Q4 and 2021 Q4 without controlling for potential confounding factors that could explain differences between the two periods (e.g., COVID-19 level).

Large-scale Deployment in All Hospitals: Based on the positive feedback and benefits from the pilot implementation, HHC decided to expand the program not just to physicians but also to medical service teams, case management, and nursing unit leadership. They deployed the tool to more floor units and included more physician champions in the study. HHC progressively rolled out the software in more units of HH, as well as other hospitals, with a focus on medicine and cardiology (some of such units also contain surgical service), because such units cover most of the discharges and would benefit most from using the predictions. Section B.3.3 of the Appendix provides details on the progressive deployment at HHC, including information on 15 units where our tool is fully deployed (with start dates between July 11, 2022 to January 15, 2023) and 12 units where the incorporation has not been completed (as of April 15, 2023).

3.6.3 Empirical Effect on LOS Reduction

We analyze the impact of using our tool on patient length of stay. We consider two groups of units from Table B.9 in the Appendix: a treatment group of 15 HHC units that fully used our tool between January 15 and April 15, 2023, and a control group of 12 units that had not yet fully incorporated the tool as of April 15, 2023. We collect discharge data for patients whose exit status is neither expired nor hospice since 2021 and compute the LOS of each patient as the difference between admission order time and discharge time. We exclude patients whose recorded admission order time is after their recorded discharge time. The outcome of interest is the average LOS of patients discharged from each unit group. This convention (i.e., assigning patients to their discharge unit only) is aligned with the conventions used by HHC for performance evaluation and with the fact that our tool can mostly impact patient LOS by helping physicians anticipate and plan the discharge process.

To estimate the treatment effect, we use the Difference-in-Differences (DiD) technique (Abadie 2005, Bertrand et al. 2004), which compares the average change in LOS among patients in the treatment group to that of patients in the control group over time. We control for similar patient population fixed effects between the two groups, as they covered units of the same level of care (general level) and specialty (medicine and cardiology). We also control for the time non-stationarity effect within the year on LOS by focusing on the January 15 - April 15 period of the past three years.

We present the results in Figure 3-11, which shows the average LOS (in the number of days with decimal points) over the three-month period from 2021 to 2022 and 2023. The control group did not use the tool throughout, whereas the treatment group had varying degrees of tool usage during the three periods. HHC reported no tool usage from January 15 – April 15, 2021, partial tool usage in 5 units due to the pilot program in January 15 – April 15, 2022, and full deployment of the tool in all 15 treatment units on January 15 – April 15, 2023. We assumed that the difference in LOS over time would have been the same between the two groups if the tool had not been used throughout the periods, based on the parallel trend assumption from DiD. We thus

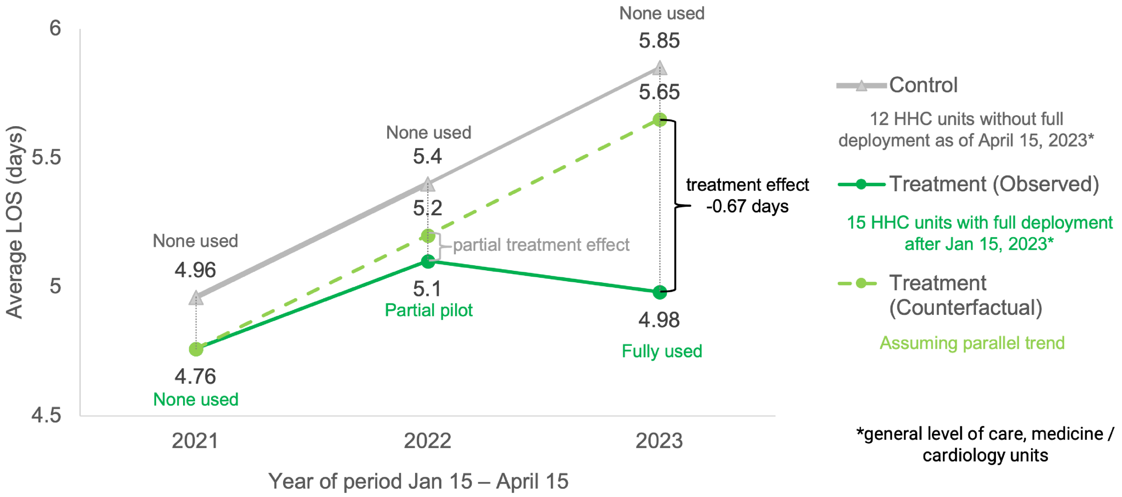


Figure 3-11: Difference-in-differences Analysis for Treatment Effect on Length of Stay.

imputed the counterfactual average LOS for the treatment group if there had been no treatment (in light green dashes).

The LOS of the control group showed a steady (approximately linear) increase, rising from 4.96 to 5.4 and eventually reaching 5.85 over three periods. Between 2021 and 2022, the treatment group's LOS increased from 4.76 to 5.1, which was in line with the parallel trend but slightly lower by 0.1 days, potentially due to the pilot's partial treatment effect. After full deployment, the treatment group's LOS dropped to 4.98 in 2023, while the control group's LOS continued to rise from 4.96 to 5.85. The difference between the parallel counterfactual and actual treatment group's LOS resulted in an estimated benefit of reducing the average patient length of stay by 0.67 days. This benefit was more significant than the estimation from the pilot evaluation at HH due to several reasons, such as the tool's increased utilization in more units and hospitals and its deeper integration into the daily clinical workflow, resulting in a higher potential for LOS reduction. Our analysis incorporated control variables such as time period of the year, level of care, and specialty of units, and it was multi-center, which was absent from the simple before-and-after comparison for the pilot.

However, our analysis holds several assumptions and limitations. For instance, this was a descriptive analysis that assumed a parallel trend of LOS between the two equivalent patient populations, controlling only basic time and unit variables.

Table 3.6: Projection of Financial Benefits from LOS Reduction.

Deployment coverage	15 treatment units	12 control units
LOS reduction (days)	0.67 (estimated)	0.67 (extrapolated)
Patients per year	49,424	24,565
Patient days saved	33,114.08	14,448.55
Average new LOS	4.98	5.22
Proportion of beds backfilled	50%	50%
Additional patients	3324.71	2767.92
Average CM per patient	$10,796	$10,796
Total CM increase	$35,893,515.18	$14,941,240.02

Additionally, there could be other confounding factors affecting LOS, such as patient demographics and other efforts to reduce LOS in HHC. Moreover, even though a small portion of physicians in the control units had access and used the predictions, we deemed these units as the control group, which could lead to a more conservative estimation of the effect. Finally, heavy-tailed LOS distributions could be significantly influenced by patient outliers with overly long LOS.

3.6.4 Projected Contribution Margin Increase

Table 3.6 outlines our projections for the estimated impact on HHC's contribution margin as a result of the LOS reduction. By reducing the average LOS by 0.67 days among patients in the 15 treatment units (49,424 annually), we can save 33,114.08 patient days. Assuming a conservative scenario where only half of the beds would be backfilled, we estimate that this would make room for an additional 3,324.71 patients per year, with an average CM of $10,796 per patient. This leads to a projected total annual CM increase of $35,893,515.18 in these units. With the expansion of the tool to the 12 control units, we expect an additional increase of $14,941,240.02 in CM. Overall, upon full utilization among these 27 units, we anticipate a financial benefit of $50.83 million per year, with the potential for further growth as the tool is extended to more units across different levels of care and specialties.

3.6.5 Limitations and Future Work

Although we tried to incorporate as many relevant variables as possible, some information considered as predictive by the medical team is still unavailable and unaccounted for by the current models. This includes medically-related variables (e.g., Clinical Institute Withdrawal Assessment for Alcohol scale score) as well as social and operational issues that are usually recorded in nursing and case management notes (e.g., goals of care, palliative discussion, pending authentication to the skilled nursing facility, and pending equipment delivery to home for patients whose discharge disposition plan is to go home with service). The team is working on obtaining these additional features and notes, where the latter would require supplemental Institutional Review Board protocol, text de-identification, and privacy management.

Like many medical institutions, HHC uses a commercial EMR system, from which we extract data using our own data pipeline (from Figure 3-3). Due to the technical complexity of this integration, data is refreshed once a day only and with an 8-hour delay. Physicians, however, identified that some relevant information might become available between the extraction time (midnight) and the time the data extracts become available (8 a.m.), information that the model currently cannot take into account. A more frequent update with shorter delays is an area of potential improvement, especially for ICU predictions. As of today, despite extensive efforts, we have not been able to lift the technical bottlenecks in the hospital data system that generate these long delays. Dynamic Ideas and HHC are currently working on running the models directly within the hospital EMR system (instead of conducting data extracts and running the models on a dedicated AWS server), which would allow for near-to-real-time data updates.

In future work, we will continue expanding the deployment of the tool and evaluating its impact on LOS and other patient/hospital outputs, as more data is collected over longer time periods. Followed by the connection between higher discharge predictions and lower readmission risks, another direction is to develop and integrate patient-level readmission risk predictions in production to enhance patient safety in practice.

3.7 Concluding Remarks

As part of a collaboration between Hartford HealthCare, MIT, and Dynamic Ideas, we develop a system of machine learning models predicting short-term discharge, mortality and discharge disposition, and ICU risk. After three years of iterative development and validation, the final models aggregate a variety of clinical and nonclinical data about the daily status of each patient, achieve state-of-the-art predictive accuracy, and are well-calibrated.

In particular, the 48-hour discharge predictions generated by the analytics program provide 6.3%-29.6% higher distinguishing power (measured by AUC) than the ones currently obtained from the healthcare providers directly; the combination of the two (human and algorithmic) predictions can achieve 12.1%-29.4% higher precision or identify 10%-28.7% more discharge opportunities. Furthermore, the predicted discharge probabilities uncover a negative representation of 30-day and 7-day readmission risks, unintentionally capturing discharge safety. Patients discharged with green alerts have statistically significantly lower readmission rates compared with those without green alerts, suggesting another potential use of discharge predictions to enhance patient safety.

In addition to being accurate, these models have been deployed in seven hospitals and are being used by hundreds of doctors, nurses, and case managers at HHC. The close and long-term partnership with key stakeholders at HHC enabled a broad adoption of the tool by medical providers and a deep integration within their daily workflow. As a result, HHC experienced first-hand benefits to shorten the length of stay, decrease the cost of care, facilitate the education of patients and families regarding discharge, enhance patient safety, and improve the overall patient experience. Empirically, we observe a reduction in average patient length of stay by 0.67 days and project an annual contribution margin increase of $35.89 million dollars.

The successful deployment in multiple centers of a representative U.S. hospital network also demonstrates the significant potential to scale the framework to other healthcare systems, in the U.S. and around the world.

Chapter 4

Data-driven Interpretable Policy Construction Personalizes Mobile Health Intervention

4.1 Introduction

Mobile health (mHealth) applications play an increasingly important role in helping people improve their lives. In part, mHealth applications do this by tracking individuals' behaviors and contextual data and sending interventions via smart devices such as smartphones and wearables to encourage healthy behaviors. Applications target a variety of health behaviors including substance use (Gustafson et al. 2014, Tofighi et al. 2019, Boumparis et al. 2019), chronic disease management (Hamine et al. 2015), stress management (Battalio et al. 2021), and physical activity (Yom-Tov et al. 2017, Klasnja et al. 2019). This work concerns an mHealth app, HeartSteps, which is designed to promote physical activity and reduce sedentary behavior in individuals with hypertension (Liao et al. 2018, Klasnja et al. 2019). In the second and third trials of HeartSteps (Liao et al. 2018), HeartSteps V2/V3, whenever a user is sedentary for more than 40 minutes, the HeartSteps app may send an anti-sedentary message with the goal of interrupting prolonged sedentary episodes. The messages are sent using an algorithm (Liao et al. 2018) that uniformly samples each user's eligible sedentary times, with the goal of sending an average of 1.5 messages per day per user. There was no effort to optimize delivery, so as to reduce the sedentary behavior. Our goal is to learn an interpretable policy for delivering anti-sedentary messages with the objective of reducing the time it takes for users to become non-sedentary with less message delivery.

Learning a policy using clinical trial data, such as the data from HeartSteps V2/V3, poses a number of challenges. The policy needs to be learned in a data-poor environment with a small number of users. In this study, we use data from 51 users who participated in the HeartSteps V2/V3 trials. The data was collected via the users' smartphones, a wrist-worn activity tracker, and pre-study surveys. The data is noisy, with many missing or imprecise measurements. Such data environments require careful data imputation and processing steps, as well as intelligently pooling data during policy learning. Further challenges are related to the nature of policy to be learned. First, to maintain user engagement with the intervention, treatments cannot be sent overly frequently, since over-treatment can lead to user disengagement (Nahum-Shani et al. 2018); this includes users becoming insensitive to the messages and, in more drastic cases, deleting the application from their smartphone. Second, in the context of HeartSteps, the learned policy, along with additional analyses of the data and reviews of the current literature on mHealth interventions, will be used to formulate how best to improve the effectiveness of the HeartSteps application in future implementations. Thus, it is critical that the learned policy be interpretable. The scientific team should be able to use the policy to derive useful conclusions regarding the importance of different contextual variables in optimizing the delivery of the messages.

Our work aims to overcome the above challenges in learning such a policy. We approximate the environment by a contextual bandit environment for each user; that is, we assume that the effect of the anti-sedentary messages is momentary, not lasting more than 2 hours on user behaviors. Our approach deals with the potential longer-term impact of messages (i.e., disengagement) by constraining the number of messages that can be delivered per day. We modify doubly robust estimation (Dudík et al. 2011) to estimate counterfactual outcomes for use in batch off-policy evaluation. For batch off-policy learning, we consider two methods. First, we derive a threshold-based approach, similar to approaches used across queueing systems (Lin and Kumar 1984), trading (Bertrand and Papavasiliou 2019), and optimal stopping (Baumann et al. 2020) fields, due to the simplicity and ability to trade off between exploration and exploitation (Sang et al. 2020). Second, we generalize the Optimal Policy Trees (Amram et al. 2022) method to accommodate constraints. We name the proposed policy Optimal Policy Trees + (OPT+), which will be used, along with qualitative studies, other analyses, and the evolving literature, to inform the further development of the HeartSteps mobile intervention.

Our contributions are three-fold:

1. We develop two innovative batch off-policy learning methods. First, we design a threshold-based method, that learns personalized thresholds with the feasibility to be used as a stand-alone policy. Second, we develop Optimal Policy Trees +, which builds on the thresholds and Optimal Policy Trees (OPT) to learn interpretable decision tree-based policies.

2. We develop and test an end-to-end batch off-policy learning approach for the mobile health pipeline. Using HeartSteps V2/V3 data we demonstrate that our method is able to improve the effectiveness of the anti-sedentary messages significantly while reducing the number of messages to send. In particular, OPT+ has an edge in interpretability and stability over the stand-alone threshold-based method without sacrificing performance.

3. The resulting decision tree-based policy is informative for further development of the HeartSteps app. The decision tree policy identifies key variables for optimizing the effect of the anti-sedentary messages, including certain user traits, as well as more dynamic information. Some variables were previously recognized as important whereas other variables are new and require validation. See further discussions in Section 4.7.1.

The outline of the chapter is as follows. We review related work in Section 4.2 and formulate the problem framework in Section 4.3. Next, we describe the methodology for policy evaluation in Section 4.4 and for policy learning in Section 4.5. We then present computational results in Section 4.6. Finally, we discuss our recommendations, limitations and future work in Section 4.7 before stating conclusions in Section 4.8.

4.2 Related Work

We relate our work to the broader literature from mobile health, batch off-policy learning, and tree-based policies.

4.2.1 Computational Methods in Mobile Health

Much machine learning work in mobile health focuses on detecting and predicting behaviors, such as activity sensing (Consolvo et al. 2008), detecting moments of stress (Hovsepian et al. 2015), clustering physical activity patterns (Fukuoka et al. 2018) and predicting user adherence (Zhou et al. 2019). Further work estimates treatment effects using data from micro-randomized trials (Bidargaddi et al. 2018, Klasnja et al. 2019) to answer questions such as "Does providing suggestions increase app engagement compared to no suggestion?" and "Do suggestions work only during certain parts of the day, like the morning or afternoon?" Despite facilitating an understanding of user behaviors, such analyses do not directly learn policies.

Policy construction methods have been developed that utilize reinforcement learning algorithms, for example in physical activity promotion (Yom-Tov et al. 2017, Zhou et al. 2018a), and in weight loss (Forman et al. 2019). Often bandit or contextual bandit algorithms are used. Some of these algorithms learn online, such as My-Behavior app for physical activity and dietary change (Rabbi et al. 2015a,b, 2017, 2018), and a context-adaptive algorithm for medical diagnosis (Tekin et al. 2014). In an off-policy setting with the goal of improving emotion regulation (Ameko et al. 2020), a contextual bandit algorithm is used for batch learning. Since they consider a simpler problem of learning a time-invariant initial policy for users, a recommender algorithm can be built directly based on doubly robust estimators and logistic regression (Ameko et al. 2020). In contrast, we consider a more complicated setting, consisting of frequent decision time steps subject to budget and other constraints, which require additional learning methodology. As discussed in the Introduction, in the HeartSteps V1/V2 studies, an algorithm was used to uniformly sample sedentary times (Liao et al. 2018). For a different HeartSteps intervention component (walking suggestions as opposed to anti-sedentary messages), an online reinforcement learning algorithm (Liao et al. 2020) was developed and implemented. Our work focuses on batch off-policy reinforcement learning to develop an interpretable policy and to provide recommendations for future studies. Anticipating future needs in policy development, we also provide a way to transition from learning the optimal policy on current users to new users.

4.2.2 Batch Off-policy Learning

Our work falls under the field of offline reinforcement learning, for which we refer the readers to the review paper (Levine et al. 2020). In the literature, there exists a variety of batch off-policy construction methods, which all use counterfactual estimation for off-policy evaluation. One commonly used evaluation method is the doubly robust method (Dudík et al. 2011), which combines modeling the outcome and inverse propensity score weighting (Horvitz and Thompson 1952) and results in an estimator with reduced variance. Among the batch off-policy learning methods, several works consider a generally similar framework to ours but have important differences. For instance, one algorithm (Athey and Wager 2021) uses a doubly robust estimator and learns a treatment policy subject to specific constraints. In their framework, the policy must belong to a pre-specified policy class, whereas our method defines our own new policy class. In another work (Sun et al. 2021), a similar problem is considered to maximize the treatment effect under budget constraints, using data collected from randomized controlled trials. Their work focuses on modeling an uncertain cost function, whereas we utilize past data points with observed costs. These approaches differ from ours in other important ways as well. They consider prioritizing between users among the population while we focus on personalizing the policy for each user in a time series setting. Their work is based on black-box methods such as random forest, whereas ours are based on thresholds and decision trees to make the policy more interpretable to behavioral scientists. Moreover, our applications to mobile health need to address the particular challenges mentioned in Section 4.1.

4.2.3 Tree-based Policies

This class of policy learning methods uses decision tree-based policies, which encourage wide practical applications due to their interpretability. A family of personalized tree algorithms is proposed to perform recursive partition on the data and minimize the defined sum of within-partition personalization impurities (Kallus 2017). Optimal Prescriptive Trees (Bertsimas et al. 2019) applies the Optimal Trees (Bertsimas and Dunn 2017) and jointly optimizes the problem for predictive accuracy in counterfactual estimations and optimality of policy outcomes. Cross-fitted Augmented Inverse Propensity Weighted Learning (Zhou et al. 2018b), with the implemented R package policytree (Sverdrup et al. 2020), applies doubly robust estimators (Dudík et al. 2011), a K-fold algorithmic structure, and a tree-formed policy.

More recently, the Optimal Policy Trees (OPT) method (Amram et al. 2022) has been developed which has several advantages over the other tree-based approaches. The counterfactual estimation is conducted separately prior to the tree construction and is not limited to the tree form of piecewise linear or piecewise constant structure. Based on the Optimal Trees (Bertsimas and Dunn 2017) structure, OPT simultaneously achieves both global optimality of the tree and high scalability. Moreover, Optimal Trees tend to be more stable than greedy decision trees due to the global optimization and automatic parameter tuning during the training process. OPT yields interpretable policies and demonstrates strong performance in benchmarking problems.

4.3 Problem Framework

In this section, we first describe the problem background and the dataset through the HeartSteps study and then formulate the problem mathematically.

4.3.1 HeartSteps Study

The HeartSteps V1 study (Klasnja et al. 2015, 2019) was a mobile health study in which sedentary users received prompts on their phones to plan physical activity and suggestions to encourage near-term activity. In the next study HeartSteps V2/V3 (Liao et al. 2018), patients with stage 1 hypertension wore a Fitbit Versa tracker and used an updated HeartSteps mobile application. One intervention uses anti-sedentary messages, messages (via phone notifications) designed to disrupt prolonged episodes of sedentary time. An illustration of the anti-sedentary Message interface is shown in Figure 4-1. At every 5-minute time step during a 12-hour window anchored by each user's typical sleep schedule, it was determined whether users were sedentary and available for treatments. Users were deemed to be sedentary if they have taken fewer than 150 steps in the past 40 minutes, and available if they had been sedentary for at least 40 minutes, no intervention message was sent in the last hour, and they had not had an extended bout of physical activity in the previous two hours. At each available sedentary time step, an algorithm (Liao et al. 2018) randomized sending a message with variable probabilities computed based on a prediction of the number of additional sedentary episodes that the user would have over the rest of the day.

4.3.2 Dataset Description and Pre-processing

We compile three data sources collected from the HeartSteps V2/V3 study: longitudinal step and heart-rate minute-level data gathered by Fitbit activity trackers, 5-minute level anti-sedentary treatments data, and baseline user characteristics gathered by surveys. To standardize between the V2 (3-month) and V3 (9-month) studies, we include up to the first 80 days after each user's treatment start date. We further

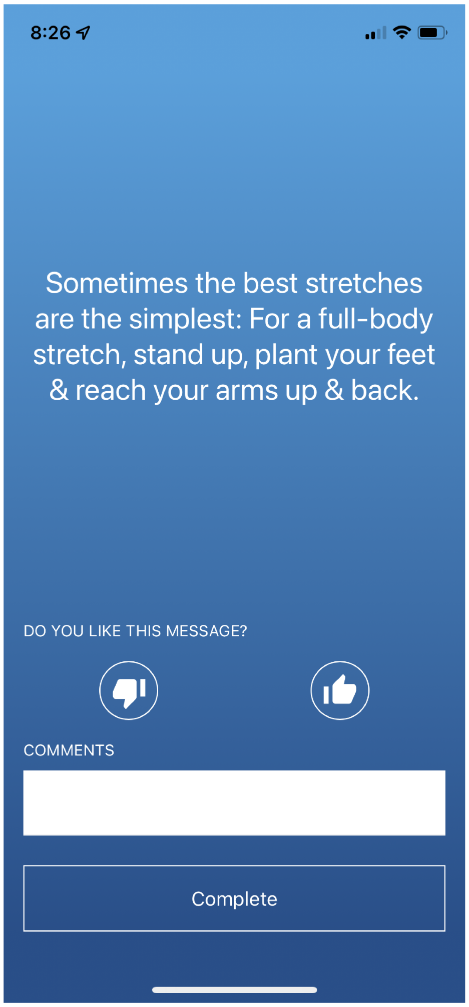

Sometimes the best stretches are the simplest: For a full-body stretch, stand up, plant your feet & reach your arms up & back.

Figure 4-1: Screen Shot of an Anti-sedentary Message from HeartSteps.

restrict to time steps with at least one measurement in the next 2 hours (to determine the outcome), with at least 5 same-day prior measurements (to construct the contextual features), and when the user is available (for treatment feasibility). Out of the original 94 users from HeartSteps V2/V3, we have access to all three data sources from 63 users, among whom 2 did not have any available time steps, and additional 10 provided less than a week of data—in both cases missingness being due to technical problems. Thus here we use the data from the remaining 51 users, including 25 V2 users and 26 V3 users. Within the 80-day duration of this study, each user can have a different number of days and 5-minute measurements for which we have recorded study data due to, for instance, non-wear of the Fitbit tracker. We only keep time steps where there are associated measurements, and impute the additional missing data (number of steps, heart rates, and user features) using the OptImpute algorithm (Bertsimas et al. 2018), a state-of-the-art imputation algorithm based on optimization and k-nearest neighbor learner.

4.3.3 Problem Formulation

We assume a contextual bandit environment for each user. For each user at each available time step, the user's current state is used to make a decision. We assume that the effect on the user of sending a message lasts at most two hours since the content of the messages suggests an activity that would last less than 1-2 minutes. The learned policy is constrained by a daily budget across the time steps. We define policy metrics at a user daily level in evaluating learned policies.

Indices and Sets. User $i \in [I]$ denotes each user in the study. Day $d \in [D_i]$ denotes each day of the study for user i. Time step $t \in [T_i]$ denotes every up to 5-min time step when a decision is to be made for user i. $[T_{id}]$ denotes the set of available time steps t on day d for user i.

States. For each user $i \in [I]$, at time step $t \in [T_i]$:

- We observe contextual information vector x_{it}, including 103 features from 5 categories: the number of steps taken and heart rate at each of the most recent 6 measured time steps, time step information such as hour of the day and day of the week, in addition to user demographics, personality traits, and routines as reported from the baseline survey.
- We access the feasibility indicator of sending a message:

$$F_{it} = \min\{I_{it}, r_{it}, N - n_{it}\} \in \{0, 1\},$$

with availability indicator $I_{it} \in \{0, 1\}$, indicator of a message was sent in the past 2 hours $r_{it} \in \{0, 1\}$, number of messages sent during the day so far $n_{it} \in \mathbb{Z}^+$, and pre-determined budget on number of messages per each user day $N \in \mathbb{Z}^+$.

- We then define the state as the tuple

$$S_{it} = \{x_{it}, F_{it}\}.$$

Decision Variables. For each user $i \in [I]$, at time step $t \in [T_i]$, we learn a policy $\pi(S_{it})$ to decide the action $A_{it} \in \{0, 1\}$ on whether to send a message. From the data, we know the historical actions $A_{it}^{\text{hist}} \in \{0, 1\}$ with the probability to send treatment $p_{it}^{\text{hist}} \in [0, 1]$ by the current algorithm (Liao et al. 2018).

Constraints. The learned policy is subject to the following constraints. Treatments can only be sent at the user's available times. At most one treatment can be sent every two hours. The total number of messages to each user on each day cannot exceed the budget N. The above can be expressed using one feasibility constraint:

$$A_{it} \leq F_{it}, \quad \forall i \in [I], t \in [T_i]. \tag{4.1}$$

Since the actions affect the feasibility of the following decision times, optimization is challenging as the constraint is intertwined with the policy we aim to learn.

Outcomes per User Time Step. For each user $i \in [I]$, at time step $t \in [T_i]$, we define the outcomes given state S_{it} and action A_{it} as follows. We define the proximal activity outcome to be the number of minutes the user takes to change from being sedentary to being non-sedentary. We observe partial outcomes y_{it}^{hist} from historical data, and then impute the outcomes $\hat{y}_{it}^1(x_{it}), \hat{y}_{it}^0(x_{it})$ under context x_{it} if a message is sent and if a message is not sent, respectively, using the procedure in Section 4.4.2. Sometimes the duration of the number of minutes until no longer sedentary is censored due to missing step count data. We redefine the duration to be the duration until step counts are observed and the user is not sedentary. We truncate this duration with a maximum cap to $y^{\max} = 120$ minutes. We then define the treatment effect of sending a message to be the difference between the outcome with a treatment and the outcome without a treatment:

$$\hat{\delta y}_{it} = \hat{y}^1_{it}(x_{it}) - \hat{y}^0_{it}(x_{it}). \tag{4.2}$$

Metrics per User Day. We use two metrics to evaluate policies. For each user $i \in [I]$, on each day $d \in [D_i]$, we define the user daily cost of the learned policy π for user i on day d as the total number of minutes of sedentary time in the day: $C_{id}(\pi)$. We define the user daily message count to be the number of messages sent to the user i during day d under the learned policy:

$$N_{id}(\pi) = \sum_{t \in [T_{id}]} \pi(S_{it}).$$

4.4 Policy Evaluation Method

4.4.1 Split of Training, Validation, and Testing Sets

We split the data points into train, validation, and test sets, in correspondence with the study structure and design. For existing users who continue the study, we can learn the policy from data from their previous days and apply the policy to their later days. For new users who will enter the study later, we can learn the policy from pooling existing users to apply it for all days for the new users. Therefore, we conduct a systematic split in two folds. First, we split the users from V2 and V3 respectively into returning users, new users for validation, and new users for testing. Out of the 25 total users from HeartSteps V2, we select 80% users as returning users, approximately 10% as new users for the validation set, and the remaining approximately 10% as newer users for the test set. Users are sorted and split based on their study start date. We apply the same splits for the 26 total users from HeartSteps V3, and combine them with the V2 splits accordingly, thereby maintaining the same ratio between V2 and V3 users in each set. For each of the returning users, we split their days approximately into 80% training, 10% validation, and 10% testing sets. The days are sorted and split based on the date. This two-fold split simulates the

Table 4.1: Summary Statistics of the Splits.

Set split	Training set		Validation set		Testing set	
User group	Return	New	Return	New	Return	New
# users	45	0	45	6	45	5
% users	80%	0%	80%	11%	80%	9%
# user days	2692	0	314	272	154	272
% user days	68%	0%	8%	7%	6%	11%

circumstances in real life, where we need to learn from past user days to apply for future user days. Table 4.1 summarizes the statistics for each group of users in each resulting set split, including the number and percentage of users and user days.

The resulting splits of data points are used in the experiments as follows. We train methods on the training days of returning users (64% of all user days). We then validate the method for both returning and new users in the validation set. Finally, we apply and evaluate the method for both returning and new users in the testing set.

4.4.2 Counterfactual Outcome Estimation

One of the main challenges in this problem is to estimate the counterfactual outcomes. For each user at each time step, exactly one out of the two counterfactual outcomes (corresponding to the realized treatment) is observed. Thus, it is unknown in the historical data what the outcome would have been if the alternative treatment was implemented. For future users and days, the policy will also be based on estimating the outcome under each treatment option. To impute the counterfactuals for our problem, we adapt the doubly robust estimation procedure as follows.

Propensity Score. The propensity scores p_{it}^{hist} (i.e., the treatment probabilities) are in observed in the dataset.

Outcome Prediction. We develop two machine learning models to predict the outcome with and without a message, $f(x_{it}, 1)$ and $f(x_{it}, 0)$ for user i at time t, respectively. We train XGBoost regression models using the training set, with a parameter grid including the maximum depth of the trees ranging from 2 to 7 and the number of rounds from 10 to 300 validated using the validation set. We then apply the two models to the entire data for policy use. In parallel, we train two other XGBoost models independently using the testing set only. The predictions will be set aside during the policy training and application and used at the end as the ground truth to evaluate the policy outcomes at all time steps, including the ones when we observed the outcomes based on historical actions. This leads to a fair evaluation without information leaking when we train and evaluate the policy.

Recall our assumption that the effect of the message on the outcome (i.e., the user's decision to get up and move around) will not last more than two hours from when the message is sent. Since our outcome is the number of minutes until the user becomes non-sedentary, one complication in our setting is that the treatment affects not only the current time step but the steps before and after the treatment as well. These time steps could add noise to the model and confuse the treatment effects. To address this problem, when training the outcome prediction models, we exclude the data points where there is both no treatment at the current time step and there was at least one treatment within 2 hours before or after the current time step. These data points will be included back as potential decision points in the treatment stage.

Doubly Robust Outcome Estimation. We obtain the doubly robust estimated outcome for each data point with and without a message sent, respectively:

$$\tilde{y}_{it}^{1}(x_{it}) = f(x_{it}, 1) + A_{it}^{\text{hist}} \cdot (y_{it}^{\text{hist}} - f(x_{it}, 1))/p_{it}^{\text{hist}},$$

$$\tilde{y}_{it}^{0}(x_{it}) = f(x_{it}, 0) + (1 - A_{it}^{\text{hist}}) \cdot (y_{it}^{\text{hist}} - f(x_{it}, 0))/(1 - p_{it}^{\text{hist}}).$$

It has been proven that the estimated policy outcome is accurate if at least one of the propensity score or outcome estimators is accurate, thus the name doubly-robust (Dudík et al. 2011). In our case, since the propensity score is known, this results in an unbiased estimate of the counterfactual outcome. We apply an additional trimming step to keep the outcome in the range $y \in [0, y^{\max}]$ to obtain the final doubly robust estimates $\hat{y}_{it}^{a}(x_{it})$.

4.4.3 Policy Evaluation Metrics.

Given the metrics per user day from Section 4.3.3, we now define two overall policy evaluation metrics, the average daily cost across all users:

$$\overline{C(\pi)} = \frac{1}{I}\sum_{i\in[I]}\frac{1}{D_i}\sum_{d\in[D_i]} C_{id}(\pi), \tag{4.3}$$

and the average daily number of messages sent across all users:

$$\overline{N(\pi)} = \frac{1}{I}\sum_{i\in[I]}\frac{1}{D_i}\sum_{d\in[D_i]} N_{id}(\pi). \tag{4.4}$$

4.5 Policy Learning Method

We take a three-step approach to solve the problem. In step 1, we follow Section 4.4.2 to estimate outcomes under a policy. In step 2, given the estimated outcomes, we develop a threshold-based decision rule, which can be used either as a stand-alone policy or as one step of our proposed policy. In step 3, we apply Optimal Policy Trees and combine with the previous steps, integrating into our proposed OPT+ policy to select the treatment that gives minimal estimated costs.

4.5.1 Threshold-based Decision Rule

We describe our algorithms to apply and learn a threshold-based decision rule. Given the budget N, we learn for each user i a threshold vector $\Gamma_i \in \mathbb{R}^N$, consisting of threshold values for each of the N messages to be sent during a day. Algorithm 2 uses a threshold to send messages at the first N feasible time steps when the expected treatment effect of sending a message reaches below user i's threshold for the n^{th} message of the day. For each user, the threshold to send the n^{th} message of the day can be different among $n \in [N]$. To trade off between the sedentary outcome and the number of messages sent, we consider varying $N \in \{1, 2\}$.

The process to learn the thresholds is described in Algorithm 3, where D^i_{train} denotes the number of training days for user i, and I_{train} denotes the number of users in the training set. We consider a set of different threshold parameters $\Gamma \in G_1 \times \cdots \times G_N$ to select from, where $G_1 = \cdots = G_N = \{-120, -115, \ldots, -5, 0\}$ is a set with cardinality 25 ranging from -120 to 0 with 5 minutes apart. For each returning user i, we learn a personalized threshold $\Gamma^\star_i$ from that user's historical data. For the new users, we pool all existing data from returning users and learn a best fixed threshold $\Gamma^\star_{fixed}$ for this pooled group of users. Then each new user i in the testing set uses this learned fixed threshold $\Gamma^\star_i = \Gamma^\star_{fixed}$ to be applied in Algorithm 2. This addresses the challenge of no historical record for the new users.

Algorithm 2 Threshold-based decision tule

```
Given budget $N$, threshold $\Gamma_i \in \mathbb{R}^N$ for user $i$, initiate at the beginning of day $d$
for $t \in [T_{id}]$ do
    Estimate outcomes $\hat{y}^1_{it}(x_{it})$ and $\hat{y}^0_{it}(x_{it})$.
    Obtain treatment effect $\widehat{\delta y}_{it}$ from (4.2).
    if $\widehat{\delta y}_{it} \leq \Gamma_i^{n_{it}+1}$ and $F_{it} = 1$ then
        $\pi^{\mathrm{Thr}}(S_{it}) \leftarrow 1$.          ▷ Send a message if threshold reached and feasible
    else
        $\pi^{\mathrm{Thr}}(S_{it}) \leftarrow 0$.
    end if
end for
```

4.5.2 Optimal Policy Trees + Policy

As discussed in Section 4.2.3, the Optimal Policy Trees (OPT) (Amram et al. 2022) is a methodology for learning optimal tree-based policies directly from data. OPT consists of two steps, estimating outcomes and learning an interpretable decision tree policy. For the outcome estimation step, we follow our procedure described in Section 4.4.2 to obtain $\hat{y}^1(x)$ and $\hat{y}^0(x)$. In the second tree training step, it takes contexts x and estimated outcomes, $\hat{y}^1(x)$ and $\hat{y}^0(x)$, as input, and outputs the policy $\hat{\pi}^{OPT}(x)$. The same structure of inputs and outputs follows when OPT is applied to new data. Due to its performance, scalability, and interpretability, we incorporate OPT in the method we propose. However, OPT has key limitations to be directly

Algorithm 3 Threshold learning

```
Given budget $N$ and threshold parameter set $G_1 \times \cdots \times G_N$, initiate
for $i \in [I_{train}]$ do
    for $\Gamma \in G_1 \times \cdots \times G_N$ do                                ▷ Parameter index
        for $d \in [D^i_{train}]$ do
            Apply Algorithm 2 with parameter $\Gamma$ to obtain policy $\pi^{\mathrm{Thr}}(S_{it}), \forall t \in [T_{id}]$.
            Obtain metrics $C_{id}(\pi^{\mathrm{Thr}})$, $N_{id}(\pi^{\mathrm{Thr}})$.
        end for
    end for
    $\Gamma^\star_i \leftarrow \arg\min_{\Gamma} \frac{1}{D^i_{train}} \sum_{d \in [D^i_{train}]} C_{id}(\pi^{\mathrm{Thr}})$.        ▷ Personalized threshold
end for
$\Gamma^\star_{fixed} \leftarrow \arg\min_{\Gamma} \frac{1}{I_{train}} \sum_{i \in [I_{train}]} \frac{1}{D^i_{train}} \sum_{d \in [D^i_{train}]} C_{id}(\pi^{\mathrm{Thr}})$.        ▷ Fixed threshold
```

applicable to our setting. Therefore, we propose OPT+ with several adaptations to build on top of OPT, including personalized policy application, budget soft constraint, and validation process. We elaborate on each of the three aspects below.

Constrained Personalized Policy. First, the construction of the tree does not support our constrained personalization setting. Since the policy also affects the feasibility constraints at each time step, the feasibility of sending a message at each time step cannot be encoded prior to training the tree. Thus, at time steps when it is not feasible to send a message due to mandatory practical constraints described in Section 4.3.3, for example after a recent treatment within the past two hours, OPT might still recommend a message due to a sizable estimated benefit in the outcome. Moreover, by construction, each leaf of the tree recommends the option that works on average the best among all data points in the leaf. While this intelligently pools user decision points into buckets, the derived action might not give the best estimated outcome for each particular user time step. To address these problems, we propose Algorithm 4. We dynamically update the feasibility indicator F_{it} and enforce the associated constraints in (4.1). At a feasible time step, OPT+ sends a treatment if both the threshold-based decision rule from Algorithm 2 and OPT recommend a treatment. This also leads to a balance between pooling among data in each leaf of OPT and ensuring the particular personal treatment effect up to the threshold. Compared to the threshold method, OPT+ is designed to be more restrictive in sending a message and do so with interpretability.

Algorithm 4 Optimal Policy Trees + policy

```
Given OPT tree, budget $N$, learned threshold $\Gamma_i \in \mathbb{R}$ for user $i$, initiate at the beginning of day $d$, initiate
for $t \in [T_{id}]$ do
    Apply OPT to obtain $\hat{\pi}^{\text{OPT}}(S_{it})$.
    Apply Algorithm 2 to obtain $\pi^{\text{Thr}}(S_{it})$ with $\Gamma_i$.
    if $\pi^{\text{Thr}}(S_{it}) = 1$ and $\hat{\pi}^{\text{OPT}}(S_{it}) = 1$ and $F_{it} = 1$ then
        $\pi^{\text{OPT}}(S_{it}) \leftarrow 1$.                ▷ Send message given consensus
    else
        $\pi^{\text{OPT}}(S_{it}) \leftarrow 0$.
    end if
end for
```

Budget Soft Constraint. Second, the training of OPT does not support direct incorporation of the budget constraint. Even though we incorporate the budget to Algorithm 4 as a hard constraint, it might not be optimal to train an OPT indifferent between sending and not sending a message given the same estimated outcome. To address this issue, we adapt the tree training process to incorporate the budget as a soft constraint by adding penalization of the outcome when training OPT. A parameter $\lambda \in \mathbb{R}^+$ is introduced to control the trade-off between the outcome and the number of messages sent. We define the penalized outcome by adding λ in the case of a message sent:

$$\hat{y}_{it}^{\text{fixed}}(x_{it}, A_{it}) = \hat{y}_{it}(x_{it}) + \lambda \cdot A_{it},$$

and train OPT with $x, A^{\text{hist}}, \hat{y}^{\text{fixed}}(x, A^{\text{hist}})$. To personalize the budget penalization, we also consider learning a user-specific λ_i. We utilize the personalized thresholds Γ_i learned from Algorithm 3 in the case of $N = 1$, with the intuition that Γ_i captures the relative unit outcome changes between users. We then define

$$\lambda_i = \Gamma_i \cdot \alpha, \quad \hat{y}_{it}^{\text{personalized}}(x_{it}, A_{it}) = \hat{y}_{it}(x_{it}) + \lambda_i \cdot A_{it},$$

where $0 < \alpha < 1$ is a discount factor to further tune the trade off between outcome and budget. Finally we train OPT with $x, A^{\text{hist}}, \hat{y}^{\text{personalized}}(x, A^{\text{hist}})$. This leads to two variations of the algorithm with $\hat{y}^{\text{fixed}}$ and $\hat{y}^{\text{personalized}}$.

Extended Validation Procedure. Third, the validation procedure of OPT is designed to optimize for the sum of outcomes among all user time steps given the policy tree. This is not sufficient for us, since our algorithm contains other policy components and targets user daily metrics. Thus, we adapt the validation process to select the best parameter combination. Let D^i_{val} denote the number of validation days for user i, and I_{val} denotes the number of users in the validation set. For each variation, we apply Algorithm 5 to tune and select the parameters of the following types and ranges. For personalized penalization of the budget constraint for tree training, we consider a list of α values ranging from 0 to 0.005 with 0.001 apart; for fixed penalization, we consider a list of λ values ranging from 0 to 3 with 0.5 apart. We consider OPT of maximum depth 5 and depth 7 and automatically tune the complexity parameter of the tree. In addition, since the best threshold from the stand-alone Algorithm 3 might not be the best for Algorithm 5, we retune the thresholds with a warm start selection by using a list of threshold values selected for at least one user from Algorithm 3.

Algorithm 5 Optimal Policy Trees + policy learning

```
Given personalized penalization option, initiate a set of   $\alpha$ or $\lambda$, OPT maximum depths, and threshold parameters $\Gamma$ , initiate
for each personalized penalization option, OPT depth and $\alpha$ or $\lambda$ do
    Train OPT with $x, A^{\text{hist}}, \hat{y}^{\text{fixed}}\ (x, A^{\text{hist}})$ or $\hat{y}^{\text{personalized}}(x, A^{\text{hist}})$.
    for each parameter $\Gamma$ do
        for $i \in [I_{val}], d \in [D^i_{val}]$ do
            Apply Algorithm 4 to obtain policy $\pi^{\text{OPT}}(S_{it})$ for parameter combination for all $t \in [T_{id}]$.
            Obtain metrics $C_{id}(\pi^{\text{OPT}})$ and $N_{id}(\pi^{\text{OPT}})$.
        end for
    end for
end for
Select the parameter combination which gives the best average daily metrics among validation set users.
```

4.6 Computational Results

We conduct experiments to demonstrate and compare the benefits of our method. We split the data into train, validation, and test sets as described in Section 4.4.1. We apply the methods in Section 4.5 and evaluate on the testing set. For each user on each day, we compute the two metrics defined in Section 4.3.3: sedentary time and message count. We evaluate the performance of both the threshold-based decision rule as a stand-alone policy and our integrated proposed method, compared with the current policy. We further present the output Optimal Policy Tree + for interpretation and stability analysis. The computational experiments are conducted in Julia programming language on a Macbook Pro with 2.3 GHz Quad-Core Intel Core i7 processor.

Comparison of Methods. For the threshold-based decision rule as a stand-alone policy, we implement 4 variations of the policy, with a budget of 1 and 2 messages and with personalization turned on and off. For the Optimal Policy Trees + policy, we implement 2 variations with a budget of 2 messages and personalized penalization turned on and off. For each variation of each method, we validate the best selected parameters, and then compute the two metrics defined in Section 4.3.3: the number of minutes in sedentary time and the number of messages sent, for each user day in the testing set. Since each user has a different number of available days in the testing set, to give each user equal weight, we compute an average daily metric among the available days for each user. Given the metrics per user, we then compute the mean from (4.3) and (4.4) and the standard deviation of metrics among the users for each method. The results for each method are summarized in Fig. 4-2, where each scatter point represents the mean metric with 95% confidence interval bars. The plot demonstrates the resulting trade-offs between the sedentary time and message count in the testing set.

For an average user, the current policy results in 148.02 minutes of daily sedentary time with 1.62 messages sent. In comparison to the current policy, significant improvements in both metrics are achievable from all variations of threshold-based

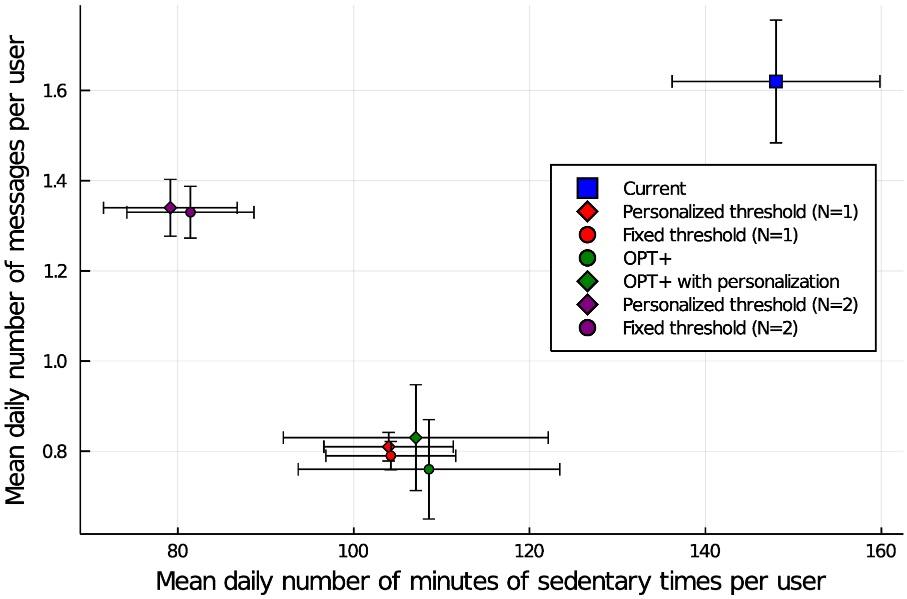


Figure 4-2: Method Comparison (Average ±95% Confidence Intervals).

decision rule and OPT+. For an average user, the stand-alone fixed threshold-based decision rule with budget 1 reduces the average daily sedentary time to 103.98 minutes with 0.79 messages, and the OPT-based policy with a fixed threshold reduces the average daily sedentary time to 108.56 minutes with 0.76 messages. These two variations have comparable performance, whereas OPT+ has the edge of interpretability. As another option, the threshold method with a budget of 2 can reduce sedentary times more if the decision maker is willing to achieve higher improvement by sending more messages than the other variations. We note that OPT+ is designed to be more restrictive about sending a message (since both OPT and the threshold need to agree to send) and provide more interpretability. We observe that personalization does not bring a benefit over fixed thresholds.

To further compare the OPT+ and threshold methods which send similar numbers of messages, we show in Fig. 4-3 the number of returning and new users for whom each method gives the least sedentary time in the testing set. The majority of users achieve the best sedentary time reduction from the fixed threshold and OPT+ methods, with similar portions between the two among both returning and new users.

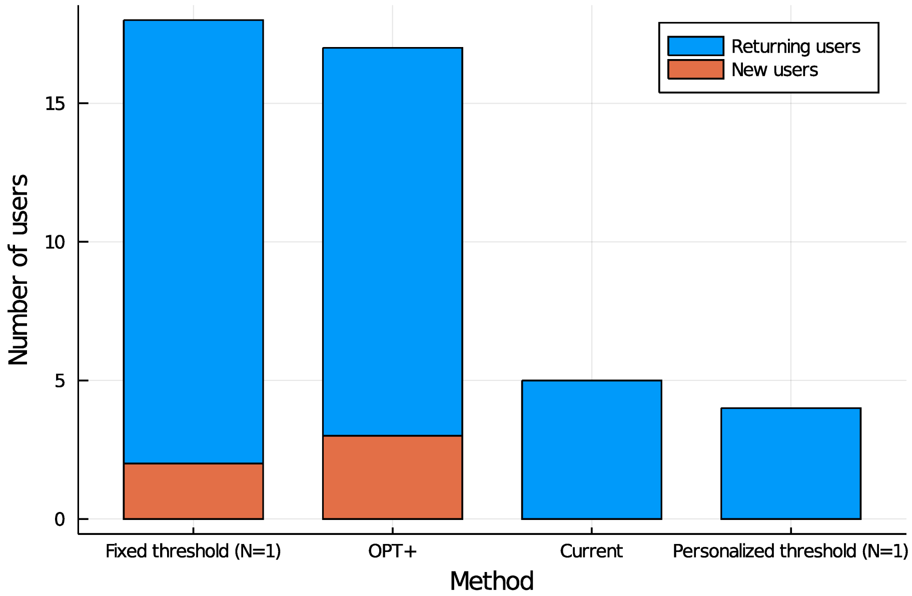


Figure 4-3: Number of Users whom Each Method Gives Least Sedentary Time on.

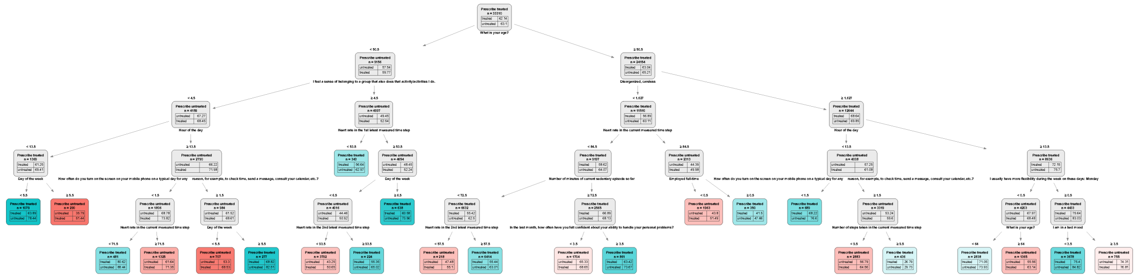


Figure 4-4: Optimal Policy Trees.

Output Optimal Policy Tree. The final Optimal Policy Trees is shown in Fig. 4-4 to be applied for users in the testing set. The tree partitions users and time steps into different subgroups with estimated heterogeneous treatment effects between them. For each user time step, we follow the logic of the decision tree based on user and contextual information to reach to a leaf node, similar to other decision trees. At each leaf node, we prescribe one of the actions: treated (sending a message) or untreated (not sending a message). The intensity of the red or green coloring of the leaves is proportional to the estimated outcome difference; the stronger the shade of the color, the more confident our recommendation is. Each node of the tree presents summary statistics among training data points that fall under the node, where the sample size of each budget is denoted in “$n =$”. It also shows a table of average estimated (penalized) outcomes under each treatment, which results in the prescribed treatment of each node.

Stability of the Optimal Policy Trees. Moreover, we analyze the stability of the Optimal Policy Trees, namely how the tree output changes with respect to the randomness and parameters involved in the tree training process. The stability issue is especially important since the output tree is critical in providing interpretability and insights for mHealth applications, and a stable tree can enhance the decision makers' trust in the algorithm. The final tree uses 15 variables in the splits. To investigate the stability of the tree, we train 99 additional alternative trees with slightly different parameters and compare if similar variables are identified by the alternative trees. Table 4.2 shows the proportion of trees among the 100 which identified each of the 15 variables. The top 7 variables appear in at least half of these 100 trees, which supports the stability of these variables and confidence in interpretation and guidance for scientific study. Stability analysis that compares the importance value (in addition to occurrence in the trees) of the variables in the 100 trees shows similar results, which we omit here for brevity.

Overall, we recommend using the OPT+ policy. Our proposed method is beneficial for sedentary users by sending a limited number of messages. It provides stable interpretability to decision makers without sacrificing performance, facilitating the use of the policy in real life.

4.7 Discussions and Future Work

In this section, we discuss how our overall work draws recommendations for the HeartSteps study. We further discuss the limitations of our study and directions for future work.

Table 4.2: Stability of Variables Used in OPT.

Variable	Proportion
Day of the week	100%
Heart rate in the current measured time step	89%
Hour of the day	81%
Heart rate in the 1st latest measured time step	74%
What is your age?	69%
Number of minutes of current sedentary episode so far	59%
Disorganized; careless	50%
Sense of belonging to group that do same activities	43%
Number of steps taken in the current measured time step	29%
Heart rate in the 2nd latest measured time step	18%
Daily frequency of turning on phone screen	14%
Employed full-time	10%
I am in a bad mood	9%
Confident to handle personal problems	4%
More flexibility on Monday	4%

4.7.1 Recommendations for HeartSteps

As initial evidence for the value of interpretable mHealth policies, the Optimal Policy Tree shown in Fig. 4-4 yields several insights that validate the usefulness of our approach and can inform future research on HeartSteps.

Concordance with Previous Research Findings. Prior research has found that individuals' sedentary time—as well as their ability to be active—is closely tied to their daily routines, such as meal times, and time spent at work. This connection of sedentary behavior and daily routines is reflected in the variables frequently identified by the OPT-based policy, day of the week and hour of the day, providing initial evidence for its ecological validity and supporting the intuition that if users have more flexibility during a part of the day (e.g., having fewer meetings), they are more able to get up and move when they receive an anti-sedentary message. Other studies found similar patterns (Bidargaddi et al. 2018, Klasnja et al. 2019, Fukuoka et al. 2018), providing a further ecological-validity check for the tree.

Informing New Theorizing. Besides reflecting known behavioral patterns, the Optimal Policy Tree also uncovered a number of unexpected moderators of intervention response. Understanding why these variables may matter requires additional theorizing and could advance our understanding of the processes that underlie behaviors and behavior changes. By uncovering such factors, the tree provides a data-driven and transparent tool to guide theorizing, which is especially useful at fast time scales (minute-, hour-, and day-level) for which we do not have existing behavioral theories. Among the interesting results unearthed by the tree are the following:

1. Among the variables that describe a user's current state (as opposed to more stable characteristics), heart rate played a significant role in 2 of the 7 most commonly found features. Given that all of the data used by the tree are for available time points only (i.e., a person has been sitting for at least 40 minutes), the heart rate is not related to the current physical activity. Since it is known that heart rate correlates with stress, we propose one possible hypothesis: that people are more likely to respond to an anti-sedentary message and get up and move around when they are feeling stressed. Physical activity is known to help with stress management, so it is possible that the prospect of getting up and moving away from the current task is more appealing when the task is stressful and when the person needs some time to regroup. In future studies, we can test this hypothesis by including intraday measures of stress and analyzing whether reported or detected stress levels moderate the response to anti-sedentary messages.

2. In addition to time-varying contextual features, the policy contains a number of variables about stable user characteristics, obtained from baseline surveys, that do not strictly relate to physical activity. These include user age and personality characteristics such as low levels of conscientiousness. While most past studies have focused on learning a policy from the data about the current context (e.g., the number of minutes of sedentary time up to that point), this finding suggests that user trait information can also be useful for forming policies.

Further group analyses will need to be done for user groups who are similar on personality-related trait measures as well as physical activity-related traits (e.g., self-efficacy for activity and commitment to being active) to discover what traits may matter for other types of interventions. In addition, if such trait characteristics are repeatedly found to influence the response to momentary interventions, additional theorizing will be needed to understand how and why those trait matter for moment-to-moment decisions about whether to be active.

Future Improvements for HeartSteps. Finally, our results suggest potential improvements for future versions of the HeartSteps intervention itself.

1. Given the promising computational results, the policy can be considered for use in the next version of the intervention. In particular, people with certain characteristics and routines are more inclined to be responsive to anti-sedentary messages when given our policy, so if these characteristics could be obtained during the study intake, they could be used to selectively provide this policy to individuals who are most likely to benefit from it.

2. On the other hand, user groups with certain traits do not benefit as much from the current treatment regime. This finding suggests that other policies may need to be developed for these user groups or that they may require a different type of policy altogether.

3. Lastly, since heart rate is a key predictive feature despite sizable amounts of missing values, this can encourage future efforts to collect more complete and accurate heart rate data to further improve outcomes.

4.7.2 Limitations and Future Work

The current work has several limitations that can guide future work. We were only able to use data (with partially known outcomes) from a subset of 51 users out of 94 total HeartSteps V2/V3 participants, which could create possible bias. More adequate data would enhance our validation scope. For policy learning, our method constructs a deterministic policy that makes binary decisions for intervention provision, whereas extending the method to send probabilistic treatments could facilitate off-policy analyses after the use of our policy. One possibility is to adapt the current binary voting regime into stochastic treatments with voting probabilities among multiple OPTs and thresholds, although the ensembling procedure could sacrifice the extent of interpretability. Finally, our work provides a proof-of-concept and demonstrates benefits from the past data from HeartSteps. This suggests the potential benefit if we were to implement our method in future HeartSteps studies online. Although our framework is developed and tested in a batch setting, it has the potential to be extended for use in an online setting. With online data collection from returning and new users, we can use our pre-learned policy as a warm start, and retrain the algorithms and trees regularly (e.g., daily or weekly) to incorporate more information. If such online extensions were found to be stable, their incorporation in the next version of the HeartSteps intervention may both improve its effectiveness and generate further data that can guide theorizing about behavior-change processes.

4.8 Conclusions

In this chapter, we develop and test two innovative batch off-policy learning methods for personalized mobile health applications. The first method learns thresholds and can be used as a stand-alone policy. Optimal Policy Trees + incorporates the learned threshold and adapts OPT to a budget-constrained setting. Experiments on HeartSteps V2/V3 data demonstrate significant improvement in the effectiveness of anti-sedentary messages with lower budgets using our methods as compared to the current practice. In particular, OPT+ identifies important variables with interpretability and stability without sacrificing performance. Insights uncovered from the decision tree can guide new theorizing on the factors that influence behavior-change processes at the intraday level, thus advancing our understanding of the dynamics of human health behavior.

Chapter 5

Conclusions

In this thesis, we improve healthcare operations from models to implementation. The main contributions of this thesis are two-fold: 1) We develop methodologies from optimization, machine learning, and policy learning to support strategic, tactical, and operational decision making for multiple healthcare systems. 2) We extensively deploy software in daily production at a comprehensive scale in a representative large hospital network. Not only do our models achieve state-of-the-art performances, but their implementation also makes tangible impacts in practice.

The two-phase optimization model developed in Chapter 2 for nurse staffing at the Emergency Department results in a less costly schedule with improved patient coverage and higher nurse satisfaction, while the benefits are flexible to adapt to changes over time. Meanwhile, the machine learning models developed in Chapter 3 enable accurate prediction of various patient operational characteristics for all inpatients across seven hospitals. Moreover, the interpretable batch off-policy learning methods developed in Chapter 4 reduce sedentary behavior more effectively for an mHealth app.

Our end-to-end software, implemented in Chapters 2 and 3, leverages optimization and machine learning to support systematic resource allocation and patient evaluation for over 400 direct users and more stakeholders at Hartford Healthcare. The adoption of our tool results in cost savings, better patient care coverage, and increased nurse satisfaction, reducing the average patient length of stay by 0.67 days and enhancing patient safety while reducing manual labor of healthcare workers. Our tool's successful implementation at HHC led to a projected \$36.62 million annual revenue uplift, including \$0.73 million cost saving from nurse staffing and \$35.89 million contribution increase from LOS reduction. As the deployment continues to expand in more hospitals and units, we expect to generate even more substantial financial benefits in the upcoming years.

Overall, the thesis demonstrates the potential of transforming the future of healthcare operations with models and implementation. From our collaboration with healthcare stakeholders throughout the development and deployment, we gained and shared insights on our first-hand challenges and successful takeaways, which we hope can enlighten future efforts in connecting the bridge further between research and practice.

Looking ahead, we envision the future expansion of the Holistic Hospital Optimization (H2O) framework to encompass four key elements: 1) Holistic data: Utilizing multiple input modalities (e.g., image, language, tabular, and time series data) to solve healthcare problems, as demonstrated in the unified Holistic AI in Medicine (HAIM) framework (Soenksen et al. 2022). 2) Holistic models: Leveraging the interactions of different methodologies, such as incorporating large language models to enhance patient outcome machine learning models in production (Villalobos Carballo et al. 2022). 3) Holistic decision making: Integrating resource allocation, patient outcome prediction, and digital intervention into a unified system (Bertsimas et al. 2023b), which aims to turn the imagination of the patient journey from the thesis introduction into an anecdote. 4) Holistic implementation: Expanding the successful deployment from a multi-center, representative U.S. hospital network to other healthcare systems worldwide.

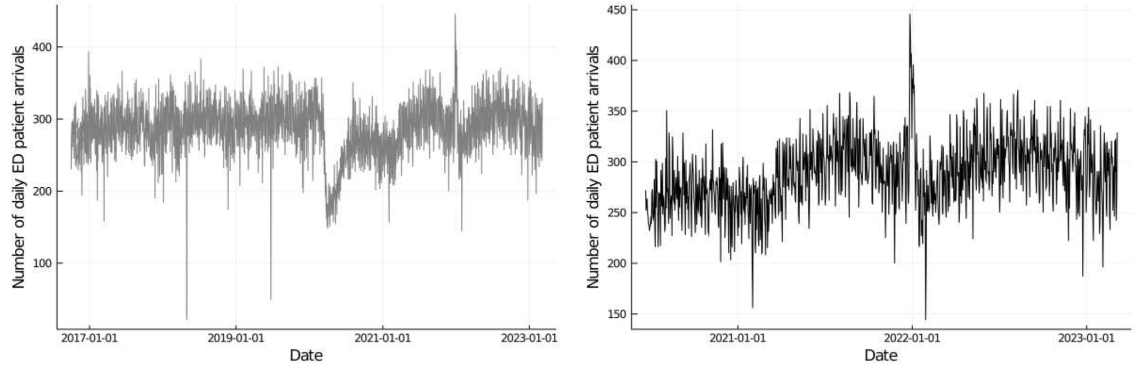
Appendix A

Appendix to Chapter 2

A.1 Demand Trends and Patterns

Figure A-1a displays the number of ED patient arrivals for each day since late 2016. Before March 2020, the ED volume was about 250-350 patient arrivals per day with small fluctuations. Since the beginning of COVID-19, the number of arrivals to the ED dropped significantly and then kept growing back through 2021 to a level slightly lower than before, with 200–300 arrivals per day. As the pandemic progressed through different stages, people were recovering from the COVID-19 isolation period, which brought the ED volume back to 250–350 arrivals per day, while having larger fluctuations over time compared with pre-COVID-19. To reduce noise from the pre-COVID-19 period as well as the time when the demand significantly dropped right after COVID-19, we use data starting June 22, 2020 for all analysis, which we highlight in Figure A-1b. These trends demonstrate the varying ED demand over time and suggest potential differences in the benefits of our approach to the ED during periods of low versus high demand.

For each pod type, Figure A-2 shows patterns for the number of patients in each pod type at each hour of the day (midnight to midnight) and for each day of the week (Monday to Sunday). In each of the plots, the solid curves represent the average value with error bars as standard deviations over the 27-week periods from June 22 to December 6, 2020. For main pods, we have a much higher demand in the afternoons and



(a) From November 7, 2016 to March 6, 2023. (b) From June 22, 2020 to March 6, 2023.

Figure A-1: ED Historical Volume Over Time.

evenings. Demand at these busier times often exceeds room capacity. During these periods, we observe higher demand on Mondays to Wednesdays, lower on Thursdays and Fridays, and further lower demand on weekends. In general, demand in the main pods has the highest variability among other pods. For the red pod, we have on average 1-2 patients and higher demand in the afternoons and evenings. It has much higher demand on Friday compared with other days of the week. For the purple pod, we have slightly higher demand during the nights and slightly higher demand on weekdays than on weekends. For iTrack, we also observe patterns of higher demand in afternoons, evenings, and weekdays. These variabilities and patterns suggest the possibility of adjusting the staffing corresponding to the demand to have better operations at the ED.

A.2 Schedule Illustration

In this Appendix section, we provide a more detailed illustration of the recommended schedule for November 2 - December 13, 2020 obtained from robust (a) approach from Section 2.4.1, including the relationship with patient demand patterns, insufficiency breakdown by pod types, and a complete recommended schedule.

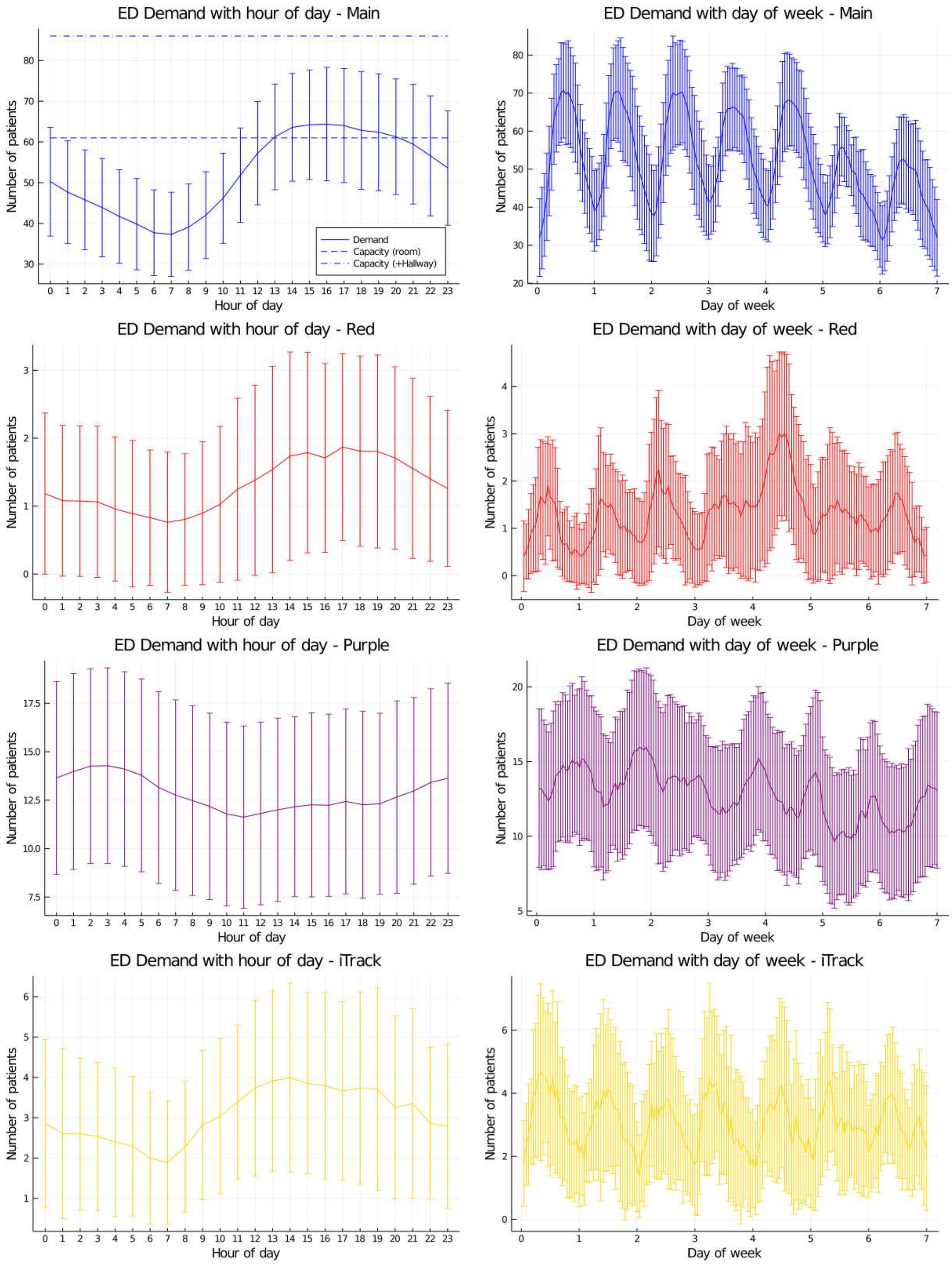
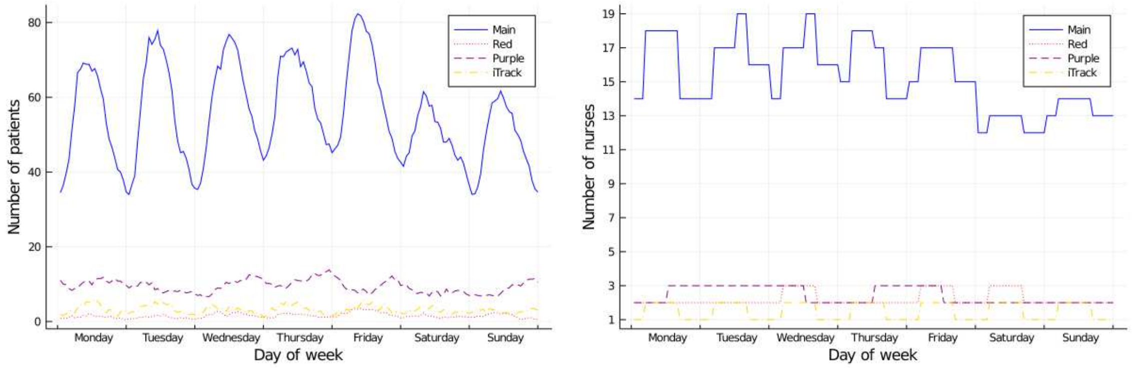


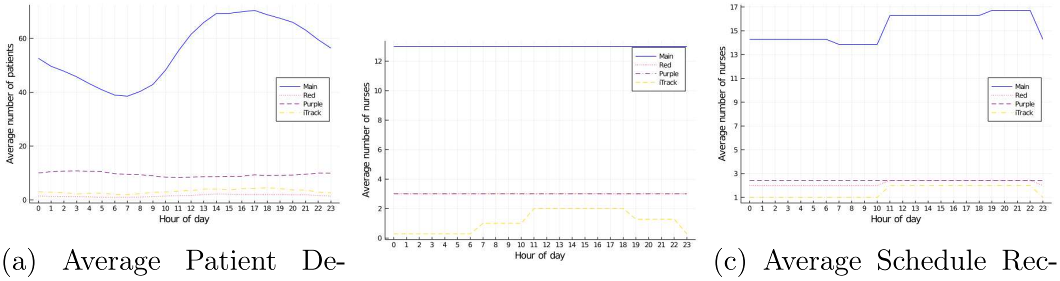
Figure A-2: Hour-of-day and Day-of week Demand Patterns by Pod Type (Average ± Standard Deviations from June 22 to December 6, 2020).



(a) Average Patient Demand.

(b) Nurse Schedule Recommendation.

Figure A-3: Day-of-week Patterns of Demand and Staffing.



(a) Average Patient Demand.

(b) Current Daily Schedule.

(c) Average Schedule Recommendation.

Figure A-4: Hour-of-day Patterns of Demand and Staffing.

Patterns of demand vs staffing. We illustrate that our recommendation matches staffing with demand patterns and results in a more cost-effective schedule. In this period, the average number of patient stays over 6 weeks and the number of nurses recommended to staff each week in each pod type from Monday to Sunday (starting from 7 am to 7 am each day) are shown in Figure A-3a and Figure A-3b. For the main pods and the purple pod, as patient demands tend to be higher on weekdays than weekends, we staff more nurses on weekdays than weekends. For the red pod, we have a particularly higher demand on Friday when we also staff more nurses. Both demand and staffing for iTrack have less variability throughout the days of the week.

For intra-day patterns, we show the average number of patients over 6 weeks in each pod type at each hour of the day (from hours starting at 0:00 to 23:00) in Figure A-4a. For main, red, and iTrack pods, we have higher demands in the afternoons and evenings, where the variability is highest for main pods. For purple pod, demand is more constant throughout the day with slightly higher demand during nighttime. Currently, every day applies the schedule shown in Figure A-4b for the number of nurses working in each position at each hour of the day. Only staffing at iTrack adapts to the variability in demand with more staffing in afternoons and evenings and less at night. However, in red and main pods where we also observe variabilities in demand (especially in main pods), staffing remains constant throughout the day and week. In contrast, the recommended schedule shown in Figure A-4c is consistent with the demand patterns by having smoother staffing in the purple pod and more staffing from 11 am–11 pm for other pods. Even though the schedule is generated with information only prior to the period, it can capture most of the day-of-week and hour-of-day demand patterns in the period prospectively, which justifies the robustness of our approach. By matching staffing with demand patterns, we reduce the staffing cost by 7.40%.

Recommended schedule. We show the complete recommended schedule for the period in Table A.1, where we specify the number of nurses assigned to each of the three shifts at each of the 8 positions on each day every week. We note that the position configuration was different in 2020 from now. We break the main pods down into blue, green, and orange pods. This aggregate schedule can lead to multiple feasible schedules for individual nurse assignments, which is flexible to incorporate nurse types and individual nurse preferences for final schedule implementation.

Table A.1: Recommended Schedule (Shift 1: 7a-7p, 2: 11a-11p, 3: 7p-7a).

Position	CNL			Triage			Blue			Green			Orange			Red			Purple			iTrack		
Shift	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Mon	2	0	2	2	1	2	5	1	5	4	2	4	5	1	5	2	0	2	2	0	3	1	1	1
Tue	2	0	2	2	1	2	5	1	5	4	1	5	5	1	6	2	0	2	3	0	3	1	1	1
Wed	2	0	2	2	1	2	5	1	5	4	1	5	5	1	6	2	1	2	3	0	2	1	1	1
Thu	2	0	2	2	1	2	5	1	5	5	1	4	5	1	5	2	0	2	2	0	3	1	1	1
Fri	2	0	2	2	1	2	5	1	5	5	0	5	5	1	5	2	1	2	3	0	2	1	1	1
Sat	2	0	2	2	1	2	4	0	4	4	0	4	4	1	4	2	1	2	2	0	2	1	1	1
Sun	2	0	2	2	1	2	4	1	4	4	0	4	5	0	5	2	0	2	2	0	2	1	1	1

Appendix B

Appendix to Chapter 3

B.1 Feature Processing

In this Appendix section, we elaborate the feature processing of the data, and in particular, changes and efforts taken to expand models from one hospital to seven hospitals.

Full List of Curated Features. We create various features divided into 6 groups.

1) Current conditions:

   - Information extracted from admission, discharge, transfer orders, and events (e.g., department, service, whether in ICU).
   - Current status such as DNR and Nothing by Mouth (NPO, meaning inability to eat or drink).
   - Others such as dialysis and oxygen (O2) device.

2) Lab results (e.g., albumin, white blood cell count):

   - Latest measurements.
   - Delta variables (difference of the current day's lab result from the previous day's).

- Normal range indicators (1 if the lab value is within the normal range, 2 if outside the normal range, and 0 if the lab value is missing).
- Distance between the lab value and its normal range, i.e., min(value-lower bound, 0)+max(value-upper bound, 0).
- Number of abnormal lab results in the past 24 hours.

3) Clinical measurements (e.g., temperature, respiratory rate, heart rate):

    - Latest measurement.
    - Highest measurement in the past 24 hours.
    - Average value in the past 24 hours.
    - Critical indicators (whether or not the value is critical with respect to the critical range provided by doctors).

4) Time series summary statistics of operational variables:

    - Number of days in the ICU and in hospitalization since admission.
    - Number of days until the next scheduled surgery, the number of days since the last surgery, and total time spent in surgery since admission.
    - Pending results (whether or not MRI/CT/ECHO/etc. is pending, and the number of labs pending).
    - Numbers of notes, orders, and medications in the last 24 hours and since admission.
    - Number of attending physicians in the last 24 hours.

5) Patient information prior to current admission:

    - Age at admission and patient type.
    - Number of days since the previous admission, and LOS of the previous admission.

6) Auxiliary operational variables which are not patient-specific:

- Current date-related variables such as day of the week and weekend indicator.
- Ward-related statistics such as census, utilization, and daily discharges.
- Hospital-related statistics (e.g. number of hospital admissions in the past 24 hours).

Missing Feature Imputation. Like most clinical data, our data extracts contain a large portion of missing data. We adopt different imputation techniques to deal with different types of missing features. First, with the hospital's help, we leverage information about how the data are collected, recorded, and computed to develop a rule that deterministically imputes a subset of variables. For some binary variables in group 1), a missing value indicates that the patient does not have this current condition. For example, DNR is missing when the patient did not sign a DNR, NPO is missing when the patient does not have an NPO constraint, and IV is missing when the patient is not on IV. Thus we fill the categorical values accordingly for such binary variables. For some multi-class categorical variables in group 1), missing values are deemed as a separate category indicating that it is missing. For example, missing O2 device, and diagnosis are imputed as "NA Value" to indicate the patient has no O2 device and has no diagnosis. For the range columns of lab results in group 2) (e.g. normal range for bilirubin), we first extract the non-missing record for each admission event and backfill on all dates for this admission event. Here we assume that the normal range column has unique input for each admission event. If all records for the admission event are missing, we fill the entries with the most frequent category of this range column. For clinical measurements in group 3), we capture cases when a value is not measured, which could reflect additional information, by imputing the missing value with a special value such as -1. For most summary statistics of operational variables in group 4, we fill the count with 0 if there is no record found. For example, if there is no pending lab and no medications in the record, then the count of pending labs and the number of medications are computed as 0. For some other counting variables in groups 4) and 5), imputing with 0 would confuse with a count of 0 due to their meanings. For example, the number of days until the next scheduled surgery is 0 if the future surgery date is the current day. Thus we impute a missing future surgery date, which means the patient does not have surgery scheduled, with the value -1 instead of 0. For the number of days since the last surgery and since the previous admission, no record of the last surgery and admission is approximated with a long time ago by imputing with 9999. After our clinical rule-based imputation step, other variables such as laboratory results and vital devices also have a high missing percentage. We drop columns with more than 50% missing values in the dataset and then impute missing values of the remaining features using the OptImpute (Bertsimas et al. 2018), an imputation method based on optimization and k nearest neighbors (Peterson 2009). We note that for features in group 3), we impute the original lab results and clinical measurements before computing the delta and distance to the normal range.

String Parsing. Several columns in the data extracts require string parsing to obtain numerical values. The normal range features are stored in the format of "lower bound - upper bound", ">upper bound", or "<lower bound" as strings. Hence, we first split these strings into two float formatted numbers as upper bound and lower bound features. Several variables such as the last RASS score and pain score take string formats containing both the numerical score and an explanation of the score, such as "0 → alert and calm", "-4 → deep sedation", "10 → hurts worst", etc. We parse such string columns to extract only the numerical score as a continuous variable.

Categorical Variable Encoding. The feature space contains several categorical variables and requires encoding before passing into the imputation and modeling process. Some features must be encoded in a specific way, for example, the feature DNR is encoded as 1 if the entry is "DNR (DO NOT RESUSCITATE)", and 0 otherwise. For other categorical variables, such as current department, current service, and oxygen device, we use a label encoder to encode each column separately.

Differences between hospitals. Since the electronic medical records of the seven hospitals are unified in HHC, most variables have consistent forms across hospitals and thus did not require additional processing to scale from one hospital to others. However, each hospital has some different conventions and some variables required specific processing for each hospital. For example, hospitals have different ways of naming departments and levels of care. Departments of intensive care level are named ICU in HH, CH, and MMC, named CCU in BH, and Critical Care in HOCC, WH does not have an ICU, and SV have specific names for ICU units. These differences were identified by inspection of the data and consultation with the hospital network. Thus we modified the way to compute ICU-related variables such as whether or not a patient is in the ICU and the number of days in the ICU, as well as to compute the ICU-related prediction targets for each hospital differently. The hospital is set to be a parameter to address the differences in the feature processing part. Another major difference is that multi-class categorical variables can take very different sets of values from one hospital to another. This could be due to differences in both patient populations and in ways of recording data by different staff members. In previous models for HH only, some multi-class categorical variables, such as service, diagnosis, insurance, etc., were manually encoded based on their critical levels where similar values of variables are grouped together. For example, O2 devices are converted into numerical values ranging from 1 (most critical devices) to 7 (room air or no device). Such encoding was done with discussions with doctors at HH based on their clinical knowledge. However, since the seven hospitals have different categories, manual categorization would need to be done for each hospital individually. To make the process more efficient and scalable, we decided to replace such manual encoding with label encoding, where we have one encoder for each categorical variable for each hospital. The encoders are saved for each hospital for consistent use in production. Similarly, we train, save,

Table B.1: Summary of Data Size.

Hospital	BH	CH	HH	HOCC	MMC	SV	WH
# Patients	15,493	7,956	105,184	20,011	15,576	11,624	4,838
# Admissions	23,354	12,822	171,072	29,490	21,612	15,319	6,924
# patient days	106,662	52,931	879,357	139,542	90,924	79,615	26,184

and apply separate OptImpute imputers for each hospital.

B.2 Supplementary Results

In this Appendix section, we present additional results of the data and experiments.

B.2.1 Data Summary Statistics

We report here the summary statistics of the data after the process of inclusion, exclusion, and splits from the Machine Learning Modeling section. For each hospital, the number of patients, admissions, and patient days in the union of training, validation, and testing sets are summarized in Table B.1. In total, we use data from 180,682 patients, 280,593 admissions, and 1,375,215 patient days in HHC. The data sizes vary across hospitals, where HH, the hospital with the largest data size, has over 33 times of patient days than WH, the hospital with the smallest data size. After filtering per target, the numbers of data points, i.e., patient days of the remaining training, validation, and testing sets combined are shown in Table B.2. Mortality and discharge-related targets keep the majority of the data points, whereas ICU targets have fewer data points largely due to the split of patients in versus not in ICU, especially leaving ICU predictions have a small portion of data because a very small portion of patients are in ICU.

We also report the proportions observed of each prediction task outcome in the testing set. For mortality and discharge disposition, since the target outcome is the outcome at the end of the stay, we compute the proportions of patients for each of the three discharge disposition classes for each hospital in Table B.3. The proportions are similar between hospitals, except that BH and WH have a higher

Table B.2: Number of Data Points (Patient Days) for Each Prediction.

Hospital Prediction	BH	CH	HH	HOCC	MMC	SV	WH
Mortality	104,552	51,542	865,954	134,684	88,999	76,618	25,823
Discharge Disposition	104,552	51,542	865,954	134,684	88,999	76,618	25,823
Discharge 24hr	105,451	52,021	869,468	135,592	89,474	77,178	25,937
Discharge 48hr	105,225	51,851	868,563	135,297	89,320	77,032	25,888
Enter ICU 24 hr	75,585	34,378	592,264	92,700	63,876	59,701	No ICU
Leave ICU 24 hr	7,808	5,792	115,997	17,499	5,526	4,590	
Enter ICU 48 hr	56,679	24,166	454,803	69,097	47,491	47,569	
Leave ICU 48 hr	7,312	5,336	109,443	16,332	5,079	4,360	

Table B.3: Proportion of Patient Admissions for Each Discharge Disposition in Testing Set.

Outcome Class	BH	CH	HH	HOCC	MMC	SV	WH
Expired/hospice	6.08%	7.47%	5.63%	7.17%	6.31%	7.41%	5.12%
Home w/o service	60.44%	45.26%	50.04%	50.16%	52.51%	52.08%	61.97%
With service	33.48%	47.27%	44.33%	42.67%	41.18%	40.51%	32.91%

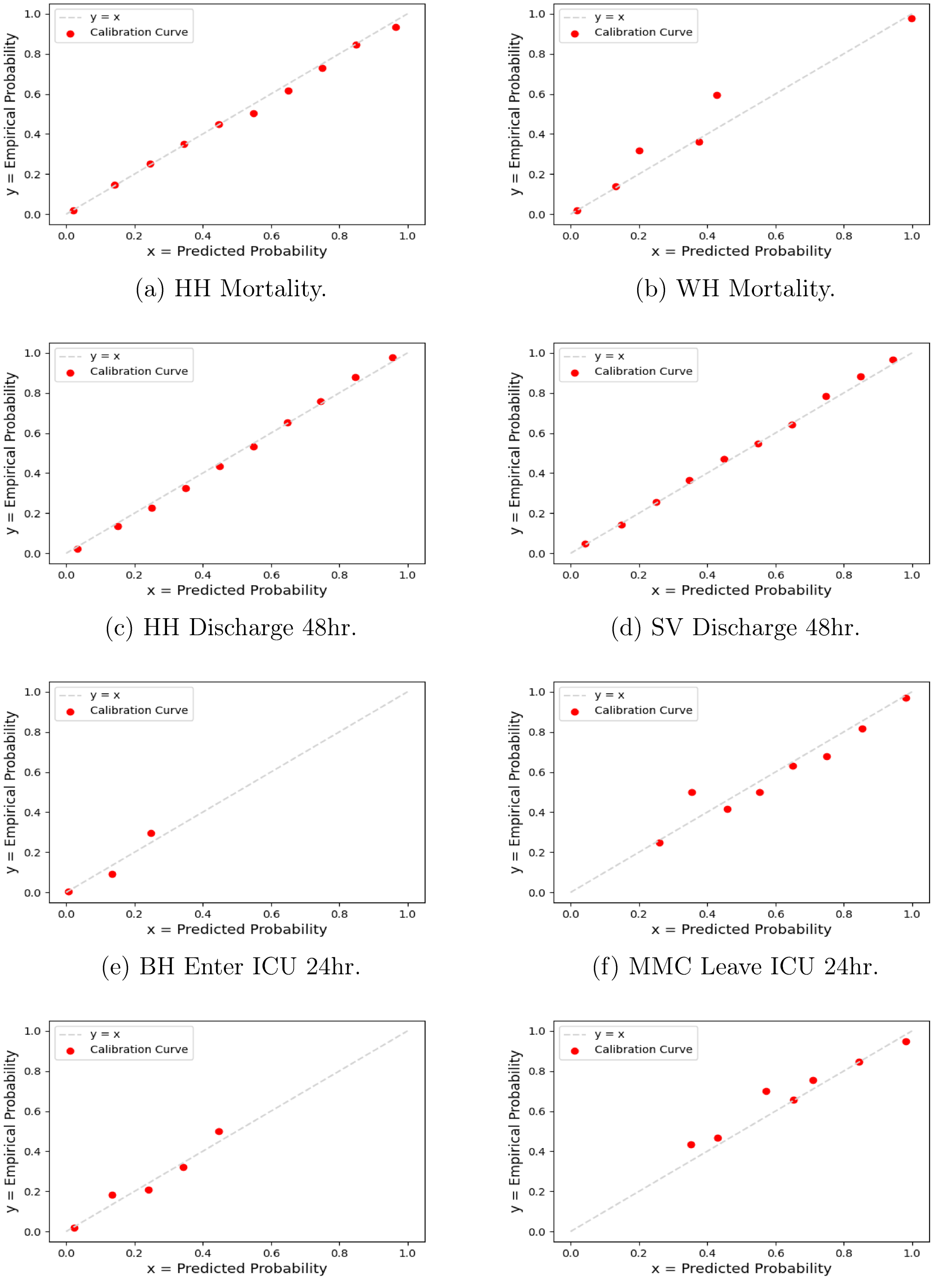
proportion of patients discharged to the home without service category compared to the other hospitals. For the discharge and ICU next 24-hr and 48-hr targets, since the outcome depends on the date, we compute the proportions of patient days that have a positive outcome for each prediction task in each hospital in Table B.4. Compared with the other six hospitals, WH has a significantly higher proportion of positive discharge 24-hr / 48-hr outcomes, as WH tends to treat less critical patients without the presence of an ICU. Entering ICU predictions have highly imbalanced classes, as less than 3% of patients enter the ICU in the next 24 and 48 hours. The proportion of patients leaving ICU range from 77.41% to 88.92% for 24 hours and from 64.19% to 80.60% between the six hospitals, likely due to the different composition of the patient population and congestion level in each hospital.

Table B.4: Proportion of Patient Days with Positive Target Outcomes in Testing Set.

Outcome	BH	CH	HH	HOCC	MMC	SV	WH
Discharge 24h	19.81%	20.56%	17.94%	18.61%	21.08%	17.44%	25.75%
Discharge 48h	36.54%	37.60%	32.95%	34.49%	38.05%	31.58%	46.12%
Enter ICU 24h	0.93%	1.35%	1.62%	1.53%	0.69%	0.90%	No ICU
Leave ICU 24h	83.34%	88.92%	81.04%	84.06%	85.69%	77.41%	
Enter ICU 48h	1.62%	2.52%	2.78%	2.48%	1.23%	1.46%	
Leave ICU 48h	73.16%	80.60%	69.89%	74.35%	76.40%	64.19%	

B.2.2 Assessment of Model Calibration

We evaluate the proper calibration of all our models on the second half of the testing set, which was not used during calibration. This evaluation is performed using calibration curves, which compare the probabilities predicted by the calibrated model vs the empirical probabilities of the data. Following DeGroot and Fienberg (1983), Niculescu-Mizil and Caruana (2005), we generate these curves by bucketing the probabilities predicted by the calibrated model into 10 uniform bins. Within each bin/bucket, we compare the empirical probability and the average predicted probability (or classification score). We then plot the resulting points (we only include points for bins with at least 10 observations, to reduce noise). As observed from sample calibration curves in Figure B-1, the points mostly fall near the diagonals, indicating that the final models are indeed well calibrated.



(a) HH Mortality.

(b) WH Mortality.

(c) HH Discharge 48hr.

(d) SV Discharge 48hr.

(e) BH Enter ICU 24hr.

(f) MMC Leave ICU 24hr.

(g) HOCC Enter ICU 48hr.

(h) CH Leave ICU 48hr.

Figure B-1: Calibration Curves for Predictions.

B.2.3 Out-of-Sample Evaluation by Department at Hartford Hospital

In this section, we evaluate the prediction models and the alerts for each of the departments at Hartford Hospital. In Table B.5 we present the out-of-sample AUCs for all prediction models. We then evaluate the alerts (green and red) for each of the departments at HH. In Tables B.6, B.7, and B.8, we report the out-of-sample accuracy, precision, and recall of the green and red alerts, respectively. In all tables, an NA entry in the table means that there is at most one class (or two classes in the case of discharge disposition) present in the specific department.

Table B.5: AUC Metrics by Department at Hartford Hospital.

Department	Mortality	Disch.	Disch.	Disch.	Enter	Leave	Enter	Leave
		Dispo.	24hr	48hr	ICU	ICU	ICU	ICU
					24hr	24hr	48hr	48hr
BLISS 11 STEP DOWN	0.824	0.779	0.882	0.852	0.769	NA	0.754	NA
NORTH 9 STEP DOWN	0.845	0.849	0.856	0.823	0.769	NA	0.738	NA
CONKLIN 3	0.934	0.850	0.791	0.781	0.814	NA	0.752	NA
BLISS 9 ICU	0.921	0.831	0.952	0.929	NA	0.878	NA	0.876
CONKLIN 2	0.905	0.871	0.790	0.782	0.793	NA	0.755	NA
CENTER 9 ICU	0.864	0.838	0.922	0.927	NA	0.868	NA	0.861
BLISS 8	0.888	0.884	0.831	0.834	0.817	NA	0.743	NA
BLISS 7 EAST	0.873	0.850	0.805	0.797	0.792	NA	0.755	NA
CONKLIN 4	0.888	0.851	0.781	0.772	0.761	NA	0.740	NA
HIGH 12	0.873	0.865	0.797	0.777	0.833	NA	0.752	NA
CONKLIN 5	0.872	0.852	0.778	0.761	0.788	NA	0.749	NA
BLISS 6 MATERNITY	NA	NA	0.914	0.938	0.471	NA	NA	NA
BONE AND JOINT 4	0.959	0.863	0.819	0.812	0.728	NA	0.728	NA
CENTER 10	0.865	0.811	0.783	0.762	0.857	NA	0.800	NA
BLISS 9 STEP DOWN	0.849	0.826	0.841	0.814	0.757	NA	0.733	NA
CENTER 8 ICU	0.811	0.794	0.910	0.916	NA	0.885	NA	0.880
BLISS 5 EAST	0.920	0.882	0.804	0.787	0.791	NA	0.711	NA
BLISS 9 EAST	0.829	0.781	0.780	0.765	0.831	NA	0.759	NA
EMERGENCY	0.848	0.796	0.782	0.746	0.825	NA	0.784	NA
BLISS 10 EAST	0.820	0.777	0.763	0.736	0.863	NA	0.819	NA
BLISS 10 STEP DOWN	0.852	0.801	0.871	0.859	0.767	NA	0.748	NA
BLISS 11 ICU	0.837	0.837	0.870	0.881	NA	0.902	NA	0.892
NORTH 9	0.842	0.779	0.791	0.791	0.838	NA	0.808	NA
BONE AND JOINT 5	NA	NA	0.801	0.802	0.391	NA	0.365	NA
NORTH 8	0.931	0.902	0.787	0.791	0.822	NA	0.755	NA
NORTH 11	0.883	0.867	0.785	0.777	0.839	NA	0.804	NA
BLISS 10 ICU	0.879	0.823	0.876	0.918	NA	0.865	NA	0.848
NORTH 10	0.841	0.794	0.781	0.757	0.905	NA	0.862	NA
BLISS 7 ICU	0.858	0.835	0.889	0.894	NA	0.847	NA	0.848
CENTER 11	0.990	0.941	0.787	0.831	0.912	NA	0.780	NA
BLISS 7 STEP DOWN	0.810	0.813	0.838	0.833	0.778	NA	0.755	NA
LABOR DELIVERY6	NA	NA	0.825	0.726	0.947	NA	0.470	NA
BLISS 6 NURSERY	NA	NA	0.718	0.742	NA	NA	NA	NA
EMERGENCY OBS	0.932	0.878	0.693	0.702	0.817	NA	0.807	NA
IR	0.957	0.907	0.779	0.839	0.858	NA	0.643	NA
MAIN PACU	0.826	0.908	0.804	0.920	0.934	NA	NA	NA
CONKLIN 1	0.895	0.849	0.687	0.661	0.988	NA	0.210	NA
MAIN OR	0.733	0.742	0.822	0.806	0.725	NA	0.729	NA
CARDIAC CATH LAB	0.722	0.587	0.930	0.667	0.826	NA	0.929	NA
BJI PERIOP SVC	0.953	0.818	0.969	0.814	0.872	NA	0.713	NA
EP LAB	NA	NA	NA	NA	NA	NA	NA	NA
GI ENDOSCOPY	NA	NA	NA	NA	NA	NA	NA	NA
BJI PRE/POST	NA	NA	NA	NA	NA	NA	NA	NA
CENTER 12	0.851	0.845	0.781	0.770	0.781	NA	0.740	NA
NORTH 12	0.881	0.848	0.780	0.774	0.824	NA	0.723	NA
BLISS NORTH 2 ICU	0.833	0.836	0.891	0.871	NA	0.736	NA	0.723
BLISS NORTH 3 ICU	0.826	0.724	0.899	0.928	NA	0.857	NA	0.851

Table B.6: Accuracy, Precision and Recall by Department for the Previous Green Alert at Hartford Hospital.

Department	Accuracy	Precision	Recall
BLISS 11 STEP DOWN	0.959	0.447	0.130
NORTH 9 STEP DOWN	0.883	0.526	0.172
CONKLIN 3	0.745	0.617	0.452
BLISS 9 ICU	0.969	0.542	0.152
CONKLIN 2	0.764	0.588	0.402
CENTER 9 ICU	0.968	0.745	0.154
BLISS 8	0.752	0.712	0.722
BLISS 7 EAST	0.729	0.628	0.585
CONKLIN 4	0.702	0.689	0.376
HIGH 12	0.715	0.673	0.472
CONKLIN 5	0.704	0.637	0.511
BLISS 6 MATERNITY	0.908	0.936	0.954
BONE AND JOINT 4	0.734	0.692	0.756
CENTER 10	0.705	0.633	0.568
BLISS 9 STEP DOWN	0.851	0.579	0.257
CENTER 8 ICU	0.978	0.600	0.045
BLISS 5 EAST	0.718	0.659	0.633
BLISS 9 EAST	0.694	0.697	0.670
EMERGENCY	0.810	0.499	0.170
BLISS 10 EAST	0.710	0.581	0.456
BLISS 10 STEP DOWN	0.858	0.696	0.378
BLISS 11 ICU	0.982	0.375	0.129
NORTH 9	0.733	0.651	0.576
BONE AND JOINT 5	0.886	0.907	0.969
NORTH 8	0.716	0.675	0.683
NORTH 11	0.723	0.641	0.543
BLISS 10 ICU	0.960	0.580	0.150
NORTH 10	0.711	0.604	0.528
BLISS 7 ICU	0.975	0.289	0.118
CENTER 11	0.749	0.682	0.765
BLISS 7 STEP DOWN	0.930	0.385	0.106
LABOR DELIVERY 6	0.738	0.680	0.337
BLISS 6 NURSERY	0.920	0.920	1.000
EMERGENCY OBS	0.647	0.693	0.629
IR	0.592	0.769	0.465
MAIN PACU	0.760	0.833	0.500
CONKLIN 1	0.660	0.418	0.509
MAIN OR	0.729	0.669	0.480
CARDIAC CATH LAB	0.682	0.714	0.500
BJI PERIOP SVC	0.750	0.700	0.467
CENTER 12	0.724	0.636	0.472
NORTH 12	0.725	0.662	0.462
BLISS NORTH 2 ICU	0.835	0.689	0.449
BLISS NORTH 3 ICU	0.965	0.650	0.197

Table B.7: Accuracy, Precision and Recall by Department for the New Green Alert at Hartford Hospital.

Department	Accuracy	Precision	Recall
BLISS 11 STEP DOWN	0.951	0.321	0.205
NORTH 9 STEP DOWN	0.877	0.478	0.325
CONKLIN 3	0.723	0.542	0.650
BLISS 9 ICU	0.967	0.457	0.200
CONKLIN 2	0.735	0.504	0.605
CENTER 9 ICU	0.967	0.611	0.248
BLISS 8	0.734	0.653	0.830
BLISS 7 EAST	0.713	0.573	0.758
CONKLIN 4	0.715	0.630	0.585
HIGH 12	0.706	0.596	0.682
CONKLIN 5	0.686	0.570	0.710
BLISS 6 MATERNITY	0.905	0.922	0.967
BONE AND JOINT 4	0.712	0.639	0.859
CENTER 10	0.676	0.562	0.747
BLISS 9 STEP DOWN	0.841	0.505	0.404
CENTER 8 ICU	0.977	0.500	0.104
BLISS 5 EAST	0.699	0.597	0.792
BLISS 9 EAST	0.685	0.644	0.805
EMERGENCY	0.787	0.422	0.328
BLISS 10 EAST	0.669	0.502	0.650
BLISS 10 STEP DOWN	0.850	0.591	0.547
BLISS 11 ICU	0.981	0.370	0.183
NORTH 9	0.705	0.573	0.750
BONE AND JOINT 5	0.887	0.897	0.984
NORTH 8	0.685	0.601	0.845
NORTH 11	0.702	0.573	0.721
BLISS 10 ICU	0.957	0.484	0.238
NORTH 10	0.673	0.526	0.727
BLISS 7 ICU	0.971	0.259	0.191
CENTER 11	0.722	0.622	0.880
BLISS 7 STEP DOWN	0.924	0.373	0.234
LABOR DELIVERY 6	0.714	0.555	0.513
BLISS 6 NURSERY	0.920	0.920	1.000
EMERGENCY OBS	0.650	0.646	0.788
IR	0.761	0.825	0.767
MAIN PACU	0.800	0.778	0.700
CONKLIN 1	0.572	0.368	0.709
MAIN OR	0.722	0.598	0.679
CARDIAC CATH LAB	0.636	0.583	0.700
BJI PERIOP SVC	0.682	0.526	0.667
CENTER 12	0.708	0.566	0.671
NORTH 12	0.702	0.569	0.666
BLISS NORTH 2 ICU	0.839	0.635	0.623
BLISS NORTH 3 ICU	0.962	0.500	0.273

Table B.8: Accuracy, Precision and Recall by Department for the Red Alert at Hartford Hospital.

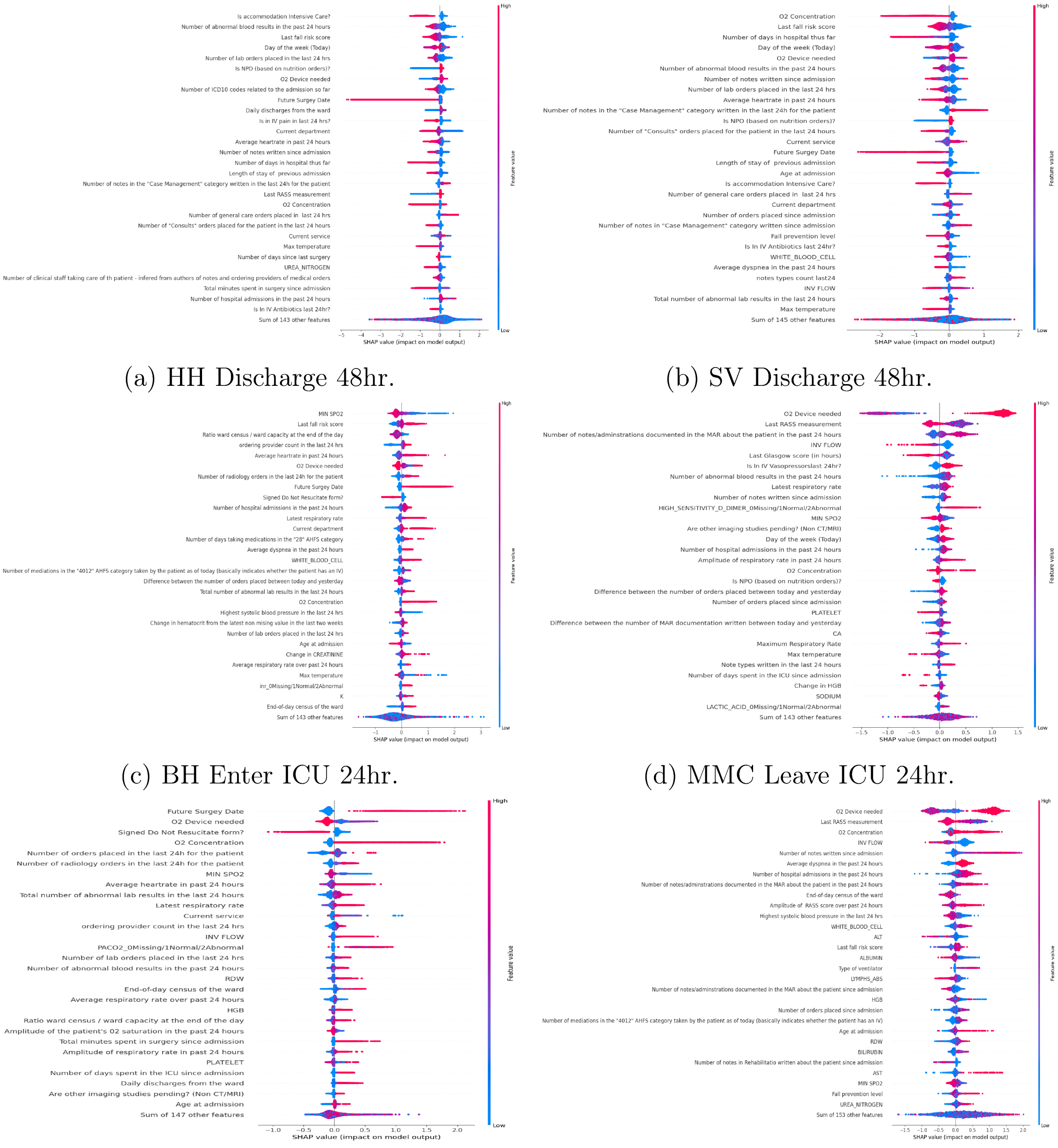
Department	Accuracy	Precision	Recall
BLISS 11 STEP DOWN	0.705	0.450	0.813
NORTH 10	0.950	0.398	0.315
CONKLIN 5	0.913	0.387	0.499
NORTH 11	0.922	0.348	0.546
BLISS 7 ICU	0.668	0.516	0.925
BONE AND JOINT 4	0.988	0.320	0.295
BLISS 7 STEP DOWN	0.667	0.355	0.803
NORTH 9	0.934	0.208	0.445
BLISS 8	0.969	0.431	0.310
BLISS 9 ICU	0.792	0.506	0.908
CONKLIN 4	0.894	0.469	0.631
BLISS 10 STEP DOWN	0.816	0.373	0.635
CENTER 11	0.968	0.945	0.890
BLISS 5 EAST	0.970	0.519	0.548
CENTER 10	0.953	0.334	0.356
BLISS 10 ICU	0.761	0.567	0.850
CONKLIN 3	0.914	0.599	0.737
CENTER 8 ICU	0.746	0.399	0.709
BLISS 9 EAST	0.988	0.089	0.121
CENTER 9 ICU	0.767	0.532	0.798
NORTH 9 STEP DOWN	0.860	0.353	0.579
CONKLIN 2	0.849	0.604	0.758
EMERGENCY	0.914	0.425	0.425
HIGH 12	0.915	0.331	0.494
EMERGENCY OBS	0.971	0.440	0.208
BLISS 9 STEP DOWN	0.868	0.207	0.567
BLISS 11 ICU	0.688	0.597	0.910
BLISS 10 EAST	0.932	0.248	0.292
NORTH 8	0.964	0.532	0.414
BLISS 7 EAST	0.874	0.438	0.608
IR	0.930	0.167	1.000
MAIN PACU	0.800	0.286	1.000
MAIN OR	0.987	0.250	0.222
CONKLIN 1	0.959	0.286	0.400
NORTH 12	0.902	0.407	0.492
CENTER 12	0.907	0.366	0.469
BLISS NORTH 2 ICU	0.891	0.267	0.646
BLISS NORTH 3 ICU	0.769	0.176	0.760

B.2.4 SHAP Summary Plots

In Figure B-2, we present SHAP summary plots for discharge and ICU predictions for 6 different hospitals. We analyze Figure B-2a and Figure B-2b for discharge 48-hr models. Some major clinical discharge barriers are identified, such as intensive care, fall risk score, NPO, O2 device, O2 concentration, and future surgery date, which are used as top variables in both models' and doctors' predictions. In addition to the clinical variables, the models use a variety of operational variables that also have significant contributions to the predictions. A lot of these variables are time-series variables such as counting the number of abnormal blood results, lab orders, ICD codes, days in hospitalization, notes, etc. in the past 24 hours or since admission. Others are not patient-specific; for example, the day of the week and daily discharges from the ward often affect discharge probabilities as well.

Compared with models, doctors focus mainly on the clinical aspects of the patients. The models learn that discharge depends on a combination of clinical and operational characteristics of the patients as well as affected by the hospital's operational status, which is the case in practice as well. Moreover, by comparing the plots between different hospitals in Figure 3-9, we observe that models for different hospitals find common important variables.

For entering ICU, the importance of the future surgery date and the service aligns with known hospital protocols; for example, patients with scheduled cardiology surgery typically go to the ICU after surgery. Many of the significant features identified are similar to those for mortality prediction as entering ICU is strongly correlated with increased mortality risk. For leaving ICU, top features include some oxygen-related variables, such as O2 device, O2 concentration, and SPO2, as well as other clinical variables such as RASS measurement and inverse flow.



(a) HH Discharge 48hr.

(b) SV Discharge 48hr.

(c) BH Enter ICU 24hr.

(d) MMC Leave ICU 24hr.

(e) HOCC Enter ICU 48hr.

(f) CH Leave ICU 48hr.

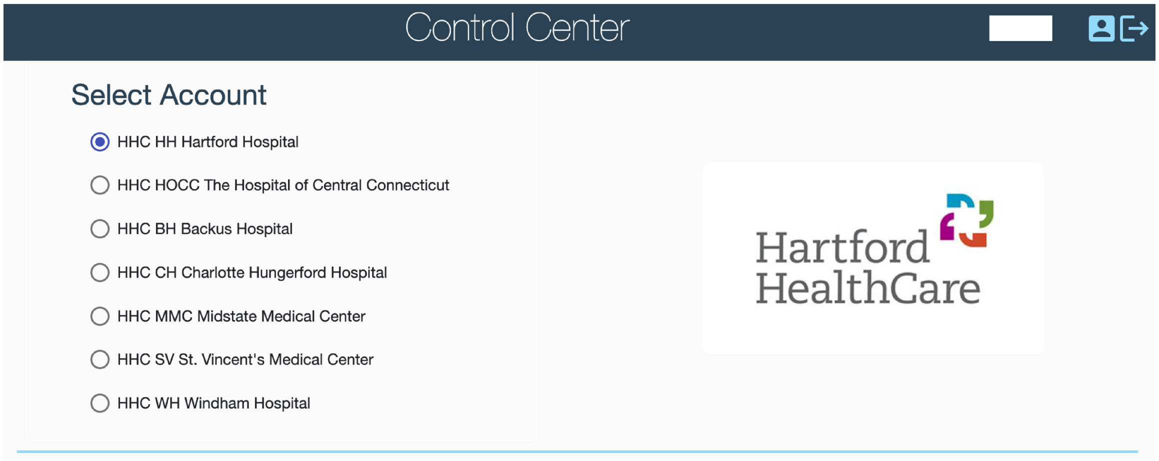
Figure B-2: SHAP Summary Plots for Discharge and ICU Predictions.

B.3 Supplementary Information on Implementation and Impact

In this Appendix section, we present additional information and results of the implementation and impact.

B.3.1 Screenshot of the Software Tool

Figure B-3 presents the log-in page of our tool where each doctor or medical staff can select their hospital.


Control Center

Select Account

● HHC HH Hartford Hospital

○ HHC HOCC The Hospital of Central Connecticut

○ HHC BH Backus Hospital

○ HHC CH Charlotte Hungerford Hospital

○ HHC MMC Midstate Medical Center

○ HHC SV St. Vincent's Medical Center

○ HHC WH Windham Hospital

Figure B-3: User Interface to Enter Software.

B.3.2 Estimation of the Financial Benefits from Pilot

We compare the lengths of stay of patients whose attending discharge physician is one of the four physician champions from the five pilot units in Q4 2020 (277 such patients) vs. Q4 2021 (351 patients). Compared with Q4 2020, the average LOS was reduced by 0.35 days (from 5.84 to 5.49) in Q4 2021. Given that HH usually has a waitlist for inpatient admission, the assumption is that the beds made available thanks to the reduction in LOS would be immediately filled, which would not result in cost savings but a revenue increase. Had the tool been available in Q4 2020, these 277 patients would have been discharged 0.35 days earlier, resulting in savings of

96.95 patient days for the quarter, extrapolated to 387.80 patient days for one year. At an average LOS of 5.84 in Q4 2020, the additional 387.80 patient days available translate into 65.89 additional patients. At an average contribution margin of $10,796 per patient, HH estimated an annual contribution margin increase of $711,348.44 as a result of the pilot implementation.

B.3.3 Staggered Roll-Out of Our Tool

Table B.9 presents information and deployment progress of all units that satisfy the level and specialty of care criterion in the seven hospitals. During the second half of 2022, some units in four hospitals (BH, CH, HOCC, and HH) fully incorporated the software in their daily clinical decision process since the start dates indicated in the table. Two more hospitals (MMC and WH) began adopting the tool in several units on January 15, 2023 as the next phase of implementation. By April 15, 2023, 15 units across six hospitals had fully integrated the predictions in their review process, where unit leads review the predictions with the provider team daily and adjust decisions accordingly. Smaller hospitals with standard procedures of progression rounds with unit leads (BH, CH, and WH) deployed the tool in most of their eligible units, while other hospitals (HH, HOCC, MMC) are still in the process of rolling out more units. Since SV does not have regular progression rounds where the medical team conducts a structured daily patient review process, a streamlined integration of our tool in their progression rounds is ongoing work. As of April 15, 2023, 12 other units (with an NA Start Date) had not officially integrated the daily process deeply, but individual physicians still have the option to access and use the predictions.

Table B.9: Unit Deployment Progress Information.

Hospital	Unit	Start Date	Specialty	Capacity
HH	HH CONKLIN 2	NA	Medicine/Oncology	27
HH	HH CONKLIN 4	9/13/22	Medicine	25
HH	HH CONKLIN 5	7/11/22	Medicine	47
HH	HH BLISS 7 EAST	8/23/22	Medicine	17
HH	HH BLISS 10 EAST	NA	Cardiology	14
HH	HH CENTER 10	NA	Cardiology	26
HH	HH CENTER 12	7/11/22	Medicine	26
HH	HH NORTH 10	NA	Cardiology	27
HH	HH NORTH 12	7/11/22	Medicine	20
BH	BH A3 MEDSURG	8/23/22	Medicine/Surgical	30
BH	BH E4 Cardiology	8/23/22	Cardiology	28
CH	CH FOURTH FLOOR	8/23/22	Medicine/Surgical	28
CH	CH FIFTH FLOOR	8/23/22	Medicine/Surgical	29
HOCC	HOCC EAST 2	NA	Medicine/Observation	12
HOCC	HOCC WEST 2	NA	Medicine	15
HOCC	HOCC NORTH 3	1/15/23	Medicine	24
HOCC	HOCC NORTH 4	10/22/22	Medicine/Cardiology	28
HOCC	HOCC NORTH 5	8/23/22	Medicine/Stroke	30
MMC	MMC PAVILION D	NA	Medicine	28
MMC	MMC PAVILION E	1/15/23	Medicine	28
SV	SV 6 NORTH	NA	Cardiology	20
SV	SV 6 SOUTH	NA	Cardiology	20
SV	SV 9 NORTH	NA	Medicine	22
SV	SV 10 NORTH	NA	Medicine	29
WH	WH 4 SHEA EAST	1/15/23	Medicine/Surgical	30
WH	WH 4 SHEA NORTH	1/15/23	Medicine/Surgical	12
WH	WH GREER	NA	Medicine/Surgical	23

Bibliography

A. Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 2005.

M. A. Ahmed and T. M. Alkhamis. Simulation optimization for an emergency department healthcare unit in Kuwait. *European Journal of Operational Research*, 198(3): 936–942, 2009.

L. H. Aiken, S. P. Clarke, D. M. Sloane, J. Sochalski, and J. H. Silber. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *JAMA*, 288(16): 1987–1993, 2002.

L. H. Aiken, D. M. Sloane, L. Bruyneel, K. Van den Heede, P. Griffiths, R. Busse, M. Diomidous, J. Kinnunen, M. Kózka, E. Lesaffre, et al. Nurse staffing and education and hospital mortality in nine european countries: A retrospective observational study. *The Lancet*, 383(9931):1824–1830, 2014.

M. K. Ameko, M. L. Beltzer, L. Cai, M. Boukhechba, B. A. Teachman, and L. E. Barnes. Offline contextual multi-armed bandits for mobile health interventions: A case study on emotion regulation. In *Fourteenth ACM Conference on Recommender Systems*, pages 249–258, 2020.

M. Amram, J. Dunn, and Y. D. Zhuo. Optimal policy trees. *Machine Learning*, 111(7): 2741–2768, 2022.

B. Y. Ang, S. W. S. Lam, Y. Pasupathy, and M. E. H. Ong. Nurse workforce scheduling in the emergency department: A sequential decision support system considering multiple objectives. *Journal of Nursing Management*, 26(4):432–441, 2018.

S. Ö. Arik and T. Pfister. TabNet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6679–6687, 2021.

S. Athey and S. Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

A. Awad, M. Bader-El-Den, and J. McNicholas. Patient length of stay and mortality prediction: A survey. *Health Services Management Research*, 30(2):105–120, 2017a.

A. Awad, M. Bader-El-Den, J. McNicholas, and J. Briggs. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *International Journal of Medical Informatics*, 108:185–195, 2017b.

B. Bardak and M. Tan. Improving clinical outcome predictions using convolution over medical entities with multimodal learning. *Artificial Intelligence in Medicine*, 117: 102112, 2021.

S. L. Battalio, D. E. Conroy, W. Dempsey, P. Liao, M. Menictas, S. Murphy, I. Nahum-Shani, T. Qian, S. Kumar, and B. Spring. Sense2Stop: A micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention. *Contemporary Clinical Trials*, 109: 106534, 2021.

C. Baumann, H. Singmann, S. J. Gershman, and B. von Helversen. A linear threshold model for optimal stopping behavior. *Proceedings of the National Academy of Sciences*, 117(23):12750–12755, 2020.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

S. L. Bernstein, D. Aronsky, R. Duseja, S. Epstein, D. Handel, U. Hwang, M. McCarthy, K. John McConnell, J. M. Pines, N. Rathlev, et al. The effect of emergency department crowding on clinically oriented outcomes. *Academic Emergency Medicine*, 16(1):1–10, 2009.

G. Bertrand and A. Papavasiliou. Reinforcement-learning based threshold policies for continuous intraday electricity market trading. In *2019 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5. IEEE, 2019.

M. Bertrand, E. Duflo, and S. Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1): 249–275, 2004.

D. Bertsimas and D. den Hertog. *Robust and adaptive optimization*. Dynamic Ideas, 2022.

D. Bertsimas and J. Dunn. Optimal classification trees. *Machine Learning*, 106:1039–1082, 2017.

D. Bertsimas, C. Pawlowski, and Y. D. Zhuo. From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research*, 18 (196):1–39, 2018.

D. Bertsimas, J. Dunn, and N. Mundru. Optimal prescriptive trees. *INFORMS Journal on Optimization*, 1(2):164–183, 2019.

D. Bertsimas, G. Lukin, L. Mingardi, O. Nohadani, A. Orfanoudaki, B. Stellato, H. Wiberg, S. Gonzalez-Garcia, C. L. Parra-Calderón, K. Robinson, M. Schneider, B. Stein, A. Estirado, L. a. Beccara, R. Canino, M. D. Bello, F. Pezzetti, A. Pan, and The Hellenic COVID-19 Study Group. COVID-19 mortality risk assessment: An international multi-center study. *PlOS One*, 15(12):e0243262, 2020.

D. Bertsimas, D. den Hertog, and J. Pauphilet. Probabilistic guarantees in robust optimization. *SIAM Journal on Optimization*, 31(4):2893–2920, 2021a. doi: 10.1137/21M1390967.

D. Bertsimas, J. Pauphilet, J. Stevens, and M. Tandon. Predicting inpatient flow at a major hospital using interpretable analytics. *Manufacturing & Service Operations Management*, 24(6):2809–2824, 2021b.

D. Bertsimas, J. Pauphilet, and B. Van Parys. Sparse classification: A scalable discrete optimization perspective. *Machine Learning*, 110(11):3177–3209, 2021c.

D. Bertsimas, P. Klasnja, S. Murphy, and L. Na. Data-driven interpretable policy construction for personalized mobile health. In *2022 IEEE International Conference on Digital Health (ICDH)*, pages 13–22. IEEE, 2022.

D. Bertsimas, L. Na, J. Pauphilet, A. Silver, P. Veronneau, N. Vogt, A. Haddad-Sisakht, and L. Raison. Optimizing emergency department nurse scheduling: Models and implementation at Hartford Hospital. *To be submitted*, 2023a.

D. Bertsimas, K. Villalobos Carballo, L. Na, and J. Pauphilet. Holistic optimization for a large hospital network. *In preparation*, 2023b.

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. URL `https://doi.org/10.1137/141000671`.

N. Bidargaddi, D. Almirall, S. Murphy, I. Nahum-Shani, M. Kovalcik, T. Pituch, H. Maaieh, V. Strecher, et al. To prompt or not to prompt? A microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. *JMIR mHealth and uHealth*, 6(11):e10123, 2018.

N. Boumparis, M. H. Schulte, and H. Riper. Digital mental health for alcohol and substance use disorders. *Current Treatment Options in Psychiatry*, 6(4):352–366, 2019.

P. Brucker, R. Qu, and E. Burke. Personnel scheduling: Models and complexity. *European Journal of Operational Research*, 210(3):467–473, 2011.

C. W. Chan, M. Huang, and V. Sarhangian. Dynamic server assignment in multiclass queues with shifts, with applications to nurse staffing in emergency departments. *Operations Research*, 69(6):1936–1959, 2021.

T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794. Association for Computing Machinery, 2016.

T.-L. Chen and C.-C. Wang. Multi-objective simulation optimization for medical capacity allocation in emergency department. *Journal of Simulation*, 10(1):50–68, 2016.

M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 307:72–77, 2018.

A. Clark, P. Moule, A. Topping, and M. Serpell. Rescheduling nursing shifts: Scoping the challenge and examining the potential of mathematical model based tools. *Journal of Nursing Management*, 23(4):411–420, 2015.

S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, et al. Activity sensing in the wild: A field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1797–1806, 2008.

M. Covino, C. Sandroni, M. Santoro, L. Sabia, B. Simeoni, M. G. Bocci, V. Ojetti, M. Candelli, M. Antonelli, A. Gasbarrini, and F. Franceschi. Predicting intensive care unit admission and death for COVID-19 patients in the emergency department using early warning scores. *Resuscitation*, 156:84–91, 2020.

M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

M. A. Draeger. An emergency department simulation model used to evaluate alternative nurse staffing and patient population scenarios. In *Proceedings of the 24th Conference on Winter Simulation*, pages 1057–1064, 1992.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

I. Dunning, J. Huchette, and M. Lubin. Jump: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017. doi: 10.1137/15M1020575.

E. Elder, A. N. Johnston, and J. Crilly. Systematic review of three key strategies designed to improve patient flow through the emergency department. *Emergency Medicine Australasia*, 27(5):394–404, 2015.

E. M. Forman, S. G. Kerrigan, M. L. Butryn, A. S. Juarascio, S. M. Manasse, S. Ontañón, D. H. Dallal, R. J. Crochiere, and D. Moskow. Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? *Journal of Behavioral Medicine*, 42(2):276–290, 2019.

Y. Fukuoka, M. Zhou, E. Vittinghoff, W. Haskell, K. Goldberg, A. Aswani, et al. Objectively measured baseline physical activity patterns in women in the mped trial: Cluster analysis. *JMIR Public Health and Surveillance*, 4(1):e9138, 2018.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023. URL `https://www.gurobi.com`.

D. H. Gustafson, F. M. McTavish, M.-Y. Chih, A. K. Atwood, R. A. Johnson, M. G. Boyle, M. S. Levy, H. Driscoll, S. M. Chisholm, L. Dillenburg, et al. A smartphone application to support recovery from alcoholism: A randomized clinical trial. *JAMA Psychiatry*, 71(5):566–572, 2014.

M. Hamid, R. Tavakkoli-Moghaddam, F. Golpaygani, and B. Vahedi-Nouri. A multi-objective model for a nurse scheduling problem by emphasizing human factors. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 234(2):179–199, 2020.

S. Hamine, E. Gerth-Guyette, D. Faulx, B. B. Green, and A. S. Ginsburg. Impact of mHealth chronic disease management on treatment adherence and patient outcomes: A systematic review. *Journal of Medical Internet Research*, 17(2):e3951, 2015.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260): 663–685, 1952.

K. Hovsepian, M. Al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar. cstress: Towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 493–504, 2015.

Y. Hu, C. W. Chan, and J. Dong. Prediction-driven surge planning with application in the emergency department. *Submitted*, 2021.

Y. Hu, K. D. Cato, C. W. Chan, J. Dong, N. Gavin, S. C. Rossetti, and B. P. Chang. Use of real-time information to predict future arrivals in the emergency department. *Annals of Emergency Medicine*, 2023.

Interpretable AI, LLC. Interpretable AI documentation, 2022. URL `https://www.interpretable.ai`.

S. F. Jencks, M. V. Williams, and E. A. Coleman. Rehospitalizations among patients in the medicare fee-for-service program. *New England Journal of Medicine*, 360(14): 1418–1428, 2009.

M. Jin, M. T. Bahadori, A. Colak, P. Bhatia, B. Celikkaya, R. Bhakta, S. Senthivel, M. Khalilia, D. Navarro, B. Zhang, T. Doman, A. Ravi, M. Liger, and T. Kass-hout. Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv preprint arXiv:1811.12276*, 2018.

N. Kallus. Recursive partitioning for personalization using observational data. In *International Conference on Machine Learning*, pages 1789–1798. PMLR, 2017.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 2017.

D. L. Kellogg and S. Walczak. Nurse scheduling: From academia to implementation or not? *INFORMS Journal on Applied Analytics*, 37(4):309–399, 2007.

S.-H. Kim, C. W. Chan, M. Olivares, and G. J. Escobar. Association among ICU congestion, ICU admission decision, and patient outcomes. *Critical Care Medicine*, 44 (10):1814–1821, 2016.

P. Klasnja, E. B. Hekler, S. Shiffman, A. Boruvka, D. Almirall, A. Tewari, and S. A. Murphy. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S):1220, 2015.

P. Klasnja, S. Smith, N. J. Seewald, A. Lee, K. Hall, B. Luers, E. B. Hekler, and S. A. Murphy. Efficacy of contextually tailored suggestions for physical activity: A micro-randomized optimization trial of HeartSteps. *Annals of Behavioral Medicine*, 53(6):573–582, 2019.

D. Koekkoek, K. B. Bayley, A. Brown, and D. L. Rustvold. Hospitalists assess the causes of early hospital readmissions. *Journal of Hospital Medicine*, 6(7):383–388, 2011.

P. Kouvelis and G. Yu. *Robust discrete optimization and its applications*, volume 14. Springer Science & Business Media, 2013.

C. Leineweber, H. S. Chungkham, R. Lindqvist, H. Westerlund, S. Runesdotter, L. S. Alenius, C. Tishelman, et al. Nurses' practice environment and satisfaction with schedule flexibility is related to intention to leave due to dissatisfaction: A multi-country, multilevel study. *International Journal of Nursing Studies*, 58:47–58, 2016.

S. Levine, A. Kumar, G. Tucker, and J. Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

P. Liao, W. Dempsey, H. Sarker, S. M. Hossain, M. Al'Absi, P. Klasnja, and S. Murphy. Just-in-time but not too much: Determining treatment timing in mobile health. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–21, 2018.

P. Liao, K. Greenewald, P. Klasnja, and S. Murphy. Personalized Heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.

G. Lim and A. Mobasher. Robust nurse scheduling problem. In *Proceedings of the 2011 Industrial Engineering Research Conference*, page 1. Institute of Industrial and Systems Engineers (IISE), 2011.

W. Lin and P. Kumar. Optimal control of a queueing system with two heterogeneous servers. *IEEE Transactions on Automatic control*, 29(8):696–703, 1984.

E. F. Long and K. S. Mathews. The boarding patient: Effects of ICU and hospital occupancy surges on patient flow. *Production and Operations Management*, 27(12): 2122–2143, 2018.

S. Maman. *Uncertainty in the demand for service: The case of call centers and emergency departments*. PhD thesis, Technion–Israel Institute of Technology, 2009.

K. Mathews, M. Durst, C. Vargas-Torres, A. Olson, M. Mazumdar, and L. Richardson. Effect of emergency department and ICU occupancy on admission decisions and outcomes for critically ill patients. *Critical Care Medicine*, 46(5):720–727, 2018.

M. Mees, J. Klein, L. Yperzeele, P. Vanacker, and P. Cras. Predicting discharge destination after stroke: A systematic review. *Clinical Neurology and Neurosurgery*, 142:15–21, 2016.

M. Mohammadian, M. Babaei, M. Amin Jarrahi, and E. Anjomrouz. Scheduling nurse shifts using goal programming based on nurse preferences: A case study in an emergency department. *International Journal of Engineering*, 32(7):954–963, 2019.

L. Na, K. Villalobos Carballo, J. Pauphilet, A. Haddad-Sisakht, D. Kombert, M. Boisjoli-Langlois, A. Castiglione, M. Khalifa, H. Pooja, B. Stein, and D. Bertsimas. Hartford HealthCare improves hospital operations with patient outcome predictions. *To be submitted*, 2023.

I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy. Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462, 2018.

J. Needleman, P. Buerhaus, S. Mattke, M. Stewart, and K. Zelevinsky. Nurse-staffing levels and the quality of care in hospitals. *New England Journal of Medicine*, 346: 1715–1722, 2002.

A. Niculescu-Mizil and R. A. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632, 2005.

A. Orfanoudaki, E. Chesley, C. Cadisch, B. Stein, A. Nouh, M. J. Alberts, and D. Bertsimas. Machine learning provides evidence that stroke risk is not linear: The non-linear framingham stroke risk score. *PlOS One*, 15(5):e0232414, 2020.

A. Orfanoudaki, A. Giannoutsou, S. Hashim, D. Bertsimas, and R. C. Hagberg. Machine learning models for mitral valve replacement: A comparative analysis with the society of thoracic surgeons risk score. *Journal of Cardiac Surgery*, 37(1):18–28, 2022.

C. S.-Y. Park, M. Kabak, H. Kim, S. Lee, and G. G. Cummings. No more unimplementable nurse workforce planning. *Contemporary Nurse*, 58(2-3):237–247, 2022.

L. E. Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

M. Rabbi, M. H. Aung, M. Zhang, and T. Choudhury. MyBehavior: Automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 707–718, 2015a.

M. Rabbi, A. Pfammatter, M. Zhang, B. Spring, T. Choudhury, et al. Automated personalized feedback for physical activity and dietary behavior change with mobile phones: A randomized controlled trial on adults. *JMIR mHealth and uHealth*, 3(2): e4160, 2015b.

M. Rabbi, M. Hane Aung, and T. Choudhury. Towards health recommendation systems: An approach for providing automated personalized health feedback from mobile data. In *Mobile Health*, pages 519–542. Springer, 2017.

M. Rabbi, M. S. Aung, G. Gay, M. C. Reid, T. Choudhury, et al. Feasibility and acceptability of mobile phone–based auto-personalized physical activity recommendations for chronic pain self-management: Pilot study on adults. *Journal of Medical Internet Research*, 20(10):e10147, 2018.

A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, and J. Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(18), 2018.

Z. Ramsey, J. S. Palter, J. Hardwick, J. Moskoff, E. L. Christian, and J. Bailitz. Decreased nursing staffing adversely affects emergency department throughput metrics. *Western Journal of Emergency Medicine*, 19(3):496–500, 2018.

P. Rerkjirattikal, R. Singhaphandu, V.-N. Huynh, and S. Olapiriyakul. Job-satisfaction enhancement in nurse scheduling: A case of hospital emergency department in Thailand. In *Integrated Uncertainty in Knowledge Modelling and Decision Making: 9th International Symposium, IUKM 2022, Ishikawa, Japan, March 18–19, 2022, Proceedings*, pages 143–154. Springer, 2022.

A. Reuther, J. Kepner, C. Byun, S. Samsi, W. Arcand, D. Bestor, B. Bergeron, V. Gadepally, M. Houle, M. Hubbell, M. Jones, A. Klein, L. Milechin, J. Mullen, A. Prout, A. Rosa, C. Yee, and P. Michaleas. Interactive supercomputing on 40,000 cores for machine learning and data analysis. In *2018 IEEE High Performance Extreme Computing Conference*. IEEE, 2018.

K. C. Safavi, T. Khaniyev, M. Copenhaver, M. Seelen, A. C. Z. Langle, J. Zanger, B. Daily, R. Levi, and P. Dunn. Development and validation of a machine learning model to aid discharge processes for inpatient surgical care. *JAMA Network Open*, 2 (12):e1917221–e1917221, 2019.

S. Saghafian, G. Austin, and S. J. Traub. Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering*, 5(2):101–123, 2015.

S. N. Saleh, A. N. Makam, E. A. Halm, and O. K. Nguyen. Can we predict early 7-day readmissions using a standard 30-day hospital readmission risk prediction model? *BMC Medical Informatics and Decision Making*, 20(227):1–7, 2020.

K. Sang, P. M. Todd, R. L. Goldstone, and T. T. Hills. Simple threshold rules solve explore/exploit trade-offs in a resource accumulation search task. *Cognitive Science*, 44(2):e12817, 2020.

K. Şenol, B. Saylam, F. Kocaay, and M. Tez. Red cell distribution width as a predictor of mortality in acute pancreatitis. *The American Journal of Emergency Medicine*, 31(4): 687–689, 2013.

L. R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. Villalobos Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, and D. Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digital Medicine*, 5(149), 2022.

M. Soni and R. Gopalakrishnan. Significance of RDW in predicting mortality in COVID-19—an analysis of 622 cases. *International Journal of Laboratory Hematology*, 43(4):O221–O223, 2021.

E. Štrumbelj and I. Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41:647–665, 2014.

S. Subudhi, A. Verma, A. B. Patel, C. C. Hardin, M. J. Khandekar, H. Lee, D. McEvoy, T. Stylianopoulos, L. L. Munn, S. Dutta, and R. K. Jain. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *npj Digital Medicine*, 4(87), 2021.

H. Sun, S. Du, and S. Wager. Treatment allocation under uncertain costs. *arXiv preprint arXiv:2103.11066*, 2021.

E. Sverdrup, A. Kanodia, Z. Zhou, S. Athey, and S. Wager. policytree: Policy learning via doubly robust empirical welfare maximization over trees. *Journal of Open Source Software*, 5(50):2232, 2020.

A. C. Svirsko, B. A. Norman, D. Rausch, and J. Woodring. Using mathematical modeling to improve the emergency department nurse-scheduling process. *Journal of emergency nursing*, 45(4):425–432, 2019.

C. Tekin, O. Atan, and M. Van Der Schaar. Discover the expert: Context-adaptive expert selection for medical diagnosis. *IEEE Transactions on Emerging topics in Computing*, 3(2):220–234, 2014.

B. Tofighi, C. Chemi, J. Ruiz-Valcarcel, P. Hein, L. Hu, et al. Smartphone apps targeting alcohol and illicit substance use: Systematic search in in commercial app stores and critical content analysis. *JMIR mHealth and uHealth*, 7(4):e11831, 2019.

D. E. Twigg, Y. Kutzer, E. Jacob, and K. Seaman. A quantitative systematic review of the association between nurse skill mix and nursing-sensitive patient outcomes in the acute care setting. *Journal of Advanced Nursing*, 75(12):3404–3423, 2019.

D. Van Hulst, D. den Hertog, and W. Nuijten. Robust shift generation in workforce planning. *Computational Management Science*, 14:115–134, 2017.

K. Villalobos Carballo, Y. Ma, L. Na, L. Boussioux, C. Zeng, L. R. Soenksen, I. Fuentes, and D. Bertsimas. TabText: A flexible and contextual approach to tabular data representation. *arXiv preprint arXiv:2206.10381*, 2022.

A.-Y. Wang, H.-P. Ma, W.-F. Kao, S.-H. Tsai, and C.-K. Chang. Red blood cell distribution width is associated with mortality in elderly patients with sepsis. *The American Journal of Emergency Medicine*, 36(6):949–953, 2018.

K. Wang, W. Hussain, J. R. Birge, M. D. Schreiber, and D. Adelman. A high-fidelity model to predict length of stay in the neonatal intensive care unit. *INFORMS Journal on Computing*, 34(1):183–195, 2022.

H. Wiberg, P. Yu, P. Montanaro, J. Mather, S. Birz, M. Schneider, and D. Bertsimas. Prediction of neutropenic events in chemotherapy patients: A machine learning approach. *JCO Clinical Cancer Informatics*, 5:904–911, 2021.

T. I. Wickert, P. Smet, and G. V. Berghe. The nurse rerostering problem: Strategies for reconstructing disrupted schedules. *Computers & Operations Research*, 104:319–337, 2019.

E. Yom-Tov, G. Feraru, M. Kozdoba, S. Mannor, M. Tennenholtz, and I. Hochberg. Encouraging physical activity in patients with diabetes: Intervention using a reinforcement learning system. *Journal of Medical Internet Research*, 19(10):e7994, 2017.

B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.

Z. Zhao, A. Chen, W. Hou, J. M. Graham, H. Li, P. S. Richman, H. C. Thode, A. J. Singer, and T. Q. Duong. Prediction model and risk scores of ICU admission and mortality in COVID-19. *PlOS One*, 15(7):e0236618, 2020.

H. Zhou, P. R. Della, P. Roberts, L. Goh, and S. S. Dhaliwal. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: An updated systematic review. *BMJ open*, 6(6):e011060, 2016.

M. Zhou, Y. Mintz, Y. Fukuoka, K. Goldberg, E. Flowers, P. Kaminsky, A. Castillejo, and A. Aswani. Personalizing mobile fitness apps using reinforcement learning. In *CEUR Workshop Proceedings*, volume 2068. NIH Public Access, 2018a.

M. Zhou, Y. Fukuoka, K. Goldberg, E. Vittinghoff, and A. Aswani. Applying machine learning to predict future adherence to physical activity programs. *BMC Medical Informatics and Decision Making*, 19(1):1–11, 2019.

Z. Zhou, S. Athey, and S. Wager. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*, 2018b.

T. Zhu, L. Luo, X. Zhang, Y. Shi, and W. Shen. Time-series approaches for forecasting the number of hospital daily discharged inpatients. *IEEE Journal of Biomedical and Health Informatics*, 21(2):515–526, 2015.