Empirical Investigations Into The Impact Of Step-Down Unit Admission And Health Information Technology On Patient Outcomes

Suparerk Lekwijit

University of Pennsylvania

Follow this and additional works at: https://repository.upenn.edu/edissertations

Part of the Databases and Information Systems Commons, Health and Medical Administration Commons, and the Operational Research Commons

Recommended Citation


This paper is posted at ScholarlyCommons. https://repository.upenn.edu/edissertations/4016
For more information, please contact repository@pobox.upenn.edu.
Empirical Investigations Into The Impact Of Step-Down Unit Admission And Health Information Technology On Patient Outcomes

Abstract
This dissertation in operations management studies healthcare operations using large-scale data sets and econometric methods. In chapter one, we examine the benefits of step-down units (SDU), which provide an intermediate level of care for semi-critically ill patients. Using data from 10 hospitals, we estimate the clinical and operational impact of SDU care and find that SDU care significantly improves health outcomes among post-intensive care unit patients. However, its benefits are less certain for patients admitted from the emergency department. In chapter two, we utilize data from a field experiment and examine the efficacy of a connected health system that aimed to reduce readmissions through improved medication adherence. Patients in our study received electronic pill bottles that tracked medication adherence and received different types of feedback when they were non-adherent. We find that patients were more likely to become adherent when receiving high levels of intervention involving personalized feedback and when the intervention is escalated quickly and consistently. We also find that long-term adherence to two crucial heart medications reduces readmission risk. Additionally, we develop a dynamic readmission risk-scoring model and use simulation to show that, when using an intervention strategy that prioritizes high-risk patients, we can significantly reduce readmissions while using the same effort level. In chapter three, we investigate the impact of video visit availability at a large academic medical center on in-system care utilization. Through difference-in-differences, we find that video visit availability is associated with increases in overall care utilization and in-person care utilization within the system. As a whole, this dissertation provides healthcare providers and managers with data-driven insight towards how to maximize effectiveness and efficiency in healthcare delivery through changes in the care structure and health technology adoption.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Operations & Information Management

First Advisor
Christian Terwiesch

Keywords
Empirical Operations Management, Health Information Technology, Healthcare Operations Management, Telehealth

Subject Categories
Databases and Information Systems | Health and Medical Administration | Operational Research

This dissertation is available at ScholarlyCommons: https://repository.upenn.edu/edissertations/4016
EMPIRICAL INVESTIGATIONS INTO THE IMPACT OF STEP-DOWN UNIT ADMISSION AND HEALTH INFORMATION TECHNOLOGY ON PATIENT OUTCOMES

Suparerk Lekwijit

A DISSERTATION

in

Operations, Information and Decisions

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

Christian Terwiesch, Andrew M. Heller Professor; Professor of Operations, Information and Decisions

Graduate Group Chairperson

Nancy R. Zhang, Ge Li and Ning Zhao Professor of Statistics

Dissertation Committee

Hummy Song, Assistant Professor of Operations, Information and Decisions

Carri W. Chan, Associate Professor of Business, Columbia Business School
EMPIRICAL INVESTIGATIONS INTO THE IMPACT OF STEP-DOWN UNIT ADMISSION AND HEALTH INFORMATION TECHNOLOGY ON PATIENT OUTCOMES

© COPYRIGHT

2021

Suparerk Lekwijit

This work is licensed under the Creative Commons Attribution NonCommercial-ShareAlike 4.0 License

To view a copy of this license, visit

https://creativecommons.org/licenses/by-nc-sa/4.0/us/
Dedicated to Mom and Dad
ACKNOWLEDGMENT

First of all, I would like to thank my PhD advisor, Professor Christian Terwiesch, who has given me tremendous support and guidance throughout my PhD career. You made my PhD experience memorable (in the best possible way), enjoyable (a word that almost no one uses to describe their PhD experience), and rewarding (not only did I learn a lot but I also got a degree and a job!). Without you, this would not have been possible. Thank you, Christian, for always looking out for me and wanting the best for me. I am fortunate to have you as my advisor.

I would like to thank the rest of my dissertation committee, Professor Hummy Song and Professor Carri Chan. Professor Hummy Song has been a mentor, a teacher, and a friend who has inspired and motivated me to be the best teacher and researcher I can be. I would like to thank Professor Carri Chan for being my early mentor, introducing me to healthcare operations, and paving me the path to PhD. Thank you, Hummy and Carri! I am grateful to have both of you as my mentors.

My warm thanks to everyone in the OID Department, including all past and present PhD coordinators, Professor Noah Gans, Professor Serguei Netessine, and Professor Xuanming Su. Most importantly, my PhD experience would not be complete without my PhD friends, especially those in my cohort—Aman, Canberk, Park, and Sam. You guys are amazing! I would also like to thank Arielle, Chloe, Dawson, Edward, Erika, Joseph, Kaps, Jiding, Lesley, Ozge, Pia, and Rowena. You all made coming to work each day a joy (and gave me distractions from work that I needed).

On the clinical front of my research, I would like to thank Professor Kevin Volpp, Professor David Asch, and Professor Krisda Chiyachati for supporting me and making our research possible.

There are so many other people that have inspired, motivated, and supported me directly and indirectly throughout the years—family, friends, and other important people in my life. I would name all of you here if I could, but this would be longer than the rest of my dissertation.

Lastly, to my future self who may be reading this, be proud of yourself. Do you remember that you wrote this during a pandemic? What a weird time, but you made it through!
ABSTRACT

EMPIRICAL INVESTIGATIONS INTO THE IMPACT OF STEP-DOWN UNIT ADMISSION AND HEALTH INFORMATION TECHNOLOGY ON PATIENT OUTCOMES

Suparerk Lekwijit

Christian Terwiesch

This dissertation in operations management studies healthcare operations using large-scale data sets and econometric methods. In chapter one, we examine the benefits of step-down units (SDU), which provide an intermediate level of care for semi-critically ill patients. Using data from 10 hospitals, we estimate the clinical and operational impact of SDU care and find that SDU care significantly improves health outcomes among post-intensive care unit patients. However, its benefits are less certain for patients admitted from the emergency department. In chapter two, we utilize data from a field experiment and examine the efficacy of a connected health system that aimed to reduce readmissions through improved medication adherence. Patients in our study received electronic pill bottles that tracked medication adherence and received different types of feedback when they were non-adherent. We find that patients were more likely to become adherent when receiving high levels of intervention involving personalized feedback and when the intervention is escalated quickly and consistently. We also find that long-term adherence to two crucial heart medications reduces readmission risk. Additionally, we develop a dynamic readmission risk-scoring model and use simulation to show that, when using an intervention strategy that prioritizes high-risk patients, we can significantly reduce readmissions while using the same effort level. In chapter three, we investigate the impact of video visit availability at a large academic medical center on in-system care utilization. Through difference-in-differences, we find that video visit availability is associated with increases in overall care utilization and in-person care utilization within the system. As a whole, this dissertation provides healthcare providers and managers with data-driven insight towards how to maximize effectiveness and efficiency in healthcare delivery through changes in the care structure and health technology adoption.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ACKNOWLEDGMENT</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td>LIST OF ILLUSTRATIONS</td>
<td>xii</td>
</tr>
<tr>
<td>PREFACE</td>
<td>xiii</td>
</tr>
<tr>
<td>CHAPTER 1 : Assessing the Impact of Service Level when Customer Needs are Uncertain: An Empirical Investigation of Hospital Step-Down Units</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Setting and Data</td>
<td>9</td>
</tr>
<tr>
<td>1.3 ICU Cohort: Econometric Approach</td>
<td>16</td>
</tr>
<tr>
<td>1.4 ICU Cohort: Results</td>
<td>22</td>
</tr>
<tr>
<td>1.5 ED Cohort: Econometric Approach</td>
<td>27</td>
</tr>
<tr>
<td>1.6 ED Cohort: Results</td>
<td>34</td>
</tr>
<tr>
<td>1.7 Conclusions and Managerial Insights</td>
<td>42</td>
</tr>
<tr>
<td>1.8 Appendix: Supplementary Tables</td>
<td>46</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>48</td>
</tr>
<tr>
<td>2.2 Literature Review</td>
<td>52</td>
</tr>
<tr>
<td>2.3 Study Setting</td>
<td>57</td>
</tr>
</tbody>
</table>
LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Distribution of Units Preceding the SDU</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Capacity of Various Inpatient Units in Terms of Number of Beds</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>Summary Statistics of Patient Demographics</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>Summary Statistics of Patient Outcomes: Mean (Number of Observations or Standard Deviation for Continuous Variables)</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>Effect of ICUBusy and SDUBusy on ICU LOS</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>Estimated Effect of SDU Admission Following ICU Discharge ((\gamma)) on Patient Outcomes and Correlation Between Error Terms ((\rho)) for the Admission Decision and Patient Outcomes: (N = 11,058)</td>
<td>23</td>
</tr>
<tr>
<td>7</td>
<td>Proportions of ED Patients Who Are Routed to the ICU, SDU, and Ward When (SDUBusy = 1) vs. (SDUBusy = 0) and Results of (t)-Tests That Compare the Difference in Routing Proportions</td>
<td>31</td>
</tr>
<tr>
<td>8</td>
<td>Routing Statistics of Patients for Different Severity Groups</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>Summary Statistics of Patient Demographics for ED Cohort by Severity Classification</td>
<td>31</td>
</tr>
<tr>
<td>10</td>
<td>Summary Statistics of Patient Outcomes for ED Cohort by Severity Classification: Mean (Number of Observations or Standard Deviation for Continuous Variables)</td>
<td>32</td>
</tr>
<tr>
<td>11</td>
<td>Estimated Effect of SDU Admission Following the ED ((\gamma)) on Patient Outcomes for Low-Severity Patients and Correlation Between Error Terms ((\rho)) for the Admission Decision and Patient Outcomes: (N = 42,385)</td>
<td>36</td>
</tr>
<tr>
<td>12</td>
<td>Estimated Effect of SDU Admission Following the ED ((\gamma)) on Patient Outcomes for High-Severity Patients and Correlation Between Error Terms ((\rho)) for the Admission Decision and Patient Outcomes: (N = 2,803)</td>
<td>37</td>
</tr>
</tbody>
</table>
TABLE 34: Estimated Effect of Instrumental Variables on Short-Term Medication Adherence ................................. 99

TABLE 35: Estimated Effect of Instrumental Variables on Long-Term Medication Adherence ................................. 99

TABLE 36: Estimated Effect of Medication Adherence on Readmission ................................................................. 100

TABLE 37: Results of Endogeneity Tests ................................................................................................................. 100

TABLE 38: Distribution of Visit Types .................................................................................................................... 108

TABLE 39: Demographics of Employees and Non-Employees .................................................................................. 109

TABLE 40: Demographics of Employees and Non-Employees by Distance Group
Before Matching ...................................................................................................................................................... 113

TABLE 41: Demographics of Employees and Non-Employees by Distance Group
After Matching ......................................................................................................................................................... 114

TABLE 42: Effects of Video Visit Availability on Monthly Care Utilization Estimated Using Standard Difference-in-Differences ................................................................. 115

TABLE 43: Effects of Video Visit Availability on Monthly Care Utilization Estimated Using Difference-in-Differences on Matched Samples .............................................................................. 116
LIST OF ILLUSTRATIONS

FIGURE 1: Types of Admission Decisions ................................................. 10
FIGURE 2: Data Selection ................................................................. 13
FIGURE 3: Proportions of ED Patients Who Are Routed to the ICU, SDU, and Ward
When $SDU_{Busy} = 1$ vs. $SDU_{Busy} = 0$ ................................................. 30
FIGURE 4: Percentage of (a) Low- or (b) Medium-Severity Patients Admitted to the
SDU from the ED When the ICU Is Busy ($ICU_{Busy} = 1$) or Not ($ICU_{Busy} = 0$) for Varying Levels of Severity as Measured by Deciles of $\hat{Tx}^*_i$, Given Patients Are Classified as Low Severity: $\hat{Tx}^*_i \leq t_1$ .... 33
FIGURE 5: Analysis Outline ................................................................. 54
FIGURE 6: Feedback Distribution over Time .......................................... 64
FIGURE 7: Examples of Escalation Patterns from Six Different Non-Adherence
Sequences .................................................................................. 65
FIGURE 8: Baseline Probability of Becoming Adherent Again ($H_0(t)$) ............. 73
FIGURE 9: Yield Obtained Using Different Readmission Intervention Approaches .... 84
FIGURE 10: Precision Obtained Using Different Readmission Intervention Approaches .... 84
FIGURE 11: Simulation Flow ............................................................. 86
FIGURE 12: Numbers of Manual Interventions and Readmissions for Different
Intervention Strategies ................................................................ 87
FIGURE 13: Yield Obtained Using Different Readmission Intervention Approaches:
Varying the Definition of Patient-Day at Risk ....................................... 95
FIGURE 14: Precision Obtained Using Different Readmission Intervention
Approaches: Varying the Definition of Patient-Day at Risk ......................... 96
FIGURE 15: Patients’ Care Options Before and After Video Visits Were Introduced ... 103
FIGURE 16: Monthly Visit Frequency over Time ..................................... 112
PREFACE

The broad goals of this dissertation are to understand how healthcare providers and managers can make operational improvements to achieve better health and operational outcomes and to provide actionable, data-driven insight towards how to maximize effectiveness and efficiency in healthcare delivery. This dissertation focuses on two broad research areas within the healthcare operations literature. First, in chapter one, we study inpatient operations and examine the impact of step-down unit care on patient outcomes. Second, in chapters two and three, we seek to understand the impact of health information technology on outpatient operations and ways providers provide care and connect with patients, as well as determine how providers can effectively adopt such technologies.

In chapter one, "Assessing the Impact of Service Intensity on Customers: An Empirical Investigation of Hospital Step-Down Units", we study a service setting where servers have different capabilities and customers have varying needs. One common way this occurs is when servers are hierarchical in their skills or in the level of service they can provide. Much of the literature studying such systems relies on an understanding of the relative costs and benefits associated with serving different customer types by the different levels of service. In this chapter, we focus on estimating these costs and benefits in a complex healthcare setting where the major differentiation among server types is the intensity of service provided. Step-down units (SDUs) were initially introduced in hospitals to provide an intermediate level of care for semi-critically ill patients who are not sick enough to require intensive care but not stable enough to be treated in the general medical/surgical ward. One complicating factor is that the needs of customers is sometimes uncertain—specifically, it is difficult to know a priori which level of care a particular patient needs. Using data from 10 hospitals from a single hospital network, we take a data-driven approach to classify patients based on severity and empirically estimate the clinical and operational outcomes associated with routing these patients to the SDU.

In chapter two, "Evaluating the Efficacy of Connected Healthcare: An Empirical Examination of Patient Engagement Systems and Their Impact on Readmission", we study a provider-initiated
connected healthcare model, which is a form of health delivery that connects patients and providers through connected health devices and allow providers to monitor patient behavior and proactively intervene before an adverse event occurs. Unlike the costs, the benefits of connected healthcare in improving patient behavior and health outcomes are usually difficult to determine. In this chapter, we examine the efficacy of a connected health system that aimed to reduce readmissions through improved medication adherence. Specifically, we study 1,000 patients with heart disease who received electronic pill bottles that tracked medication adherence. Patients who were non-adherent received active social support that involved different types of feedback such as text messages and calls. By integrating data on adherence, intervention, and readmission, we aim to (1) investigate the efficacy of connected healthcare in promoting medication adherence, (2) examine the relationship between medication adherence and readmission, and (3) develop a dynamic readmission risk-scoring model that considers medication adherence and use the model to better target non-adherent patients.

In chapter three, "The Impact of Video Visit Availability on Care Utilization", we study a patient-initiated connected healthcare model known as video visits or tele-visits. Video visits allow patients to remotely connect at any time of day or night with their providers through a computer, smart phone, or tablet. This chapter investigates the impact of video visit availability at a large academic medical center on in-system care utilization using difference-in-differences and discusses the next steps in our work towards understanding how video visit availability impacts the choices that patients make when they are sick.
CHAPTER 1: Assessing the Impact of Service Level when Customer Needs are Uncertain: An Empirical Investigation of Hospital Step-Down Units

Joint work with Carri W. Chan, Linda V. Green, Lijian Lu, and Gabriel Escobar

1.1. Introduction

Hospitals are responsible for the largest component of national health care expenditures and are therefore under pressure from government and private payers to become more cost efficient (Centers for Medicare & Medicaid Services 2016). Traditionally, inpatient care at hospitals had been defined by two levels of care: Intensive care units (ICUs) and general medical/surgical wards (wards). With one nurse per one or two patients, ICUs provide the highest level of care and are very costly to operate, with annual costs in the U.S. between $121 and $263 billion (i.e., 17.4%-39% of total hospital costs; Coopersmith et al. 2012). In an effort to mitigate critical care costs, Step-down units (SDUs), sometimes called transitional care or intermediate care units, have been used to provide an intermediate, third level of care for semi-critically ill patients who are not severe enough to require intensive care but not stable enough to be treated in the ward. SDUs typically have one nurse per three to four patients and are generally less expensive to operate than ICUs primarily due to lower nurse-to-patient ratios. On the other hand, SDUs are more expensive than general wards where there are, generally, about 6 patients per nurse. With the use of SDUs becoming more widespread, it is of growing importance for hospital administrators and healthcare providers to have a better understanding of the benefits and best practices associated with using this intermediate level of care.

At a conceptual level, the hospital and ICU/SDU/ward system can be thought of as a general service system with three levels of service and heterogeneous customers. The levels are nested, in the sense that the lowest level (ward) has the least capabilities and can only provide service to a subset of customers (patients); the second level (SDU) can provide service to the lowest level customers plus additional customers with greater needs; and the highest level (ICU) can provide service
(theoretically) to all customers. Due to higher staffing levels as well as specialized equipment, higher levels of care are more costly to provide. It is of interest to understand whether such a structure is beneficial and, if so, how to best utilize the different levels of service. This is more challenging when there is uncertainty concerning which customers are best served at each level, making it very difficult to evaluate the cost-benefit tradeoffs. The ultimate goal is to understand effective management of such a service system, including capacity management of each level of service, when and how to route customers, as well as how to classify customers and identify their needs for the different levels of service.

There has been a considerable amount of research into capacity management of service systems and the development of routing policies to different service types (e.g., Wallace and Whitt 2005, Gurvich et al. 2008). Such issues have been studied in various service settings including call-centers (e.g., Gans et al. 2003), hospitals (e.g., Armony et al. 2018, Best et al. 2015), cloud-computing (e.g., Maguluri et al. 2012), among many others. A common assumption in these works is a general understanding of the relative costs and benefits associated with different customer groups receiving service from the various server types. Yet, in some contexts these relative costs and benefits may not be known. Specifically, the needs of customers may be uncertain prior to starting service. In this work, our goal is to gain an understanding of how best to use different levels of service to serve customers with uncertain needs by empirically examining how different customer groups are impacted by being served at differing levels. We examine this question in a healthcare context–the SDU.

There is a lack of consensus in the medical community surrounding the use of SDUs as well as a lack of substantive evidence concerning their effectiveness. Still, many hospitals have SDUs and others are considering introducing these units. Even within a single hospital, the use of SDUs is generally not standardized. Therefore, it is very important to understand their value and how they can best be used. This paper examines whether or not SDUs are associated with improved operational and/or clinical outcomes for different types of patients. In this context, the aforementioned costs and benefits are not necessarily financial in nature. For instance, they can correspond to deteriorations
or improvements in patient outcomes. Such analysis can provide insights into how the nested levels of care structure could be used to treat patients with differentiated service requirements and potentially lower hospital operating costs without sacrificing patient outcomes. Given the increasing pressures for hospitals to reduce costs and improve quality, such insights can be very valuable to hospital administrators. More broadly, this analysis may also provide insights into the analysis and management of other service systems with different levels of care (e.g., call centers).

To the best of our knowledge, our work is the first to conduct a multi-hospital study to empirically examine the role of an SDU for patients who are discharged from the ICU as well as those who are admitted from the Emergency Department (ED). Our analyses are based on recent data from Kaiser Permanente Northern California, an integrated health care delivery system serving 3.6 million members that operates 21 hospitals, some of which do and some of which do not have SDUs. The cohort and type of data we employ have been described in previous studies (see Escobar et al. 2013 and Kim et al. 2015, among others). Our data source is based on nearly 170,000 hospitalizations in a total of 10 hospitals over a course of one and half years. Each of the 10 hospitals in our study has an ICU and SDU, though the number of beds in each of the units varies across hospitals.

There are a number of challenges which arise when trying to understand the impact of SDU care on patient outcomes. One challenge is that there are limited studies regarding its efficacy and, more specifically, which patients can be safely admitted to the SDU (Nasraway et al. 1998). While there is some evidence that some ICU patients who are at low risk of needing life support could be given less intensive care in an SDU with no impact on outcomes (e.g., Zimmerman et al. 1995), there is also evidence that some critical care patients who are treated in SDUs or general wards instead of the ICU are worse off (e.g., Simchen et al. 2004). As such, it seems that there are patients who may benefit from being cared for in an SDU rather than in a general ward, while others who are treated in an SDU rather than an ICU may suffer adverse consequences. An important empirical challenge is to be able to classify patients in order to accurately assess the impact of SDU admission on patient outcomes. To that end, we initially segregate patients who are candidates for SDU care into two broad groups: those who are discharged from the ICU and those who are admitted to an inpatient
unit from the ED. Taking a data-driven approach, we then stratify patients from the ED into high- and low-severity groups.

In developing an understanding of SDUs, we face an important estimation challenge. The SDU admission decision may be affected by health factors which are known to the physician at the time of the decision, but are unobservable in the data. For instance, a patient’s physical appearance (i.e., whether he/she appears ashen or pale) may provide evidence of early shock. Thus, a physician may determine that, despite relatively stable vital signs and lab scores, a patient who is pale and sweating will benefit from SDU care relative to being sent to the general medical ward. But because the patient is more critical than the average ward patient, he/she is also more likely to have worse outcomes. Similarly, it may be more appropriate to admit a patient to the ICU if he is cognitively impaired and not lucid. Thus, patients who are admitted to the SDU instead of the ICU may be healthier by unobservable measures. Ignoring this potential endogeneity could result in biased estimates. To address this challenge, we utilize an instrumental variable approach to identify the desired effects.

Our empirical findings suggest that SDU care is associated with substantial improvements in various patient outcomes for patients discharged from the ICU as well as low-severity patients being admitted from the ED. However, we find that SDU admission is associated with worse outcomes for high-severity patients coming from the ED. Our results suggest that when SDUs are used as originally intended, as intermediary units for post-ICU care, they may result in improved outcomes relative to ward care. However, if hospital administrators wish to expand the use of SDUs beyond post-ICU care, it is important to be able to classify which patients should or should not be treated in the SDU. More generally, our findings highlight the importance of being able to accurately classify customers and to quantify the (dis)utility associated with different service capabilities when considering routing decisions.

The rest of the paper is organized as follows. We conclude this section with a brief summary of related papers in the literature. In Section 1.2, we introduce our study setting and describe our data, including the two patient cohorts we study. In Section 1.3, we describe our econometric model for
our first cohort of patients—those being discharged from the ICU. The estimation results for this cohort is provided in Section 1.4. Section 1.5 describes how we partition patients who are admitted from the ED into high and low-severity patients and then discusses the econometric model we use for these patients. Results for these patient types are provided in Section 1.6. Section 1.7 provides concluding remarks as well as discussions for future research.

1.1.1. Literature Review

Our work is related to existing literature in both the operations management and medical communities. Within the operations literature, our work is related to three streams of research: 1) management of general service systems, 2) management of healthcare operations, and 3) empirical analysis of healthcare operations.

There has been a large body of literature examining how to route customers to servers with different skill sets (see the survey article Gans et al. (2003) and the references therein). Research in this area has considered customer prioritization (e.g., Mandelbaum and Stolyar 2004, Gurvich and Whitt 2009), customer routing (e.g., Bell and Williams 2001, Tezcan and Dai 2010), and staffing (e.g., Wallace and Whitt 2005, Gurvich and Whitt 2010). Additionally, there have been a number of works studying service settings with different levels of service. In call-centers, one can consider human servers as providing more intense and costly service than chat-room or automated response systems (e.g., Gans et al. 2003, Tezcan and Behzad 2012, Luo and Zhang 2013, Tezcan and Zhang 2014). Maglaras and Zeevi (2005) considers pricing, admission control, and the design of a mechanism to relay congestion information in a system where servers can provide either a guaranteed service rate or a best-effort service rate. In call center settings, VIP customers often require a higher level of service than the typical customer, raising questions on how to route customers to various servers (e.g., Gans et al. 2003). Such features also arise in healthcare settings including the SDU we study in this paper. Chan et al. (2013) considers how to prioritize burn-injured patients for treatment in hospitals with burn-units which provide the specialized, intense therapies (e.g., skin grafting
surgeries) required for severely burned patients versus other hospitals with less intense treatment capabilities.

The nested structure of the different levels of care we examine in the hospital setting bears similarities to the gate-keeper literature (e.g., Shumsky and Pinker 2003, Hasija et al. 2005, Lee et al. 2012) where the specialist is able to provide services the gate-keeper is not able to. However, in contrast to this literature, in our setting, the lowest level of service does not make the decision to route customers to higher levels of service as in the gate-keeper literature. The nested structure is also related to the classic toll-booth problem considered in Edie (1954) as certain lanes can serve all types of vehicles, while others can only serve a subset of them (e.g., Green 1985). Rather than having a central planner making routing decisions for customers whose needs may be unknown to him/her, in the toll-booth problem, the customers know their needs and self-direct to servers.

There are a number of papers which utilize stochastic modeling and queueing approaches to study resource allocation in hospital settings (e.g., Mandelbaum et al. 2012, Shi et al. 2016, Huang et al. 2015, Huh et al. 2013, Barz and Rajaram 2015). In all of these works, the focus is on admitting patients with heterogeneous needs to different units within the same level of care. That is, servers are interchangeable. In contrast, our work considers the impact of admitting patients to different levels of care. In doing so, we are able to capture heterogeneous service requirements of customers (patients) as well as the various levels of service (care).

There has been a growing body of work in healthcare operations management using mathematical models to manage heterogeneous patients in systems with differentiated server types. Best et al. (2015) examines how to determine the amount of flexibility allowed in hospital wings in order to minimize costs associated with lack of access to care. Dai and Shi (2017) uses an approximate dynamic programming approach to determine how to allocate patients to primary and non-primary units. Armony et al. (2018) uses fluid and diffusion models to determine allocation among expensive resources (ICU beds) that can be used to treat all patient types rather than cheaper resources (SDU beds) that can only treat a subset of patients. An underlying assumption in all of these works is that, in addition to a patient’s type, the relative costs (i.e., degradation of patient outcomes) to treat that
patient in different types of units are known. Our aim is to provide a framework to classify patients as well as to provide rigorous, quantitative estimates of the outcomes for patients treated in an SDU.

As we take an empirical approach to quantify the costs/benefits of treating patients in the SDU, our work is closely related to papers in the empirical operations management literature, especially those focused on healthcare settings. Jerath et al. (2015) empirically estimates how customers’ service needs impact their preferences to use different types of service channels when interacting with a health insurance call center. In hospital settings, Stowell et al. (2013), Kim et al. (2015), and Kuntz et al. (2019) take an empirical approach to explore the impact of admitting patients to different types of hospital units on patient outcomes. While these works highlight the undesirability of ‘off-placement’, Wang et al. (2019b) explicitly considers how information on hospital (server) quality needs to be patient-specific. As such, while hospitals are capable of treating all different types of patients, which is similar to the SDU, the costs/benefits associated with being treated at a specific hospital are quite varied. Unfortunately, it is not always possible to treat patients at the most appropriate hospital or hospital unit. Congestion is a common reason for this lack of access to care. There have been a number of studies examining the impact of congestion and lack of access to care on patient outcomes (e.g., Kc and Terwiesch 2012, Kuntz et al. 2015, Berry Jaeker and Tucker 2017). Batt and Terwiesch (2017) and Freeman et al. (2016) empirically examine how less or more skilled servers can be used to treat some patients during congested periods. In a similar vein, we examine how treating different patient types in an SDU, which is a higher level of care than the ward, but lower than the ICU, impacts their outcomes.

There is a lack of consensus within the medical community about the role of the SDU. Those who advocate the use of SDUs see them as an alternative to either maintaining larger ICUs or jeopardizing patient care due to premature, demand-driven, discharge of patients from ICUs to general care units. As the name suggests, the initial role of SDUs was to serve as a transition for patients after being discharged from the ICU. In practice, SDUs are often used to treat other patients, for example, those who might have gone to an ICU but were blocked because the ICU was full. In general, the use of SDUs has evolved without substantial evidence as to their benefits and
what their role should be. On one hand, some studies argue that SDUs are a cost-effective approach to treat patients by providing a safe and less expensive environment for patients who are not quite sick enough to require treatment in the ICU, but not quite stable enough to be treated in the ward. Without an SDU, most of these patients end up being cared for in the ICU. Byrick et al. (1986) suggests that the use of the SDU could alleviate ICU congestion by reducing ICU length-of-stay (LOS) without increasing mortality rates. This reduction is possible because patients do not have to reach as high a level of stability to be discharged to an SDU rather than to a general medical-surgical ward. Other studies that have shown the cost-effectiveness of an SDU include Harding (2009), Stacy (2011), and Tosteson et al. (1996). On the other hand, a survey of studies on SDUs raises doubts about these benefits and argues that there is not enough evidence of cost-effectiveness (Keenan et al. 1998). While we do not explicitly consider the cost-effectiveness of SDUs (due to lack of detailed financial data), our study provides some insight into these questions by providing rigorous and robust estimates to the effectiveness of SDUs for patients of varying types. At a high-level, one can project ordinal cost estimates due to the lower (higher) staffing levels in the SDU versus the ICU (ward). From a methodological standpoint, our study differentiates itself in that the majority of these studies are conducted exclusively within a single hospital, whereas our study utilizes data from 10 different hospitals. Additionally, rather than conducting a before-and-after study, which may be limited by the inability to control for temporal changes such as staffing changes or closures of nearby hospitals, we utilize an instrumental variable approach to identify the impact of different care pathways (going to the SDU versus ward following ICU discharge as well as going to the SDU versus ward or ICU upon hospital admission from the ED). Our multi-center study provides compelling evidence that there are some patients for whom SDU care is associated with improved clinical outcomes, while there are others for whom SDU care is associated with worse clinical outcomes. As such, our results suggest that it would be of value for the medical community to focus more attention on developing an understanding of which patients would or would not benefit from SDU care at hospitals of varying patient mix and resource availability. More broadly, our results suggest that one must be prudent when introducing multiple levels of service in service systems.
with highly heterogeneous customers as there can be substantial variation in the costs and benefits associated with (incorrectly) routing customers to these servers.

Our estimation approach utilizes an instrumental variable which is based on an operational measure—congestion in an inpatient unit—as has been done in Kim et al. (2015) and Kc and Terwiesch (2012), among others. While the general methodology is similar, the question we are considering is wholly different. The aforementioned works focus on the ICU, while our focus is on the SDU. From an operational standpoint, it is of value to develop an understanding of how servers with lower costs due to lower staffing levels (SDUs) may be used to serve heterogeneous customers. Additionally, from the viewpoint of clinicians and hospital administrators, these units are fundamentally different in their use and role. As a customer’s type and, subsequently, his service requirements are not always observable to managers of the service system, it can be challenging to estimate the costs and benefits associated with being served by particular server types. This challenge arises in the SDU setting because they serve as the site of intermediate care between the ICU and the ward; that is, there are risks of adverse consequences in admitting a patient to the SDU who actually needs ICU care, as well as benefits to admitting patients who might be too sick for the ward. As such, we first take a data-driven approach to help classify customers (patients) before estimating the impact of SDU care on patient outcomes.

1.2. Setting and Data

We utilize patient data from 10 hospitals from Kaiser Permanente Northern California\footnote{This project was approved by the Kaiser Permanente Northern California Institutional Review Board for the Protection of Human Subjects, which has jurisdiction over all study hospitals, and the Columbia University Institutional Review Board for the Protection of Human Subjects.}, containing 165,948 hospitalizations over a course of one and a half years. We note that even within the Kaiser Permanente Northern California system, there is no consensus on how to use SDUs. Thus, some hospitals have SDUs, while others do not.

Our data contains operational and patient level information. Operational level information includes every unit to which a patient is admitted during his hospital stay along with the date and time
of admission and discharge for each unit. Our objective in this work is to understand the impact of service by flexible servers (SDU care) on heterogeneous customer (patient) types. Table 1 summarizes the distribution of where patients come from immediately preceding their SDU visit. Over 78% of patients in the SDU come from the ED or ICU. As such, our analysis will focus on these two patient cohorts. Specifically, we will focus on how transfer to the SDU impacts patients who are admitted to an inpatient unit from the ED as well as patients who are discharged from the ICU to lower levels of care. Figure 1 depicts these two transfer decisions that will be the heart of our empirical investigation. Given the contrasting routes to the SDU of these patients, it is reasonable to assume the impact of SDU care may differ substantially and our objective is to rigorously estimate the treatment effect of SDU care for these heterogeneous patient types.

Table 1: Distribution of Units Preceding the SDU

<table>
<thead>
<tr>
<th>Unit Preceding SDU</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED</td>
<td>60.93%</td>
</tr>
<tr>
<td>ICU</td>
<td>17.11%</td>
</tr>
<tr>
<td>Ward</td>
<td>13.88%</td>
</tr>
<tr>
<td>Post-Anesthesia Recovery Unit (PAR)</td>
<td>4.25%</td>
</tr>
<tr>
<td>Operating Room (OR)</td>
<td>3.58%</td>
</tr>
<tr>
<td>Other/Unknown</td>
<td>0.25%</td>
</tr>
</tbody>
</table>

Figure 1: Types of Admission Decisions

For each inpatient unit in each hospital, we use these patient flow data to derive hourly occupancy levels and we define its capacity as the maximum occupancy level over the time horizon of our study. Table 2 summarizes the capacity for each of the different levels of inpatient care in each hospital. While each level of care may have further divisions based on specific services, e.g., medical versus
surgical ICU, clinicians and administrators at the study hospitals indicate that it is widely accepted practice at their hospitals to consider the boundaries as somewhat fluid in the sense that if a medical service patient requires ICU care, but there are no medical ICU beds available, he will likely be cared for in the surgical ICU. We observe substantial heterogeneity across these hospitals; the SDU capacity varies from 11 to 32 beds and the number of ICU beds in a given hospital ranges from one half to twice the number in the SDU.

Table 2: Capacity of Various Inpatient Units in Terms of Number of Beds

<table>
<thead>
<tr>
<th>Hosp</th>
<th>ICU</th>
<th>SDU</th>
<th>Ward</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>24</td>
<td>61</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>25</td>
<td>76</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>14</td>
<td>77</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>19</td>
<td>76</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>24</td>
<td>78</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
<td>19</td>
<td>124</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>20</td>
<td>145</td>
</tr>
<tr>
<td>8</td>
<td>26</td>
<td>27</td>
<td>110</td>
</tr>
<tr>
<td>9</td>
<td>31</td>
<td>11</td>
<td>188</td>
</tr>
<tr>
<td>10</td>
<td>32</td>
<td>32</td>
<td>100</td>
</tr>
</tbody>
</table>

Our dataset also contains information about patient characteristics such as age, gender, admitting diagnosis and three different severity scores. One score (LAPS2) is based on lab results taken 72 hours preceding hospital admission and the second (COPS2) is based on comorbidities, such as diabetes, that may complicate patient recovery. These severity scores are assigned at hospital admission and are not updated during the hospital stay (more details on these scores can be found in [Escobar et al., 2008, 2013]). The third severity score is the simplified acute physiology score 3 (SAPS3), which is a common severity score used exclusively for ICU patients (see, e.g., [Strand and Flaat, 2008, Mbongo et al., 2009, Christensen et al., 2011]).

1.2.1. Data Selection

Since we study two different transfer decisions (from the ED and from the ICU), we form two separate patient cohorts: an ICU Cohort and an ED Cohort. Our data selection process is depicted in Figure 2. Because we use the patient flow data to determine the occupancy level (and capacity)
for each unit, we first restrict both of our cohorts to the 12 months in the center of the 1.5 year
time period in order to avoid censored estimates. A patient’s admission category is defined as
a combination of whether or not they were admitted through the ED, and whether they were
admitted to a medical or surgical service resulting in 4 categories: ED-medical, ED-surgical,
non-ED-medical, or non-ED-surgical. We primarily focus on patients who are admitted via the ED
to a medical service for two major reasons. First, this group is the largest, consisting of about 60%
of the patients treated in these hospitals, and is similar to the cohort considered in Kim et al. (2015).
Second, the care pathways of surgical patients tend to be fairly standardized (e.g., Gustafsson et al.
patients, which is the larger of the two surgical groups. In contrast, the care pathways of ED-medical
patients are more variable. It is this variability we will leverage in our identification strategy (see
Sections 1.3 and 1.5).

ICU Cohort

Many SDUs are designed as true ‘step-down units’, where patients can only be admitted following
ICU discharges (e.g., Eachempati et al. 2004). Moreover, the ICU is the second most frequent unit
from which SDU patients are transferred. Thus, our first cohort considers patients discharged from
the ICU to either the SDU or ward. To form the ICU Cohort, we consider patients who are admitted
to the ICU at least once during their hospital stay. For each patient, we focus on the initial ICU
admission within each hospitalization. We exclude patients who die in the ICU or are discharged
directly home from the ICU, since there is no decision about whether to route these patients to the
SDU or ward following ICU discharge.2

ED Cohort

Over 60% of SDU patients are admitted from the ED. For these patients, we consider the ED to
inpatient unit admission decision. The three possible units a patient can be admitted to are the ICU,

2We consider analysis including these patients in our robustness checks.
the SDU, or the Ward. We exclude the less than 5% of ED-medical patients who go directly to the Operating Room (OR) or Post-Anesthesia Recovery unit (PAR) from the ED.

Table 3 provides some summary statistics of these two cohorts. The SDU introduces a third level of care that, ideally, will be used to treat moderate to low-severity patients, but not high-severity patients. Our goal is to understand how service in this unit impacts quality of service, as measured by patient outcomes across different patient types. In doing so, we can gain a better understanding of the costs and benefits associated with utilizing a three levels of care structure to provide service to heterogeneous customers.
Table 3: Summary Statistics of Patient Demographics

<table>
<thead>
<tr>
<th>Variable</th>
<th>ED Cohort</th>
<th>ICU Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Age</td>
<td>67.68</td>
<td>17.53</td>
</tr>
<tr>
<td>Male</td>
<td>0.47</td>
<td>0.50</td>
</tr>
<tr>
<td>LAPS2</td>
<td>74.70</td>
<td>37.35</td>
</tr>
<tr>
<td>COPS2</td>
<td>46.18</td>
<td>44.21</td>
</tr>
<tr>
<td>SAPS3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED LOS (hrs)</td>
<td>1.46</td>
<td>2.20</td>
</tr>
<tr>
<td>Total LOS (hrs)</td>
<td>108.89</td>
<td>162.71</td>
</tr>
<tr>
<td>ICU LOS (hrs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOS before ICU (hrs)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: LAPS2 is a severity score based on lab results taken 72 hours preceding hospital admission. COPS2 is a severity score based on comorbidities. SAPS3 is a severity score used for ICU patients.

1.2.2. Patient Outcomes

We consider four patient outcomes: (1) in-hospital death (Mortality), (2) remaining hospital length-of-stay (HospRemLOS), (3) hospital readmission (HospReadm), and (4) ICU readmission (ICUReadm) for ICU patients.

The outcome HospRemLOS is defined as the remaining time spent in the hospital following the transfer decision. Thus, for patients in the ED Cohort, this will be their total inpatient LOS; for patients in the ICU Cohort, this will be the remaining time spent in the hospital following ICU discharge.

HospReadm2w is defined as hospital readmission within two weeks after leaving the hospital (e.g., see Doran et al. (2013) and Ouanes et al. (2012) which use these durations). In calculating hospital readmission rates, we exclude patients with in-hospital death. We also do robustness checks for different time windows for hospital readmission.

Following Brown et al. (2013) which aims to define reasonable time windows for ICU readmission, we consider ICUReadm2d (ICUReadm5d) which indicate ICU readmission within two (five) days following ICU discharge. This measure is studied only for the ICU Cohort. We also do robustness checks for different time windows for ICU readmission.
Table 4 summarizes these patient outcomes for the two cohorts.

Table 4: Summary Statistics of Patient Outcomes: Mean (Number of Observations or Standard Deviation for Continuous Variables)

<table>
<thead>
<tr>
<th>Outcome</th>
<th>ED Cohort</th>
<th>ICU Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICU</td>
<td>SDU</td>
</tr>
<tr>
<td></td>
<td>Mean (N/SD)</td>
<td>Mean (N/SD)</td>
</tr>
<tr>
<td>Mortality</td>
<td>0.12 (8,630)</td>
<td>0.04 (14,832)</td>
</tr>
<tr>
<td>HospRemLOS (days)</td>
<td>6.67 (11.51)</td>
<td>4.23 (5.89)</td>
</tr>
<tr>
<td>ICUReadm - 2 weeks</td>
<td>0.12 (7,629)</td>
<td>0.11 (14,269)</td>
</tr>
<tr>
<td>ICUReadm - 5 days</td>
<td>N/A</td>
<td>0.04 (3,832)</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

1.2.3. Hypotheses

As there are various flows of patients into the SDU, we expect the impact of admission to the SDU to vary across different patient types. In particular, there is evidence that SDU care may improve or degrade patient outcomes (e.g., [Zimmerman et al., 1995; Simchen et al., 2004]). Thus, we hypothesize that the SDU is beneficial or detrimental depending on patient type and severity—it will help moderate- to low-severity patients, but hurt high-severity patients. More formally, we outline our hypotheses below.

As SDUs were initially developed with the intent to provide a ‘step-down’ from the ICU, we expect that ICU clinicians use SDUs appropriately:

**Hypothesis 1 (ICU patients)** *Patients discharged from the ICU will have better outcomes (lower mortality and readmission rates and shorter LOS) if admitted to the SDU rather than the ward.*

For patients admitted from the ED, the impact of SDU care is likely to be more nuanced. Specifically, this is a highly heterogenous group. We will describe how we partition patients into low, medium, and high-severity groups in Section 1.5. The majority of patients admitted to the hospital from the ED do not go to the ICU ([Kim et al., 2015]). Thus, we expect that for most patients (i.e., low- and medium-severity patients), being treated in the SDU will either improve or have no impact on their outcomes. On the other hand, the sickest patients should be admitted to the highest level of care, so being admitted to the SDU is likely to result in worse outcomes. Note that in the
following, we assume that low-severity patients are rarely admitted to the ICU while high-severity patients are rarely admitted to the ward.

**Hypothesis 2 (Low-Severiry ED patients)** Low-severity patients admitted from the ED will have no worse, and possibly better outcomes (lower mortality and readmission rates and shorter LOS), if admitted to the SDU rather than the ward.

**Hypothesis 3 (Medium-Severiry ED patients)** Medium-severity patients admitted from the ED will have no worse, and possibly better outcomes (lower mortality and readmission rates and shorter LOS), if admitted to the SDU rather than the ward. On the other hand, they will have have no better, and possibly worse outcomes, if admitted to the SDU rather than the ICU.

**Hypothesis 4 (High-Severiry ED patients)** High-severity patients admitted from the ED will have worse outcomes (higher mortality and readmission rates and long LOS) if admitted to the SDU rather than the ICU.

1.3. ICU Cohort: Econometric Approach

We begin by explicitly stating our fundamental research question for the ICU cohort: Following ICU discharge, is SDU care associated with better patient outcomes than those for patients receiving ward care and, if so, what is the magnitude of the improvement? By exploring these questions, we will develop some insight into the value of differentiated levels of service (i.e., SDU versus ward) for one customer type (ICU patients). In Section 1.5, we expand our analysis to understand the impact of this level of service on additional patient types, providing insights into the role of customer differentiation.

1.3.1. Econometric Challenge: Endogeneity

Our objective is to utilize retrospective patient data to determine if ICU patients who are transferred to the SDU have better outcomes than those transferred to the ward. Because we are using retrospective data, an estimation challenge arises due to the fact that the routing decision following
ICU discharge is likely correlated with patient outcomes. To highlight this challenge, we start with
the following reduced form model for hospital LOS:

$$\log(HospRemLOS_i) = \beta X_i + \gamma ADMITSDU_i + \nu_{h(i)} + \epsilon_i$$

(1.1)

where $X_i$ is a vector of control variables including patient characteristics (e.g., age) and seasonal
factors (e.g., admission time of day), $ADMITSDU_i$ is an indicator variable that equals 1 if patient
$i$ is transferred directly to the SDU following ICU discharge, $h(i)$ is the hospital where patient $i$ is
treated, $\nu_{h(i)}$ is the hospital fixed effect and $\epsilon_i$ denotes the error term. See Table 14 in the Appendix
for more details on control variables. While we include controls for patient severity, unobservable
patient severity measures may be correlated with both $HospRemLOS$ and $ADMITSDU$. That
is, sicker patients are more likely to be transferred to the SDU than the ward, but are also more
likely to have bad outcomes. As such, our estimates for $\gamma$ may be biased and we may erroneously
conclude that going to the SDU hurts patients. To overcome this potential endogeneity bias, we
utilize an identification strategy using Instrumental Variables (IVs).

**Instrumental Variable**

A valid instrument should be 1) correlated with the endogenous variable, $ADMITSDU_i$, and 2) unrelated to the unobservable factors captured in $\epsilon_i$ which affect patient outcomes. We propose to use congestion in the SDU one hour before the ICU discharge as an IV. In particular, we define $SDUBusy_i$ as an indicator variable that equals one when the number of available beds in the SDU one hour prior to patient $i$’s discharge from the ICU is less than or equal to two, and zero otherwise. On average, about 11% patients are discharged from the ICU when the SDU is busy ($SDUBusy = 1$), though this varies quite a bit across hospitals (see Table 15).

When controlling for various patient characteristics in a Probit regression model, we also find at
the 0.1% significance level that when the SDU is busy, patients are less likely to go to the SDU.
In particular, we estimate that, on average, 21.14% percent of patients are routed to the SDU if

\[\text{We also do a number of robustness checks by considering different specifications of } SDUBusy_i.\]
\(SDUBusy = 1\) and this percentage increases to 35.91% if \(SDUBusy = 0\). Namely, a congested SDU is predicted to result in a 47% reduction in the likelihood of the SDU admission. Hence, condition 1 is satisfied.

We now consider Condition 2 and consider whether \(SDUBusy_i\) is uncorrelated with unobservable factors in patient outcomes captured in \(\epsilon_i\). Since we cannot examine unobservable measures, we use patient severity, \(SAPS3\), as a proxy for those unobservable factors. In particular, we perform a two-sample Kolmogorov-Smirnov test (see Gibbons and Chakraborti [2011] for details) to test the hypothesis that the distribution of \(SAPS3\) for patients who are discharged from ICU when \(SDUBusy = 1\) is not statistically different to that when \(SDUBusy = 0\). The p-value for the combined Kolmogorov-Smirnov test is 0.136. Thus, we cannot reject the null hypothesis and believe that patients who are discharged from the ICU when \(SDUBusy = 1\) are statistically similar to patients who are discharged from the ICU when \(SDUBusy = 0\). For completeness, we also check this for the LAPS2 score, which is assigned at the time of hospital admission. The p-value of the combined Kolmogorov-Smirnov test is 0.334.

Kc and Terwiesch [2012] demonstrates that ICU congestion could result in early discharge, which could, in turn, affect the routing decision of ICU patients. While ICU congestion has been used as an IV in a number of hospital studies (e.g., Kc and Terwiesch [2012], Kim et al. [2015]), we find that ICU congestion is not a valid IV. This is because the impact of ICU congestion does not exhibit a consistent effect on routing post-ICU patients, i.e., a congested ICU could result in both a higher and a lower percentage of patients being admitted to the SDU depending on a patient’s severity score. Moreover, we find that the ICU congestion is correlated with a patient’s \(SAPS3\) and LAPS2 score.

We also considered using a number of additional instrumental variables. Specifically, we considered a measure of the average severity of other patients in the ICU, a measure of how the discharged patient compares to the severity of other patients in the ICU, and a measure of severity for the most recently discharged ICU patient. We find that all of these measures are correlated with the \(SAPS3\)
and LAPS2 scores, suggesting they may also be correlated with unobservable measures of severity, thereby invalidating these variables as potential instruments.

### 1.3.2. Econometric Model

**Continuous outcome models**

We now present our estimation model for our continuous outcome, \( \text{HospRemLOS} \). Since the ICU to SDU routing decision, \( \text{ADMITSDU}_i \), is a binary variable, we model the ICU discharge decision via a latent variable model.

\[
\text{ADMITSDU}^*_i = X_i \theta + \alpha \text{SDUBusy}_i + \omega_{h(i)} + \xi_i, \\
\text{ADMITSDU}_i = \begin{cases} 
1 \{\text{ADMITSDU}^*_i > 0\}, 
\end{cases} \\
\log(\text{HospRemLOS}_i) = X_i \beta + \gamma \cdot \text{ADMITSDU}_i + \delta \cdot \text{AvgOccVisited}_i + \nu_{h(i)} + \varepsilon_i, 
\]

(1.2)

where \( \text{ADMITSDU}^*_i \) is a latent variable which represents the propensity towards SDU admission; \( X_i \) is a vector of control variables for patient information; \( \omega_{h(i)} \) is the hospital fixed effect; and, \( \xi_i \) represents unobservable factors that affect the routing at ICU discharge. For the outcome equation, \( \nu_{h(i)} \) is the hospital fixed effect; and \( \varepsilon_i \) captures unobservable factors that affect patient outcomes.

Because congestion during a patient’s hospital stay could impact the patient’s outcomes (see Kuntz et al. 2015 and Kc and Terwiesch 2012), we also control for the daily average occupancy level, denoted as \( \text{AvgOccVisited}_i \), patient \( i \) experiences for all inpatient units s/he is admitted to after leaving the ICU and before leaving hospital. We also conduct robustness checks for different specifications of occupancy during the stay, as well as with such a control excluded. Kim et al. (2015) provides additional discussion regarding the necessity of such a control.

The error terms \( (\xi_i, \varepsilon_i) \) in (1.2) may be correlated to model the endogeneity between the routing decision at ICU discharge and the patient outcome. We assume that \( (\xi_i, \varepsilon_i) \) follows a Standard Bivariate Normal distribution with correlation coefficient \( \rho \). This model can be jointly estimated
using a treatment effect model via Full Maximum Likelihood Estimation (FMLE) \cite{Greene2012}. A likelihood ratio test of null $\rho = 0$ can be used to test the presence of endogeneity.

**Discrete outcome models**

For the binary outcomes ($Mortality$, $HospReadm$, $ICUReadm$), we modify Equation (1.2) by replacing the continuous patient outcome with a probit model. Specifically, we have:

\[
ADMITSDU_i^* = X_i \theta + \alpha SDUBusy_i + \omega_{h(i)} + \xi_i,
\]

\[
ADMITSDU_i = 1 \{ADMITSDU_i^* > 0\},
\]

\[
y_i^* = X_i \beta + \gamma \cdot ADMITSDU_i + \delta \cdot AvgOccVisited_i + \nu_{h(i)} + \varepsilon_i,
\]

\[
y_i = 1 \{y_i^* > 0\}
\]

where $y_i^*$ is a latent variable which represents the propensity for the outcome. Similar to before, we assume that $(\xi_i, \varepsilon_i)$ follows a Standard Bivariate Normal distribution with correlation coefficient $\rho$. This Bivariate Probit model can be jointly estimated via FMLE \cite{Cameron1998, Greene2012}. The presence of endogeneity can be tested through a likelihood ratio test of null $\rho = 0$.

For ICU readmission, we modified $AvgOccVisited_i$ to be the daily average occupancy level that patient $i$ experiences in all inpatient units s/he is admitted to between two consecutive ICU admissions.

1.3.3. **Impact of Congestion on ICU LOS**

\cite{Kc2012} found evidence that when ICUs are highly congested, current ICU patients may be demand-driven discharged, in order to accommodate incoming demand of more severe patients. \cite{Kim2015} found that patients admitted to a medical service from the ED do not seem to be susceptible to such demand-driven discharges. While we look at a similar group of patients to \cite{Kim2015}, one potential concern is that we only consider patients treated in
hospitals with SDUs, while Kim et al. (2015) includes hospitals with SDUs as well as those without. Thus, it is possible that the presence of an SDU makes it more likely for medical patients who were admitted to the hospital via the ED and are being treated in the ICU to be demand-driven discharged; thus, making it possible that these types of discharges occur in our dataset. A patient who is demand-driven discharged is by definition, discharged earlier than under ordinary circumstances and therefore more critical than if he were discharged later at a more appropriate time. So such a patient is more likely to be admitted to the SDU, but also more likely to have bad outcomes. If this were the case, this could cause a downward bias of our results.

To check this, we estimated the following reduced form model:

$$\log(\text{ICULOS}_i) = \eta X_i + \kappa \text{ICUBusy}_i + v_i$$

(1.4)

to explore whether ICU LOS is reduced when the ICU is busy. We estimate \(\kappa\) to be \(-0.05\) with standard error 0.04. Thus, consistent with Keenan et al. (1998) and Kim et al. (2015), we do not find evidence that patients are demand-driven discharged. To dig a little deeper, we examined whether the SDU congestion had an impact on whether patients are demand-driven discharged. To do this, we enhance our regression model to include a measure of SDU congestion:

$$\log(\text{ICULOS}_i) = \eta X_i + \kappa \text{ICUBusy}_i + \phi \text{SDUBusy}_i + \psi (\text{ICUBusy}_i \times \text{SDUBusy}_i) + v_i$$

(1.5)

In particular, we would expect demand-driven discharges to be most common when the ICU is busy and the SDU is not. Table 5 summarizes these results with the base case of both the ICU and SDU not being busy (81.5% of time). We find that the coefficients have very large standard errors and are not statistically significant. While it is possible that lack of statistical power is the reason we do not find evidence to support the hypothesis that a busy ICU may result in demand-driven discharges, we find that our sample size would need to be larger than 350,000 for the estimated coefficients to be statistically significant when using the approach in Gelman and Hill (2006).
Table 5: Effect of ICUBusy and S Dubusy on ICU LOS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ICUBusy</th>
<th>S Dubusy</th>
<th>Estimate (SE)</th>
<th># Observations: Total = 11,058</th>
</tr>
</thead>
<tbody>
<tr>
<td>κ</td>
<td>1</td>
<td>0</td>
<td>-0.057 (0.040)</td>
<td>855</td>
</tr>
<tr>
<td>φ</td>
<td>0</td>
<td>1</td>
<td>-0.039 (0.039)</td>
<td>1,056</td>
</tr>
<tr>
<td>ψ</td>
<td>1</td>
<td>1</td>
<td>-0.034 (0.096)</td>
<td>136</td>
</tr>
</tbody>
</table>

Notes: Standard error in parentheses. + (p < 10%), *(p < 5%), **(p < 1%), *** (p < 0.1%).

Our IV analysis is based on the evidence that a busy SDU decreases the likelihood of SDU admission. However, it is also possible that patients may stay longer in the ICU when the SDU is busy, making them more stable upon discharge from the ICU and potentially biasing our results. To test this hypothesis, we ran the reduced form model in Equation (1.4), but with S Dubusy as an explanatory variable. We find the coefficient for S Dubusy to be -0.02 with standard error 0.03. This is consistent with the results in Table 5 which suggests that the relationship between a busy SDU and ICU LOS is not statistically significant. As an additional check, we ran a hazard rate model to examine the impact of S Dubusy after controlling for patient characteristics, seasonality, and hospital fixed effects. Again, we see that a busy SDU does not have a statistically significant effect on the likelihood of ICU discharge. Thus, we do not find evidence to support that the busy-ness of the SDU impacts ICU LOS.

1.4. ICU Cohort: Results

We start by exploring the impact of SDU care on patients being discharged from the ICU. Because we jointly estimate the SDU admission decision and patient outcomes, using FMLE, the impact of S Dubusy may vary slightly for different outcomes. That said, we observe that the differences are very minor. For illustrative purposes, we note that the coefficient for the impact of S Dubusy in the Mortality model is −0.5110 with standard error 0.0503 and p-value < 0.1%.

As we are primarily interested in estimating the causal effects of SDU admission on patient outcomes, we report only the coefficient of SDU admission on the patient outcomes, i.e., γ in (1.2) and (1.3). Table 6 summarizes the relationship between SDU admission right after ICU discharge and patient outcomes. The sign of SDU admission is negative and statistically significant in all
Table 6: Estimated Effect of SDU Admission Following ICU Discharge ($\gamma$) on Patient Outcomes and Correlation Between Error Terms ($\rho$) for the Admission Decision and Patient Outcomes: $N = 11,058$

<table>
<thead>
<tr>
<th>Outcome</th>
<th>With IV</th>
<th>Without IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\gamma$ (SE)</td>
<td>Predicted Outcome</td>
</tr>
<tr>
<td></td>
<td>$\hat{P}_{SDU\ Busy=0}$</td>
<td>$\hat{P}_{SDU\ Busy=1}$</td>
</tr>
<tr>
<td>Mortality</td>
<td>-0.60** (0.22)</td>
<td>8.24%</td>
</tr>
<tr>
<td>log(HospRemLOS)</td>
<td>-0.35*** (0.10)</td>
<td>3.77%</td>
</tr>
<tr>
<td>ICUReadm2d</td>
<td>-0.51** (0.20)</td>
<td>5.22%</td>
</tr>
<tr>
<td>ICUReadm5d</td>
<td>-0.51** (0.18)</td>
<td>8.18%</td>
</tr>
<tr>
<td>HospReadm2w</td>
<td>-0.43* (0.21)</td>
<td>14.02%</td>
</tr>
</tbody>
</table>

Notes: Standard error in parentheses. + ($p < 10%$), * ($p < 5%$), ** ($p < 1%$), *** ($p < 0.1%$).

Predicted outcome: $\hat{P}_{SDU\ Busy=0}$ - Average predicted outcome if the SDU was never busy
$\hat{P}_{SDU\ Busy=1}$ - Average predicted outcome if the SDU was always busy.
Predicted HospRemLOS (days) is shown instead of log(HospRemLOS).

Outcome measures, suggesting that routing an ICU discharge to the SDU is associated with improved patient outcomes. To get a rough estimate of the magnitude of the effects we have estimated, we also use our estimation results to predict patient outcomes under two extreme scenarios: (i) the SDU has ample capacity and is never busy (referred to as $\hat{P}_{SDU\ Busy=0}$) versus (ii) the SDU is always busy ($\hat{P}_{SDU\ Busy=1}$). There are some patients who are stable enough that even if there were ample capacity in the SDU, they would not be admitted to the SDU. Instead they would be sent to the ward, irrespective of the SDU bed availability. Using the first stage of our models, we estimate the likelihood of SDU admission depending on whether the SDU is busy or not, and then use our second stage results to estimate the likely patient outcome. We find that, on average, availability of SDU care is associated with significant improvements in patient outcomes: the relative reduction is 17% in the likelihood of in-hospital death, 0.28 fewer remaining hospital days, 18% (17%) in the likelihood of ICU readmission within 2 (5) days, and 8% in the likelihood of hospital readmission within 2 weeks. Based on the size of our cohort, we estimate eliminating SDU busy-ness would translate into annual savings of 187 lives, 3,096 hospital days, 128 (182) ICU readmissions in 2 (5) days, and 137 hospital readmissions in 2 weeks (weekly savings of 3.6 lives, 59.5 hospital days, 2.5 (3.5) ICU readmissions, and 2.6 hospital readmissions) aggregated across the 10 hospitals.

Our empirical findings also suggest strong evidence of an endogeneity bias between the routing following ICU discharge and patient outcomes. The $p$-value of the likelihood ratio test with null
hypothesis that the correlation between the two error terms in our model $\rho = 0$ is small, as seen in Table 6, implying a strong correlation between the routing at ICU discharge and patient outcomes. Ignoring this endogeneity tends to result in underestimates of the benefit of SDU care and could result in a qualitatively different insight; see the column titled “Without IV”.

1.4.1. Robustness Checks

We now describe a number of robustness checks for our main results. First, we tried different specifications of control variables. Recall that, some of our control variables – age, severity scores (LAPS2, COPS2, SAPS3), length-of-stay at ICU, and length-of-stay before ICU admission – are modeled as spline variables to account for their possible non-linear effects on the ICU to SDU routings and patient outcomes. We repeated the analysis with different specifications, including changing the number of cutoffs and values of these cutoffs. Our results are qualitatively similar to these changes.

The second robustness check we did is with respect to specifications for the congestion experienced by a patient during the hospital visit ($Avg\text{OccVisited}$ in Equations (1.2) and (1.3)). We considered specifications which exclude this control as well as ones that examine the maximum occupancy in any unit during a patient’s hospital stay. All specifications yield similar results to those reported in Table 6.

Another factor which could be impacting our results is “do not resuscitate (DNR)” orders, which are patients’ end-of life wishes not to undergo Cardiopulmonary resuscitation (CPR) or advanced cardiac life support if their heart were to stop or they were to stop breathing. In speaking with intensivists, we learned it is possible that patients with DNRs are more likely to be sent to the ward, but also may be more likely to die, resulting in an overestimate of the effect of SDU care. Unfortunately, we do not have access to patients’ DNR status, so cannot control for this. That said, DNR orders only represent 9% of ICU patients [Jayes et al., 1993], so this is likely to affect only a small percentage of patients. Additionally, there is evidence that DNR orders do not change the quality of care [Baker et al., 2003]. We do not expect DNR orders to impact our results for hospital
readmission since we exclude patients who died in hospital in this model. For the LOS models, we also considered the robustness of our results to including patients with in-hospital death. We find that our results are very robust.

We also considered alternative specifications for the length of time window for readmission. For ICU readmission, we varied the time window of the ICU readmission from time of ICU discharge from 2 to 7 days and also during any time frame during the same hospital stay. Only the results for ICU readmission within 2 days were statistically significant, though the sign of the coefficient was negative in all models. For hospital readmission, we consider hospital readmission within 1 week, 2 weeks, and 30 days after a patient is discharged from the hospital. We found that while SDU admission is associated with lower hospital readmission risk, the effect is weaker when the elapsed time between two consecutive hospital stays is longer.

**Definition of our IV**

We also consider various definitions of a busy SDU. First, we considered different cutoffs for the number of available beds, ranging from one bed to four. On average, the percentage of patients, who are discharged from the ICU when the SDU is congested, varies from 34% to 3% when the cutoff is decreased from four beds to one (Table 15). The capacity of the SDU was defined as the maximum occupancy level over the 12-month time horizon in our study. While capacity changes in the hospitals we study are very rare, we also allow for the bed capacity (defined as the maximum occupancy level) to change over time. Specifically, we define a time-varying capacity as the maximum occupancy level over three non-overlapping 4-month periods during the total 12-month time horizon.

Note that while we find our IV to be statistically significant based on various definitions of bed capacity, it can be very challenging to accurately determine the number of beds available in a unit. This is because capacity depends on multiple factors including the number of physical beds, but also the number of nurses and physicians available to staff them. As such, we also considered alternative measures of SDU congestion based on percentiles of the SDU occupancy level. We did
this using a binary variable indicating whether the occupancy level exceeds a threshold percentile as well as a piece-wise linear spline to potentially model non-linear effects of SDU congestion on the SDU admission decision. Next, we considered different time lines for when SDU congestion was measured: 1 hour (main specification), 2 hours and 6 hours before ICU discharge.

We find that the results for the mortality and LOS models to be very robust to the various specifications with the coefficients all negative and all with the same order of magnitude. All coefficient estimates have a $p$-value $< 0.05$, with most having a $p$-value $p < 0.001$. Interestingly, the coefficient estimates for the ICU and hospital readmission models are all negative; however, the statistical significance of the coefficient estimates varies substantially, with some specifications indicating a $p$-value $< 0.001$ and others not being significant, even at the level of $p$-value $< 0.1$. Thus, while our mortality and LOS results are quite robust, the readmission results do not seem to be.

**Patients discharged out of the hospital**

In all of our analysis for the ICU cohort, we focused on patients who were discharged from the ICU to the SDU or ward. While the majority of patients (83.82%) go to one of these units, a number of patients are actually discharged directly out of the hospital from the ICU (see Table 16 in the Appendix). Not surprisingly, patients who are discharged out of the hospital directly from the ICU appear to be healthier (lower severity scores and younger) than those admitted to an inpatient unit following ICU discharge. We find that if we include all patients who are discharged alive from the ICU (to the SDU, ward or out of the hospital) instead of just those discharged to the SDU or ward, the busy-ness of the SDU still has a statistically significant effect on the likelihood of SDU admission following ICU discharge. In this analysis, we found evidence that patients are 1.77% more likely to be discharged home alive when the SDU is busy ($p < 0.05$). As such, excluding these patients from our analysis may bias our estimates to make the SDU seem more beneficial than it is because the patients who end up staying in the hospital are sicker and SDU treatment likely benefits them more. With this in mind, we re-ran our ICU and hospital readmission models including patients discharged out of the hospital alive. Note that we do not examine our mortality
and LOS models as these are inpatient outcomes, and a patient who is discharged out of the hospital alive will, by definition, have $\text{death}_i = 0$ and $\text{HospRemLOS} = 0$. We find that the coefficient estimates for our readmission models are negative, but not statistically significant. This is consistent with our other specifications which suggest that the readmission results are not very robust.

1.5. ED Cohort: Econometric Approach

In this section, we study the routing decision regarding the ED Cohort. We aim to empirically estimate how SDU admission immediately following transfer from the ED affects patient outcomes, comparing to ED patients who are transferred to the ICU or ward. Here, a similar estimation challenge arises. Routing decisions are associated with patient severity and, thus, with patient outcomes.

Kim et al. (2015) examined this problem in the context of admitting patients to the ICU from the ED. In that paper, the goal was to estimate the impact of admitting a patient to the highest level of care, i.e., the ICU versus elsewhere. In contrast, our objective is to understand the impact of admitting patients to an intermediary level of care, the SDU. In contrast to the ICU case, it is possible that the impact of SDU care could be positive, neutral or even negative. For instance, high-severity patients who should be admitted to the ICU, but are instead admitted to the SDU may experience worse outcomes as a result. On the other hand, SDU care may have no impact or even benefit low-severity patients who would traditionally be cared for in the ward. There are limited objective standards for who should be treated in the ICU (see Task Force of the American College of Critical Care Medicine, Society of Critical Care Medicine 1999 and Kim et al. 2015), let alone for the SDU (Nasraway et al. 1998). Thus, such categorizations of patients are likely to be highly varied across different physicians. As such, we take a data-driven approach to stratifying patients by severity. Such an approach could be useful in other service settings where precisely defining a customer’s type is a challenging, but necessary, step toward determining the costs and benefits associated with service by different server types.
1.5.1. Severity Categorization

In order to estimate the impact of SDU care for patients admitted from the ED, we categorize patients based on their severity and study each severity group separately. Specifically, we aim to identify a ‘low-severity’ cohort, for which the decision is to admit patients to either the ward or SDU, and a ‘high-severity’ cohort for which the decision is to admit to either the SDU or ICU. One can also consider a ‘medium-severity’ cohort whose patients can be admitted to any one of the three levels of care. Certainly, it seems reasonable to expect the decision to admit a patient to the SDU will have a different impact on patients of varying severity.

We begin by considering how patient level characteristics influence whether a patient is admitted to the ICU, SDU or ward from the ED. Specifically, we use our data to estimate an Ordered Probit regression model using only patient characteristics.

\[ Tx_i = \begin{cases} 
Ward, & \text{if} \ Tx_i^* \leq t_1 \\
SDU, & \text{if} \ t_1 < Tx_i^* \leq t_2 \\
ICU, & \text{if} \ t_2 < Tx_i^* 
\end{cases} \]

and

\[ Tx_i^* = X_i' \theta + \xi_i, \quad (1.6) \]

where \( X_i' \) is a vector of control variables for patient characteristics and \( \xi_i \) represents unobservable factors.

We use the observed latent variable \( \hat{Tx}_i^* = X_i' \theta \) to define each patient’s severity. Intuitively, \( \hat{Tx}_i^* \) is a linear transformation of patient characteristics into a single continuous variable which can be interpreted as a measure for the desired amount of care for the patient. The larger the value of \( \hat{Tx}_i^* \), the more likely the patient will be routed to higher level units, e.g., the ICU; the lower the value, the more likely a patient will be routed to the ward.

We differentiate patient severity groups by partitioning the \( \hat{Tx}_i^* \) space with thresholds. In theory, \( t_1 \) and \( t_2 \) from (1.6) partition the \( Tx_i^* \) space into patients who will be routed to the Ward, SDU, and ICU, so that patients with \( Tx_i^* \leq t_1 \) could be classified as low-severity patients and patients with \( Tx_i^* > t_2 \) could be classified as high-severity patients. However, because we do not observe \( \xi_i \),
we are only able to observe an estimate $\hat{T}_x^*$, instead of $Tx^*_i$. Thus, some patients with $\hat{T}_x^* \leq t_1$ will be routed to the SDU, or even the ICU. Similarly, patients with $\hat{T}_x^* > t_2$ may be routed to the SDU or ward. Increasing $t_2$ will increase the proportion of patients with $\hat{T}_x^* > t_2$ who are routed to the ICU and simultaneously decrease the proportion who are routed to the ward. Similarly, decreasing $t_1$ will increase the proportion of patients with $\hat{T}_x^* \leq t_1$ being admitted to the ward and decrease the proportion being admitted to the ICU. Of course, this also comes at the cost of reducing the number of patients which satisfy these two criteria. Thus, we define the cutoffs to balance increasing the proportion of patients in the high (low) severity group who are routed to the ICU (ward) versus maintaining large enough patient cohorts to allow for meaningful statistical analysis.

The tradeoff we are concerned with is close thresholds lead to increases in patient spill-over into the high or low-severity groups, resulting in patients who do not comply with our instrument, versus far thresholds which reduce sample sizes, resulting in less statistical power. We use a data-driven approach and find that setting thresholds at the $95^{th}$ and $60^{th}$ percentiles of the distribution of $\hat{T}_x^*$ seen in the data achieve this delicate balance. In Section 1.6 we discuss robustness checks using different thresholds.

We expect that when the SDU is congested, patients will be less likely to be admitted (see Section 1.3). Thus, we examine where patients are admitted when the SDU is busy, defined as done in Section 1.3. Figure 3 shows the proportion of high and low-severity patients admitted to each unit, while Table 7 summarizes these results. Note that the ICU and SDU congestion have a correlation coefficient of 0.08, so the busyness of the ICU does not factor substantially into these results. Specifically, we ran $t$-tests comparing the proportion of patients admitted to each level of inpatient unit when the SDU is busy versus not busy. As we can see, when the SDU is busy, low-severity patients will be rerouted to the ward ($p < 0.001$), rather than the ICU ($p = 0.327$). Conversely, when the SDU is busy, high-severity patients tend to be rerouted to the ICU ($p = 0.002$), rather than the ward ($p = 0.212$). These results are suggestive that these severity categorizations are reasonable for our purposes.
Note that one can also define a ‘medium severity’ group as patients with $t_2 \leq \hat{T}_{x^*_i} < t_1$. While a busy SDU does decrease the likelihood of SDU admission (Figure 3(b) and Table 7), the challenge with this cohort is that some patients who are discouraged from being admitted to the SDU will be admitted to the ICU, while others will be admitted to the general ward. Certainly, being bumped to a higher versus lower level of care will have a substantial impact on patient outcomes. As seen in Figure 3(b) and Table 7, there is a heterogenous effect of $SDUBusy$ on these patients, whereas high-severity patients are consistently bumped up to the ICU and low-severity patients are consistently bumped to the ward. This suggests that within the medium-severity group a mix of high- and low-severity patients are being admitted to the SDU, so SDU admission can be beneficial or detrimental. By grouping these patients together into a medium severity classification, we cannot tease out the true impact of SDU admission. Still, for completeness, we will include results for this group of patients.

![Graphs showing proportions of ED patients routed to ICU, SDU, and ward](image)

(a) High-Severity Patients  
(b) Medium-Severity Patients  
(c) Low-Severity Patients

Figure 3: Proportions of ED Patients Who Are Routed to the ICU, SDU, and Ward When $SDUBusy = 1$ vs. $SDUBusy = 0$

As summarized in Table 8, for the high-severity group, 54.9% are admitted to the ICU, 20.76% to the SDU, and 24.35% to the ward. For low-severity patients 4.65%, 14.73% and 80.62% are admitted to the ICU, SDU, and ward, respectively. We can see that even with our classifications, some high (low) severity patients will still be admitted to the ward (ICU). In order to focus on the impact of SDU admissions on patient outcomes, we exclude high (low) severity patients who are
Table 7: Proportions of ED Patients Who Are Routed to the ICU, SDU, and Ward When $SDUBusy = 1$ vs. $SDUBusy = 0$ and Results of $t$-Tests That Compare the Difference in Routing Proportions

<table>
<thead>
<tr>
<th>Severity</th>
<th>ICU</th>
<th>SDU</th>
<th>Ward</th>
<th>ICU</th>
<th>SDU</th>
<th>Ward</th>
<th>p-value of $t$-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Severity</td>
<td>4.62%</td>
<td>15.97%</td>
<td>79.14%</td>
<td>4.94%</td>
<td>5.23%</td>
<td>89.93%</td>
<td>0.327 &lt; 0.001 &lt; 0.001</td>
</tr>
<tr>
<td>Medium Severity</td>
<td>16.90%</td>
<td>31.05%</td>
<td>52.04%</td>
<td>21.84%</td>
<td>12.64%</td>
<td>65.51%</td>
<td>0.013 &lt; 0.001 0.001</td>
</tr>
<tr>
<td>High Severity</td>
<td>53.60%</td>
<td>22.39%</td>
<td>24.01%</td>
<td>63.38%</td>
<td>10.06%</td>
<td>26.56%</td>
<td>0.002 &lt; 0.001 0.212</td>
</tr>
</tbody>
</table>

Notes: Severity thresholds, $t_1$ and $t_2$, defined by $95^{th}$ and $60^{th}$ percentiles of $\hat{T}_{x_i}$.

routed to the ward (ICU). For the medium-severity group, we consider patients admitted to all three levels of care as it is not clear whether the ICU or ward is the ‘more desirable’ unit if the SDU is not available. Tables 9 and 10 report summary statistics of patient demographics and outcomes for each severity group.

Table 8: Routing Statistics of Patients for Different Severity Groups

<table>
<thead>
<tr>
<th>Unit following the ED</th>
<th>Low Severity</th>
<th>Medium Severity</th>
<th>High Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Percentage</td>
<td>Frequency</td>
</tr>
<tr>
<td>ICU</td>
<td>2,067</td>
<td>4.65</td>
<td>4,529</td>
</tr>
<tr>
<td>SDU</td>
<td>6,549</td>
<td>14.73</td>
<td>7,514</td>
</tr>
<tr>
<td>Ward</td>
<td>35,836</td>
<td>80.62</td>
<td>13,885</td>
</tr>
</tbody>
</table>

Notes: Severity thresholds, $t_1$ and $t_2$, defined by $95^{th}$ and $60^{th}$ percentiles of $T_{x_i}$.

Table 9: Summary Statistics of Patient Demographics for ED Cohort by Severity Classification

<table>
<thead>
<tr>
<th>Variable</th>
<th>Low Severity</th>
<th>Medium Severity</th>
<th>High Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Mean 67.48</td>
<td>SD 18.57</td>
<td>Min 18</td>
</tr>
<tr>
<td>Male</td>
<td>0.43</td>
<td>0.49</td>
<td>0</td>
</tr>
<tr>
<td>LAPS2</td>
<td>59.48</td>
<td>26.89</td>
<td>0</td>
</tr>
<tr>
<td>COPS2</td>
<td>41.96</td>
<td>41.21</td>
<td>0</td>
</tr>
<tr>
<td>ED LOS (hrs)</td>
<td>1.38</td>
<td>1.99</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Notes: LAPS2 is a severity score based on lab results taken 72 hours preceding hospital admission.
COPS2 is a severity score based on comorbidities. SAPS3 is a severity score used for ICU patients.
Severity thresholds, $t_1$ and $t_2$, defined by $95^{th}$ and $60^{th}$ percentiles of $\hat{T}_{x_i}$.

1.5.2. IV Justification

We are again faced with the econometric challenge of endogeneity bias. Our econometric model is very similar to that of (1.2) and (1.3). The main difference is that for low (high) severity patients, $ADMITSDU_i$ is equal to 1 if the patient is admitted to the SDU and 0 if to the ward (ICU).
Table 10: Summary Statistics of Patient Outcomes for ED Cohort by Severity Classification: Mean (Number of Observations or Standard Deviation for Continuous Variables)

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Low Severity</th>
<th>High Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDU</td>
<td>Ward</td>
</tr>
<tr>
<td></td>
<td>Mean (N/SD)</td>
<td>Mean (N/SD)</td>
</tr>
<tr>
<td>Mortality</td>
<td>0.02 (6,549)</td>
<td>0.02 (35,836)</td>
</tr>
<tr>
<td>HospRemLOS (days)</td>
<td>3.97 (5.85)</td>
<td>3.95 (5.21)</td>
</tr>
<tr>
<td>HospReadm - 2 weeks</td>
<td>0.10 (6,431)</td>
<td>0.10 (35,258)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Medium Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDU Mean (N/SD)</td>
</tr>
<tr>
<td>Mortality</td>
<td>0.04 (7,514)</td>
</tr>
<tr>
<td>HospRemLOS (days)</td>
<td>4.20 (5.15)</td>
</tr>
<tr>
<td>HospReadm - 2 weeks</td>
<td>0.12 (7,202)</td>
</tr>
</tbody>
</table>

Notes: Severity thresholds, \( t_1 \) and \( t_2 \), defined by 95\(^{th}\) and 60\(^{th}\) percentiles of \( Tx_i \).

For medium-severity patients, \( ADMITSDU_i \) is equal to 1 if the patient is admitted to the SDU and 0 if to the ward or ICU. Detailed descriptions of the covariates are shown in Table 14 in the Appendix. Similarly, we also control for \( AvgOccVisited_i \), i.e., the daily average occupancy level patient \( i \) experiences for all inpatient units s/he is admitted to after leaving the ED and before leaving hospital.

Similar to our models for the ICU Cohort, we consider using \( SDUBusy_i \) as an instrumental variable. Additionally, we consider using \( ICUBusy_i \) as an instrument as Kim et al. (2015) found that it is a good instrument when studying patients who are or are not admitted to the ICU, which is similar to our high-severity group. Specifically, we define \( SDUBusy_i \) (\( ICUBusy_i \)) as an indicator variable that equals one when the number of available beds in the SDU (ICU) one hour prior to patient \( i \)’s transfer from the ED is less than or equal to two, and zero otherwise. On average, the proportions of patients who are transferred from the ED when the SDU is busy and the ICU is busy are approximately 12\% and 6\%, respectively.

As discussed previously, in order for a variable to be a valid instrument, it has to be 1) correlated with the endogenous variable, \( ADMITSDU_i \), and 2) unrelated to the unobservable factors which affect patient outcomes. As seen in Table 7 when the SDU is busy, patients are less likely to be
admitted to the SDU. However, we find that ICU congestion does not appear to have a monotonic effect on SDU admission for low- or medium-severity patients. Specifically, we observe in Figure 4 that when we partition the (a) low- or (b) medium-severity patients into deciles of $\hat{Tx}_i^*$, ICU congestion increases the percentage of SDU admissions for some patients, while it has no effect or even decreases the percentage of SDU admissions for other patients. Therefore, we conclude that $ICUBusy_i$ is not a valid instrument for low- or medium-severity patients. We see these effects more concretely when we analyze a Probit regression model, which controls for various patient characteristics and operational controls. We find with 0.1% significance level that SDU congestion reduces the likelihood of SDU admission for both low, medium and high-severity patients, and that ICU congestion increases the chance of SDU admission for only high-severity patients. The impact of ICU congestion for low- and medium-severity patients is not statistically significant.

![Figure 4: Percentage of (a) Low- or (b) Medium-Severity Patients Admitted to the SDU from the ED When the ICU Is Busy (ICUBusy = 1) or Not (ICUBusy = 0) for Varying Levels of Severity as Measured by Deciles of $\hat{Tx}_i^*$, Given Patients Are Classified as Low Severity: $\hat{Tx}_i^* \leq t_1$](image)

We next examine whether our instruments are correlated with observable measures of severity. We again perform a two-sample Kolmogorov-Smirnov test to test the hypothesis that the distribution of LAPS2 is not statistically different when $SDUBusy = 1$ ($ICUBusy = 1$) from that when $SDUBusy = 0$ ($ICUBusy = 0$). For low-severity patients, the p-value for the Kolmogorov-Smirnov test is 0.135, thus, we conclude that patients who leave the ED when
SDUBusy = 1 are statistically similar to those who leave the ED when SDUBusy = 0. For medium-severity patients, the p-value for the Kolmogorov-Smirnov test is 0.120. For high-severity patients, the p-values are 0.141 and 0.358 for SDUBusy and ICUBusy, respectively. Therefore, our models for low- and medium-severity patients use SDUBusyi as an instrument, while both SDUBusyi and ICUBusy are used in the models for high-severity patients.

**Additional Instruments**

Apart from the congestion in the ICU and the SDU, we also consider other potential behavioral IVs discussed in [Kim et al. (2015)](https://doi.org/10.1016/j.jclinepi.2015.05.012). The first factor is RecentDischargeSDU, which accounts for the number of all SDU discharges in the 3-hr window before patient i’s admission to the first inpatient unit. The second behavioral factor, RecentAdmissionSDU, accounts for the number of SDU admissions in the 3-hr window before patient i’s admission to the first inpatient unit. To define RecentDischargeSDU and RecentAdmissionSDU, we normalize the number of discharges or admissions by the SDU capacity of each hospital. The third factor, LastAdmitSeveritySDU, measures the severity of the last patient admitted to the SDU from the ED. We also consider RecentDischargeICU, RecentAdmissionICU, LastAdmitSeverityICU, which are defined the same way but instead involve the ICU. Most of these variables demonstrate a heterogeneous impact on the SDU admission decision; for instance, amongst low-severity patients, RecentAdmissionSDU will increase the likelihood of SDU admission, while it will decrease likelihood for other patients. We find that only RecentAdmissionICU is a valid instrument and is valid only for high-severity patients. However, we do not include this as a third IV for high-severity patients in our main specifications because the results are similar.

### 1.6. ED Cohort: Results

We now present our main results for our ED cohort on the impact of SDU admission on patients being admitted to an inpatient unit from the ED. We start with the two patient cohorts for which the routing decision is more straight-forward (low- and high-severity patients). Then, for completeness, we include the results for the medium-severity patients.
1.6.1. Low Severity

For low-severity patients, a busy SDU is associated with a decrease in likelihood of SDU admission. For the mortality model, the coefficient on $SDUBusy_i$ is $-0.5117$ with standard error $0.0376$ and $p$-value $< 0.1\%$. The results are similar for the other patient outcome models.

Table [11] summarizes our results. We find that SDU care may benefit low-severity patients. Specifically, we find that SDU care is associated with lower mortality rate and shorter hospital remaining length-of-stay, as seen in the negative sign of SDU admission coefficient. We also present the predicted patient outcomes under two extreme scenarios: (i) the SDU is never busy ($\hat{P}_{SDUBusy=0}$) and (ii) the SDU is always busy ($\hat{P}_{SDUBusy=1}$). Our results indicate that, on average, availability of SDU care is associated with a reduction in mortality by 3.2% and 29 minutes of hospital remaining length-of-stay. We note that the estimated marginal effects are quite small as there is a substantial proportion (83%) of low-severity patients who will not be admitted to the SDU even when the SDU is not busy. As such, the outcomes for these patients will be agnostic to whether the SDU is busy, since they will be admitted to the ward either way. Thus, our estimates are only for the remaining 17% of low-severity patients whose routing from the ED is dictated by the state of the SDU. We do not find a statistically significant relationship between SDU care and the likelihood of hospital readmission within 2 weeks. Based on the size of the low severity cohort, we estimate eliminating SDU busy-ness would translate into annual savings of 34 lives and 979 hospital days (weekly savings of 0.65 lives and 18.8 hospital days) aggregated across the 10 hospitals.

1.6.2. High Severity

We find that a busy SDU is associated with a decrease in likelihood of SDU admission for high-severity patients, while a busy ICU is associated with an increase in likelihood of SDU admission. For the mortality model, the coefficient on $SDUBusy_i$ is $-0.6325$ with standard error $0.1043$ and $p$-value $< 0.1\%$; for $ICUBusy_i$, the coefficient is $0.4072$ with standard error $0.1352$ and $p$-value $< 0.1\%$. The results are similar for the other patient outcome models. Table
Table 11: Estimated Effect of SDU Admission Following the ED (γ) on Patient Outcomes for Low-Severity Patients and Correlation Between Error Terms (ρ) for the Admission Decision and Patient Outcomes: N = 42,385

<table>
<thead>
<tr>
<th>Outcome</th>
<th>With IV</th>
<th></th>
<th>Without IV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>γ (SE)</td>
<td>Predicted Outcome</td>
<td>ρ (SE)</td>
<td>Test</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ˆPSDUBusy=0</td>
<td>ˆPSDUBusy=1</td>
<td></td>
</tr>
<tr>
<td>Mortality</td>
<td>-0.55** (0.28)</td>
<td>2.16%</td>
<td>2.23%</td>
<td>0.33+ (0.17)</td>
</tr>
<tr>
<td>log(HospRemLOS)</td>
<td>-0.20*** (0.04)</td>
<td>2.92</td>
<td>2.94</td>
<td>0.18*** (0.03)</td>
</tr>
<tr>
<td>HospReadm2w</td>
<td>-0.13 (0.12)</td>
<td>-</td>
<td>-</td>
<td>0.09 (0.07)</td>
</tr>
</tbody>
</table>

Notes: Standard error in parentheses. + (p < 10%), * (p < 5%), ** (p < 1%), *** (p < 0.1%).
Low severity threshold, t1, defined by 60th percentile of ˆTx.
Predicted outcome: ˆPSDUBusy=0 - Average predicted outcome if the SDU was never busy
ˆPSDUBusy=1 - Average predicted outcome if the SDU was always busy.
Predicted HospRemLOS (days) is shown instead of log(HospRemLOS)

[12] summarizes the impact of SDU admission after ED transfer on the various patient outcomes for these patients.

For high-severity patients, being admitted to the SDU appears to be associated with worse outcomes, as seen in the sign of SDU admission coefficient, which is positive and statistically significant in all outcome measures. We again use our estimation results to predict the patients outcomes under two cases: (i) the SDU is never busy and (ii) the SDU is always busy. Our results suggest that being admitted to the hospital when the SDU is busy is associated with substantial degradation in patient outcomes. SDU bed availability is, on average, associated with an increase of 12.1% in in-hospital deaths, a 50.8% increase in hospital readmissions within 2 weeks, and .35 additional days in the hospital. As with the ICU cohort, the marginal effects estimates are based on the estimated treatment effect, which is averaged across all patients who comply with the instruments. Thus, one should interpret our results as demonstrating substantive and rigorous evidence to the statistical significance and direction of the treatment effect. Again, we see evidence of a correlation between the SDU admission decision and patient outcomes with the estimated correlation coefficient ρ being statistically different than 0. Based on the size of this cohort, we estimate eliminating SDU busy-ness would translate into annual increases of 71 deaths, 981 more hospital days, and 173 additional hospital readmissions within 2 weeks (weekly increases of 1.37 deaths, 18.9 hospital days, and 3.33 readmissions) aggregated across the 10 hospitals.
The results for LOS and hospital readmissions are consistent with Kim et al. (2015). Interestingly, we find that being admitted to the SDU is associated with an increase in mortality risk, while Kim et al. (2015) did not find an impact of non-ICU care on mortality. One potential explanation is that Kim et al. (2015) considered all patients admitted from the ED to a medical service, while we stratify our analysis to focus on only the high-severity patients. As such, the results of Kim et al. (2015) may be distorted as SDU care may improve mortality risk for some patients within their cohort while also degrading mortality risk for other patients, thereby cancelling each other out. In contrast, since we focus on patients who are more likely to be admitted to the ICU (i.e., 54.90% compared to 11% in Kim et al., 2015), we are able to provide a cleaner estimate.

Table 12: Estimated Effect of SDU Admission Following the ED ($\gamma$) on Patient Outcomes for High-Severity Patients and Correlation Between Error Terms ($\rho$) for the Admission Decision and Patient Outcomes: $N = 2,803$

<table>
<thead>
<tr>
<th>Outcome</th>
<th>With IV</th>
<th>Without IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\gamma$ (SE)</td>
<td>Predicted Outcome</td>
</tr>
<tr>
<td><strong>Mortality</strong></td>
<td>0.75* (0.33)</td>
<td>23.64%</td>
</tr>
<tr>
<td>$\log(HospRemLOS)$</td>
<td>0.45*** (0.12)</td>
<td>6.22</td>
</tr>
<tr>
<td>$HospReadm_{2w}$</td>
<td>1.27** (0.40)</td>
<td>18.28%</td>
</tr>
</tbody>
</table>

Notes: Standard error in parentheses. * ($p < 10\%$), ** ($p < 5\%$), *** ($p < 1\%$), **** ($p < 0.1\%$).

High severity thresholds, $t_2$, defined by 95th percentile of $T_{x_i}$.
Predicted outcome: $\hat{P}_{SDUBusy=0}$ - Average predicted outcome if the SDU was never busy.
$\hat{P}_{SDUBusy=1}$ - Average predicted outcome if the SDU was always busy.
Predicted $HospRemLOS$ (days) is shown instead of $\log(HospRemLOS)$

1.6.3. **Medium Severity**

We now consider the impact of SDU admission on medium-severity patients. For these patients, a busy SDU is associated with a decrease in likelihood of SDU admission. For the mortality model (when comparing admission to the SDU versus ICU or ward), the coefficient on $SDUBusy_i$ is $-0.5503$ with standard error 0.0377 and $p$-value < 0.1%. The results are similar for the other patient outcome models.

Because of the aforementioned substantial heterogeneity within the medium severity cohort, when we run our models on this population, the results are not statistically significant and our instruments do not seem to be able to address potential endogeneity biases (see Table 13). The only result that is
statistically significant is the remaining hospital LOS when considering whether a patient is admitted to the SDU versus ICU or ward. We’re not sure what to make of this result due to the substantial heterogeneity of this group, as discussed before. We find that when the SDU is busy, patients are more likely to be rerouted – lower severity patients tend to go to the ward, while higher severity patients tend to go to the ICU. One possible explanation for the statistically significant effect on hospital LOS is that the high-severity patients who are admitted to the SDU are less sick than those who are not (and instead are sent to the ICU). Combining this with the lower severity patients who benefit from SDU care results in a statistically significant effect. Unfortunately, because of limitations in our data and the presence of unobservable factors, it is difficult to accurately assess the severity of the patients in this particular cohort, so we cannot be sure what is driving this result.

Table 13: Estimated Effect of SDU Admission Following the ED (\( \gamma \)) on Patient Outcomes for Medium-Severity Patients and Correlation Between Error Terms (\( \rho \)) for the Admission Decision and Patient Outcomes

<table>
<thead>
<tr>
<th>Outcome</th>
<th>With IV</th>
<th>Without IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \gamma ) (SE)</td>
<td>( \rho ) (SE)</td>
</tr>
<tr>
<td>SDU vs. Ward: ( N = 21,399 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mortality</td>
<td>-0.19 (0.25)</td>
<td>0.13 (0.15)</td>
</tr>
<tr>
<td>log(HospRemLOS)</td>
<td>-0.02 (0.13)</td>
<td>0.10 (0.07)</td>
</tr>
<tr>
<td>HospReadm2w</td>
<td>0.32 (0.25)</td>
<td>-0.16 (0.15)</td>
</tr>
<tr>
<td>SDU vs. ICU: ( N = 12,043 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mortality</td>
<td>0.07 (0.20)</td>
<td>-0.18 (0.13)</td>
</tr>
<tr>
<td>log(HospRemLOS)</td>
<td>-0.26 (0.08)</td>
<td>0.10 (0.06)</td>
</tr>
<tr>
<td>HospReadm2w</td>
<td>-0.15 (0.19)</td>
<td>0.09 (0.12)</td>
</tr>
<tr>
<td>SDU vs. ICU or Ward: ( N = 25,928 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mortality</td>
<td>-0.23 (0.19)</td>
<td>0.10 (0.11)</td>
</tr>
<tr>
<td>log(HospRemLOS)</td>
<td>-0.21** (0.07)</td>
<td>0.17** (0.05)</td>
</tr>
<tr>
<td>HospReadm2w</td>
<td>-0.01 (0.25)</td>
<td>0.03 (0.15)</td>
</tr>
</tbody>
</table>

Notes: Standard error in parentheses. \( + (p < 10\%), * (p < 5\%), ** (p < 1\%), *** (p < 0.1\%) \).
Medium severity thresholds, \( t_1 \) and \( t_2 \), defined by 95\textsuperscript{th} and 60\textsuperscript{th} percentiles of \( \hat{T}_{x_i} \).

While it is challenging to consider this medium-severity group and to interpret the results of our estimations for these patients, we highlight that this does not change the fundamental result of our work. In particular, we find that sending high-severity patients from the ED to the SDU when the ICU is busy can have substantial adverse consequences. On the other hand, for some low-severity
patients care in the SDU may be beneficial with regard to some patient outcomes. These results indicate the need for further study of the use of the SDU for patients originating in the ED.

1.6.4. Robustness Checks

We now describe a number of robustness checks for our ED cohort. Due to the challenges associated with the medium-severity patients, we focus our attention on the robustness of our results for the low- and high-severity patients. Our initial robustness checks are very similar to those we did with the ICU cohort. We tried different specifications of control variables, different time-lines for hospital readmission (within 1 week, 2 weeks, and 30 days after a patient is discharged from the hospital), and different ways to control for congestion during a patient’s stay (max occupancy and no control). We also varied the definition of our instrument(s) by varying the number of available beds from one to four, using indicator variables for whether the percentile occupancy level was above a threshold, a spline variable for the percentile occupancy level, as well as a time-varying capacity measure based on three 4-month periods. For our LOS models, we also considered specifications including patients with in-hospital death.

Low-Severity Patients

For the low-severity patients, we found the results for LOS were very robust in sign, magnitude and statistical significance for all of these different specifications, including when we include patients with in-hospital death. The sign and magnitude for the in-hospital mortality results were also very robust. When the IV of SDU busy was based on very limited bed availability (i.e., $\leq 1$ free beds or occupancy above the 95th percentile), the coefficient estimates were not statistically significant. In these instances, we cannot reject the null hypothesis that the correlation between the two error terms in our model is 0, i.e., $\rho = 0$. This suggests that in these instances, the instrument is not able to adequately address the endogeneity biases in our data. We consistently found no statistically significant association between SDU admission and hospital readmissions within 1 week, 2 weeks, or 30 days.
High-Severity Patients

For the high-severity patients, we found the results for HospRemLOS were very robust in sign, magnitude and statistical significance for all of these different specifications. The mortality results were also quite robust. However, when defining ICUBusy or SDUBusy with a relatively low occupancy level (e.g., 80th percentile), the statistical significance of the coefficients can drop to p-value < 10% or in some rare instances, is no longer statistically significant even at the 10% level. In these instances, we cannot reject the null hypothesis that the correlation between the two error terms in our model is 0, i.e., \( \rho = 0 \). This suggests that in these instances, the instrument is not able to adequately address the endogeneity biases in our data. Similarly, the hospital readmission results are always consistent in terms of sign and magnitude for these different specification. However, there are some instances when the results are not statistically significant. These instances correspond to when we cannot reject the null hypothesis that \( \rho = 0 \). This happens most frequently with the 1 week time to hospital readmission. We also found that all of our results were robust to including an additional instrumental variable based on the number of recent admissions to the ICU (RecentAdmission\(_i^{ICU}\)).

For the LOS models, we also considered the robustness of our results to including patients with in-hospital death. When including patients with in-hospital mortality in the high-severity ED Cohort, the sign of \( \gamma \) is negative (−1.83) and statistically significant at the p-value < .01 level. This raises questions as to the robustness of our LOS results for the high-severity group. However, we believe the main results as reported are more likely to be aligned with the true effect direction and size of SDU admission as it has been well established in the medical literature to exclude patients with in-hospital death for LOS models (e.g., Rapoport et al. [1996], Norton et al. [2007]).

Severity categorizations

In our severity categorizations for the ED Cohort, we took a data-driven approach and used thresholds on \( \hat{T}_{x_i}^* \) to partition the patients into low- and high-severity groups. We varied the thresholds for these categorizations from the 45th to 85th percentile for low-severity patients and
from the 90th to 97th percentile for high-severity patients. We then examined the robustness of our estimation results to these different thresholds.

**Low-Severity Patients**

As with our main specification, we do not find statistically significant results for the hospital readmission models. We find that the results for $HospRemLOS$ is very robust in magnitude and statistical significance to all of the different specifications of the low severity threshold. While the mortality results are robust to lowering the threshold, which reduces the sample size, we lose statistical significance when increasing the threshold above the 60th percentile. This may be because as the sample size is increased, there are (moderately) high-severity patients whose mortality risk may suffer with SDU admission are included in the cohort. When examining the LOS results more closely, we see that as the threshold is increased, the magnitude of the coefficient decreases, suggesting that the low severity cohort is including more patients for which SDU care is detrimental. Moreover, we cannot reject the null hypothesis that the correlation between our error terms is 0 ($\rho = 0$), which suggests that as the threshold increases, there are more non-compliers included in the cohort, making the instrument ineffective to address the endogeneity issues.

**High-Severity Patients**

In our readmission models, increasing the threshold for high-severity patients above the 95th percentile results in the regressions not converging. This is likely because the size of the cohort is being made smaller and smaller, and there are not enough samples to solve the FMLE optimization. These results suggest that the hospital readmission results are not very robust.

On the other hand, the $HospRemLOS$ results are quite robust to changes in the threshold. Similar to our observations for the low-severity patients, we see that as the threshold decreases, the magnitude of the coefficient decreases. This may be because low-severity patients who benefit from SDU care are entering into the high severity cohort as the threshold is decreased. A similar argument can be made for the mortality results. We find that when the threshold for high-severity patients is less than the 93rd percentile, the mortality and LOS results are no longer statistically significant.
About 10% of the ED-medical patients are admitted to the ICU from the ED. Additionally, our admission model in Equation (1.6) incorporates an unobservable term $\xi_i$, such that if the observed latent variable $\hat{T}_{x_i}$ plus $\xi_i$ is above the threshold, the patient will be routed to the ICU. As the threshold gets closer to the 90th percentile, there will be more spill over of patients for which SDU care is beneficial (instead of detrimental).

1.7. Conclusions and Managerial Insights

This paper studies the role of different levels of service for customers with uncertain needs. We examine this in a hospital setting where step-down units (SDUs) can be used to treat a variety of patients with very different repercussions. We consider fundamental questions regarding the SDU: Does admitting a patient to the SDU improve or degrade patient outcomes? What is the magnitude of these effects? And, how does it vary across different types of patients? Our work represents an important first step towards answering these questions. We find that while the answer for patients discharged from the ICU (its original purpose) is fairly clear, for those admitted from the ED, it is quite nuanced – some patients will benefit, while others will not. Moreover, the impact of SDU care can be substantial, so it is essential to be able to carefully identify which patients are appropriate for SDU care. These findings suggest that while different levels of service may be used to serve multiple customer types, the costs and benefits associated with each level of service can be highly heterogeneous due to the different and sometimes uncertain needs of customers.

There are a number of opportunities for future work. Our empirical analysis relies on the variation in patient routings following ICU discharge or following admission from the ED due to SDU and/or ICU capacity constraints. Consequently, our estimates fundamentally apply to patients whose SDU admission comply with our instrumental variables. As such, it is not possible to make any statements about the impact of SDU care for patients whose care pathway is invariant to SDU (or ICU) bed availability. While it is difficult to extrapolate our results to make inferences on the precise magnitude of the effect of the SDU on individual patients, our results demonstrate strong evidence as to the directional impact of an SDU. Because SDUs go in and out of favor at individual hospitals,
there may be opportunities for natural experiments to make such inferences without requiring an instrumental variable analysis. Alternatively, at a hospital system such as Kaiser Permanente, it might be possible to conduct a controlled randomized trial by randomizing which hospitals have SDUs. Of course, such a study would require substantial buy-in from hospital administrators and staff. Our empirical setting focuses on patients admitted to the hospital via the ED to a medical service. A number of studies in the medical literature consider the impact of SDUs on surgical patients (e.g., Eachempati et al. 2004). The impact of SDU congestion is likely very different for surgical patients, where surgical procedures and schedules often dictate the precise care pathway for these patients. Hence, an alternative identification strategy is likely needed. In other service settings where experimentation is less costly (e.g., call-centers), randomized experiments may be a feasible approach to providing unbiased estimates of costs and benefits for different customers. That said, as routing of customers to servers in call-centers is often done by computers, the likelihood of biases due to unobservable factors may be lower, so an IV approach or randomized experiment may not be necessary in these settings.

Our approach to classifying patients could be used in other service settings where customers’ needs are uncertain. For instance, in an increasing number of healthcare settings—including EDs, critical care, primary care and oncology, among others—Physician Assistants (PAs) and Nurse Practitioners (NPs) are used as lower cost alternatives to physicians (e.g., Hooker and McCaig 2001, Naylor and Kurtzman 2010, Hinkel et al. 2010, Doan et al. 2011, Gershengorn et al. 2011, Green et al. 2013, Gershengorn et al. 2016). PAs and NPs are trained in some, but not all, of the skills of physicians, raising important questions as to which patient types and tasks can be safely and effectively handled by these healthcare professionals rather than by a physician. In such instances, a data-driven approach, such as the one taken in this paper, could be utilized to classify patients. It could also be useful in other service settings in which servers have different skill levels, such as call-centers or repair facilities, where customers are heterogeneous, but their needs are not known a priori. With the growing availability of customer information (e.g., demographics, spending habits, etc.), a data-driven approach to customer segmentation may be useful.
From a stochastic modeling point of view, there are a number of directions that could build upon this work. From the healthcare operations management standpoint, it would be interesting to study optimal control policies regarding where to transfer patients from the ED or following ICU discharge in the presence of an SDU. This would provide a system-level view that would capture the potential benefits of an SDU, including externalities on other patients, beyond the estimates of individual patients estimated in this work. This would complement the growing body of work which examines how to make patient transfer decisions from the ED as well as inpatient units (e.g., Mandelbaum et al. 2012, Barz and Rajaram 2015, Samiedaluie et al. 2017, Dai and Shi 2017, Kilinc et al. 2016). Additionally, one could consider how to determine the capacity of the SDU relative to the ICU and general ward given patient mix and arrival rates. One factor which would significantly impact this decision is whether to restrict use of the SDU to be a true step-down versus allowing admission of patients from non-ICU units, such as the ED. This work quantifies the impact of lack of access to care for various patient types and could be used to set performance benchmarks or to calibrate a cost minimization framework when determining bed capacity (e.g., Yankovic and Green 2011, Yom-Tov and Mandelbaum 2014, Best et al. 2015, Armony et al. 2018). Analysis of these questions can also provide insights into how to utilize nested levels of service and routing policies in other types of service settings with heterogeneous customers (e.g., call-centers (Gans et al. 2003), retail stores, restaurants, etc.). For instance, it would be interesting to understand what factors such as number of customer types, differences in customer demand, service times, and costs impact the optimal number of levels of service and the optimal capacity to allocate to each level.

In understanding the benefits of the nested structure, an interesting tradeoff arises where increasing the number of levels reduces pooling benefits and may increase delays or reroutings. On the other hand, increasing the number of levels of care allows for more specialization that may result in efficiencies that reduce service times and improve outcomes. In many nested service systems, including the hospital situation studied here, it would be interesting to examine the potential tradeoffs between pooling and efficiency, similar in spirit to the work in Song et al. (2015) and how this would impact the allocation of servers (e.g., beds) to different levels of care (e.g., Best et al. 2015).
Acknowledgments: We thank Marla Gardner, John Greene, and Benjamin Turk for their help in preparing the data, along with the staff in the Division of Research and hospitals in Kaiser Permanente Northern California for their time and invaluable contributions to this research. The work by Carri W. Chan and Gabriel Escobar was supported in part by NSF/AHRQ grant number CMMI-1233547. Dr. Escobar was also supported by The Permanente Medical Group, Inc., and Kaiser Foundation Hospitals, Inc.
1.8. Appendix: Supplementary Tables

Table 14: Control Variables for Patient Characteristics and Hospital Care

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>ICU Cohort</th>
<th>ED Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Dummy variable: male = 1; female = 0</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Age</td>
<td>Continuous variable: coded as piecewise linear spline variables with knots at its $50^{th}$ and $80^{th}$ percentiles (65 and 81)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>LAPS2</td>
<td>Laboratory-based Acute Physiology Score: measures physiologic derangement at admission and is mapped from 14 laboratory test results such as arterial pH and white blood cell count obtained 72 hours preceding hospitalization to an integer value that ranges from 0 to 262 in our data set (higher scores indicate poorer condition); coded as piecewise linear spline variables with knots at its $50^{th}$ and $80^{th}$ percentiles (94 and 134)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>COPS2</td>
<td>Comorbidity Point Score: measures the chronic illness burden and is based on 41 comorbidities, such as diabetes, to which patients are categorized using outpatient and inpatient data from the 12 months preceding hospitalization; ranges from 0 to 267 in our data set, a higher score indicates a higher comorbid illness burden; coded as piecewise linear spline variables with knots at its $50^{th}$ and $80^{th}$ percentiles (33 and 87)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SAPS3</td>
<td>Simplified Acute Physiology Score: measures the severity of illness and predict vital status at hospital discharge based on ICU admission data. SAPS3 score is associated with each ICU admission and is calculated based on data obtained within one hour of ICU admission; ranges from 14 to 100 in our data set; coded as piecewise linear spline variables with knots at its $50^{th}$ and $80^{th}$ percentiles (52 and 61)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Admitting diagnosis</td>
<td>A way of classifying ICD9 codes: this clinical classification system was developed by HCUP and buckets ICD9’s into about 200 groups. A further grouping of the variable HCUP developed by Gabriel Escobar to condense the HCUP grouping into 38 groups so it could be used in a similar fashion as PRICOND3.</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Seasonality</td>
<td>Month/day-of-week/time-of-day: category variable for each month and day-of-week. For time-of-day, we use category variables for nurse shifts happening three times a day at 7am, 3pm, and 11pm.</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Previous unit</td>
<td>Category variable to track inpatient unit a patient is admitted to immediately before ICU admission.</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>LOS before ICU</td>
<td>Continuous variable that is the total length-of-stay (hours) prior to the ICU admission; measures how long a patient has been in hospital before being admitted to the ICU; coded as piecewise linear spline variables with knots at its $50^{th}$ and $80^{th}$ percentiles (2 and 31).</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ICU LOS</td>
<td>Continuous variable that is the length-of-stay (hours) at the first ICU; measures how long a patient has been in the ICU; coded as piecewise linear spline variables with knots at its $50^{th}$ and $80^{th}$ percentiles (38 and 83).</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ED LOS</td>
<td>Continuous variable that is the length-of-stay (hours) at the first ED; measures how long a patient has been in the ED.</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

Notes: To account for potential non-linear effects of some of the variables used to control for patient severity, we code them as piecewise linear spline variables.
Table 15: ICU Cohort: Percentage of Patients Who Are Discharged from ICU When SDU Is Busy

<table>
<thead>
<tr>
<th>Hosp</th>
<th>SDU Size</th>
<th>( \leq 1 )</th>
<th>( \leq 2 )</th>
<th>( \leq 3 )</th>
<th>( \leq 4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>0.93</td>
<td>3.57</td>
<td>7.80</td>
<td>12.17</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>0.66</td>
<td>2.95</td>
<td>7.54</td>
<td>12.46</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>0.56</td>
<td>7.94</td>
<td>24.29</td>
<td>45.63</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>3.17</td>
<td>12.68</td>
<td>27.07</td>
<td>41.59</td>
</tr>
<tr>
<td>5</td>
<td>24</td>
<td>0.28</td>
<td>1.54</td>
<td>3.93</td>
<td>7.87</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>0.82</td>
<td>3.34</td>
<td>6.76</td>
<td>15.37</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>0.00</td>
<td>2.84</td>
<td>16.74</td>
<td>36.77</td>
</tr>
<tr>
<td>8</td>
<td>27</td>
<td>2.81</td>
<td>9.34</td>
<td>18.80</td>
<td>31.74</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>9.76</td>
<td>37.72</td>
<td>63.94</td>
<td>80.34</td>
</tr>
<tr>
<td>10</td>
<td>32</td>
<td>0.34</td>
<td>2.66</td>
<td>6.19</td>
<td>12.71</td>
</tr>
<tr>
<td>All hosp</td>
<td></td>
<td>2.52</td>
<td>10.64</td>
<td>21.70</td>
<td>34.00</td>
</tr>
</tbody>
</table>

Table 16: Post-ICU Location

<table>
<thead>
<tr>
<th>Unit</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ward</td>
<td>7,226</td>
<td>54.77</td>
</tr>
<tr>
<td>SDU</td>
<td>3,832</td>
<td>29.05</td>
</tr>
<tr>
<td>Death in ICU</td>
<td>985</td>
<td>7.47</td>
</tr>
<tr>
<td>Out of hospital (alive)</td>
<td>1,150</td>
<td>8.72</td>
</tr>
<tr>
<td>Total</td>
<td>13,193</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 17: Effect of SDU Admission Following the ED (\( \gamma \)) on HospRemLOS When Including Patients with In-Hospital Death

<table>
<thead>
<tr>
<th>Cohort</th>
<th>( \gamma ) (SE)</th>
<th>Predicted Outcome</th>
<th>( \rho ) (SE)</th>
<th>Test ( \rho = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED Cohort - High Severity</td>
<td>-1.83** (0.07)</td>
<td>1.33% 8.31% (( P_{ICU} ))</td>
<td>0.81*** (0.02)</td>
<td>0.00</td>
</tr>
<tr>
<td>ED Cohort - Low Severity</td>
<td>-0.20*** (0.04)</td>
<td>2.38 2.92 (( P_{Ward} ))</td>
<td>0.18*** (0.03)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Notes: Standard error in parentheses. + (p < 10%), * (p < 5%), ** (p < 1%), *** (p < 0.1%).

Predicted outcome: \( P_{SDU} \) - Average predicted outcome if all patients could be routed to the SDU and \( P_{ICU} \) (\( P_{Ward} \)) if no SDU and everyone is routed to the ICU (Ward).

Predicted HospRemLOS (days) is shown instead of \( \log(\text{HospRemLOS}) \)

Joint work with Christian Terwiesch, David A. Asch, and Kevin G. Volpp

2.1. Introduction

More than ever, technology is bringing businesses and customers closer together. In the past, consumers and their banks would interact only sporadically in a retail branch. Now, online and mobile banking connect the two parties almost seamlessly, creating a delivery model known as connected banking. Similarly, automotive insurance companies used to interact with the drivers they insured only at predefined touch points, such as policy renewals or claims events. Now, sensors and mobile apps continuously monitor driving behavior. Such data enable a connected insurance business model by providing safety feedback to drivers (and parents) as well as informing future underwriting decisions.

The domain that we study in this paper is healthcare. In the past, before the possibility of being connected technologically to the healthcare system, patients made decisions related to diet, exercise regimen, or medication adherence without the involvement of care providers (Asch et al. 2012). Now, connected health devices, such as smart pill bottles, connected scales, and wearable health trackers, are capable of transmitting clinically relevant data to information systems that can then relay this information to patients and providers. The promise of connected healthcare is that such information allows healthcare providers to monitor health-related behavior and to proactively intervene rather than waiting for an adverse event. Moreover, it has been suggested that patients adjust their behavior in response to feedback, becoming more adherent in taking their medication and more active in their lifestyle.

Active patient engagement and feedback systems have the potential to reduce the number of preventable readmissions and ultimately reduce healthcare costs (Chouvarda et al. 2015). Despite
this enormous upside, however, there is limited evidence for the effectiveness of connected healthcare and little guidance on how to effectively implement it (Caulfield and Donnelly 2013). One of the crucial health behaviors that may affect the readmission risk and have been the target for improvement is medication adherence. Although the costs of implementation are usually straightforward to estimate, the benefits of engagement systems to improve medication adherence and the impact of higher adherence on readmission are oftentimes ambiguous. Moreover, it is unclear how to effectively engage with non-adherent patients in order to minimize the number of readmissions given limited resources.

The aims of our research are to (1) investigate the efficacy of connected feedback systems in promoting medication adherence among patients, (2) examine the relationship between patient medication adherence and readmission, and (3) develop a readmission risk-scoring model that takes into account medication adherence and use the model to better target non-adherent patients. Establishing how connected health systems causally affect clinical outcomes can be done only in a large-scale randomized controlled trial. Running such an intervention requires overcoming multiple challenges. First, one needs to recruit a large cohort of patients, equip them with connected devices, and train them how to use the devices. Second, for patients to benefit from connected healthcare, new delivery processes have to be designed and implemented. Third, one has to track patients in the study not only during their usage of the connected healthcare devices, but also for a long period so that the full scope of the clinical outcomes can be observed.

Our study is part of a larger research project that overcame these challenges by implementing one of the largest clinical interventions related to connected healthcare that has been conducted so far. This study, known as the HeartStrong study, enrolled 1,453 patients for a two-arm randomized clinical trial with a year-long intervention that aimed to improve medication adherence and reduce the number of readmissions after myocardial infarction. The 1,000 patients in the intervention group received a compound intervention integrating wireless pill bottles that can electronically track openings and transmit them to the care team, lottery-based incentives, and social support that involved feedback systems such as automated messaging, manual messaging, and phone calls,
delivered to either the patients themselves or their enlisted partner at increasing levels of escalation. Patients were then tracked for 12 months, capturing any future hospital readmissions.

The results of the study were published in one of the leading medical journals (Volpp et al. 2017) and did not show a statistically significant difference in readmission between the control group and the intervention group. The lack of statistically significant difference begs the question of why patients who were better connected to the healthcare system did not benefit in the form of better outcomes. Did the various reminder and feedback interventions improve medication adherence? Did higher rates of adherence reduce the likelihood of readmission? Did we sufficiently allocate our capacity to the patients who were at high risk of readmission? Without a more micro-level theory, one can only speculate.

The contribution of this study is to present such a micro-level operational model. To evaluate the effectiveness of the connected health system, the HeartStrong study focused on the variation between the control group and the intervention group and did not use any data related to behavioral interventions and daily medication adherence. In contrast, the present study exploits variation within the intervention group, leveraging the micro-level data capturing patient behavior and provider actions at the patient-day level. At this micro level, hundreds of observations per patient (did patient \(i\) receive an intervention \(j\) on day \(t\)?) were analyzed and linked with adherence behavior (did patient \(i\) take medication \(k\) on day \(t\)?) and clinical outcomes (was patient \(i\) readmitted on day \(t\)?). This allowed us to establish the following novel contributions:

1. **Showing the impact of escalation strategies on medication adherence**: By studying the effect of different levels and dynamics of intervention escalation, we show that patients are significantly more likely to become adherent when (a) they or their partners receive high levels of intervention that involve personalized feedback like phone calls and manual messages and (b) the intervention is escalated quickly and consistently. Our results imply that, in order to effectively make a previously non-adherent patient adherent again, one should start calling the patient as soon as he or she becomes non-adherent. Receiving a personal phone call immediately after the first day of non-adherence more than doubles the probability of
becoming adherent again. Then, if non-adherence continues, one should escalate to manual messages and calls to the patient’s partner successively on the following days.

2. *Showing the impact of medication adherence on readmission:* We explore how consistency in medication use in the past affects the readmission probability in the present. Specifically, we show that, for patients with cardiovascular disease, long-term adherence to statins and \( \beta \)-blockers is associated with a 51% reduction in the odds of being readmitted on any given day, thus better medication adherence is strongly associated with reduced risk of readmission.

3. *Developing a dynamic readmission risk-scoring model and using it to better target non-adherent patients:* Using multi-layer perceptron, we develop a dynamic readmission risk-scoring model that includes patient-day level medication adherence as predictors. We evaluate the model primarily using the area under the ROC curve and find that (a) the model outperforms a baseline model that does not include medication adherence and (b) the model outperforms other models developed using different machine learning methods. Moreover, using counterfactual analysis, we apply the dynamic readmission risk-scoring model to our setting and show that, when using an intervention strategy that prioritizes patient-days with highest readmission risks, we obtain 10% fewer readmissions than we would obtain without considering readmission risk while maintaining the same level of effort.

Our connected healthcare setting holds great significance because reducing readmission rates of patients with heart failure is a national priority (Bradley et al. [2013]). Nearly 20% of Medicare beneficiaries discharged from the hospital are readmitted within 30 days, and these readmissions have been estimated to cost the country more than $24 billion (Agency for Healthcare Research and Quality [2014]). Since a significant part of the readmission risk can be attributed to patients’ lack of adherence to their prescribed medications, real-time adherence monitoring and reminder systems hold particular promise.

The rest of the paper is organized as follows. We provide a brief summary of related papers in the literature in Section 2.2. In Section 2.3 we introduce our study setting. We present the study on
the impact of escalation strategies on adherence behavior and its estimation results in Section 2.4. The study on the impact of medication adherence on readmission is presented in Section 2.5. We develop a dynamic readmission-risk scoring model and apply it to our setting in Section 2.6. Section 2.7 provides concluding remarks as well as discussions for future research.

2.2. Literature Review

The adoption of remote monitoring and feedback has long been prevalent outside of the healthcare industry. Researchers in operations management have been interested in examining remote monitoring, diagnosis, and feedback in manufacturing and other service settings. In manufacturing, the focus is usually on monitoring people, processes, and machines. Information technology is often adopted to reduce the cost of capturing, storing, and transmitting data to all members of the supply chain as well as to enable useful feedback functions such as fault notification, remote counseling, and real-time online help and process intervention. In service industries, remote monitoring and feedback became popular thanks to the call for digitalization in services (Tan and Netessine 2019) and the rapid growth of teleservices, which allow customers to receive service virtually at their convenience and, at the same time, allow behavioral data to be collected and shared with service providers. An increasing number of studies explore the prospects of utilizing such technologies in popular service domains such as retail and banking, through the use of e-commerce and online banking (e.g., Campbell and Frei 2010, Moon et al. 2018). In these service settings, real-time information sharing and customer interaction can benefit service providers by helping them target customers with specific price promotions, detect fraudulent activities, improve customer retention, and increase market share. This leads to “connected strategies” that allow firms to build continuous relationships with customers by having frequent, low-friction interactions with them and to address their needs as or even before those needs arise (Siggelkow and Terwiesch 2019). Apart from their use with customers, remote monitoring and feedback are also increasingly utilized to oversee internal service operations such as monitoring employee theft and productivity (e.g., Pierce et al. 2015).
More recently, the applications of remote monitoring and feedback have expanded into the healthcare community. For healthcare workers, remote electronic monitoring can be put in place to help increase workers’ behavioral compliance, such as hand hygiene (Staats et al. 2017). For patients, telehealth practices have been increasingly adopted to facilitate efficient management of health and wellness, allowing health personnel to proactively connect with patients through health IT and provide patients with real-time health and behavioral assessment, diagnosis, interventions, and consultation (Kvedar et al. 2014). As more researchers and practitioners are interested in learning how health IT may transform healthcare practices, we see a growing body of operations management literature that explores the impact of various long-standing forms of health IT and telehealth, such as home monitoring and e-visits (e.g., Rajan et al. 2019, Bavafa et al. 2018, Angst et al. 2011, Devaraj et al. 2013).

In our study, we focus on connected healthcare that provides remote monitoring and feedback for patients with heart disease. A number of past studies investigate the benefits of connected health, or more broadly, telemonitoring, in patients with chronic conditions (e.g., Watson et al. 2009, Trappenburg et al. 2008), with many focusing on heart conditions (e.g., Chaudhry et al. 2010, Cleland et al. 2005, Maeng et al. 2014). Although past studies, which primarily rely on small- to medium-scale randomized and non-randomized control trials, suggest that telemonitoring may be an effective strategy for disease management in heart failure patients, the evidence base is inconclusive and quite limited (Chaudhry et al. 2007). Moreover, there exists little guidance on how to effectively implement such a system.

Our study aims to (1) examine the efficacy of connected health systems in improving medication adherence, (2) examine the relationship between medication adherence and readmission, and (3) develop a dynamic readmission risk-scoring model that takes medication adherence as input and use it to better target non-adherent patients. As shown in Figure 5, we address our first two aims in two studies, Study A and Study B. To address the third aim, we use machine learning to develop a dynamic risk-scoring model and investigate the benefit of utilizing the model in our intervention delivery through counterfactual simulations, which link intervention to readmission by treating
medication adherence as a mediator. Our work joins an increasing number of studies in operations that focus on remote monitoring and compliance (e.g., Staats et al. 2017, Jonasson et al. 2020), applications of data analytics in healthcare (e.g., Wang et al. 2019a), chronic disease management (e.g., Jonasson et al. 2017), and behavioral healthcare analysis (e.g., Ibanez et al. 2018), particularly analysis of patient behavior (e.g., Liu et al. 2018).

In the first study (Study A), we investigate the effectiveness of different escalation strategies in promoting medication adherence behavior. The lack of medication adherence has been a serious problem in managing chronic diseases around the world. Numerous studies have shown that patients with chronic illnesses adhere to their prescribed medications only 50% to 60% of the time, and the total direct and indirect cost estimates for non-adherence range from $100 billion to $300 billion each year (Bosworth et al. 2011). Until now, traditional behavioral interventions targeting medication adherence have produced only modest success (e.g., Ho et al. 2009, Nieuwlaat et al. 2014). Furthermore, the evidence regarding the effectiveness of health IT interventions to improve adherence is fairly thin (Bosworth et al. 2011). Bosworth et al. (2011) has found that almost all of the interventions that were effective were complex, including combinations of convenient care, information, reminders, self-monitoring, reinforcement, counseling, family therapy, and crisis intervention. The finding is in line with Ho et al. (2009), which concludes that multimodal interventions have been more successful than unimodal interventions, which rely on methods like reducing the number of daily doses of medications or packaging medications into special containers. These findings consistently suggest that connected healthcare, which can deliver multimodal interventions at reasonable cost, is a promising care model.
To the best of our knowledge, no one has attempted to study how patients’ medication adherence behavior changes in response to different escalation strategies. Prior studies that study the impact of adherence intervention on medication adherence primarily rely on randomized controlled trials and before-and-after comparisons. They work with macro-level intervention information (e.g., always receiving reminders vs. never receiving reminders) and approximate aggregate adherence data (e.g., self-reported adherence or prescription fill rates in the past three months). These studies do not have the granularity of data that we have available and therefore do not examine directly how patients respond to different escalation levels and dynamics. In contrast, we conduct our analysis at the patient-day level by utilizing micro-level intervention and adherence data, which allow us to examine the relationship between different escalation strategies and patients’ immediate changes in behavior (or lack thereof). Micro-level data recorded over time are commonly used in healthcare operations studies, especially to analyze workers’ behavior and performance (e.g., Gurvich et al. 2019, Berry Jaeker and Tucker 2017).

In the second study (Study B), we examine the impact of adherence to two crucial medications, statins (a class of lipid-lowering medications; also known as HMG-CoA reductase inhibitors) and β-blockers (a class of medications that are particularly used to manage cardiac arrhythmias; also written as beta-blocker) on the risk of readmission. We focus on readmission as the outcome variable for two main reasons (Kansagara et al. 2013). First, the ability to assess the readmission risk helps physicians target the delivery of care, especially resource-intensive interventions, to patients who are at highest risk for readmission. Second, readmission rates are often used as a care quality measure (e.g., Kim et al. 2015, Chan et al. 2019) and a quality metric for healthcare providers. Being able to identify drivers of readmission is important for hospitals. Since 2012, the Centers for Medicare and Medicaid Services (CMS) has publicly reported readmission rates and planned to lower reimbursement to hospitals with high risk-standardized readmission rates.

The relationship between adherence and health outcomes is recognized but understudied. DiMatteo et al. (2002) conducted a meta-analysis and proposed that the relationship was under-investigated because the effect of adherence on outcomes was often taken for granted. Nevertheless, several
studies have attempted to determine the impact of medication adherence on health outcomes such as rehospitalization and mortality for patients with chronic diseases such as coronary heart disease, diabetes, and AIDS (e.g., McDermott et al. 1997, Yu et al. 2010, Han et al. 2014). However, the difficulties of coming up with effective experimental study designs and having an accurate measure of adherence have made it difficult for researchers to establish a clear causal relationship between medication adherence and outcomes. Most previous studies rely on traditional measures of adherence such as self-reports, physician reports, and prescription fill rates. They cannot accurately measure medication adherence (Lam and Fresco 2015) and obtain medication adherence information directly prior to each readmission. In contrast, our study utilizes medication adherence data from electronic pill bottles. This allows us to analyze the probability of readmission at the patient-day level using reliable adherence information leading up to each day using an econometric approach, as opposed to traditional cross-sectional comparisons.

Finally, in our third analysis, we develop a dynamic readmission risk-scoring model that includes medication adherence as predictors and use counterfactual simulations to examine the benefit of using the model in our intervention delivery. Instead of alerting every patient at pre-specified moments of non-adherence, our new strategy directly targets the patients (and the patient-days) with highest readmission risks. Many researchers have explored various ways data analytics can improve healthcare delivery (e.g., Raghupathi and Raghupathi 2014, Wang et al. 2019a). However, to the best of our knowledge, none of them has studied the application of predictive analytics in connected health delivery. Moreover, although many studies attempt to develop readmission prediction models (e.g., Min et al. 2019, Shulan et al. 2013), they mostly aim to predict the probability of readmission within a fixed period after hospital discharge (e.g., probability of readmission within 30 days). Unlike our study, most prior studies do not have post-discharge patient-day level covariates, barring them from obtaining patient-day level readmission risk.
2.3. Study Setting

This study builds on a national two-arm randomized controlled trial program we conducted from 2013 to 2016. We collected data from 1,000 patients who enrolled in the program and received a 12-month connected healthcare intervention that aimed to reduce repeated cardiovascular events through improved medication adherence. Our original study was one of the first studies to deploy connected health devices on a large scale. The study aimed to improve upon earlier intensive case management efforts, which were mostly unwieldy and expensive, by using a simpler, cheaper, and more scalable approach. By leveraging more connected technology, we were able to deploy a novel, proactive model of chronic disease management. We also created new workforce roles and shifted some intensive case management roles from physicians to non-physician providers, which can benefit both physicians and the hospital (Powell et al. 2012).

Our study was approved by the Institutional Review Board of the University of Pennsylvania. Prior to the study, patients were recruited by University of Pennsylvania research staff from 2013 through 2015 and observed for one year. Eligible patients were:

(a) 18 to 80 years old.

(b) Admitted as hospital inpatients for one to 180 days.

(c) Discharged home with a primary diagnosis code of acute myocardial infarction (AMI).

(d) Not suffering from dementia.

(e) Not enrolled in other research studies incorporating wireless pill bottles.

Patients could enroll up to 60 days after discharge. To recruit patients, we initially contacted them through letters and phone calls. Of 19,678 potentially eligible patients contacted, 18,169 declined, could not be reached, or were ineligible upon further inspection, leaving an enrolled sample of 1,509. The most common reasons cited by the patients who declined were: (a) not being interested, (b) not wanting to change the current system, and (c) having privacy concerns. All enrolled patients
received $25 for participation. Patients in our sample were insured with five large US insurers or with Medicare fee-for-service at the University of Pennsylvania Health System.

We randomized patients in a 2:1 ratio of intervention:usual care using permuted block randomization stratified by insurance provider to balance the allocation across provider groups. After excluding patients who withdrew or did not have post-enrollment medical claims, 975 patients from the intervention group remain in our final patient cohort. Patients in the intervention group received an additional $25 for activating wireless pill bottles (Vitality GlowCaps), which electronically monitored bottle openings with a small remote device that plugged into a wall outlet and transmitted cellular signals to the care team, with no home wireless network, computer, or special setup required. Adherence information transmitted from those pill bottles was recorded on an electronic platform that we primarily utilized to facilitate real-time monitoring and feedback. Our connected healthcare platform, which was developed with funding from the National Institutes of Health, is a flexible and secure web-based infrastructure that consists of a portal that can be linked to various peripheral connected health devices, such as scales, glucometer, and pill bottles. The platform is capable of automating the delivery of feedback and communicating back to patients using email, text messaging, and interactive voice recording.

Patients in the intervention group were assigned an engagement advisor (EA) for the duration of their participation. The EAs assisted patients with pill bottle setup and troubleshooting, monitored the patient’s daily medication adherence on the electronic platform, provided manual forms of feedback, and served as a resource for patients struggling to stay adherent. All intervention patients were asked to enlist a potential support partner, usually a friend or family member. The role of the support partner was to receive information about the patient’s adherence through the platform and EAs, and provide support and encouragement using their preferred format and content. Support partners also had an account on the electronic platform, allowing them to automatically receive alerts through email, text message, or automated phone call. We gave intervention patients access to social work resources as well as five main forms of adherence feedback with increasing degrees of intensity: (1) automated message to patient, (2) automated message to partner, (3) phone call to
patient, (4) manual message to patient, and (5) phone call to partner. We discuss different forms of intervention in further detail in Section 2.4.2 in conjunction with Study A. Patients in the control group received usual care for the duration of the study and had no further contact with EAs or study staff. The results of our randomized controlled analysis are described in Volpp et al. (2017). We did not find statistically significant differences between study arms in time-to-first-rehospitalization for a vascular event or death, or total number of repeated hospitalizations.

To investigate further if and how the connected health systems worked, this paper focuses on the patients who received the intervention. These intervention patients had within-person variation in their adherence behavior and variation in how the intervention was escalated. The data we use for our micro-level analysis contains both operational and patient-level information. Operational information includes daily data for each patient indicating whether:

(a) Each of the five forms of intervention was delivered.

(b) The patient opened each of his or her pill bottles.

(c) The patient was readmitted.

We also have information on the time between patients’ initial discharge and study enrollment and on whether and when he or she was readmitted before enrollment. In addition to operational information, our data set contains information about patient characteristics such as age, gender, Medicare enrollment, Patient Health Questionnaire-2 (PHQ-2) score, and baseline Elixhauser comorbidity score. The PHQ-2, which is used to screen depression, inquires about the frequency of depressed mood and anhedonia over the past two weeks (McManus et al. 2005, Kroenke et al. 2003). A PHQ-2 score can range from 0 to 6, with higher scores translating to higher chances of having depression. The second score, Elixhauser comorbidity score, is evaluated using up to 12 months of pre-enrollment data based on 31 individual conditions identified from diagnoses in hospital and physician data and can take positive or negative values (van Walraven et al. 2009). These scores were assigned once and not updated during the study period. Table 18 provides the summary statistics of patient demographics for all patients in our cohort.
Table 18: Summary Statistic for Patient Demographics

<table>
<thead>
<tr>
<th>Patient characteristic</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>61.26</td>
<td>10.41</td>
<td>23</td>
<td>80</td>
</tr>
<tr>
<td>Female</td>
<td>0.34</td>
<td>0.47</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Medicare</td>
<td>0.44</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PHQ-2 depression score</td>
<td>1.26</td>
<td>1.63</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Baseline Elixhauser score</td>
<td>6.46</td>
<td>10.03</td>
<td>-14</td>
<td>44</td>
</tr>
<tr>
<td>Time from discharge to enrollment (days)</td>
<td>40.52</td>
<td>12.36</td>
<td>8</td>
<td>60</td>
</tr>
</tbody>
</table>

We describe the data in more detail in Table 27 in the Appendix. Following the Management Science policy for Data and Code Disclosure, we have made our code and data available on the journal website to permit replication. Moreover, future research in connected healthcare can also benefit from the data we provide.

2.4. Study A – Impact of Remote Monitoring and Feedback Systems on Medication Adherence Behavior

We begin by examining the efficacy of the feedback systems. Specifically, we want to see how effective different escalation strategies are in turning non-adherence into adherence. By answering this question, we can develop insights into how feedback systems play a role in improving patient behavior, and how best to engage with non-adherent patients in order to make them adherent again.

2.4.1. Data and Statistics

The data used in this analysis contains a series of non-adherence sequences for each patient. Each non-adherence sequence begins on the first day without pill bottle use and lasts until the first day the patient takes all of his or her medications again. One patient could have multiple non-adherence sequences. Each entry in our data contains daily information on whether the patient opened all the pill bottles and whether he or she received each type of intervention. On average, we observe each patient for 318 days, including both adherence and non-adherence days. Our observation period is slightly shorter than the intervention period because the wireless pill bottle activation was
occasionally delayed, causing a lag time between the start of the intervention period and when the bottles were functional.

Of 975 patients, approximately 3% were always adherent and are excluded from our analysis and 2% of the patients had irregularly long periods of non-adherence. For these patients, the average number of non-adherence days within each non-adherence sequence is greater than 15, and the maximum number of non-adherence days within a sequence ranges from 68 to 334 days. We exclude these patients from our main analysis because they may have intentionally not taken some or all of their medications, or their pill bottles were not working properly. After those patients are excluded, our data contain 930 patients who exhibit reasonable variability in their adherence pattern. Table [19] provides adherence statistics of the remaining patients and all patients. In our final cohort, on average, each patient has 15.57 non-adherence sequences, each lasting a mean of 2.78 days.

Table 19: Summary of Medication Adherence Behavior

<table>
<thead>
<tr>
<th>Adherence Characteristic</th>
<th>Focus Cohort (N = 930)</th>
<th>All patients (N = 975)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of non-adherence sequences per person</td>
<td>15.57  12.39  1  73</td>
<td>14.94  12.77  0  73</td>
</tr>
<tr>
<td>Number of non-adherence days within a sequence</td>
<td>2.78  3.05  1  122</td>
<td>3.86  6.10  1  334</td>
</tr>
</tbody>
</table>

2.4.2. Escalation of Feedback

We aimed to remind patients to take medications at increasing levels of escalation. Lower-level interventions were delivered automatically through our electronic platform while higher-level interventions were delivered manually by the engagement advisors who closely monitored adherence behavior. We usually provided lower-level interventions before we escalated to a higher-level intervention and stopped delivering interventions once the patient became adherent again. We illustrate the nature of escalation in Figure 6 which shows the distribution of each type of intervention over different days within a non-adherence sequence. Each mark on the vertical axis indicates the $i^{th}$ day of non-adherence. Each of the smaller plots corresponds to each form of intervention and is displayed along with its total number of occurrences at the bottom. Each plot illustrates the percent distribution of each form of intervention over day of non-adherence and is compared across all forms of intervention, capturing the nature of escalation. For example,
we delivered 122 manual messages to patients in total, 47% of which occurred on the sixth day without pill bottle use. In contrast, we initiated automated messages to patients much earlier in the non-adherence period, oftentimes in the first few days of non-adherence. Based on the escalation pattern, we define five levels of escalation corresponding to five main forms of intervention, which are described below in increasing order of escalation.

1. *Escalation level 1 — Automated message to patient:* In the lowest level of escalation, we deliver a first automated message to tell the patient that he did not take his medications. These non-personalized messages were mostly triggered via the electronic platform in the first two days of non-adherence and were usually delivered through text message or email. We usually sent automated non-adherence messages regularly until the patient became adherent again. In total, 19,323 such messages were sent. These messages contained a standard text that wrote: “You did not take all your meds yesterday.”

2. *Escalation level 2 — Automated message to partner:* In the next level of escalation, we send an automated message to the patient’s partner. Most non-personalized automated messages to support partners were initially triggered via the platform in the first four days of non-adherence and mostly continued until non-adherence ended. Like other platform-triggered messages, they were mainly delivered through text message and email. In total, 5,768 automated partner messages were sent. The standard message wrote: “Your HeartStrong partner has missed some heart medications for the past few days and may need your support.”

3. *Escalation level 3 — Phone call to patient:* In the third level of escalation, the EAs make the first phone call to the patient after non-adherence starts. Telephone calls to patients were mostly made after the fifth day of non-adherence. The EAs directly made these calls to tell patients about their non-adherence and ask them about the reasons for why the pill bottles were not opened. In total, the EAs made 305 calls that successfully reached the patients.
4. *Escalation level 4 — Manual message to patient:* With the next escalation level, the EAs send the first manual message to the patient after non-adherence starts. Manual messages were tailor-written and sent to the patients via the electronic platform, informing them of their non-adherence and asking them to contact the care team. The EAs mostly sent these messages after the sixth day without pill bottle use. In total, 122 manual messages were sent. The content of these messages varied; for example, a manual message may write: “Our system shows that you have not opened your GlowCaps for [number of non-adherence days] days. We have been unsuccessful in [action] so we will be contacting [person we would contact, e.g., support partner] if we do not hear back from you by the end of today. Please call your HeartStrong Program Advisor, [EA name], at [EA’s phone number].”

5. *Escalation level 5 — Phone call to partner:* At the highest escalation level, the EAs first call the patient’s partner after non-adherence starts. The EAs mostly made telephone calls to patients’ support partners after the seventh day of non-adherence to inform support partners of the adherence problems and enlist their support. In total, the EAs made 46 calls that successfully reached the patients’ partners.

Although we did not deliver calls and manual messages regularly like automated messages, it was not uncommon for the patients or the patients’ partners to receive them multiple times within a non-adherence sequence. In our main analysis, we consider only the calls that successfully reached the intended recipient. Apart from these five forms of intervention, the EAs also mailed letters to patients in eight cases, and contacted the patient’s primary care provider’s office by telephone to inform them of adherence problems in three cases. We exclude those interventions from our analysis because they were very rare events and we do not know when the letters were delivered to the patients and if they were received.

As Figure 6 shows, there exists considerable variation in when we delivered the intervention. We did not deliver the intervention in a perfectly consistent manner for several reasons. For automated messages, the inconsistency is primarily because (1) the patients were sometimes unable to receive messages and (2) our electronic platform occasionally had errors that prevented it from successfully
sending messages. For manual messages and phone calls, we observe considerable variation in delivery timing due to operational constraints such as our care team having limited capacity, patients not being reachable, and the intervention not being delivered on the weekends. Moreover, due to a lack of a strict intervention protocol, we also sometimes delivered manual interventions relatively early within a non-adherence period. These limitations result in varying escalation patterns, which are illustrated in Figure 7. The figure shows six examples of the way the intervention was escalated. In each plot, the marks on the vertical axis indicate current escalation levels, with level 0 meaning no intervention has been delivered. As we can see, there exists significant variation in whether and when each escalation level was reached. Since the variation resulted mostly from operational challenges on our end, we have no reason to believe that the variation is driven by patient-specific factors. Thus, from a research design perspective, the operational challenges created quasi-experiments that we can use to estimate the effect of escalation on adherence.

Since our goal is to estimate the effect of different escalation strategies on the probability of becoming adherent again on a given non-adherence day, we first need to define what characterizes
Notes: (a) No intervention was delivered, (b) Intervention was escalated to level 1 and stopped being escalated, (c) Intervention was escalated to level 4 and stopped being escalated, (d) Intervention was consistently escalated to level 5, (e) Intervention was escalated to 5, but was slowly escalated at the beginning, (f) Intervention was quickly escalated to level 5, bypassing levels 3 and 4.

Figure 7: Examples of Escalation Patterns from Six Different Non-Adherence Sequences

the escalation leading up to that day. We denote the characteristics of escalation that patient $i$ receives on the $t^{th}$ day of non-adherence within the $j^{th}$ non-adherence sequence as $\text{Escalation}_{ijt}$. For example, $\text{Escalation}_{15,2,5}$ corresponds to the escalation that patient 15 in our sample receives on the fifth day of non-adherence in his or her second non-adherence period. Based on the varying escalation patterns we observe, we describe $\text{Escalation}_{ijt}$ using two dimensions:

1. Current level of escalation ($\text{EscLevel}_{ijt}$): This captures the intensity of escalation on non-adherence day $t$. As defined above, levels of escalation range from 0 to 5 with increasing degrees of intensity. The escalation reaches level 1 when the patient first receives an automated message and reaches level 5 when their partner first receives a phone call.

2. Dynamics of escalation: This captures the dynamics of escalation leading up to non-adherence day $t$. The escalation dynamics are further broken up into:

   (a) Overall escalation dynamics ($\text{OverallEsc}_{ijt}$): $\text{OverallEsc}_{ijt}$ is observed from the time non-adherence begins up to non-adherence day $t$. We quantify $\text{OverallEsc}_{ijt}$
as the overall rate of escalation, which is calculated by dividing $EscLevel_{ijt}$ by $t$. This captures the speed of escalation, which can also be interpreted graphically as the slope of the escalation graph measured from the beginning of the non-adherence sequence to non-adherence day $t$.

(b) Recent escalation dynamics ($RecentEsc_{ijt}$): Since $OverallEsc_{ijt}$ does not necessarily capture recent escalation dynamics, we use $RecentEsc_{ijt}$ to indicate whether the intervention has been escalated within the past three days. Specifically, we define $RecentEsc_{ijt}$ as follows.

$$RecentEsc_{ijt} = \begin{cases} 
1, & \text{if } EscLevel_{ijt} - EscLevel_{ij,t-2} > 0, \quad t \geq 3 \\
\quad \quad \quad EscLevel_{ijt} > 0, \quad t < 3 \\
0, & \text{if } EscLevel_{ijt} - EscLevel_{ij,t-2} = 0, \quad t \geq 3 \\
\quad \quad \quad EscLevel_{ijt} = 0, \quad t < 3 
\end{cases}$$ (2.1)

In the following section, we incorporate these escalation characteristics into the econometric model that we present and explain how we identify the effect of different escalation strategies on the probability of becoming adherent again.

2.4.3. Econometric Model

To study the effectiveness of the feedback, we conduct discrete-time survival analysis where we view each non-adherence period as a spell that starts when the patient misses the medications and ends when the patient resumes taking the medications again. Discrete-time survival models are often used to model events that happen in truly discrete time and events that happen in continuous time but are observed in discrete intervals, i.e., interval-censored events. In our case, we observe whether the patient resumed taking the medications daily. Many researchers use continuous-time survival models with discrete-time data and interval-censored data because continuous-time models are easier to implement. In contrast, we use a discrete-time model for two reasons. First, we have
many ties in the time when patients became adherent again, especially on the first few days of non-adherence. According to Chalita et al. (2002), one should use a discrete-time model when the proportion of ties is greater than 0.25, which is the case in our data. Second, discrete-time models are less error-prone to implement since our data contain time-varying escalation characteristics that change at high frequency.

To conduct discrete-time survival analysis, we use a complementary log-log (cloglog) model. The cloglog model, which is also known as a discrete-time proportional hazards model, is a mathematically exact time-aggregated version of the continuous-time Cox proportional hazards model (Allison 1982, Prentice and Gloeckler 1978, Jenkins 1995). Like the Cox model, the cloglog model makes no assumption regarding the nature of the hazard function, allowing the baseline probability of becoming adherent on the \( t \)th day of non-adherence to take any distributional form. We conduct our survival analysis at the patient-day level where the discrete-time survival function, \( S(t | \text{Escalation}_{ijt}, X_{ijt}) \), is the probability that patient \( i \) remains non-adherent for at least \( t \) days within non-adherence sequence \( j \); and the discrete-time hazard function, \( H(t | \text{Escalation}_{ijt}, X_{ijt}) \), is the probability that patient \( i \) becomes adherent again on non-adherence day \( t \) within non-adherence sequence \( j \) given that he or she has been non-adherent up until that point. \( \text{Escalation}_{ijt} \) and \( X_{ijt} \) are a vector of escalation characteristics and a vector of control variables, respectively.

Based on the proportional hazards framework, the survival function takes the following form:

\[
S(t | \text{Escalation}_{ijt}, X_{ijt}) = S_0(t)\exp(E\text{scalation}_{ijt}\beta + X_{ijt}\delta)\nu_i
\]  

(2.2)

where \( S_0(t) \) the baseline survival function. Since each patient could have multiple non-adherence sequences, we allow observations within each patient to be correlated by introducing a patient-specific frailty, \( \nu_i \), which is an unobserved quantity that is log-normally distributed with a mean of one and a variance of \( \sigma_{\nu}^2 \). A frailty, or a latent random effect, is the same for each patient and is used to describe unobserved heterogeneity among patients (Meyer 1990). Our model is known as a shared-frailty model, which is the survival-analysis analog to random-effects regression
models. From Equation 2.2, we obtain a similar relationship for the complement of the hazard function:

\[ 1 - H(t|Escalation_{ijt}, X_{ijt}) = (1 - H_0(t)) \cdot \exp(Escalation_{ijt} \cdot \beta + X_{ijt} \cdot \delta) \cdot \nu_i \]  

(2.3)

where \( H_0(t) \) is the baseline hazard on non-adherence day \( t \). We can rewrite Equation 2.3 in the form of a cloglog model as follows (Kalbfleisch and Prentice 2002, Allison 1982).

\[
cloglog(H(t|Escalation_{ijt}, X_{ijt})) = \ln(-\ln(1 - H(t|Escalation_{ijt}, X_{ijt}))) = \alpha_t + Escalation_{ijt} \cdot \beta + X_{ijt} \cdot \delta + \mu_i
\]

(2.4)

where \( cloglog(\cdot) \) is a complementary log-log link function, and \( \alpha_t \) is a constant for non-adherence day \( t \) representing the baseline hazard and is equal to \( cloglog(H_0(t)) \). The log frailty \( \mu_i \), or \( \ln(\nu_i) \), is analogous to random effects in standard regression models and is assumed to be i.i.d., \( N(0, \sigma^2_\mu) \).

We can determine whether it is necessary to include the frailty by testing the hypothesis that the proportion of total variance contributed by the patient-heterogeneity variance \( (\sigma^2_\mu) \) is equal to zero.

Similar to the Cox model, we can interpret the exponent of a coefficient as a hazard ratio, which captures a proportional shift in the hazard due to a unit change in the associated covariate given all other factors, including the frailty, being equal.

The main treatment variable in our model is \( Escalation_{ijt} \), which consists of \( EscLevel_{ijt} \), \( OverallEsc_{ijt} \), and \( RecentEsc_{ijt} \). We model \( EscLevel_{ijt} \) as a factor variable to allow for a non-linear effect. Additionally, we include \( X_{ijt} \) to control for seasonality and other factors that potentially influence medication adherence. Specifically, \( X_{ijt} \) consists of the following.

1. **Seasonality**: We control for two main seasonality factors: (1) weekends and holidays and (2) days since enrollment. To control for weekends and holidays, we include a binary indicator that equals one if the observation falls on a weekend or a federal holiday. Additionally, since Staats et al. (2017) suggests that individuals’ compliance may vary by how long they
have been monitored, we control for the number of days the patient had been enrolled in the program.

2. **Financial outcomes**: Our original study used lottery-based incentives to promote medication compliance in addition to providing real-time intervention. Each day, the electronic platform randomly selected a lottery number, which was then compared with the patient’s assigned lottery number. If the numbers matched under certain criteria and the patient was adherent the previous day, the patient would receive a monetary prize. Adherent patients were told if they won, and non-adherent patients were told if they would have won had they been adherent, allowing us to leverage regret aversion and anticipated regret. To control for lottery outcomes, we include $\text{Regret}_{ijt}$, which is a binary variable that equals 1 if the patient was eligible to win but did not receive the prize due to non-adherence.

3. **Patient characteristics**: Random-effects models allow us to include time-invariant patient characteristics. We include age, gender, PHQ-2 score, baseline Elixhauser score, and whether or not the patient is enrolled in Medicare.

4. **Variation in practices among EAs**: Since the EAs possibly had varying intervention delivery practices that may in turn affect patients’ adherence behavior, we include EA fixed effects to control for this.

In our analysis, we aim to separately identify the effect of escalation level and the effect of escalation dynamics. We are able to do so because the correlations among $\text{EscLevel}$, $\text{OverallEsc}$, and $\text{RecentEsc}$ are low. Specifically, we find that the magnitudes of correlation among them are all below 0.5. Moreover, we also aim to disentangle the effect of escalation from the effect of time. Although we control for the baseline hazard rate on non-adherence day $t$, high correlations between escalation characteristics and non-adherence day may prevent us from reliably identifying the underlying baseline hazard rates and the effect of escalation. To address this, we obtain the correlation coefficients between (1) non-adherence day and each $\text{EscLevel}$, (2) non-adherence day
and OverallEsc, and (3) non-adherence day and RecentEsc, and find that the magnitudes of the correlation coefficients are below 0.62, with most being below 0.1.

Even though we do not have a high correlation problem, it is still possible that the effect of escalation level depends on when the escalation level is reached. For example, patients may be less sensitive to escalation level 4 that occurs on non-adherence day 6 than one that occurs on non-adherence day 3 because when patients reach the sixth day of non-adherence, they are naturally more non-adherent and are less likely to respond to intervention. If this were to be true, the fact that higher levels of escalation tend to occur later within a non-adherence sequence could lead to a downward bias in the estimated effect of escalation level. This would also violate the proportionality assumption of the discrete-time hazard model, which requires the effect of escalation level to be identical in every non-adherence day \( t \). To address this, we will test the proportionality hazard assumption by adding the interaction terms between escalation levels and non-adherence day [Singer and Willett 2003, Therneau and Grambsch 2000].

2.4.4. Results

Table 20 summarizes the relationship between escalation characteristics and the probability of becoming adherent again in terms of hazard ratios. As we are primarily interested in estimating the effect of feedback escalation, we only report the effects of EscLevel, OverallEsc, and RecentEsc. We do not find a statistically significant relationship between the lowest escalation levels that involve only automated messages and the probability of returning to adherence. However, higher escalation levels that involve more customized feedback and a greater degree of personal involvement were associated with greater patient responsiveness. Holding everything else constant, the probability of becoming adherent increased by 28% when the patient started receiving personal phone calls, and increased slightly higher when the patient started receiving manual messages. More remarkably, when the patient’s partner started receiving phone calls, the probability of becoming adherent again increased by more than 35%. Additionally, we find that patients were more likely to become adherent again when the intervention was escalated quickly and when they received a
recent escalation. When the patient received an escalation in the past three days, the probability of becoming adherent again increased by 32%. To check for robustness, we also consider voicemails in addition to calls that successfully reached the intended recipient. Although the estimated effects of escalation level 3 (call to patient) and escalation level 5 (call to partner) are slightly smaller after including voicemails, the findings are qualitatively similar.

To further investigate if the escalation levels that only involve automated messages are ineffective in making the patients adherent again or if we simply do not find an effect because of a low signal-to-noise ratio, we perform equivalence tests [Harms and Lakens 2018, Rogers et al. 1993]. The first step of the equivalence test is to specify the smallest effect size of interest (SESOI), which is the effect size that we consider too small to be meaningful. To determine what the SESOI should be, we first hypothetically assume that patients did not receive any intervention. Then, we determine by what percentage, on average, the hazard rates would have to increase for the patients to be more likely than not to become adherent when they were currently less likely than not to become adherent. We find the average minimum increase to be 2.882%. Using 2.882% as the SESOI, we perform two one-sided t-tests to determine if we can reject the hypothesis that the effect size is larger than or equal to the SESOI. The test results suggest that we can reject the null hypothesis for both escalation level 1 (p-values < 0.001) and escalation level 2 (p-values < 0.05). Therefore, the effects of escalation levels 1 and 2, if they exist, are likely to be of negligible significance.

Table 20: Estimated Effect of Escalation on the Probability of Becoming Adherent

<table>
<thead>
<tr>
<th>Variable</th>
<th>Hazard Ratio (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EscLevel</strong></td>
<td></td>
</tr>
<tr>
<td>Level 0 – Intervention not yet delivered (base case)</td>
<td>1.01 (0.00)</td>
</tr>
<tr>
<td>Level 1 – Automated message to patient</td>
<td>1.01 (0.01)</td>
</tr>
<tr>
<td>Level 2 – Automated message to partner</td>
<td>1.28*** (0.05)</td>
</tr>
<tr>
<td>Level 3 – Phone call to patient</td>
<td>1.29** (0.10)</td>
</tr>
<tr>
<td>Level 4 – Manual message to patient</td>
<td>1.36** (0.16)</td>
</tr>
<tr>
<td>Level 5 – Phone call to partner</td>
<td></td>
</tr>
<tr>
<td><strong>OverallEsc</strong></td>
<td>1.18* (0.09)</td>
</tr>
<tr>
<td><strong>RecentEsc</strong></td>
<td>1.32*** (0.12)</td>
</tr>
</tbody>
</table>

*Notes: Standard error in parentheses. *(p < 5%), **(p < 1%), *** (p < 0.1%).
To determine whether the effect of escalation level depends on non-adherence day, we compare a model that includes interaction terms between escalation levels and non-adherence day to a model that does not. We find that interaction coefficients are not statistically significant. Moreover, by conducting a deviance goodness-of-fit test, we find that the model that allows for a time-dependent effect does not provide a better fit than the proportional hazards model. This suggests that there is insufficient evidence to indicate that the effect of escalation level depends on non-adherence day, and, therefore, the fact that higher levels of escalation tended to occur later in a non-adherence sequence should not bias the estimates. Furthermore, to determine whether patient-specific frailties are necessary, we conduct a likelihood ratio test to evaluate the null hypothesis that the proportion of the total variance contributed by the patient-heterogeneity variance \( \sigma^2_p \) is zero. The \( p \)-value for the likelihood ratio test is less than 0.001. Thus, we can reject the null hypothesis and conclude that patient-specific frailties are important.

Figure 8 shows the baseline hazard for each non-adherence day \( t \), which translates to the probability that the patient becomes adherent again on the \( t \)th day of non-adherence given that all covariates are zero. As we can see, the baseline hazard tended to decrease over time, suggesting that patients were less likely to become adherent again the longer they had been non-adherent. We report the full estimation results except for the baseline hazards and patient-specific frailties in Table 28 in the Appendix. In contrast to Staats et al. (2017), which finds that employees were less likely to be compliant the longer they had been monitored, we do not find that the probability of becoming adherent changed the longer the patient had been enrolled in the program. One possible reason for this difference is that, in our study, we not only monitored patients’ adherence but also proactively reminded patients to take medications and provided financial incentives to promote adherence throughout the study period. Due to these constant stimulations, patients’ likelihood of becoming adherent was likely unaffected by how long they had been in the program. This finding potentially implies that compliance monitoring alone may not be enough to maintain compliance—one also needs to consistently provide patients (or employees) with behavioral feedback and/or incentives to promote compliance.
The results from our study suggest that, in order to effectively make a previously non-adherent patient adherent again, we should start providing personalized feedback as quickly and as consistently as possible. Specifically, in our setting, we should start calling the patient immediately after the first day of non-adherence and, if non-adherence continues, escalate to manual messages and calls to the patient’s partner successively on the following days. When combining the effect of escalation level and escalation dynamics, we find that receiving a personal phone call on the second day of non-adherence, a manual message on the third day of non-adherence, and a partner call on the fourth day of non-adherence more than doubled the probabilities of becoming adherent during those days.

2.5. Study B – Impact of Medication Adherence on Readmission

To investigate how behavior change improved health outcomes, we explore the effect of medication adherence on the likelihood of readmission. Specifically, we aim to examine how consistency in medication use in the past affects the readmission probability in the present.

2.5.1. Data and Statistics

We utilize a panel data set that includes (1) daily data for each patient indicating whether he or she opened each of the pill bottles and (2) daily data of patient readmissions. Consistent with most
prior work, we consider as readmissions all-cause inpatient hospitalizations and observation stays, but not emergency room visits. These criteria are also frequently used when hospital performance is assessed. Table 21 provides the distribution of the number of readmissions between the initial discharge and the end of the one-year study period. As the figure shows, approximately 36% of the patients in our cohort were readmitted at least once between their initial hospital discharge and the end of the study period. Apart from medication adherence, patient characteristics may also contribute to the risk of readmission. Table 22 provides summary statistics of patient demographics for all patients in our cohort along with a breakdown by whether or not the patients were readmitted between their initial discharge and the end of the study period.

Table 21: Distribution of the Number of Readmissions between Initial Discharge and the End of Study Period

<table>
<thead>
<tr>
<th>Number of Readmissions</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>64.41%</td>
</tr>
<tr>
<td>1</td>
<td>20.00%</td>
</tr>
<tr>
<td>2</td>
<td>7.69%</td>
</tr>
<tr>
<td>3</td>
<td>3.49%</td>
</tr>
<tr>
<td>4</td>
<td>2.15%</td>
</tr>
<tr>
<td>5 and above</td>
<td>2.26%</td>
</tr>
</tbody>
</table>

Table 22: Summary Statistics of Patient Demographics by Readmission Profile

<table>
<thead>
<tr>
<th>Variable</th>
<th>Not Readmitted (N = 628)</th>
<th>Readmitted (N = 347)</th>
<th>All (N = 975)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>std</td>
<td>min</td>
</tr>
<tr>
<td>Age</td>
<td>61.03</td>
<td>10.35</td>
<td>23</td>
</tr>
<tr>
<td>Female</td>
<td>0.31</td>
<td>0.46</td>
<td>0</td>
</tr>
<tr>
<td>Medicare</td>
<td>0.39</td>
<td>0.49</td>
<td>0</td>
</tr>
<tr>
<td>PHQ-2 depression score</td>
<td>1.10</td>
<td>1.53</td>
<td>0</td>
</tr>
<tr>
<td>Baseline Elixhauser score</td>
<td>5.11</td>
<td>9.12</td>
<td>-14</td>
</tr>
</tbody>
</table>

In our study, we analyze the impact of adherence to two medications: statins (which lower cholesterol) and β-blockers (which lower the heart rate). Each pill bottle the patients received was dedicated to either one of the medications. Statins are a class of drugs that help lower cholesterol levels in the blood. β-blockers are usually prescribed to patients with high blood pressure. Statins and β-blockers are generally well-tolerated and nearly universally recommended to patients following heart attack. Although we were able to track adherence to each of the two medications separately, we consider adherence to statins and β-blockers combined because the
correlation between adherence to the two medications is high. For these medications, we are interested in studying the impact of short- and long-term adherence on readmission. We define short-term adherence as adherence measured over the three days leading up to any given day, and long-term adherence as adherence measured over the 120 days leading up to any given day. We drop the first 120 days of the study for each patient because they do not have long-term adherence information. The adherence measures are binary, taking the value of 1 when the average adherence within a given time window is at least 80% and taking the value of 0 otherwise. The 80% cutoff is widely used in medication adherence literature (Burnier 2019). Given that patients took one dosage of each medication per day, this adherence cutoff implies that they had to take medications for three out of three days to be adherent short-term.

Our main consideration in determining the appropriate lengths of short- and long-term adherence windows was the correlation between short- and long-term adherence. We find that the larger the difference between the lengths of short- and long-term adherence windows is, the lower the correlation between short- and long-term adherence will be. We use three days for short-term adherence and 120 days for long-term adherence because the correlation between 3-day and 120-day adherence is only 0.398. Of course, the correlation would be even lower if we used a longer-length long-term adherence window, e.g., 150 days. However, we would have to drop more observations that do not have long-term adherence information. We believe that, by using three and 120 days, we can balance the need to minimize the correlation with the preservation of information. Our choices of adherence windows are also supported by the LASSO regression, which suggests that 3-day adherence and 120-day adherence are the most contributive adherence measures.

To ensure that dropping the first 120 days of the study does not cause a sample selection issue, we examine whether adherence behavior varied with how long the patients had been in the program. First, we consider the effect of the number of days since enrollment, which is a control variable in Study A, on the probability of becoming adherent. The results indicate that the effect of the number of days since enrollment is not statistically significant (p-value = 0.409). Second, we consider the average adherence rates before and after day 120. We find that the average adherence rates
before and after day 120 were 0.810 and 0.805, respectively. We also find that the difference is not statistically significant ($p$-value = 0.102). These results suggest that there exists insufficient evidence to indicate that patients’ adherence behavior was different in the first 120 days. Therefore, dropping the first 120 days of the study should not cause a sample selection issue.

Table 23 shows average short- and long-term adherence to statins and $\beta$-blockers leading up to days patients were readmitted and average adherence leading up to days patients were not readmitted. We find that average short- and long-term adherence was lower leading up to days patients were readmitted. For both adherence measures, we assume that patients were prescribed a medication from the time of the first fill until the end of the study period. To reduce errors in measuring adherence, we exclude patients whose communication history indicated that they had difficulties setting up their pill bottles or that their pill bottles did not work as intended. We also remove days that patients stayed in the hospital and assume that they were adherent on those days because they were under hospital care.

Table 23: Average Short- and Long-Term Adherence to Statins and $\beta$-blockers

<table>
<thead>
<tr>
<th>Variable</th>
<th>Prior to Day with No Readmission</th>
<th>Prior to Day with Readmission</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-day statin and $\beta$-blocker adherence</td>
<td>0.885</td>
<td>0.802</td>
</tr>
<tr>
<td>120-day statin and $\beta$-blocker adherence</td>
<td>0.880</td>
<td>0.746</td>
</tr>
</tbody>
</table>

2.5.2. Econometric Model

To study the effect of medication adherence on the risk of readmission, we conduct discrete-time survival analysis where we view each healthy period as a spell that starts when the patient is discharged from the hospital and ends when the patient is readmitted. Instead of using the cloglog link function, we use the logit link function, which is another common link function for discrete-time hazard models (Allison 1982, Cox 1972). The logit function allows us to incorporate patient fixed effects, which account for all observable (both available and unavailable in our data) and unobservable patient-specific factors that influence the patient’s likelihood of readmission and do not vary over the span of the study period. Incorporating patient fixed effects ensures that our estimates are immune to an omitted variable bias due to unobservable time-invariant factors that are
correlated with both medication adherence and readmission. For example, patients who consistently feel unwell may be more likely to adhere to their medications and, because they are likely in poorer health, are also more likely to be readmitted. On the other hand, one can also argue that healthy patients who are less likely to be readmitted tend to have higher motivation and thus are likely to be more adherent. This is known in the literature as the healthy-user effect (Shrank et al. 2011). Patient fixed effects adjust for these potential time-invariant attributes that are also confounding variables.

Although we do not have a high proportion of ties in the failure time and could potentially use a continuous-time model for convenience, we use a discrete-time model because it allows us to extend the model to account for time-varying confounders afterwards. We conduct our analysis at the patient-day level where the discrete-time hazard function, \( H(t|\text{Adherence}_{ijt}, Z_{ijt}) \), is the probability that patient \( i \) is readmitted on the \( t^{th} \) day since last discharge within healthy period \( j \). We specify our model as follows:

\[
\logit(H(t|\text{Adherence}_{ijt}, Z_{ijt})) = \alpha_t + \text{Adherence}_{ijt} \beta + Z_{ijt} \delta + \gamma_i \tag{2.5}
\]

where \( \text{Adherence}_{ijt} \) is a vector of short- and long-term adherence measures; \( \alpha_t \) is the logit transformation of the baseline hazard (\( \logit(H_0(t)) \)); and \( \gamma_i \) is patient \( i \)’s fixed effect. \( Z_{ijt} \) is a vector of control variables, which include (1) a binary variable indicating whether the observation falls on a weekend or a holiday, and (2) the number of previous readmissions, i.e., the number of times the patient had been readmitted prior to the present day. Unlike our original randomized controlled study, this model allows us to evaluate the probability of readmission by leveraging within-patient intertemporal variation in medication adherence.

The fixed-effects logit model is a discrete-time equivalent of the stratified Cox model, i.e., Cox model with each patient treated as a separate stratum (Allison 1996, Allison and Christakis 2006). As opposed to the cloglog and Cox models where the exponent of a coefficient is simply a hazard ratio, the exponent of a coefficient from the logit model is an odds ratio of hazard rates (Cox 1972). However, this converges to a hazard ratio in our case because the probabilities of readmission are generally very small.
To fit the fixed-effects logit model, we use a conditional logistic approach where we group data by patient and calculate the likelihood relative to each patient group, i.e., a conditional likelihood is used (Chamberlain 1980). Through this approach, $\gamma_i$’s are not directly estimated since the conditional likelihood does not directly involve the patient fixed effects. We estimate the model using robust-cluster standard errors, i.e., standard errors adjusted for clustering of observations within patients, to allow a patient’s observations to be correlated. In all our analyses, we check for robustness by varying the length of short-term and long-term adherence windows and by varying the adherence threshold. Specifically, we consider 5 and 7 days as the short-adherence window, and 80 and 100 days as the long-term adherence window. For the adherence threshold, we also consider using 60%, 70%, and 90%.

2.5.3. Results

Table 24 summarizes the relationship between short- and long-term medication adherence and readmission. The odds ratio for long-term adherence is statistically significant and less than one, suggesting that adhering to statins and $\beta$-blockers in the long term was associated with a reduced likelihood of readmission. Specifically, we observe that taking statins and $\beta$-blockers at least 80% of the required amount during the past 120 days was associated with a 51% reduction in the odds of readmission. This translates to a similar-sized reduction in the risk of readmission since the probability of readmission is generally very small. Given that patient fixed effects are zero, we find that the average marginal effect of long-term adherence is -0.373%, which is quite significant considering that the probability of readmission is usually lower than 1%.

<table>
<thead>
<tr>
<th>Adherence Measure</th>
<th>Odds Ratio (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-day statin and $\beta$-blocker adherence</td>
<td>0.75 (0.29)</td>
</tr>
<tr>
<td>120-day statin and $\beta$-blocker adherence</td>
<td>0.49** (0.12)</td>
</tr>
</tbody>
</table>

Notes: Standard error in parentheses. *($p < 5%$), **($p < 1%$), ***($p < 0.1%$).

We report the full estimation results in Table 29 in the Appendix and check for robustness when using different specifications of adherence. As shown in Tables 30 and 31 in the Appendix, the
results are robust when we vary the adherence threshold between 60% and 90% and when we vary the short-term adherence window between 3 and 7 days, and the long-term adherence window between 80 to 120 days.

In our analysis, we use patient fixed effects to account for time-invariant unobservable patient attributes. However, there may also exist unobservable time-varying factors that influence both readmission and adherence behavior. For example, patients may be more likely to be adherent at times when they feel unwell or at times when they are healthier and have higher motivation. Although researchers usually view the healthy-user effect as a person-specific time-invariant phenomenon, we do not want to rule out the possibility that it varies with time. To address this, we use an instrumental variable (IV) approach to control for potential endogeneity. As we discuss in detail in the Appendix B, we use intervention as IVs for medication adherence and conduct the analysis using a subset of patients who normally had relatively low adherence and needed intervention to stay adherent. Using the IV approach, we find the results to be quite similar to what we obtain using patient fixed effects. Specifically, we find that the effect of long-term adherence is statistically significant while the effect of short-term adherence is not. For long-term adherence, we find the average marginal effect on the probability of readmission to be -0.340%.

2.6. Readmission Risk Prediction and Its Application in Connected Healthcare

Being able to predict the risk of readmission can help health professionals to effectively intervene with non-adherent patients. We aim to use machine learning to develop a model that predicts readmission risk using recent medication adherence in addition to static risk factors. By utilizing adherence information leading up to each day, one can obtain readmission risk at the patient-day level and use this information when intervening with non-adherent patients. As a care team usually handles a large number of patients and has a limited intervention capacity, it would be helpful for them to be able to identify patients (and patient-days) that are at risk for an imminent readmission and would therefore be prime candidates for the care team’s attention.
In this section, we aim to (1) develop a dynamic readmission risk-scoring model that uses medication adherence as predictors (Section 2.6.1), (2) compare its performance to a baseline model that excludes medication adherence and only contains baseline risk factors such as patient characteristics (Section 2.6.2), and (3) apply the dynamic readmission risk-scoring model to our connected health setting to better target non-adherent patients (Section 2.6.3).

2.6.1. Model Development

Most traditional readmission prediction models only predict the probability of readmission within a fixed period after discharge using static predictors captured at discharge. Unlike previously developed models, we want to develop a model that predicts patient-day level risk using recent medication adherence in addition to static predictors and compare its performance to a baseline model that excludes medication adherence. In particular, we consider the following models:

1. **Baseline risk-scoring model**: The baseline risk-scoring model only includes the following baseline risk factors:

   (a) **Patient characteristics**: These include age, gender, PHQ-2 score, baseline Elixhauser score, and whether or not the patient is enrolled in Medicare.

   (b) **Seasonality**: Seasonality factors include a categorical variable for month of the year and a binary variable indicating whether the observations falls on a weekend or a holiday.

   (c) **Number of previous readmissions**: This is the number of times the patient had been readmitted prior to the present day.

   (d) **Day since last discharge**: This is the number of days since the patient was last discharged.

2. **Dynamic risk-scoring model**: The dynamic risk-scoring model includes the baseline risk factors presented earlier as well as short- and long-term adherence to statins and β-blockers as defined in Study B.
To train our models, we consider five machine learning methods: (1) logistic regression, (2) decision tree\(^1\), (3) random forest\(^1\), (4) support vector machine\(^2\), and (5) multi-layer perceptron\(^3\). These classification methods are relatively well-known and are commonly used in the readmission prediction literature. To prepare the data for model training, we standardize all continuous variables to prevent some machine learning algorithms from putting excessive weight on features with large values. Furthermore, we address class imbalance by using synthetic minority over-sampling technique (SMOTE) (Chawla et al. 2002). Since there exist significantly more patient-days without readmission than patient-days with readmission, the machine learning algorithms may bias toward the majority class. To overcome this issue, SMOTE balances the data by creating synthetic observations using \(k\)-nearest neighbors. In order to create a synthetic observation, SMOTE finds the \(k\) nearest neighbors of each minority observation, selects one of them, and calculates linear interpolations to create a new minority observation in the neighborhood. Research has shown that SMOTE is superior to random oversampling, which is known to increase the likelihood of overfitting.

In evaluating the models, we use a five-fold cross validation at the patient level instead of the patient-day level in order to avoid label leaking, i.e., having observations with similar characteristics in both training and validation sets. We use SMOTE within each cross-validation fold after removing the validation sample so that we create synthetic data by interpolating only observations that will not be used for validation.

2.6.2. Model Evaluation

As a main performance metric, we consider an average area under the ROC curve (AUC) across all five cross-validation folds. Table 25 shows average AUCs along with their standard deviations for the baseline risk models and dynamic risk models that we develop using different machine learning methods. Based on the AUCs, we find that dynamic risk-scoring models outperform baseline risk-scoring models for all classification methods. Furthermore, we find that multi-layer

---

\(^1\)With Gini impurity criterion.
\(^2\)With radial basis function (RBF) kernel.
\(^3\)With rectified linear unit activation function.
perceptron, which is a class of artificial neural networks (ANN), outperforms non-ANN methods that we consider. The average AUC for the dynamic multi-layer perceptron model is 0.941. This means that, if we take two observations, one with readmission and one without readmission, the model can correctly predict which observation is which 94.1% of the time. Using the dynamic multi-layer perceptron model, we find that the predicted probabilities of readmission on a given day range from 0.01% to 1.98%, with an average of 0.16%. This average is reasonable given that readmission occurred on approximately 0.14% of all patient-days.

Table 25: Average Areas Under the ROC Curve (AUC) for All Models

<table>
<thead>
<tr>
<th>Method</th>
<th>Average AUC (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline Model</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>0.740 (0.015)</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.709 (0.051)</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.817 (0.020)</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>0.767 (0.074)</td>
</tr>
<tr>
<td>Multi-layer perceptron</td>
<td>0.843 (0.017)</td>
</tr>
</tbody>
</table>

To further compare the baseline risk model and the dynamic risk model, we focus on the models developed using multi-layer perceptron and consider a hypothetical scenario where the care team has a limited capacity to intervene with patients to prevent hospital readmissions, e.g., sending medical professionals to patients’ homes. We suppose that the care team can target only $c\%$ of all patient-days using one of the three approaches: (1) using no information from a predictive model, (2) using information from the baseline risk model, and (3) using information from the dynamic risk model. If using the first approach, the care team will randomly target $c\%$ of patients each day. If using the second or third approach, the care team will target the patient when his or her predicted risk is higher than $(100 - c)^{th}$ percentile.

To evaluate the three approaches, we define patient-day at risk as patient-day with readmission or prior to readmission, which accounts for 0.28% of all patient-days, and consider the following performance measures:

$$Yield = \frac{\text{Number of patient-days at risk we target}}{\text{Number of all patient-days at risk}}$$

(2.6)
\[ \text{Precision} = \frac{\text{Number of patient-days at risk we target}}{\text{Number of all patient-days we target}} \]  

(2.7)

We calculate yield and precision using the validation sample across all cross-validation folds. As seen in Figures 9 and 10, we obtain higher yield and precision when we use the dynamic risk model as opposed to the baseline model, and when we use the baseline risk model as opposed to not using a predictive model at all. Figure 9 shows that yield increases as capacity increases. We find that, if we have the capacity to target 17% of patient-days using the dynamic model, we will be able to successfully target all patient-days at risk. In contrast, if we use the baseline model or use no information from a predictive model, we will be able to successfully target only 75.32% and 17% of the patient-days at risk, respectively. Figure 10 shows that precision generally decreases as we target more patient-days. If we have the capacity to target only 0.1% of patient-days using the dynamic model, as many as 24.56% of the patient-days we target will be patient-days at risk. In contrast, if we use the baseline model or use no information from a predictive model, only 7.02% and 0.28% of the patient-days we target will be patient-days at risk, respectively. We also vary the definition of patient-day at risk and show the results in Figures 13 and 14 in the Appendix. As expected, although yield and precision vary according to the definition of patient-day at risk, we always obtain highest yield and precision when using the dynamic risk model.

Our finding suggests that the variation in medication adherence over time provides useful information about the readmission risk and enhances the quality of prediction in addition to baseline risk factors. We note that the clinical value of adherence-based predictive models depends critically on the real time availability of adherence data.

2.6.3. Application of Dynamic Readmission Risk-Scoring Model in Connected Healthcare

Predictive analytics can benefit connected healthcare by helping to identify patients who really need to change their behavior to avoid adverse health outcomes. In our setting, the dynamic readmission risk-scoring model can help identify patients who are at the highest risk of readmission and have the greatest need to be adherent. We investigate the benefit of using the dynamic risk-scoring model...
in our setting through counterfactual simulations, which link intervention to readmission by using medication adherence as a mediator. We focus on the multi-layer perceptron model and numerically examine the impact of different intervention strategies, which do and do not consider predicted readmission risks when delivering adherence intervention. The strategies that we consider are as follows.

1. **Strategy A—Adherence-maximizing strategy**: The adherence-maximizing strategy aims to quickly make patients adherent again once they become non-adherent. We define the adherence-maximizing strategy based on the findings from Study A. The strategy involves calling the patient on the second day of non-adherence, sending a manual message on the third day of non-adherence, and calling the patient’s partner on the fourth day of non-adherence.
2. **Strategy B—Adherence-maximizing strategy with readmission risk prioritization**: We use the adherence-maximizing strategy, but the strategy is only triggered when the predicted readmission risk is greater than or equal to the $n^{th}$ percentile. We consider four values of $n$ in our main analysis: 80, 85, 90, and 95.

3. **Strategy C—No intervention**: We do not deliver any intervention.

For each intervention strategy, we are interested in four outcome metrics: (1) total number of calls and manual messages delivered, (2) average individual adherence rate, (3) number of patients who are readmitted at least once, and (4) total number of readmissions. We use the empirical numbers that we observe in our experiment as a baseline.

Our counterfactual simulations link intervention to medication adherence using the estimated proportional hazards model from Study A, and link medication adherence to readmission using the dynamic risk-scoring model. The unit of consideration in our simulation is patient $i$-day $t$. We consider 930 patients in our focus cohort. For each patient, $t$ starts from 120 until the patient’s last day in the study. We illustrate a simulation flow for patient $i$ in Figure 11 and present our simulation parameters in Table 32 in the Appendix. In Figure 11, $Adherent_{it}$ is a binary variable that equals one when patient $i$ takes all of his or her medications on day $t$ and, at the beginning of the simulation, takes the actual observed value in the data. When the patient is non-adherent, we calculate the probability that the patient will become adherent again using the proportional hazards model from Study A, which takes into account the intervention the patient receives. Based on that probability, we determine whether the patient becomes adherent again using random sampling. $PredictedReadm_{it}$ is a binary variable that equals one when patient $i$ is readmitted on day $t$. We calculate the probability of readmission using the dynamic risk-scoring model and determine whether the patient is readmitted using random sampling. If the patient is readmitted, he or she will be in the hospital for $LOS$ days, where $LOS$ follows the empirical distribution shown in Table 32.

We replicate the simulation 1,000 times and obtain the averages for the four outcomes of interest.
Notes: (1) This chart depicts a simulation flow for patient $i$. (2) $t$ starts from 120 until patient $i$’s last day in the study.

Figure 11: Simulation Flow

We report the simulation results for different intervention strategies in Table 26 and, to visualize the results, plot the total number of readmissions against the number of calls and manual messages delivered in Figure 12. We find that although Strategy A, which is the most aggressive strategy, can increase adherence to as high as 98.36%, it may not be a practical way to reduce readmissions. While it is true that high medication adherence leads to lower readmission risk, it may be unnecessary to increase medication adherence among patients who already have relatively low readmission risk and are unlikely to be readmitted. As Figure 12 shows, compared to Strategy A, Strategy B leads to a significant decrease in the total number of calls and manual messages but only a small increase in the total number of readmissions. When $n$ equals to 80 (i.e., we intervene with patients only when the predicted risk is greater than or equal to the 80th percentile), Strategy B leads to a 70% decrease in the total number of calls and manual messages but only a 2% increase in the total number of readmissions compared to Strategy A. Furthermore, we find that, compared to the actual case we observe, Strategy B with $n$ equal to 95 yields a comparable number of calls.
and manual messages but a significantly lower number of readmissions. Specifically, when using the strategy, we deliver 3% fewer calls and manual messages and obtain 10% fewer readmissions compared to what we empirically observe in the experiment. These results suggest that, by utilizing predicted risk information, the care team can effectively allocate its capacity to the patient-days that really need intervention.

Table 26: Simulation Results for Different Intervention Strategies

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Observed</th>
<th>Strategy A (n = 80)</th>
<th>Strategy B (n = 80)</th>
<th>Strategy B (n = 95)</th>
<th>Strategy C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of manual interventions</td>
<td>327</td>
<td>3453.9</td>
<td>1043.2</td>
<td>318.4</td>
<td>0</td>
</tr>
<tr>
<td>Call to patient</td>
<td>210</td>
<td>3190.5</td>
<td>960.1</td>
<td>294.0</td>
<td>0</td>
</tr>
<tr>
<td>Manual message to patient</td>
<td>83</td>
<td>239.4</td>
<td>74.8</td>
<td>21.1</td>
<td>0</td>
</tr>
<tr>
<td>Call to partner</td>
<td>34</td>
<td>24.1</td>
<td>8.2</td>
<td>3.2</td>
<td>0</td>
</tr>
<tr>
<td>Average individual adherence rate</td>
<td>85.50%</td>
<td>98.36%</td>
<td>92.03%</td>
<td>86.94%</td>
<td>81.43%</td>
</tr>
<tr>
<td>Total number of readmissions</td>
<td>329</td>
<td>281.9</td>
<td>286.3</td>
<td>297.3</td>
<td>340.0</td>
</tr>
<tr>
<td>Number of patients readmitted at least once</td>
<td>288</td>
<td>259.8</td>
<td>261.2</td>
<td>266.2</td>
<td>290.6</td>
</tr>
</tbody>
</table>

Notes: Results for Strategy B, n = 85 and 90, are show in Table 33 in the Appendix.

Figure 12: Numbers of Manual Interventions and Readmissions for Different Intervention Strategies

Our simulation findings not only highlight the importance of predictive analytics in connected healthcare but also provide a possible explanation for why we did not observe a significant reduction in readmissions in our original randomized controlled trial. Since we could deliver only a limited number of interventions and did not consider readmission risks, it is possible that we delivered many interventions to low-risk patient-days and overlooked high-risk patient-days. As Figure 12 shows,
we find that the observed number of readmissions is very close to the number of readmissions we obtain in a simulated scenario where no intervention is delivered. To further investigate this, we apply the readmission risk-scoring model to the observed data and find that more than half of the patient-days that received manual messages or calls had readmission risks below the 59th percentile. Moreover, only 11.42% of the patient-days that received manual messages or calls had readmission risks above the 95th percentile. This evidence supports the explanation that we may have not delivered enough interventions to patient-days with high readmission risk and that our intervention would have been more effective if it had been based on dynamic risk scoring.

2.7. Conclusions and Managerial Insights

This paper studies the effectiveness of a connected health system that aimed to reduce the number of readmissions through better medication adherence and examines the benefit of predictive analytics in connected healthcare. By utilizing micro-level intervention and adherence data, we find that patients are significantly more likely to become adherent again when they receive high levels of intervention that involve personalized feedback and when the intervention is escalated quickly and continuously. Our findings highlight the importance of personal involvement and speed when delivering connected health intervention, which has not been widely explored by healthcare researchers. Although the marketing and psychology literature has documented the benefit of personal involvement in changing customers’ behavior (e.g., Gordon et al. [1998]), not much has been studied in the context of healthcare.

To investigate the extent to which behavior change could improve patients’ health outcomes, we explore the effect of short- and long-term medication adherence to two crucial medications on readmission risk. We find that, for patients with cardiovascular disease, long-term adherence to statins and \( \beta \)-blockers is associated with a 51% reduction in the odds of being readmitted on any given day. Although we do not find a significant effect of short-term adherence, the lack of significance does not suggest that short-term medication adherence is not important. Rather, our finding implies that, in order to significantly benefit from adherence, patients need to consistently...
maintain their adherence over a long period. The benefit of long-term adherence is supported by the view in the medical and behavioral compliance literature, which suggests that long-term adherence to medications and treatments is crucial to maintaining good health (Sabate 2003). Most importantly, our findings emphasize the importance of connected healthcare as a tool to improve compliance in the long term by continuously connecting with patients over a long period, as opposed to traditional care where such connectivity is not possible.

In addition to studying the impact of the intervention on medication adherence and the impact of adherence on readmission, we develop a dynamic readmission-risk scoring model that considers real-time medication adherence and find that the model outperforms a baseline risk model that does not consider medication adherence. The results suggest that real-time behavior information can be useful in predicting health outcomes. Although real-time behavior data were not extensively available in the past, they are becoming more widely available thanks to an increasing availability of connected health devices.

Lastly, we examine the benefit of using the dynamic risk model in connected healthcare when the care team has a limited capacity. Using counterfactual simulations, we find that, when using an intervention strategy that prioritizes patient-days with highest readmission risk, we can achieve a significantly lower number of readmissions than we would obtain without considering readmission risk while maintaining the same amount of effort. The dynamic risk model helps the care team allocate their resources to high-risk patients who are non-adherent and have the greatest need to stay adherent. Our findings underscore the value of predictive analytics in connected healthcare. Since connected health technologies allow care providers to conveniently connect with patients, it also enables them to oversee more patients than they traditionally could. With predictive analytics, care providers can optimally allocate their limited resources to achieve the desired outcome. Predictive analytics is not only useful for connected healthcare, but also for other industries that want to implement a connected strategy and make the most out of their limited resources. For example, in financial advising, instead of contacting consumer investors and asking them to rebalance their account every time their asset allocation changes or their cash balance is low, investment firms can
use predictive models to target only the investors whose portfolio health is more likely to be affected if they do not rebalance their account.

There exist a number of opportunities for future work. To better understand the effectiveness of connected healthcare, future research could explore whether and how different patient demographics respond differently to intervention. Since it is possible that some patients are more sensitive to certain types of intervention than others, understanding these differences can help health professionals design an effective intervention strategy to target different demographics. We hope that by providing the data that we obtained as part of a connected healthcare experiment to the research community, future research can take advantage of the clinical interventions we performed. Beyond using the data that we collected, we also believe that future research should conduct a randomized controlled trial in which adherence behavior is tracked in both the control and intervention groups, but feedback is delivered only to intervention patients. This would improve our understanding on how exactly patients react to feedback and reminders. Moreover, future connected healthcare studies could conduct a randomized controlled trial that utilizes a predictive model similar to what we propose to better understand how predictive analytics can actually benefit connected healthcare.

Acknowledgments: The randomized controlled trial was funded with a grant from Center for Medicare & Medicaid Innovation (CMS), Health Care Innovation Award 1C1CMS331009.
Table 27: Data Description

<table>
<thead>
<tr>
<th>Data</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adherence information</td>
<td>Daily information on whether the patient opened each of the pill bottles.</td>
</tr>
<tr>
<td>Intervention information</td>
<td>Daily information on whether our care team delivered each type of intervention to the patient, the patient’s partner, or the patient’s provider’s office.</td>
</tr>
<tr>
<td>Financial incentive outcomes</td>
<td>Daily lottery results, indicating whether the patient was eligible to win but did not receive the prize due to non-adherence.</td>
</tr>
<tr>
<td>Readmission</td>
<td>Readmissions include all-cause inpatient hospitalizations and observation stays, but not emergency room visits. It is a dummy variable, coded 1 if readmission criteria are met and 0 otherwise.</td>
</tr>
<tr>
<td>Age</td>
<td>Patient’s age. It is a continuous variable, coded as piecewise linear spline variables with knots at its 50th and 80th percentiles to account for potential nonlinear effects.</td>
</tr>
<tr>
<td>Gender</td>
<td>Patient’s gender. It is a dummy variable, coded as 1 if the patient is female and 0 otherwise.</td>
</tr>
<tr>
<td>Medicare enrollment</td>
<td>Information on whether the patient is enrolled in Medicare, a national health insurance program in the United States. It is a dummy variable, coded as 1 if the patient is enrolled and 0 otherwise.</td>
</tr>
<tr>
<td>PHQ-2 depression score</td>
<td>Patient’s depression score from Patient Health Questionnaire-2 (PHQ-2), which inquires about the frequency of depressed mood and anhedonia over the past two weeks. The score ranges from 0 to 6 with higher scores translating to higher chances of having depression. We asked patients to complete PHQ-2 once at the time of enrollment.</td>
</tr>
<tr>
<td>Baseline Elixhauser score</td>
<td>Patient’s Elixhauser comorbidity index, which measures comorbidities of patients based on the International Classification of Diseases (ICD) diagnosis code. We assigned the score to patients once at the beginning of the study period using pre-enrollment data. The score is a continuous variable, coded as piecewise linear spline variables with knots at its 50th and 80th percentiles to account for potential nonlinear effects.</td>
</tr>
<tr>
<td>Days since enrollment</td>
<td>The number of days since enrollment.</td>
</tr>
<tr>
<td>Weekend and holiday</td>
<td>Indicator for weekend and holiday. It is a dummy variable, coded as 1 if the day falls on a weekend or a federal holiday.</td>
</tr>
</tbody>
</table>
Table 28: Full Estimation Results for Study A

<table>
<thead>
<tr>
<th>Variable</th>
<th>Hazard Ratio (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EscLevel</strong></td>
<td></td>
</tr>
<tr>
<td>Level 0 – Intervention not yet delivered</td>
<td>(base case)</td>
</tr>
<tr>
<td>Level 1 – Automated message to patient</td>
<td>1.01 (0.00)</td>
</tr>
<tr>
<td>Level 2 – Automated message to partner</td>
<td>1.01 (0.01)</td>
</tr>
<tr>
<td>Level 3 – Phone call to patient</td>
<td>1.28*** (0.05)</td>
</tr>
<tr>
<td>Level 4 – Manual message to patient</td>
<td>1.29** (0.10)</td>
</tr>
<tr>
<td>Level 5 – Phone call to partner</td>
<td>1.36** (0.16)</td>
</tr>
<tr>
<td><strong>RecentEsc</strong></td>
<td>1.32*** (0.12)</td>
</tr>
<tr>
<td><strong>Regret</strong></td>
<td></td>
</tr>
<tr>
<td>Days since enrollment</td>
<td>1.00 (0.06)</td>
</tr>
<tr>
<td>Weekend and holiday</td>
<td>0.91 (0.34)</td>
</tr>
<tr>
<td>EA 1 (base case)</td>
<td></td>
</tr>
<tr>
<td>EA 2</td>
<td>1.05 (0.07)</td>
</tr>
<tr>
<td>EA 3</td>
<td>0.87 (0.32)</td>
</tr>
<tr>
<td>EA 4</td>
<td>0.94 (0.35)</td>
</tr>
<tr>
<td>Age [&lt; 62]</td>
<td>1.00 (0.00)</td>
</tr>
<tr>
<td>Age [62-71]</td>
<td>1.02 (0.01)</td>
</tr>
<tr>
<td>Age [&gt; 71]</td>
<td>0.98** (0.00)</td>
</tr>
<tr>
<td>Female</td>
<td>1.04* (0.01)</td>
</tr>
<tr>
<td>Baseline Elixhauser score [&lt; 5]</td>
<td>0.87 (0.33)</td>
</tr>
<tr>
<td>Baseline Elixhauser score [5-15]</td>
<td>0.92 (0.34)</td>
</tr>
<tr>
<td>Baseline Elixhauser score [&gt; 15]</td>
<td>0.99* (0.00)</td>
</tr>
<tr>
<td>PHQ-2 score</td>
<td>0.41* (0.19)</td>
</tr>
<tr>
<td>Medicare</td>
<td>0.96 (0.04)</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>40,250</td>
</tr>
<tr>
<td><strong>Wald χ²</strong></td>
<td>1319.14 (p &lt; 0.1%)</td>
</tr>
</tbody>
</table>

Notes: Standard error in parentheses. * (p < 5%), ** (p < 1%), *** (p < 0.1%). αt’s are not reported.
Table 29: Full Estimation Results for Study B

<table>
<thead>
<tr>
<th>Variable</th>
<th>Odds Ratio (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-day statin and β-blocker adherence</td>
<td>0.75 (0.29)</td>
</tr>
<tr>
<td>120-day statin and β-blocker adherence</td>
<td>0.49** (0.12)</td>
</tr>
<tr>
<td>Number of previous readmissions</td>
<td>3.02*** (0.61)</td>
</tr>
<tr>
<td>Weekend and holiday</td>
<td>1.17 (0.93)</td>
</tr>
<tr>
<td>Number of observations</td>
<td>67,872</td>
</tr>
<tr>
<td>Pseudo $R^2$</td>
<td>0.283</td>
</tr>
</tbody>
</table>

Notes: Standard error in parentheses. *($p < 5\%$), **($p < 1\%$), ***($p < 0.1\%$). $\alpha_t$’s are not reported.

Table 30: Robustness Test for Study B: Varying Adherence Threshold

<table>
<thead>
<tr>
<th>Adherence threshold</th>
<th>Odds Ratio (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Short-term statin and β-blocker adherence</td>
</tr>
<tr>
<td>60%</td>
<td>0.81 (0.44)</td>
</tr>
<tr>
<td>70%</td>
<td>0.75 (0.34)</td>
</tr>
<tr>
<td>80%</td>
<td>0.75 (0.29)</td>
</tr>
<tr>
<td>90%</td>
<td>0.73 (0.29)</td>
</tr>
<tr>
<td></td>
<td>Long-term statin and β-blocker adherence</td>
</tr>
<tr>
<td>60%</td>
<td>0.54* (0.20)</td>
</tr>
<tr>
<td>70%</td>
<td>0.48** (0.13)</td>
</tr>
<tr>
<td>80%</td>
<td>0.49** (0.12)</td>
</tr>
<tr>
<td>90%</td>
<td>0.46** (0.13)</td>
</tr>
</tbody>
</table>

Notes: Standard error in parentheses. *($p < 5\%$), **($p < 1\%$), ***($p < 0.1\%$). We vary the threshold for short- and long-term adherence simultaneously.

Table 31: Robustness Test for Study B: Varying Short- and Long-Term Adherence Windows

<table>
<thead>
<tr>
<th>Adherence window</th>
<th>Odds Ratio (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Short-term statin and β-blocker adherence</td>
</tr>
<tr>
<td>3-day adherence</td>
<td>0.75 (0.29)</td>
</tr>
<tr>
<td>5-day adherence</td>
<td>0.72 (0.32)</td>
</tr>
<tr>
<td>7-day adherence</td>
<td>0.80 (0.36)</td>
</tr>
<tr>
<td></td>
<td>Long-term statin and β-blocker adherence</td>
</tr>
<tr>
<td>80-day adherence</td>
<td>0.53* (0.23)</td>
</tr>
<tr>
<td>100-day adherence</td>
<td>0.50** (0.15)</td>
</tr>
<tr>
<td>120-day adherence</td>
<td>0.49** (0.12)</td>
</tr>
</tbody>
</table>

Notes: Standard error in parentheses. *($p < 5\%$), **($p < 1\%$), ***($p < 0.1\%$).
Table 32: Simulation Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of becoming adherent again</td>
<td>Follows the proportional hazards model from Study A</td>
</tr>
<tr>
<td>Probability of readmission</td>
<td>Follows the dynamic readmission risk-scoring model</td>
</tr>
<tr>
<td>Random effect distribution for adherence model</td>
<td>$N(0,0.269)$</td>
</tr>
<tr>
<td>Readmission length-of-stay (LOS) distribution</td>
<td>$P(LOS = 1 \text{ day}) = 0.792, P(LOS = 2 \text{ days}) = 0.105, P(LOS = 3 \text{ days}) = 0.065, P(LOS = 4 \text{ days}) = 0.021, P(LOS = 5 \text{ days}) = 0.011, P(LOS = 6 \text{ days}) = 0.006$</td>
</tr>
</tbody>
</table>

Table 33: Simulation Results for Strategy B When $n$ Is 85 and 90

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Outcome</th>
<th>Strategy B ($n = 85$)</th>
<th>Strategy B ($n = 90$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total number of manual interventions</td>
<td>831.3</td>
<td>607.3</td>
</tr>
<tr>
<td></td>
<td>Call to patient</td>
<td>766.7</td>
<td>559.3</td>
</tr>
<tr>
<td></td>
<td>Manual message to patient</td>
<td>58.5</td>
<td>42.9</td>
</tr>
<tr>
<td></td>
<td>Call to partner</td>
<td>6.2</td>
<td>5.1</td>
</tr>
<tr>
<td></td>
<td>Average individual adherence rate</td>
<td>90.11%</td>
<td>88.69%</td>
</tr>
<tr>
<td></td>
<td>Total number of readmissions</td>
<td>289.0</td>
<td>295.7</td>
</tr>
<tr>
<td></td>
<td>Number of patients readmitted at least once</td>
<td>262.9</td>
<td>265.7</td>
</tr>
</tbody>
</table>
Notes: The definition of patient-day at risk in each panel is: (a) patient-day with readmission, (b) patient-day with readmission or prior to readmission (the default), (c) patient-day with readmission or up to two days prior to readmission, and (d) patient-day with readmission or up to three days prior to readmission.

Figure 13: Yield Obtained Using Different Readmission Intervention Approaches: Varying the Definition of Patient-Day at Risk
Notes: The definition of *patient-day at risk* in each panel is: (a) patient-day with readmission, (b) patient-day with readmission or prior to readmission (the default), (c) patient-day with readmission or up to two days prior to readmission, and (d) patient-day with readmission or up to three days prior to readmission.

Figure 14: Precision Obtained Using Different Readmission Intervention Approaches: Varying the Definition of Patient-Day at Risk
2.9. Appendix B: Study B – Instrumental Variable Analysis

As we explain in Section 2.5.3, there may exist unobservable time-varying factors that influence readmission and adherence behavior in addition to time-invariant factors. To address this, we use the instrumental variable (IV) approach to control for both time-invariant and time-varying unobservable factors that may be confounders.

There exist two groups of patients in our study: (1) patients who were usually able to maintain relatively high adherence—they tended to receive few or no interventions when their adherence was high and receive more interventions when their adherence was low; and (2) patients who normally had relatively low adherence and needed intervention to stay adherent—they tended to have higher adherence when they received more interventions and lower adherence when they received fewer or no interventions. For the latter group, we see the possibility of using intervention as IVs for medication adherence. The reasons why we may be able to use intervention as IVs for these patients are because (1) intervention is positively correlated with medication adherence and (2) intervention is likely uncorrelated with unobservable factors that affect readmission. Because of this observation, we focus on this cohort of 115 patients in our IV analysis. Among these patients, the average short- and long-term adherence are 0.512 and 0.544, respectively.

There exist two potentially endogenous binary variables, short-term adherence and long-term adherence. We need at least one IV for each endogenous variable. We want our IVs to be (1) correlated with medication adherence (relevance condition) and (2) uncorrelated with unobservable factors that affect readmission (exogeneity condition). We propose using the number of manual interventions the patient received prior to the present day as IVs. Specifically, for short-term adherence, the IVs that we propose are (1) the number of manual messages the patient received in the past 8 days (i.e., length of adherence window plus five days), (2) the number of phone calls the patient received in the past 8 days, and (3) the number of phone calls the patient’s partner received in the past 8 days. For long-term adherence, the IVs that we propose are (1) the number of manual messages the patient received in the past 125 days (i.e., length of adherence window plus five days),
(2) the number of phone calls the patient received in the past 125 days, and (3) the number of phone calls the patient’s partner received in the past 125 days. By using intervention as IVs, we obtain the effect of medication adherence that varies depending only on how much intervention the patient receives. We will also validate the relevance and exogeneity conditions for these IVs.

Since both readmission and medication adherence are binary variables, we use a bivariate probit model jointly estimated via Full Maximum Likelihood Estimation (FMLE) [Cameron and Trivedi 1998, Greene 2012]. We do not include patient fixed effects because patient fixed effects (or individual fixed effects in general) in a two-stage binary response model may cause an incidental parameter problem and biased estimates [Greene 2004]. However, since the IV approach already accounts for both time-invariant and time-varying unobservables that may be confounders, it is not necessary to include patient fixed effects. We specify our model for patient $i$ and day $t$ as follows:

$$\begin{align*}
\text{ShortTermAdherence}^*_{it} & = Z_{it} \omega + \theta_1 \text{ManualMessages8Days}_{it} + \theta_2 \text{PatientCalls8Days}_{it} \\
& \quad + \theta_3 \text{PartnerCalls8Days}_{it} + \varepsilon^1_{it}, \\
\text{ShortTermAdherence}_{it} & = \mathbb{1}\{\text{ShortTermAdherence}^*_{it} > 0\}, \\
\text{LongTermAdherence}^*_{it} & = Z_{it} \Omega + \Theta_1 \text{ManualMessages125Days}_{it} \\
& \quad + \Theta_2 \text{PatientCalls125Days}_{it} + \Theta_3 \text{PartnerCalls125Days}_{it} + \varepsilon^2_{it}, \\
\text{LongTermAdherence}_{it} & = \mathbb{1}\{\text{LongTermAdherence}^*_{it} > 0\}, \\
\text{Readmitted}^*_{it} & = Z_{it} \delta + \beta_1 \text{ShortTermAdherence}_{it} + \beta_2 \text{LongTermAdherence}_{it} \\
& \quad + \alpha_t + \xi_{it}, \\
\text{Readmitted}_{it} & = \mathbb{1}\{\text{Readmitted}^*_{it} > 0\}
\end{align*}$$

(2.8)

where $\text{ShortTermAdherence}_{it}$ and $\text{LongTermAdherence}_{it}$ are short- and long-term adherence as defined in Study A; $\text{Readmitted}_{it}$ is a binary variable that equals one when readmission occurs; $\alpha_t$ is fixed effects of days since last discharge; and $Z_{it}$ is a vector of control variables, which include (1) patient characteristics (age, gender, PHQ-2 score, Elixhauser score, and Medicare enrollment),
(2) a binary variable indicating weekend and holiday, (3) the number of previous readmissions, (4) days since enrollment, and (5) EA fixed effects.

In Equation 2.8, the error terms \( (\xi_{it}, \varepsilon_{1it}) \) may be correlated to model the endogeneity between the short-term medication adherence and readmission. Similarly, the error terms \( (\xi_{it}, \varepsilon_{2it}) \) may be correlated to model the endogeneity between the long-term medication adherence and readmission. We assume that \( (\xi_{it}, \varepsilon_{1it}) \) and \( (\xi_{it}, \varepsilon_{2it}) \) follow a Standard Bivariate Normal distribution with correlation coefficient \( \rho_1 \) and \( \rho_2 \), respectively. We can conduct the likelihood ratio tests of null \( \rho_1 = 0 \) and null \( \rho_2 = 0 \) to test the presence of endogeneity.

We report the first-stage results in Tables 34 and 35. We find that the effects of the numbers of interventions received on short- and long-term medication adherence are positive and statistically significant. One exception is the effect of the number of manual messages received in the past 8 days on short-term adherence, which is not significant at a 95% confidence level. However, the insignificance is not an issue since we only need at least one IV to be “relevant” to each endogenous variable. The first-stage results suggest that the relevance condition is satisfied.

We show the second-stage results in the column titled “With IVs” in Table 36. Similar to the results obtained from the fixed-effects logit model, we find that the effect of long-term adherence is statistically significant while the effect of short-term adherence is not. For long-term adherence,
the estimated coefficient of -0.24 translates to an average marginal effect of -0.34%. This marginal effect is slightly smaller than the average marginal effect we obtain using the fixed-effects logit model, which controls for only time-invariant confounders. Moreover, to determine what the results will be if we ignore the potential endogeneity, we run only the second-stage model and report the results in the column titled “Without IVs”. We find that the effects of adherence on readmission are exaggerated when we do not account for either time-invariant or time-varying confounders. Our findings suggest that patients who are healthier and are less likely to be readmitted are possibly more adherent to their medications. Table 37 shows the results of endogeneity tests, which suggest that the error terms in the first- and second-stage equations are negatively correlated. Therefore, endogeneity is present.

Table 36: Estimated Effect of Medication Adherence on Readmission

<table>
<thead>
<tr>
<th>Adherence Measure</th>
<th>With IVs</th>
<th>Without IVs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ShortTermAdherence</td>
<td>-0.12 (0.09)</td>
<td>-0.13 (0.10)</td>
</tr>
<tr>
<td>LongTermAdherence</td>
<td>-0.24** (0.11)</td>
<td>-0.33*** (0.04)</td>
</tr>
</tbody>
</table>

Notes: Standard error in parentheses. *(p < 5%), **(p < 1%), *** (p < 0.1%).

Table 37: Results of Endogeneity Tests

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value (SE)</th>
<th>Test ( \rho = 0 ) p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho_1 )</td>
<td>-0.16 (0.06)</td>
<td>0.04</td>
</tr>
<tr>
<td>( \rho_2 )</td>
<td>-0.46 (0.09)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Notes: Standard error in parentheses.

Although we cannot directly verify the exogeneity condition using the bivariate probit model, we can verify the condition using a two-stage least squares (2SLS) model by conducting a Hansen J test of overidentifying restrictions. Although the 2SLS model linearizes binary responses, many economists including Angrist and Pischke (2013) have proposed using 2SLS for nonlinear models with endogenous regressors. Many researchers also adopt a linearized two-stage model partly because it can handle individual fixed effects (e.g., Bavafa et al., 2018). We conduct our analysis using a 2SLS model and find that the results are consistent with those from the bivariate probit
model. Using a Hausman’s specification test, we find that endogeneity is present. Moreover, to verify the exogeneity condition, we conduct the test of overidentifying restrictions, which is possible because the number of IVs exceeds the number of endogenous variables. Based on the test, we find that the Hansen J statistic is 2.624 with a $p$-value of 0.623. This suggests that the IVs are uncorrelated with unobservable factors that affect readmission and, therefore, the exogeneity condition is satisfied.
3.1. Introduction

Video visits are transforming the way patients are able to connect with providers—by providing not only an additional pathway for patients to access care but also one that is more convenient and available sooner. Traditionally, to see a primary care provider in person, patients in the US need to wait 29 days on average for an available appointment (Merritt Hawkins 2017) and spend two hours on average, including travel and wait time, for a visit that lasts an average of 20 minutes (Daschle and Dorsey 2015, Ray et al. 2015). While it is well established that prompt access to care improves both patient satisfaction and health outcomes (Reichert and Jacobs 2018), only 10% of new patients are seen in person the same day (Hayhurst 2017). Video visits remove these geographical and temporal barriers by allowing patients to remotely connect at any time of day or night with their providers through electronic devices. With video visits, patients are able to see their providers not just on the same day, but within minutes, from the comfort of where they are.

In recent years, researchers have been attempting to understand the impact of video visits on various service outcomes, such as patient satisfaction, and examine why patients choose video visits over in-person visits. Many studies go on further to examine how the use of video visits affects the use of traditional in-person visits. However, these studies primarily rely on post-care surveys and simple retrospective analysis and/or focus only on a specific patient demographic, such as the elderly or a specific care specialty. There exists limited empirical research that utilizes large-scale, general care utilization data and comprehensively examines how the availability of video visits impacts overall care utilization and in-person care utilization. Moreover, to the best of our knowledge, no study has utilized both in-system and out-of-system care utilization data to investigate how the availability of video visits affects patients’ decisions regarding whether, how, and where to seek care. The impact of video visit availability on patients’ patterns of care utilization is an important empirical
question, the answer to which can help payors and providers decide whether and how to introduce and promote video visits.

Our study investigates the impact of video visit availability at a large academic medical center, the University of Pennsylvania Health System (UPHS), which has recently introduced video visits to a subset of patients who are employees. Our overarching aims are to understand how the availability of video visits for UPHS’s health system-insured employees impacts their care utilization and the choices that they make when they are sick. As shown in Figure 15, prior to the introduction of video visits, patients (employees) could choose to seek care in person, at UPHS or elsewhere, or not seek care. After the introduction of video visits, their choice set expanded to also include the option of seeking care via a video visit at UPHS. The higher convenience of video visits may make patients who would have previously used an out-of-system provider more likely to seek their care inside UPHS. It is also possible that the increased convenience leads to more care consumption as patients who were previously deterred by the inconvenience might now see a provider via video visit. Additionally, video visits may also act as a gateway leading to subsequent in-person care consumption. Without a careful empirical analysis, one can only speculate.

![Diagram showing patient care options before and after video visits introduction](image)

**Figure 15: Patients’ Care Options Before and After Video Visits Were Introduced**

Although we want to both determine the impact of video visit availability on care utilization and explain the change in care utilization, this current study serves as a first step in our work and focuses on the first question—that is, whether, and to what extent, the availability of video visits impacts in-system care utilization. In particular, we investigate how the availability of video visits
offered at UPHS impacts (1) overall care utilization at UPHS and (2) in-person care utilization at UPHS. To examine these, we utilize a unique panel data set containing in-system patient encounters, both from employees and non-employees. Our data contain over 1.5 million patient encounters collected over the course of three and a half years. Visit types include primary care physician (PCP) visits, emergency department (ED) visits, and video visits, the combination of which represents the majority of visits that took place at UPHS.

Treating non-employees as a control group, we conduct difference-in-differences analyses and find that the availability of video visits was associated with increased overall care utilization and increased in-person care utilization at UPHS. Specifically, for patients who lived closer to their place of care, video visit availability was associated with a 21% increase in overall care utilization and a 15% increase in-person care utilization. For patients who lived farther from their place of care, the effects were more prominent. Specifically, we find that video visit availability was associated with a 34% increase in overall care utilization and a 25% increase in-person care utilization. Furthermore, we see increases in both PCP utilization and ED utilization across both patient groups. These results persist after we restrict to a sample of employees who are matched to non-employees based on key patient characteristics.

The results from the current study lay the foundations for future analyses that could help us further understand the impact of video visit availability on patients’ patterns of care utilization. Since we now find that video visit availability was associated with increased overall care utilization and in-person care utilization, the next important step is to understand the reasons for such increases. This involves investigating whether patients who previously used an out-of-system provider or did not seek care were more likely to seek care inside UPHS after video visits became available, as well as whether the use of video visits generated subsequent in-person visits.

The rest of the paper is organized as follows. We provide a brief summary of related papers in the literature in Section 3.2. In Section 3.3, we introduce our study setting and describe our data. We present our empirical approach in Section 3.4 and report the results in Section 3.5. In Section 3.6, we discuss the results and the next steps in our work as well as provide concluding remarks.
3.2. Literature Review

Over the last decade, there has been a growing body of research that studies various aspects of telehealth and remote care delivery. In the medical literature, the majority of research is descriptive and conducted using post-care surveys. There exist many survey studies that evaluate patient experience and satisfaction with telehealth (e.g., Slightam et al. 2020, Ramaswamy et al. 2020) and examine why patients choose remote care over in-person care. The findings from these studies are mostly consistent and suggest that leading reasons for choosing remote care are shorter wait times and convenience (e.g., Health Industry Distributors Association 2017). To better understand patients’ preferences, several studies examine patient characteristics associated with choosing remote care and find that higher in-person care barriers are significantly associated with a higher likelihood of choosing remote care (e.g., Reed et al. 2020).

Although the majority of past research focuses on patient perception and preferences regarding telehealth, there exist an increasing number of studies that directly investigate its operational and clinical impact. These studies are done in various settings such as telemonitoring for patients with chronic conditions (Steventon et al. 2012), medication adherence monitoring and feedback (Lekwijit et al. 2020), and remote doctor consultation in EDs (Sun et al. 2020). A small number of studies focus specifically on the impact of remote care delivery on care utilization. Bavafa et al. (2018) examine the impact of e-visits (i.e., secure messaging between patient and provider) on visit frequencies and find that e-visit adoption leads to more in-person care consumption. This finding challenges a number of medical studies that argue that e-visits could substitute in-person visits, including Zhou et al. (2007) and Bergmo et al. (2005). In addition to e-visits, a small but increasing number of studies focus on video visits, another common form of remote care that more resembles in-person care, and their impact of care utilization. However, like studies on e-visits, the results have been mixed. Delana et al. (2020) find that tele-ophthalmology generates an increase in the overall visit rate, but a decrease in the in-person visit rate, suggesting that tele-ophthalmology can improve overall care access and substitute in-person care. Shah et al. (2015) study video visit use among older adults who live in a senior living community and find that video visit use is associated with a
significant decrease in ED visits related to ambulatory-care-sensitive conditions. On the other hand, Ashwood et al. (2017) examine how the use of Teladoc, a direct-to-consumer video visit platform, affects care utilization among patients with acute respiratory illnesses and find that the magnitude of video visit substitution for in-person visits is not significant. These studies do not only yield mixed results and beg further studies—most of them also study only a specific patient demographic or care specialty, or study third-party video visit platforms that are not specific to any healthcare system. To the best of our knowledge, no one has studied how video visits offered within a healthcare system impact overall and in-person care utilization within the system, without restricting to any specific patient demographic or care needs.

Our study is related not only to the work on remote care delivery but also to several streams of operations management literature beyond the context of telehealth. First, our work joins a growing number of studies on multichannel healthcare, which broadly examine how improving access to one channel of care affects care utilization in other channels. For example, Bavafa et al. (2021) studies how an increase in PCP availability affects ED utilization, Ahuja et al. (2020) examines how improved patient-provider interactions in primary care affects inpatient care utilization, and Soltani et al. (2021) studies how ED physician workload impacts post-ED care utilization. Second, since remote care delivery is an integral part of health information technology, our work is related to studies on the impact of health information technology adoption, which consider various operational outcomes such as provider productivity (e.g., Adler-Milstein and Huckman 2013) and hospital performance (e.g., Angst et al. 2011). Finally, our study compliments operations management studies that examine the impact of operational interventions in healthcare, especially the impact of changes in provider flexibility (e.g., Balasubramanian et al. 2012) and work reallocation (e.g., Deo et al. 2013), on system outcomes.

3.3. Setting and Data

This study investigates the impact of video visit availability at the University of Pennsylvania Health System (UPHS) and was deemed exempt by the Institutional Review Board of the University
of Pennsylvania. In July 2017, UPHS began to offer a video visit service known as Penn Medicine OnDemand (PMOD) to approximately 50,000 UPHS-insured employees and their adult dependents. PMOD is an on-demand, direct-to-consumer primary care service that is available around the clock, every day of the year. PMOD providers are nurse practitioners who are stationed at a centralized telemedicine center and are employed by UPHS. In addition to conducting live video visit appointments, PMOD providers also provide short-term prescriptions, order lab tests, and coordinate follow-up in-person care when necessary. To use PMOD, patients must have a UPHS online account and have a video- and audio-enabled device, such as a smartphone, tablet, or computer. If using a smartphone or tablet, patients need to download the myPennMedicine app, which is available on both Apple App Store or Google Play Store. To see a provider, patients can schedule a visit through either a desktop patient portal or the mobile app by selecting one of the time slots, which are offered every 20 minutes throughout the day, and are able to request a specific provider depending on the provider’s availability. Before the visit, patients have to check in and verify their medications and allergies. PMOD is a covered benefit under employees’ insurance plans and is therefore free for employees. There is no limit to the number of times one can use PMOD.

The goals of this study are to investigate how the availability of PMOD impacts (1) overall care utilization at UPHS and (2) in-person care utilization at UPHS, which is further broken down into (2.1) PCP utilization and (2.2) ED utilization. Since UPHS is a self-insured health system, we observe both in-system and out-of-system utilization for employees through their medical claims but only observe in-system utilization for non-employees through the electronic health record. For the purpose of this analysis, we utilize in-system, employees’ and non-employees’ visit data, which are collected from July 2016, or one year prior to the introduction of PMOD, to December 2019. Prior to our analysis, we restrict our sample based on two selection criteria. First, to ensure that we can accurately track employees’ care consumption over time, we keep only employees who were employed for the entire study period and assume that an employee was employed for the entire period if (1) his or her first visit (based on the claims record) was in 2016 or earlier and (2) his or

1 For convenience, we refer to this group of patients simply as employees.
her last visit was in 2019 or later. For consistency, we also apply these criteria to non-employees. Second, to focus on patients who were active UPHS care users, we consider only those who had two or more visits over the 42-month study period. After excluding patients who do not meet the criteria, our final patient cohort contains 13,849 employees who had a total of 105,752 PCP, ED, urgent care, and PMOD visits combined, and 213,347 non-employees who had a total of 1,434,940 PCP, ED, and urgent care visits combined. Table 38 summarizes the distribution of visit types for employees and non-employees. In both groups, over 80% of the visits were with a PCP and approximately 10% were with an ED. For employees, while only 4.5% of all visits in the data were video visits, video visits actually made up over 6% of the visits that happened after July 2017. Although there existed a small number of urgent care visits, we exclude those visits from our study. Since UPHS opened the first urgent care clinic in January 2018 after PMOD became available, we cannot identify the effect of PMOD availability on urgent care utilization. Nevertheless, the opening and subsequent expansion of urgent care clinics should not bias the estimated effects of PMOD availability on care utilization of other types because urgent care was available to both employees and non-employees, and our empirical specification, particularly time fixed effects, would be able to capture its impact on care utilization.

Table 38: Distribution of Visit Types

<table>
<thead>
<tr>
<th>Visit Type</th>
<th>Employees (N = 13,849)</th>
<th>Non-Employees (N = 213,347)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of Visits</td>
<td>Percentage</td>
</tr>
<tr>
<td>PCP</td>
<td>88,400</td>
<td>83.59%</td>
</tr>
<tr>
<td>ED</td>
<td>11,191</td>
<td>10.58%</td>
</tr>
<tr>
<td>PMOD</td>
<td>4,826</td>
<td>4.56%</td>
</tr>
<tr>
<td>Urgent Care</td>
<td>1,335</td>
<td>1.26%</td>
</tr>
<tr>
<td>Total</td>
<td>105,752</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

For each visit in the data, we have both patient-level information and visit-level information. Patient-level information, some of which can change over time, includes age, gender, patient’s zip code, patient’s health insurance plan, and Charlson Comorbidity Score, which is a weighted count of specific chronic conditions in a patient’s diagnosis history at the time of visit. Visit-level information includes visit date, visit type, service zip code, provider specialty, and primary diagnosis, which we identify based on the International Classification of Diseases, Tenth Revision (ICD-10).
each in-person visit, we determine the distance the patient was from his or her place of care using
geographical distance between the centroid of patient’s zip code and the centroid of service zip
code. We observe that a patient’s distance from his or her place of care mostly stayed consistent
over time. Specifically, almost 80% of patients had no variation in distance over time, and over 99%
of patients had a within-person coefficient of variation in distance below 0.5. We report summary
statistics of distance from the place of care as well as patient demographics in Table 39. We find that
employees are statistically different from non-employees in all considered aspects. In particular, on
average, employees were younger, more female, had a lower comorbidity score, had higher UPHS
care utilization, and lived closer to the their place of care than non-employees. We note that, to
obtain these summary statistics, we use each patient’s age in July 2016, Charlson score averaged
across all visits, and distance from the place of care averaged across all patient-months.

Table 39: Demographics of Employees and Non-Employees

<table>
<thead>
<tr>
<th>Variable</th>
<th>Employees (N = 13,849)</th>
<th>Non-Employees (N = 213,347)</th>
<th>t-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monthly UPHS visits</td>
<td>0.18 (0.15)</td>
<td>0.16 (0.13)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Age (years)</td>
<td>43.24 (13.68)</td>
<td>50.61 (17.13)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Male</td>
<td>0.33 (0.47)</td>
<td>0.39 (0.49)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Charlson score</td>
<td>0.70 (1.26)</td>
<td>1.11 (1.73)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Distance from place of care (miles)</td>
<td>16.56 (56.71)</td>
<td>8.35 (54.43)</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Note: The statistics are reported at the patient level.

3.4. Empirical Approach

We conduct difference-in-differences analyses at the patient-month level by treating employees as a
treatment group and non-employees as a control group. Since the impact of video visit availability
may depend on whether a patient lived closer to or farther from the place of care, we first stratify
patients into two groups based on each patient’s average distance from his or her place of care and
use a cutoff of seven miles, which is the median distance across all employees. Then, we estimate
the effect of video visit availability within each of the groups separately. For each group, we conduct
our analyses in two ways. First, we perform a standard difference-in-differences analysis using data
on all employees and non-employees in the group. Second, we match employees to non-employees
within the group based on several key patient characteristics in order to obtain a better control group and perform difference-in-differences analyses on this subset of patients.

Our dependent variable, $MonthlyVisits_{it}$, is the number of visits that patient $i$ had during month $t$, where $t$ ranges from 1 to 42, representing July 2016 to December 2019. Since our dependent variable is a count variable by nature and can only be a non-negative integer, we use a count regression model to appropriately reflect its distribution and define $\lambda_{it}$ as a rate parameter, i.e., the mean of the distribution of $MonthlyVisits_{it}$. To determine whether the distribution of $MonthlyVisits_{it}$ is Poisson or negative binomial, which are the two standard distributions for count variables, we conduct a test of overdispersion and find that the dispersion parameter is statistically different from zero ($p$-value < 0.001), suggesting that we should use a negative binomial model rather than a Poisson model.

In addition to determining the count distribution, another important consideration is whether there exists zero inflation in the count data. Given the nature of healthcare utilization where most patients do not seek care very often, $MonthlyVisits_{it}$ has a mean of only 0.16 and has a value of zero on most months. The large number of zeros can likely be explained in one of the two ways. It is possible that (1) there are two zero-generating processes—one generates only zeros (e.g., when the patient is healthy) and the other generates counts, some of which may be zero (e.g., when the patient is unhealthy); or (2) all zeros are generated from a single process, and we see a large number of zeros because the rate parameter happens to be very small. To investigate this, we conduct a score test where the null hypothesis is that there is no zero inflation, i.e., the latter case is true (van den Broek [1995], Deng and Paul [2005]). The test suggests that the null hypothesis cannot be rejected ($p$-value = 0.920). In other words, there exists insufficient evidence for zero inflation and, therefore, visit counts can be modeled using a standard negative binomial distribution.
3.4.1. Standard Difference-in-Differences

We specify our difference-in-differences model as:

$$\ln(\lambda_{it}) = \ln(E(MonthlyVisits_{it}|X_{it})) = \beta \cdot VideoVisitAvailable_{it} + Month_{t} + Year_{t} + Patient_{i},$$

(3.1)

where $\lambda_{it}$ is the expected value of $MonthlyVisits_{it}$ conditional on a vector of independent variables, $X_{it}$, which consists of the following.

1. $VideoVisitAvailable_{it}$: This is our main independent variable that indicates whether PMOD was available to patient $i$ during month $t$. For employees, $VideoVisitAvailable_{it}$ is equal to zero prior to July 2017 and is equal to one afterwards. For non-employees, $VideoVisitAvailable_{it}$ always remains zero.

2. $Month_{t}$ and $Year_{t}$: We include month fixed effects to control for seasonality and year fixed effects to control for care utilization trends over time.

3. $Patient_{i}$: Patient fixed effects account for all observable (both available and unavailable in our data) and unobservable patient-specific factors that influence the patient’s care consumption and do not vary over the span of the study period. In an alternative analysis, we replace $Patient_{i}$ with patient characteristics, which include age, gender, and Charlson score, and check for the results’ robustness.

Since we aim to examine the effect of PMOD availability on four types of care utilization, i.e., overall, in-person, PCP, and ED utilization, we conduct four separate analyses, each with $MonthlyVisits_{it}$ modified to capture the corresponding type of care utilization. Figure 16 illustrates the number monthly visits over time for employees and non-employees.
Figure 16: Monthly Visit Frequency over Time

3.4.2. Difference-in-Differences on Matched Samples

One potential issue with the standard difference-in-differences analysis is that non-employees as a whole may not be an appropriate control group for employees. As Table[39] shows, non-employees’ characteristics appear to be systematically different from employees’. To improve the comparison between the two groups, we match each employee to a non-employee who is the nearest neighbor based on age, gender, Charlson score, and distance from the place of care. Then, we perform the difference-in-differences analysis in Equation[3.1] on this new cohort. Although it may seem reasonable to improve the comparison by considering non-employees who had a care usage level similar to that of employees, we do not match employees to non-employees based on average
monthly visit frequency prior to the introduction of video visits because matching on outcomes in the pre-treatment period could cause regression to the mean (Daw and Hatfield 2018). Tables 40 and 41 report the demographics of patients who lived within seven miles and farther than seven miles from their place of care, before and after matching, respectively. We can see that, after matching, employees and non-employees in both distance groups are no longer statistically different in the dimensions we match on.

Since UPHS is a self-insured health system, another factor that differentiates employees from non-employees and that possibly influences care utilization patterns is health insurance plans. To obtain a better control group, we can consider only non-employees who had health plans that were comparable to employees’ health plans. However, since only a small number of employees had such health plans, we do not include this matching criterion in our main analysis but will discuss the results from this extended analysis as we present the main results.

Table 40: Demographics of Employees and Non-Employees by Distance Group Before Matching

<table>
<thead>
<tr>
<th>Variable</th>
<th>Distance ≤ 7 Miles</th>
<th>Distance &gt; 7 Miles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Employees (N = 6,834)</td>
<td>Non-Employees (N = 151,069)</td>
</tr>
<tr>
<td>Monthly UPHS visits</td>
<td>Mean (SD) 0.17 (0.14)</td>
<td>Mean (SD) 0.16 (0.14)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>Mean (SD) 42.59 (13.65)</td>
<td>Mean (SD) 50.80 (17.28)</td>
</tr>
<tr>
<td>Male</td>
<td>Mean (SD) 0.34 (0.47)</td>
<td>Mean (SD) 0.39 (0.49)</td>
</tr>
<tr>
<td>Charlson</td>
<td>Mean (SD) 0.65 (1.21)</td>
<td>Mean (SD) 1.11 (1.74)</td>
</tr>
<tr>
<td>Distance from place of care</td>
<td>Mean (SD) 3.59 (1.97)</td>
<td>Mean (SD) 2.77 (2.07)</td>
</tr>
</tbody>
</table>

3.5. Results

Table 42 presents the estimated effects of video visit availability on overall, in-person, PCP, and ED care utilization from the standard difference-in-differences model. We report the estimated effects in terms of incidence rate ratios (IRRs), which are the ratios of expected monthly visit counts if video visits were available to expected monthly visit counts if video visits were not available. For example,
Table 41: Demographics of Employees and Non-Employees by Distance Group After Matching

<table>
<thead>
<tr>
<th>Variable</th>
<th>Distance ≤ 7 Miles</th>
<th>Distance &gt; 7 Miles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Employees(N = 6,830)</td>
<td>Non-Employees(N = 6,717)</td>
</tr>
<tr>
<td></td>
<td>Mean(SD)</td>
<td>Mean(SD)</td>
</tr>
<tr>
<td>Monthly UPHS visits</td>
<td>0.17(0.14)</td>
<td>0.16(0.14)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>42.59(13.65)</td>
<td>42.73(13.61)</td>
</tr>
<tr>
<td>Male</td>
<td>0.34(0.47)</td>
<td>0.34(0.47)</td>
</tr>
<tr>
<td>Charlson</td>
<td>0.65(1.21)</td>
<td>0.62(1.17)</td>
</tr>
<tr>
<td>Distance from place of care</td>
<td>3.59(1.97)</td>
<td>3.58(1.98)</td>
</tr>
</tbody>
</table>

an IRR of 1.21 translates to a 21% increase in expected monthly visit counts, or equivalently, a 21% increase in care utilization in general as a result of video visit availability. In addition to IRRs, we also report average marginal effects (AMEs), which are average magnitudes of changes in monthly visit counts as a result of video visit availability.

Using standard difference-in-differences, we find that the availability of video visits led to increases in both overall utilization and in-person utilization and that the effects were larger for patients who lived farther from their place of care. Specifically, for patients who lived closer to their place of care, video visit availability was associated with a 21% increase in overall care utilization and a 15% increase in-person care utilization, which, on average, translates to 0.37 more visits per year in total and 0.27 more in-person visits per year.

In addition, we break down in-person utilization and find that video visit availability was associated with a 16% increase in PCP utilization and an 8% increase in ED utilization. For patients who lived farther away, we find that video visit availability was associated with a 34% increase in overall care utilization and a 25% increase in-person care utilization, which translates to 0.54 more visits per year in total and 0.42 more in-person visits per year on average.

With regard to in-person utilization, video visit availability was associated with a 21% increase in PCP utilization and an 87% increase in ED utilization. For ease of understanding, we visualize yearly average marginal effects in Figure. Although the average yearly marginal

---

2 Calculated as $0.031 \times 12$ and $0.023 \times 12$, respectively
3 Calculated as $0.045 \times 12$ and $0.035 \times 12$, respectively
effects may not seem large for an individual patient, for our hospital system with approximately 230,000 active care users, these effects amount to over 90,000 additional visits and over 70,000 additional in-person visits each year if video visits were available to everyone. We find these results to be robust when we replace the patient fixed effects with patient characteristics and when we use a linear model instead of the negative binomial model.

Table 43 reports the results obtained using difference-in-differences on matched samples. Consistent with our prior analysis, we find that video visit availability was associated with increases in both overall and in-person utilization. However, the effect sizes are slightly larger with matched samples. We find that video visit availability was associated with a 23% increase in overall utilization and an 18% increase in in-person utilization for patients who lived closer to their place of care, and a 38% increase in overall utilization and a 32% increase in in-person utilization for patients who lived farther from their place of care. In addition to matching employees and non-employees on patient characteristics, we also consider only non-employees who had health plans comparable to employees’ and find the estimation results to be similar to what we obtain without considering health plans. However, with this approach, the sample size decreases by approximately 70%.

Table 42: Effects of Video Visit Availability on Monthly Care Utilization Estimated Using Standard Difference-in-Differences

<table>
<thead>
<tr>
<th>Type of Utilization</th>
<th>Distance ≤ 7 Miles</th>
<th>Distance &gt; 7 Miles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VideoVisitAvailable</td>
<td>Mean of Dependent Var.</td>
</tr>
<tr>
<td>(1) Overall utilization</td>
<td>1.21*** (0.01)</td>
<td>0.031</td>
</tr>
<tr>
<td>(2) In-person utilization</td>
<td>1.15*** (0.01)</td>
<td>0.023</td>
</tr>
<tr>
<td>(2.1) PCP utilization</td>
<td>1.16*** (0.01)</td>
<td>0.022</td>
</tr>
<tr>
<td>(2.2) ED utilization</td>
<td>1.08** (0.03)</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Notes: Number of patient-months = 9,542,232. *(p < 5%), **(p < 1%), *** (p < 0.1%).
IRR = incidence rate ratio. SE = standard error. AME = average marginal effect.
Table 43: Effects of Video Visit Availability on Monthly Care Utilization Estimated Using Difference-in-Differences on Matched Samples

<table>
<thead>
<tr>
<th>Type of Utilization</th>
<th>Distance ≤ 7 Miles</th>
<th>Distance &gt; 7 Miles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VideoVisitAvailable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IRR</td>
<td>AME</td>
</tr>
<tr>
<td></td>
<td>Mean of Dependent Var.</td>
<td>Mean of Dependent Var.</td>
</tr>
<tr>
<td>Overall utilization</td>
<td>1.23*** (0.02)</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td>0.175</td>
<td></td>
</tr>
<tr>
<td>In-person utilization</td>
<td>1.19*** (0.02)</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>0.145</td>
<td></td>
</tr>
<tr>
<td>PCP utilization</td>
<td>1.15*** (0.02)</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>0.019</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Number of patient-months = 1,151,724. * (p < 5%), ** (p < 1%), *** (p < 0.1%). IRR = incidence rate ratio. SE = standard error. AME = average marginal effect.

Figure 17: Yearly Average Marginal Effects Estimated Using Standard Difference-in-Differences

3.6. Discussion and Conclusions

This study takes the first step towards understanding how video visit availability impacts the choices that patients make when they are sick. Through difference-in-differences analyses, we investigate the impact of video visit availability at the University of Pennsylvania Health System (UPHS) on in-system care utilization and find that the availability of video visits was associated with increases in overall and in-person care utilization within the system. These results lay an important foundation for the next step in our work, which is to understand the reasons for the changes in care utilization.
There are two main plausible explanations for the increases in care utilization that our future research can investigate. The first explanation is the substitution effect. It is possible that, when there was an option to see UPHS care providers quickly and conveniently, patients who previously used out-of-system providers or who did not seek care were more likely to seek care inside UPHS through video visits. Additionally, it is also possible that the convenience in accessing UPHS care not only generated additional utilization through video visits but also directly led to increased in-person utilization at UPHS. Since there existed an easy way to connect and communicate with UPHS providers, patients may be more likely to seek in-person care inside UPHS rather than at out-of-system hospitals that do not provide the same benefit. To investigate the substitution effect, we can utilize employee claims data and use patient choice models to examine how the availability of video visits impacts patients’ decisions regarding whether to seek care inside or outside of UPHS when they decide to seek care. In addition, we can potentially examine patients’ decisions regarding whether to seek care using sick days as a proxy for times when patients were sick but did not seek care.

The second explanation for the increases in care utilization is known as the gateway effect. Our preliminary analysis suggests that the use of video visits likely led to in-person visits shortly afterwards. Based on our data, 18% of video visits were followed by an in-person visit within seven days, while only 4% of PCP visits and 2% of ED visits were followed by another in-person visit within seven days. There exist several reasons that can potentially explain the gateway effect. It is possible that some video visits could not provide adequate care and therefore led to repetition of care. It is, however, also possible that some patients used video visits in order to get an in-person appointment with difficult-to-access specialties. In our future research, we can formally investigate the presence of the gateway effect using hazard models where the hazard rate is time until the next visit. Additionally, we can examine the causes of the effect (i.e., repetition of care vs. improvement of care access) by incorporating visit-level information of video visits and subsequent visits, such as the patient’s diagnosis and the type and specialty of the care provider.
In addition to explaining the increases in care utilization, it is also important to understand the impact of distance from the place of care on the sizes of the increases. Our future research can examine whether and how the substitution effect and the gateway effect presented themselves differently depending on how far a patient was from the hospital. Our preliminary analysis suggests that the gateway effect may not differ between those who lived closer to the place of care and those who lived farther from the place of care. Specifically, we find that the common reasons for video visits and the rates of follow-up care after a video visit are consistent across both distance groups. This implies that the larger increases in care utilization among patients who lived farther away can possibly be explained primarily by the substitution effect. In other words, video visit availability may have influenced care choices among patients who lived farther away more than it did among patients who lived close by. To investigate this, we can incorporate distance into our patient choice models and examine how distance changes the impact of video visit availability on patients’ decisions regarding whether and where to seek care.

Lastly, since we find that in-person utilization increased as a result of video visit availability, our future research can investigate how care providers were able to accommodate additional in-person visits. There exist several possibilities that we can explore—for example, care providers may have reduced the amount of time spent with each patient, admitted fewer new patients, or worked more hours to provide additional visits.


126


