ESSAYS IN HEALTHCARE OPERATIONS AND MANAGEMENT

A Dissertation
Presented to
The Academic Faculty

By

Jan Vlachy

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology

August 2021

ESSAYS IN HEALTHCARE OPERATIONS AND MANAGEMENT

Approved by:

Dr. Turgay Ayer, Advisor
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Pinar Keskinocak
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. David Goldsman
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Karthik Ramachandran
Scheller College of Business
Georgia Institute of Technology

Dr. Mehmet Ayvaci
Jindal School of Management
University of Texas at Dallas

Date Approved: January 28, 2021

To Eva Vlachá.

Za neustálou podporu zvídavosti a zvědavosti.

ACKNOWLEDGEMENTS

There is so much to be grateful for. I am grateful to everyone who has made this work possible, pleasurable, or useful. What follows is an incomplete list.

I am grateful to my committee, my advisor Turgay Ayer and the committee members Mehmet Ayvaci, David Goldsman, Pinar Keskinocak, and Karthik Ramachandran for much constructive feedback. I am also proud to acknowledge all the research collaborators who have been involved in the research that led to these dissertation chapters (Turgay Ayer, Mehmet Ayvaci, Zeynal Karaca, Srinivasan Raghunathan, and Pinar Keskinocak).

I was blessed with many other inspiring research collaborators during my PhD. Some resulted in fame and knowledge, some in my further education, grateful for all of them! These included:

- Laurie Garrow and Greg Macfarlane from the School of Civil Engineering at Georgia Tech
- George Rust, Anne Gaglioti, Junjun Xu, Peter Baltrus, Lilly Immergluck, Luceta McRoy, Khusdeep Malhotra, Shun Zhang, and others at Morehouse School of Medicine
- Andres Rodrigues Ruiz, Suzette LaRoche, Hiba Arif, Atul Vats, Larry Olson, Wendy Book, and others from Emory Healthcare and Children's Healthcare Atlanta. Also Emily Gilmore from Yale School of Medicine and other collaborators from the Critical Care EEG Monitoring Research Consortium.
- Mustafa Ozkaynak from the University of Colorado Denver
- Georgia Tech student collaborators: Qing Li and Mingyoung Jo as well as two talented undergraduate researchers, Lianyan Gu and Diem Tran.
- Lori Houghtalen for teaching collaboration

I also cannot forget many researchers in the past who steered me toward the research route, which ultimately led to this PhD. Some among these researchers, professors, and scientists are Michel Mandjes, Ton Dieker, Karma Dajani, Roberto Fernandez, Andre Henriques, David Stanovsky, Ludek Zajicek, Petr Holicky, Jan Trlifaj, Pavel Jelinek, Antonin Fejfar, and Michal Farnik.

I am grateful for all the excellent research environments that I passed through, but such environments are not created just by researchers. Many other staff members need to work hard to keep the environment productive and supportive. At Georgia Tech, these included Alan Erera, Amanda Ford, Pam Morrison, Dima Nazzal, Paul Kvam, Edwin Romeijn, and many others. And of course, thanks also to Thomas Holmquist, for always being there to fix my computer. I am also grateful to the Georgia Tech Communications Center for taking my English writing and speaking from "unintelligible" to "horrible but occasionally understandable." Special thanks to Jane Chisholm, for taking me through several Communications classes.

During the PhD summers, I was given several opportunities to put my newly-learned skills into practice during summer projects and internships. For this, I am grateful to my advisor that he allowed such distractions, with the foresight that they would benefit me in the future. And then I am grateful to my summer colleagues and friends who welcomed me enthusiastically and helped me learn. A very incomplete list of these benefactors would include Rayid Ghani, Lauren Haynes, Joe Walsh, Brian McInnis, Ivana Petrovic, Fernanda Alcala Durand, John Croley, Esther Bongenaar, Winfried Theis, Thomas Rasmussen, Ganarajan Govindarajan, Remco de Ruiter, Jorrit Van Der Togt, Philip Jonathan, Onno de Noord, Stijn Bierman, and Jose Gonzalez Martinez. And of course, great thanks to my current colleagues at Google for putting up with all my grumbling about the thesis writing process.

When speaking about practical experience, I also need to mention my two industry mentors from the Georgia Tech Student Alumni Association mentoring program: Vincent

Emanuele and Jason LaRoche. They played a pivotal role in steering my learning to use my theoretical knowledge for real-world impact.

I am also grateful to the people and organizations who also showed me the way to use those skills and that impact to actually influence what matters, to positively shape the world, now (at least a bit!) and in the future. Many of the aforementioned individuals played a role; additionally, I would like to highlight three particular organizations: Data Science for Social Good, 80,000 Hours, and Giving What We Can.

And now is the time for the friends! Thanks to Anthony Bonifonte, Qiushi Chen, Can Zhang, Zhaowei She, Caglar Caglayan, and Andrew ElHabr from my research group for creating an intellectual but fun group – also grateful for our weekly lunches. Then, many other friends from Georgia Tech (yes, I am missing so many on this list; and the order is very random): Geet Lahoti, Tony Yaacoub, Yassine Ridouane, Amy Musselman, Jeff Pavelka, Burak Kocuk, Beste Basciftci, Ezgi Karabulut, Toyya Pujol, Ethan Mark, Kevin Ryan, Na Yeon Kim, Seyma Guven, Mathias Klapp, Yuan Li, Richard Birge, Murat Yildirim, Weihong Hu, Ruilin Zhou, Evren Gul, Fang Cao, Linwei Xin, Mina Georgieva, Tonghoon Suk. And then, our Graduate Student Advisory Team: Daniel Silva, Matt Plumlee, Tonya Woods, Erin Garcia, Ben Johnson, Shanshan Cao, Luke Marshall, and, among the old guard, Vinod Cheriyan and Mallory Soldner. And then, friends from the past lives who connect me back. Geoffrey Wielingen, Stein Andreas Bethuelsen, and Danny Chan from our studies in the Netherlands. Then, from Charles University, Adela Skokova, Lida Divisova, Frantisek Zak, Gabriela Tethalova, Pavel Hajek, Marcel Sebek, Pavol Pseno, Lenka Slavikova, Hana Krulisova, and others. And another long list for GYBU, including Eva Pauliova, Alena Stojdlova, Kristina Richterova, Helena Kalaninova, Radek Doubrava, Lenka Mikulickova, Anna Janurova, Katka Cmuntova, Hanka Ruzickova, Jana Oltova, Jan Buchar, Vojtech Luhan, Lukas Kindl, Ondrej Burda, and many more.

Special thanks to the several of you who actively kept me sane during the PhD process: Prami Sengupta, Yasaman Mohammadshahi, Minkyoung Kang, Adam Raz, and Vu Pham

Quynh Lan.

Finally, I am grateful to my family in Czechia for supporting my incessant travel bug and unquestioningly helping me in pursuing all these experiences. Thank you, grandma, father, sister, and brother.

Anyway, happy that there are so many people that I can be grateful for. I still feel there are many missing. There are others who I have not mentioned for the lack of space or for failings of my memory. If you are reading this and you think you should have been acknowledged, you are probably right. In such case, please also send me an e-mail, I will get you a coffee, and we can reconnect.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The overarching objective of the research described in this dissertation is to analyze some pressing healthcare problems in various areas, ranging from health IT to payment models and healthcare operations, using mathematical and economic models. Analytical models have been at the forefront of the recent transformational efforts in healthcare (see e.g., Reid et al. (2005), Obermeyer and Emanuel (2016), Romeijn and Zenios (2008), Bates et al. (2014)). The dissertation comprises three chapters corresponding to three research projects: 1) econometric evaluation of *Health Information Exchanges*, 2) game-theoretical modeling of the *Bundled Payments* financing model, and 3) queueing algorithm design for *Bed Management* with overflows in hospitals. This dissertation has been a result of many collaborations, reflecting the wider trends in both healthcare and research. Next, we provide a more detailed overview of each of these chapters.

In Chapter 1, we study health information exchanges (HIEs), and especially their role in emergency departments (EDs). HIEs are expected to improve poor information coordination in EDs; however, whether and when HIEs are associated with better operational outcomes remains poorly understood. In this work, we study HIE and length of stay (LOS) relationship using a large dataset from the Healthcare Cost and Utilization Project consisting of about 7.4 million treat-and-release visits made to 63 EDs in Massachusetts. Overall, we find that HIE adoption is associated with a 10.2% reduction in LOS and the percentage reduction increases to 14.8% when the hospital is part of an integrated health system or to 21.0% when a patient has a previous visit to an HIE-carrying hospital. We further find that 1) teaching hospitals benefit more from HIE adoption compared with non-teaching hospitals, 2) patients with severe or multiple comorbid conditions spend less time in the ED under HIE presence. Together, these results imply that 1) HIE adoption reduces overall ED LOS, 2) wider HIE adoption would scale up the benefits for individual hospitals, 3) magnitude of the association between HIE and LOS is higher when financial incentives for

HIE adoption are stronger (e.g., integrated health systems), and 4) the size of the reduction depends on certain contextual moderating factors. Given that HIEs are a key component of healthcare delivery and ongoing reforms, we believe that our findings have important implications and may inform policymakers regarding the nationwide HIE adoption.

Chapter 2 is concerned with one of the emerging new payment models in healthcare, the bundled payments. Under the prevailing fee-for-service payments (FFS), hospitals receive a fixed payment, while physicians receive separate fees for each treatment or procedure performed for a given diagnosis. Under FFS, incentives of hospitals and physicians are misaligned, leading to large inefficiencies. Bundled payments, an alternative to FFS unifying payments to the hospital and physicians, are expected to encourage care coordination and reduce ever increasing healthcare costs. However, as hospitals differ in their relationships with physicians in influencing care (level of physician integration), it remains unclear what spectrum of physician integration will facilitate bundling. There is a lack of both academic and practical understanding of hospitals and physicians bundling incentives. Our study builds on and expands the recent Operations Management literature on alternative payment models. We formulate game-theoretic models to study (1) the impact of the level of integration between the hospital and physicians in the uptake of bundled payments, (2) cost and quality characteristics of a care context that facilitate bundling, and (3) when feasible, the consequences of bundling with respect to overall care quality and costs/savings across the spectrum of integration levels. We find that (1) hospitals with high physician integration or with low physician integration are less likely to gain from bundled payments, while the hospitals that lie in between these two cases will likely benefit the most; (2) although bundled payments are likely to decrease costs, quality may also decrease; (3) initiatives that promote quality awareness in hospitals may dampen the incentives for bundling in hospitals with independent physicians, whereas they are likely to enhance incentives for bundling in hospitals with salaried physicians. Our findings have important managerial implications for both hospitals and payers: (1) in deciding whether to enroll in bundled payments, hospitals should consider their level of physician integration, and (2) payers should be aware of and account for potential negative effects of bundling, including a possible quality reduction, or even a cost increase. Based on our findings, we expect that a widespread use of bundled payments may trigger further market concentration via hospital mergers or service-line closures.

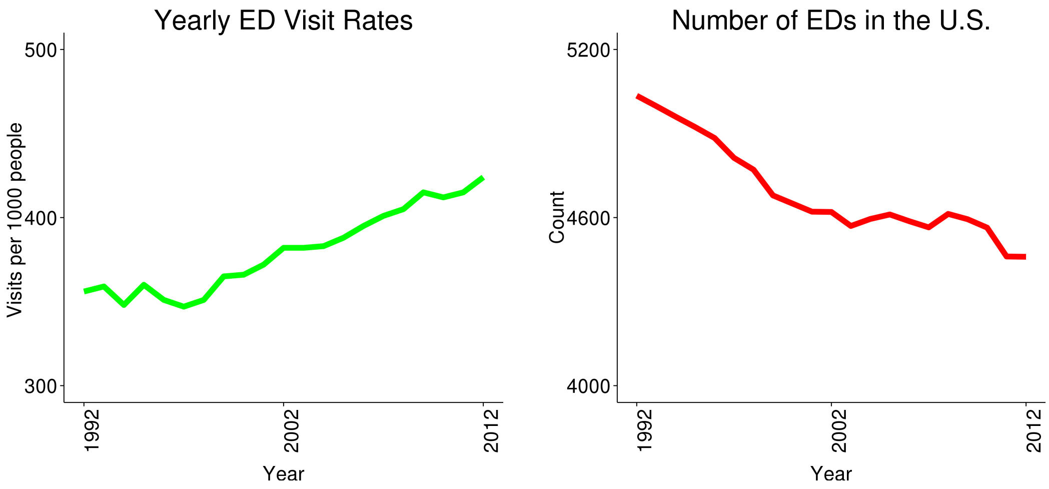
Chapter 3 focuses on the problem of patient boarding in emergency departments before they are accepted to hospital internal wards. In particular, we are interested in hospital patient-bed matching and bed pooling when ED patients who cannot be accommodated by the most appropriate internal ward can be redirected to other internal ward that can care for them. To this end, we first discuss two commonly used policies in the literature, reservation and threshold policies, proving their favorable structural properties under special conditions, and their limitations in more general settings. We then introduce our novel "Generalized Reservation and Threshold Reinforcement Learning" (GREAT-RL) policy, demonstrate how it generalizes threshold and reservation policies, and describe our reinforcement learning implementation. In an extensive numerical simulation, we show how the GREAT-RL policy outperforms the reservation and threshold policies as well as other applicable policies in the literature. The numerical results suggest substantial potential improvements in operational metrics and patient outcomes once we are able to deploy the policies in hospital practice.

CHAPTER 1
THE IMPACT OF HEALTH INFORMATION EXCHANGES ON EMERGENCY DEPARTMENT LENGTH OF STAY

1.1 Introduction

The United States have experienced a steady increase in the total number of visits to emergency departments (EDs) and a decrease in the total number of operating EDs in the past two decades. Between 1992 and 2012, annual ED visits increased faster than the growth in the U.S. population, leading to a cumulative increase of 47% in the number of ED visits. During the same years, the number of EDs decreased from 5,035 to 4,460 (a 11% decline), and hospitals reduced the total number of beds by 118,939 (a 13% decline) (American Hospital Association 2014b) as demonstrated in Figure 1.1. As a result, the average length of stay (LOS) for an ED visit has steadily been increasing over years, especially for treat-and-release patients (Pitts et al. 2012, Pines et al. 2010). Factors such as poor coordination, lack of communication between the ED and the patient's primary care provider, and unavailability of the patient's medical history at the point of care have exacerbated the growing problem of increased LOS and ED crowding nationwide (Bodenheimer 2008), leading to the so called "crowdedness epidemic" (Institute of Medicine 2006c, Pines et al. 2008a). With annual visits exceeding 130 million, EDs are currently under extreme pressure to reduce the ever-increasing LOS (Emerman 2012, Pallin et al. 2013).

LOS is a key measure of ED throughput and a marker of overcrowding (National Quality Forum 2009, Medicare 2013, Hwang and Concato 2004). ED LOS is also a critical component of ED quality assurance monitoring and is negatively correlated with patient satisfaction, length of hospital stay, morbidity, and mortality (Carr et al. 2007, Chalfin et al. 2007). Furthermore, LOS is a reasonable proxy for resource consumption and is important



(a) The number of ED visits has been rising. . .

(b) . . . while the number of EDs has been decreasing.

Figure 1.1: EDs under pressure

for planning and management of care (Pallin et al. 2013). In 2008, the National Quality Forum approved the median time from ED arrival to departure as the main quality measure related to ED LOS. Although there is no consensus on the acceptable range for ED LOS, a 90th percentile of ED LOS less than 8 hours for patients admitted to the hospital and 4 hours for those that are not admitted (called the "treat-and-release patients") have been commonly recommended and used (Fee et al. 2012, Weber et al. 2011, CAEP 2009).

Health information technology (HIT), in particular health information exchange (HIE), has the potential to fundamentally transform the delivery of healthcare (President's Council of Advisors on Science and Technology 2010). HIE is an electronic means for transferring patient records among various healthcare providers, which offers many potential benefits, including increased operational efficiency, real-time access to patients' past data, improved care quality, and reduced administrative overhead (Walker et al. 2005, Halamka 2013, Tzeel et al. 2011). In order to accelerate the adoption of HIT, the U.S. government issued the HITECH Act in 2009 (Blumenthal 2010), which led to the establishment of the "Meaningful Use" program. The Meaningful Use program aims to increase the use of electronic health records (EHRs) and HIEs to improve quality, safety, efficiency, and care coordination, and reduce health disparities (Blumenthal and Tavenner 2010). Starting in 2014, Stage 2 of the Meaningful Use Program has focused on advanced clinical processes including HIEs and required the coordination of care across delivery settings. In addition to incentive payments, the Office of the National Coordinator for HIT has committed $564 million to support state HIE programs. Mainly due to these initiatives, HIEs have recently experienced rapid growth and adoption, with the percentage of hospitals with a functional HIE surging from 41% in 2008 to 76% in 2014 (Mahajan 2016). Going forward, it seems that the role of HIEs would be even further amplified as the sharing of health information is incorporated into the Meaningful Use Stage 3 (DeSalvo and Haque 2015, Jacob 2015).

Existing qualitative research identifies several mechanisms through which the HIE coordinates care and subsequently influences LOS (cf. Rudin et al. 2011, Thorn et al. 2013, Kuperman and McGowan 2013). These mechanisms include that: 1) HIEs may reduce the number of duplicate prescriptions and repeated procedures and diagnostics (Lammers et al. 2014, Tzeel et al. 2011), and hence decrease LOS due to direct time savings, 2) by allowing access to additional information, HIEs may enhance clinicians' understanding of the underlying problem and hence improve the quality of diagnosis and treatment (Hincapie et al. 2011, Thorn et al. 2013), which may in turn lead to either an overall increase in LOS (because of the need for more thorough examination of the patient) or a decrease in LOS (because of more efficient diagnosis), and 3) HIEs may disrupt existing clinical workflow, inflict additional administrative workload, and increase information overload (Scott et al. 2005, Bhargava and Mishra 2014, Goh et al. 2011, Kuperman and McGowan 2013), and hence lead to an increase in LOS. Furthermore, these potential relationships between HIE adoption and LOS could be moderated by many factors (e.g., the type of the hospital, the load at the ED, the type of the patients served).

However, as the widespread adoption of HIEs is a recent phenomenon, compelling quantitative evidence on the relationship between HIE and operational outcomes has been very limited (Rudin et al. 2014). In particular, research into the HIE and ED LOS relationship has remained elusive primarily because of data limitations. As such, the purposes of our study are 1) to assess the interplay between HIE adoption and overall ED LOS, and 2) to examine some of the moderators of this relationship that are important in care delivery workflow. In particular, in addition to studying the overall effect, we investigate the HIE and LOS relationship in the context of the organizational setting (teaching vs. non-teaching hospitals or integrated vs. non-integrated healthcare systems), the operational environment in the ED when the patient arrived (crowded vs. non-crowded), and patient-related factors that could drive the need for information (severity level, primary diagnosis, and existence of comorbid conditions).

To address our research questions, we use 2009-2013 Healthcare Cost and Utilization Project (HCUP) State Emergency Department Databases (SEDD), which are maintained by the Agency for Healthcare Research and Quality (AHRQ) through a Federal-State-Industry partnership. This is a unique dataset which, to our knowledge, is the largest and the most comprehensive used for the purpose of testing HIE and LOS relationship in a systematic way. We obtained information about hospital characteristics (e.g., teaching status and bed size) from 2009-2013 American Hospital Association (AHA) Annual Survey Database and linked them to SEDD files using hospital identifiers. We estimated median household income of the patient's ZIP Code of residence from the 2009-2013 U.S. Census Bureau Databases and used it as a proxy for socioeconomic status, which is known to be correlated with ED LOS (Karaca and Wong 2013). We obtained information about HIE adoption using Health Information Management Systems Society (HIMSS) Analytics Database and linked it to SEDD using Medicare provider numbers. Finally, we collected hospital-reported *Meaningful Use* data from the Center for Medicare and Medicaid Services (CMS) for testing the robustness against measuring implementation rather than use.

Using our dataset, we develop multivariate regressions that control for patient and hospital characteristics to examine the relationship between the HIE adoption and LOS in the ED. Exploiting the variation in HIE adoption decisions across hospitals and years in our dataset, we use a quasi-experimental setup where hospitals are divided into control (those who never adopted HIE or had HIE during the entire study period) and treatment (these adopt HIE at some point within the study period) groups, corresponding to a difference-in-differences (DID) estimation (Wooldridge 2010). Our analyses account for possible confounders that may affect LOS both at the hospital level (e.g., hospital bed size, teaching status) and the patient level (e.g., demographics, socioeconomic factors, severity/complexity, and time indicators).

We make several key contributions in this paper. First, this research is among the first to study the link between adoption of an advanced information technology, namely HIEs, and operational efficiency in EDs, as measured by LOS. To our knowledge, the only studies that examined the value of information sharing for the ED LOS are two small-scale observational studies conducted in the context of diabetes (Speedie et al. 2014) and heart failure (Connelly et al. 2012). On the other hand, large scale studies examining the impact of HIT adoption either considered other forms of IT such as Electronic Medical Records (EMR) adoption (c.f., Lee et al. (2013a)) or analyzed the impact of HIE on other outcomes such as the imaging rate reduction (c.f., Lammers et al. (2014)). Our results suggest that, overall, HIE is associated with a 10.2% reduction in LOS for treat-and-release ED patients. By quantifying the relationship between LOS and HIE, our study can inform policy makers, hospital managers, and researchers about overall HIE and LOS relationship in the ED. Hospital managers are interested in reducing LOS in the ED because such reduction can lead to increased revenue for reasons such as increased patient satisfaction, reduced rate of patient leaving the ED without being seen due to long wait times, or higher throughput.

Second, we show that wider HIE adoption would scale up the benefits for individual hospitals. In particular, by tracking patients over time, we are able to show that the LOS reduction due to HIE significantly improves when patients go to the EDs with HIE in their subsequent visits following an initial visit to an ED with HIE. More specifically, the percentage reduction for patients visiting an ED with functional HIE increases to 21.0% when the patient has a previous visit to an HIE-carrying hospital, as compared with the overall reduction of 10.2% for an arbitrary patient. For the latter, we enhance the analysis around the overall adoption vs. LOS relationship by incorporating a use index based on hospitals' actual utilization of different HIE functionalities as reported under the Meaningful Use program and show that hospitals that effectively use HIE benefit more from HIE adoption in terms of LOS. More specifically, we find that the percentage reduction in LOS is about 22.2% with effective use, as compared with the average reduction of 10.2%.

Third, the richness of our dataset enabled us to separate the magnitude of HIE and LOS relationship under different contexts (e.g., teaching status of the hospital, crowdedness of the ED, severity/complexity of the patient, revisits, and integrated health systems). This is important because, as highlighted by Rudin et al. (2014), HIE evaluation studies thus far have not properly studied contextual factors that moderate the value of HIE. Our analyses of such factors suggest the following: 1) teaching hospitals are more likely to benefit from the HIE adoption in terms of LOS compared with non-teaching hospitals, 2) patients who arrive to crowded EDs spend more time on average, and they experience even higher LOS in EDs with HIE, 3) severe patients or patients with comorbid conditions on average spend less time in the ED under HIE presence, and 4) the magnitude of the association between HIE and LOS depends on the patient's primary disease/condition. We believe this detailed level of analysis provides deeper insights into HIE and LOS relationship in the ED.

Finally, this paper contributes to the ongoing discussion of how the Patient Protection and Affordable Care Act (ACA) might affect access, cost control, and care coordination. The comprehensive healthcare reform that Massachusetts enacted in April 2006 is widely viewed as a mini-model of the ACA implementation (Chandra et al. 2011). More specifically, the Massachusetts law required that all individuals have access to affordable coverage and mandated that they obtain health insurance, with some subsidized options available (McDonough et al. 2008). In that regard, the Massachusetts experiment may provide invaluable lessons about access, cost control, and care coordination for other states and the federal government, particularly as the ACA is being implemented. Our analyses on Massachusetts data further contribute to this discussion by estimating the impact of HIE on ED LOS, which can be translated into potential cost savings from hospitals' perspective, increased access to healthcare from patients' perspective, and better care coordination from physicians' and patients' perspective. Furthermore, our findings can inform the healthcare policymakers regarding the nationwide HIE adoption and its relation to access, cost control, and care coordination.

1.1.1 Health Information Exchange (HIE)

Early HIE attempts began in 1990s (Johnson et al. 2008). Later initiatives that are still active include the Indianapolis Network for Patient Care and Research (Overhage et al. 1995), starting in 1994, and the MidSouth e-Health Alliance (Johnson et al. 2008) in 2006. For instance, in Indianapolis, IN, Marc Overhage and William Tierney served as professors of medicine (Overhage et al. 1995). Then, they added another professional responsibility: the design and implementation of an HIE in Indianapolis. In a grant-funded project, they strove to share information among multiple organizations and to study how this sharing, this HIE, would affect care and its cost. They turned to a wide range of clinical partners across Indianapolis for collaboration, including clinicians in EDs and community health centers, pharmacists, health maintenance organization (HMO) administrators, and homeless care workers. They intended, among other objectives, to evaluate ED testing and drug prescribing. This intention indicated a shift in how healthcare information should be communicated. In fact, traditionally, EDs would wastefully test unknown patients, and patients would inaccurately disclose their drug usage.

The implementation of HIE was initially gradual. It has, however, increased in pace in the late 2000s as the HITECH Act was passed (Blumenthal 2010), and clinicians became aware of the HIE benefits (Wright et al. 2010). Such increased adoption is what the creators of HITECH Act apparently had in mind when they proposed the act that would invest heavily in HIT in the environment of rapidly developing IT throughout the US. In fact, after the HITECH adoption, the percentage of hospitals with electronic health records has grown dramatically, from 9% in 2008 to 76% in 2014, and continues to increase (Charles et al. 2015). During the same period, the number of hospitals engaging in some HIE has risen from 41% to 76% and presumably has become more sophisticated as the overall EHR infrastructure has improved as well. In 2015, HIE initiatives were further supported by CMS through the incorporation of HIE requirements in the third stage of the Meaningful Use requirements (for Medicare & Medicaid Services 2015). HIE has also benefited the shift across healthcare toward alternative payment models that require more effective care coordination, information sharing, and resource allocation (Buntin et al. 2010). For instance, Thorn et al. (2013) report that many physicians in large, urban EDs that they studied would routinely turn to HIE as part of their efforts to improve patient care, reduce repeated testing, identify hidden allergies, or disambiguate unreliable information. One of the physicians in the study even described HIE as a "life-saving tool."

Some effects of HIE are known but not many and not with certainty. Recently, two reviews thoroughly examined the research on the effects of HIE, and the authors were disappointed by the lack of high-quality evidence (Rahurkar et al. 2015, Rudin et al. 2014). Notably, past research has evaluated the impact of HIE in terms of reduced imaging and testing (Lammers et al. 2014, Ross et al. 2013, Ayabakan et al. 2013, Bailey et al. 2013), cost savings (Frisse et al. 2012, Tzeel et al. 2011, Overhage et al. 2002), patient satisfaction (Vest and Miller 2011), readmissions (Jones et al. 2011), and hospital and ED visit rates (Vest 2009). However, similar to the broader HIT literature, empirical studies assessing the impact of HIE on operational measures are lacking. Such operational measures are exactly what we are considering in this study.

1.1.2 Emergency Departments

Many people experienced an ED first when they could not find a physician's appointment. After they would finish their ED appointment, they might return home, read newspapers, watch TV, and learn about "real", nerve-wrecking ED stories. In the stories, patients would be rushed by an ambulance into the depths of the hospital, sprinting for their life, endangered by a critical cardiac condition or a deadly road accident. In fact, ED clinicians fluctuate between both of these extremes, with about 13% ED patients being non-urgent, 15% emergent, 35% urgent, and the rest somewhere in between (Institute of Medicine 2006c). Clinicians get to know diverse patients, from infants brought by their parents with bronchitis or otitis media, young adults with car injuries or severe depression, uninsured adults without other source of care, all the way to elderly with pneumonia or heart conditions (Skinner et al. 2014, DeLia and Cantor 2009). Taken together, these examples illustrate several main aspects of EDs that distinguish them from other healthcare providers: EDs serve patients who cannot obtain care elsewhere, for financial, insurance, or availability reasons, provide emergency care needs such as trauma, support physician practices, and often serve as the first point of entry for mental-health patients (Institute of Medicine 2006c).

The Institute of Medicine (2006c) recounts a case of one typical urban ED and trauma center that is chronically overburdened, serving 80 patients instead of the 40 that it was designed for. Of these patients, about one third is boarding (waiting for a bed to open an internal ward), and many others are waiting for more than seven hours. Not only is the patient care in the ED compromised, but the ED limits outside access by diverting ambulances. In this story, if five new car accident patients arrive, the ED could not care for them adequately. Eventually, ED clinicians will have weathered another busy day. Sometimes, patients suffer from nothing more than inconvenience; other times, some of the patients may suffer from adverse outcomes. Although this ED is a part of a large, urban medical center with a level 1 trauma center, its story is far from unusual.

In 2012, there were 4,460 active EDs in the US, with this number still decreasing

(American Hospital Association 2014b). Among the EDs, 52% report having more than 10% of admitted patients stay in the ED for more than six hours (Horwitz et al. 2010), 45% report sometimes diverting ambulances (Burt et al. 2006), and 90% report boarding patients before hospital admission (Rabin et al. 2012). About 39% of ED directors report that their EDs are crowded every day (Moskop et al. 2009a). Overwhelmed and crowded EDs are not unhealthy just because of the inconvenience to the patients (Pines et al. 2008b). In fact, extended length of stay and crowded hallways can cause adverse outcomes, impair care quality, worsen care access, and trigger hospital losses (Hoot and Aronsky 2008). Some EDs, such as the ones in the Boston Medical Center or Grady Health System, use industrial engineering, operations research, and other state-of-the-art approaches to prevent and alleviate the negative consequences of crowding (Institute of Medicine 2006c). But other EDs still approach crowding only reactively (Rabin et al. 2012). The rural and urban EDs that serve particularly disadvantaged patient populations are of particular concern to public health (Trzeciak and Rivers 2003).

This emergency is why the Institute of Medicine was commissioned to study and recommend how to improve the country's EDs and the entire emergency system (Institute of Medicine 2006c,b,a). "[A] growing national crisis in emergency care is brewing," states one of the reports in the introduction. More attention to EDs represents an opportunity to improve the nation's ED system. We obviously have no silver bullet, but success stories of some EDs (Rabin et al. 2012) suggest that striving for improvement is a worthy aim.

1.1.3 Literature Contributions

Our work contributes to two broad streams of research: a) impact of health IT on clinical and operational outcomes, and b) healthcare operations management. In the broad HIT literature, studies assessing impact of HIT mostly considered quality metrics (e.g., Menon and Kohli (2013), Ruben et al. (2009), DesRoches et al. (2010)), while others have considered the impact of HIT on financial and efficiency outcomes. In particular, based on aggregate economic analyses, research has focused on the impact of HIT on costs (Menon and Lee 2000, Borzekowski 2009), revenues (Menon et al. 2000, Devaraj and Kohli 2003, Kohli and Devaraj 2003, Ayal and Seidman 2009), and productivity (Lee et al. 2013b). However, several other studies have shown that these aggregate level results do not necessarily hold when more granular metrics were used (e.g., Himmelstein et al. (2010), Agha (2014), Kennebeck et al. (2012a)). As a result, despite massive investments, the benefits of HIT are not deeply understood (Jones et al. 2014).

Specifically in the HIE literature, past research has evaluated the impact of HIE in terms of reduced imaging and testing (Lammers et al. 2014, Ross et al. 2013, Ayabakan et al. 2013, Bailey et al. 2013), cost savings (Frisse et al. 2012, Tzeel et al. 2011, Overhage et al. 2002), patient satisfaction (Vest and Miller 2011), readmissions, and ED revisits (Jones et al. 2011, Shy et al. 2016), and hospital and ED visit rates (Vest 2009). For more details about the recent findings, we refer the reader to excellent reviews by Goldzweig et al. (2009) and Jones et al. (2014) on the impact of HIT and by Rudin et al. (2014) and Rahurkar et al. (2015) on the impact of HIE.

Our research also contributes to the healthcare operations management literature at the interface of IT. To date, some management studies have focused on the relationship between IT and operational outcomes in non-ED settings. For example, Goh et al. (2011) conducted an elaborate qualitative study on workflow changes during the adoption of EHRs. Lahiri and Seidmann (2012) discussed how information gaps in healthcare could disrupt downstream clinical workflow in a large radiology network. Bhargava and Mishra (2014) studied how technology adoption impacted physician productivity and found differences in the level of productivity gains among different primary care specialties and across time periods. Bavafa et al. (2013) studied the impact of e-visits, a form of telehealth, on health outcomes and physician productivity and showed that e-visits complemented traditional office visits but could not replace them. Dobrzykowski and Tarafdar (2015) analyzed how local information exchange using EMR affected the communication between physicians and patients and find that increased information exchange improves patient-provider communication. Angst et al. (2011) studied whether the sequence in which various HITs were adopted mattered in terms of overall hospital performance including LOS and found that the sequence was an important predictor of performance. While all of these studies focused on health IT and operations management interface, we are not aware of any large-scale studies that analyzed the interplay between HIE adoption and LOS in ED settings, which is the focus of our study.

1.2 Hypothesis Development

In this section, we motivate and formulate our hypotheses about the impact of HIE adoption on the ED LOS. We start with an overall HIE and LOS relationship.

As we briefly discussed in the introduction, HIE may affect LOS through several pathways. Below, we discuss how reduced redundancy, changes in decision processes, and workflow effects due to new technology adoption may play a role in the relationship between HIE and LOS.

First, recent research has shown that HIE adoption leads to a substantial reduction in redundant imaging (Lammers et al. 2014), laboratory testing, and medication ordering (Hebel et al. 2012, Speedie et al. 2014). For instance, an HIE in New York allowed clinicians to directly query patients' information and access to their laboratory and radiology reports, the most commonly used HIE functionalities, resulting in significant reductions in the number of tests ordered (Campion et al. 2013). Given that the recent increases in ED occupancy rates are associated with increased use of diagnostic and treatment procedures such as blood tests, advanced imaging, and intravenous fluids (Pitts et al. 2012), HIE adoption may lead to a reduction in LOS by reducing the number of tests ordered, thus directly saving time.

Second, additional information accessed through HIE can either increase or decrease the efficiency of the decision-making process and hence may lead to a change (in either direction) in LOS (Hincapie et al. 2011). For instance, in the presence of HIE, clinicians may access otherwise absent patient history and identify patients with chronic pain or those with possible substance abuse problems (Johnson et al. 2008, Hincapie et al. 2011). Although the relationship between more efficient decision making and LOS is difficult to isolate, Stiell et al. (2003) were able to track the prevalence of information gaps among emergency physicians and the influence of these gaps on the eventual LOS. In particular, their study showed that the presence of information gaps such as unavailable medical histories or laboratory test results was associated with more than one hour of additional LOS. On the other hand, it is also possible that HIE could exacerbate the information overload facing ED physicians and subsequently result in impaired decision making (Ash et al. 2004, Spencer et al. 2004), leading to increased service duration and hence increased LOS.

Third, similar to the impact of other information technologies on workflow (Goh et al. 2011), the presence of HIE may significantly change the clinical management of a patient in ED (Franczak et al. 2014). Further, the changes introduced by the information technologies into the processes, which in some cases cause disruptions, may not lead to the expected benefits (Dranove et al. 2014, Bhargava and Mishra 2014). Hence, the process-changing effect of HIE could be associated with increased LOS.

Finally, the well-known distinction between technology adoption and use/access may play a role in the quantification of HIE and LOS relationship. Previous research reports that the actual rates of HIE access may vary. Rudin et al. (2014), in their recent systematic review on HIE, document that the HIE access rates vary between 2% and 10%, while Halamka (2013) argues that in their institution and possibly elsewhere, the access rates might be much higher. These observations highlight that HIE use may be driven by local context and implementation factors, including practice-related factors, patient condition, past utilization, age, comorbidities, crowdedness, race/ethnicity (Yaraghi et al. 2015, Vest et al. 2011, Rudin et al. 2011, 2014). Some of the purposes for which physicians reported HIE as the most useful are better decision making, identifying substance abuse, reviewing a patient's medications, and reducing excessive imaging and testing (Thorn et al. 2013, Hincapie et al. 2011). Under the circumstances in which HIE is more beneficial, use may be more likely (e.g., see Yaraghi 2015, for a good discussion around the incentives and HIE's value generation process). As compared to use, adoption captures the overall effect similar to other large scale HIE evaluation studies (e.g., Jones et al. 2011, Lammers et al. 2014, Vest and Miller 2011). The more frequent use of HIE would increase the value derived from its adoption – positive or negative.

In summary, the aggregate effect of HIE on LOS will likely depend on many factors. As we discussed above, some of these factors may lead to an increased LOS while others may reduce it. However, interviews with practicing ED physicians suggest that the overall expectation is that HIE would most often help in reducing the LOS (Thorn et al. 2013). We also remark that the use relates to the idiosyncrasies of an actual implementation at a micro level whereas we focus on the relation between the HIE adoption and LOS at a high level. We also note that access to and use of HIE will enhance the overall observed value of adoption, capturing the benefit at an aggregate average use level. Given the discussion around expected HIE impact on LOS, we formulate

Hypothesis 1 (a): HIE adoption leads to reduced ED LOS.

Hypothesis 1 (b): Increased use of HIE leads to an increased reduction in LOS.

1.2.1 The Role of Teaching Status

There has been extensive research on how teaching and non-teaching hospitals differ in terms of care delivery, process characteristics, and outcomes (Papanikolaou et al. 2006). To summarize, in addition to being home for medical education, teaching hospitals differ from non-teaching hospitals in several ways such as providing care more often for complex patients (Koenig et al. 2003), providing specialized services such as a neonatal intensive care unit or a trauma center (American Hospital Association 2009), treating rare diseases, serving indigent patient populations, conducting biomedical research (Ayanian and Weissman 2002), and using advanced technology more often (e.g., (Jha et al. 2009)). Indeed, compared with non-teaching hospitals, teaching hospitals are known for their relentless pursuit of innovation and use of technology (American Association of Medical Colleges 2014). Given these differences, differentiating the role of HIE adoption in these two different hospital organizational structures may prove useful and may in turn provide insights for policymakers, payers, patients, and researchers.

As innovators, teaching hospitals were among the early adopters of HIE (Jha et al. 2009). Consequently, many of the initial evaluation studies were based on local or regional datasets from HIEs originating in teaching hospitals (e.g., Johnson et al. 2008, Halamka 2013, Finnell et al. 2003), and these studies have served as a de facto benchmark on expectations from HIE. However, given that teaching hospitals only account for 20% of all the US hospitals (American Hospital Association 2014a), it is important to assess the value of HIEs comprehensively in a representative cohort of hospitals, including non-teaching hospitals.

Teaching hospitals may benefit more from HIE for at least two reasons. First, teaching hospitals are usually technologically advanced, with some forms of internal information sharing infrastructure. Empirical evidence suggests that many large teaching hospitals have long been sharing information internally (Miller and Tucker 2014), which could make it easier for them to engage in and adapt to external information-sharing activity when it becomes available. Second, HIE needs to be a part of regular clinician workflow for effective use (Kuperman and McGowan 2013, Halamka 2013), which teaching hospitals are more likely to accomplish. This is because other forms of HIT are more common and are typically already part of the workflow in teaching hospitals; and this past experience may lead to more effective use of HIE, and hence a larger reduction in LOS. Therefore, we formulate

Hypothesis 2: Teaching hospitals benefit more from HIE in terms of the reduction in ED LOS, compared with non-teaching hospitals.

1.2.2 The Role of Crowdedness

Crowdedness is a major potential barrier against the effective utilization of an advanced technology especially in ED setting because of the time- and information-sensitive nature of ED-specific care (Ben-Assuli et al. 2012). This is because despite all their advantages, advanced HITs including HIEs may sometimes have disruptive effects on the workflow (Kennebeck et al. 2012b, Kuperman and McGowan 2013), which may not be tolerable in crowded EDs.

Ineffective technology under extreme workload is consistent with adaptive and anomalous behavior observed also in other service operations (Bendoly et al. 2006, Chan and Green 2013). In particular, several studies in different settings and industries have shown that as the workload or crowdedness increases, workers change their behavior. For instance, Tan and Netessine (2014) have shown that waiters tended to speed-up in crowded restaurant settings. Similarly, in healthcare, clinicians suffering from overload in crowded settings were shown to speed-up by avoiding some of the procedures that they deemed non-critical (Batt and Terwiesch 2012, Kc 2014, Kuntz et al. 2014, Powell et al. 2012). In this regard, one of the "non-critical" processes considered by the ED clinicians may be accessing and utilizing HIE, which could hence be skipped when the ED is crowded (Vest et al. 2011). Overall, this suggests that the benefit of HIE might shrink for patients arriving to a crowded ED. Hence, we formulate

Hypothesis 3: Patients visiting a crowded ED will benefit less from HIE in terms of a reduction in LOS, compared with those visiting a non-crowded ED.

1.2.3 Patient-related Factors: Severity and Complexity

Another pressing issue in the ED in regards to deriving value from HIEs is the patient-related factors at a given care instance. In exploring the patient-related factors, we are motivated by the concepts of task complexity, task urgency, and the interdependency between the two in the management literature (Campbell 1988, Reddi and Carpenter 2000, Bozarth et al. 2009). Both the complexity and the urgency of a task relate to processing of information, and affect how the task is fit for the use of information technology and how the use of information technology for the task is perceived by the userperceived as fit in realizing the benefits of information technology (Goodhue 1995). In the context of our study, treatment of a patient can be viewed as a task. Urgency of a case in the ED setting is typically determined by the severity of the case, which is measured by the Emergency Severity Index (ESI) (Gilboy et al. 2012) and is typically proxied by the mode of ED arrivals, i.e. walk-ins vs. ambulance arrivals (Rucker et al. 1997, Larkin et al. 2006). On the other hand, complexity of a case is captured by the co-existence of multiple comorbid conditions, which is commonly measured by the Charlson comorbidity index (Charlson et al. 1987, Shwartz et al. 1996).

The need for coordinated information is higher for patients with severe conditions and for those suffering from multiple comorbid conditions. Severe patients visiting EDs often require coordination of multiple specialists (Risser et al. 1999) and will be typically subject to multiple procedures (Baumann and Strout 2007), which makes the information coordination critical. Similarly, patients with multiple comorbidities (i.e., simultaneous presence of multiple conditions), who we call "complex patients", may require good information management because of potential complications that may arise due to comorbid conditions (Stiell et al. 2003). The need for coordinated information for severe or complex patients therefore makes such patients a natural test bed for assessing the value of HIE adoption in terms of a process measure such as LOS.

There is evidence that advanced information technologies improve the quality of care more for severe or complex patients in non-ED care settings (McCullough et al. 2013). However, while interviews with physicians indicate that they also consider HIE more useful for patients with multiple conditions or those that are difficult to treat (Rudin et al. 2011), the question of whether the LOS will improve also for such patients in the fast-paced ED setting remains unanswered. With better information management through coordination with HIE, we expect that the process efficiency will increase and would be amplified for severe or complex patients, who need better information coordination. Although the actual use may vary from instance to instance, overall, the adoption and therefore the actual presence of the HIE technology would be associated with an average benefit for the patients of similar severity or complexity. As such, we will test the following two hypotheses for patient-related factors moderating the HIE and LOS relationship:

Hypothesis 4a: Severe patients visiting EDs with HIE will have shorter LOS than those visiting EDs without HIE.

Hypothesis 4b: Complex patients visiting EDs with HIE will have shorter LOS than those visiting EDs without HIE.

1.3 Methods

1.3.1 Data

Our data sources included 2009-2013 HCUP Massachusetts SEDD files (Agency for Healthcare Research and Quality (2014)), AHA Annual Survey Database (American Hospital Association 2014a), U.S. Census Bureau Databases, and HIMSS Analytics Database (HIMSS 2010), and the CMS Meaningful Use (MU) database (CMS 2016). In general, the SEDD provide detailed diagnoses, procedures, and patient demographics including age, gender, race, and insurance coverage (i.e., Medicare/Medicaid, private insurance, other insurance, and uninsured). As part of the HCUP Project, AHRQ negotiates with data organizations that maintain statewide data systems to acquire hospital-based data, processes the data into research databases, and subsequently releases a subset of the data to the public with a signed data use agreement. Some data elements are considered too sensitive by these data organizations for general release to the public. However, under the terms of their agreements with AHRQ, some AHRQ staff may use these more sensitive data for analysis. For this study, the Massachusetts Division of Health Care Finance and Policy granted permission to AHRQ for internal use of the data elements, admission time and discharge time. The key variable, ED LOS, is expressed in minutes, measured as the difference between admission time and discharge time. This dataset also includes unique encrypted patient identifiers that enabled us to track them across times and institutions.

We have linked the SEDD data to the AHA databases using unique hospital identifiers and then merged the resulting files to HIMSS and MU databases using unique Medicare provider numbers. The AHA database allowed us to obtain hospital characteristics such as teaching status and bed size. We used information technology variables from the HIMSS database to identify adoption of HIE over time. After linking all these datasets and excluding the visits that were made to a few hospitals with missing Medicare identification numbers, our final database after contained almost all (about 7.4 million) treat-and-release visits made to 63 EDs in the entire state of Massachusetts.

Before closing this section, we remark that treat-and-release ED visits are ideally suited to study the association between HIE adoption and LOS because 1) treat-and-release ED visits account for about 81% of all ED visits in the U.S. (National Center for Health Statistics 2013), and 2) in contrast to ED patients admitted to the hospital, the LOS for treat-and-release patients is not much sensitive to hospital circumstances unrelated to the ED itself such as inpatient bed availability (Emerman 2012, Ding et al. 2010, McClelland et al. 2011).

1.3.2 Variables of Interest

In this section, we describe several variables of special interest in detail. All the substantive variables considered in our models are presented in Table 1.1. First, the restricted part of the dataset contains exact admission and discharge times, which allows us to compute the exact LOS expressed in minutes, the dependent variable in our model, for every single ED visit. The calculated LOS captures the sum of the waiting time and the service time for a patient (Welch et al. 2011). We apply the logarithmic transformation to LOS and use log(LOS) in our regression analysis, which is consistent with the published literature (Bartel et al. 2014,

Table 1.1: Description of variables used in analyses

Variable	Description	Source
$\log(\text{LOS})$	Logarithm of the ED length of stay, expressed in minutes, measured as the difference between admission time and discharge time	AHRQ Intramural SEDD
HIE	1 if HIE was available during the visit. See the Data section for assumptions about HIE availability.	HIMSS
Year XX	Year dummy variable for XX=10,11,12,13	AHRQ Intramural SEDD
Crowded	1 if the ED was crowded on the patient's arrival. (see the Data section for crowdedness definition)	AHRQ Intramural SEDD
CCI	1 if the patient's Charlson Comorbidity Index for the visit ≥ 2, 0 otherwise.	Calculated from SEDD
Transport	1 if the patient arrived in an ambulance or a helicopter, 0 otherwise.	AHRQ Intramural SEDD
Female	1 if the patient is female, 0 otherwise	AHRQ Intramural SEDD
Age group	Dummy-coded age in years: $< 1, 1-5, 6-11, 12-18, 19-34, 35-44, 45-54, 55-64, 65-74, 75+$	AHRQ Intramural SEDD
Race/Ethnicity	A dummy-coded categorical variable: Hispanic, non-Hispanic white, black, Asian, other	AHRQ Intramural SEDD
Insurance	A dummy-coded categorical variable for the primary expected payer: 1) Medicare, 2) Medicaid, 3) private, 4) uninsured (i.e., self-pay or no charge), 5) other	AHRQ Intramural SEDD
High Income	1 for the upper 3rd and 4th quartile of median household income of the patient's ZIP code, 0 for others (1st and 2nd quartiles)	AHRQ Intramural SEDD
Weekend	1 if admitted on weekend, 0 otherwise.	AHRQ Intramural SEDD
Monday	1 if admitted on Monday, 0 otherwise.	AHRQ Intramural SEDD
Teaching Hospital	1 for teaching hospitals, 0 for others. A hospital is considered to be a teaching hospital if it has an AMA-approved residency program, is a member of the Council of Teaching Hospitals, or has a ratio of full-time equivalent interns and residents to beds of 0.25 or higher.	AHA
Large Bedsize	1 if hospital bed is $100+$ (rural hospitals), $200+$ (urban non-teaching hospitals), or $425+$ (urban teaching hospitals). 0 otherwise.	AHA
CCS	Patient diagnosis dummy-coded through the CCS scheme.	AHRQ Intramural SEDD
MU	A weighted average of the fulfillment of five HIE-related Meaningful Use measures.	CMS MU reports

Kc 2014) and is supported by formal statistical evaluation (Marazzi et al. 1998).

Another key variable is HIE adoption, which is captured in the HIMSS database. We observe this variable on a yearly basis. Because adoption of HIE may have a delayed impact, and we observe the HIE adoption variable on a yearly basis, we use a one-year post-adoption lag to label hospitals as HIE implementing hospitals, which is a commonly used approach in the literature (Lammers et al. 2014, Appari et al. 2013, McCullough et al. 2013). For instance, if a hospital is labeled in the HIMSS database as adopting HIE in 2009, we code it as having HIE in 2010 (but not in 2009). To explore the robustness of this approach, we also conduct our analysis with a two-year post-adoption lag instead.

To capture patients' overall medical condition, we use patient comorbidity and severity. In particular, to capture a patient's comorbid conditions, we use Charlson comorbidity index (CCI), a widely-used index to quantify the presence of multiple conditions (Murray et al. 2006, Charlson et al. 1987). Also, to capture the severity of a patient's condition, we use the mode of transportation, which indicates whether patients arrived in the ED via an ambulance/helicopter or via their own means. The mode of transportation is a strong indicator of the urgency of the need for healthcare services, and is therefore a good proxy for the true severity (Rucker et al. 1997, Larkin et al. 2006). In an alternative analysis (later defined as Disease Analysis), we use the primary diagnosis code of each ED visit to identify the primary diagnosis for that visit. For this purpose, we use HCUP's clinical classification of diseases (CCS) in identifying the diagnosis codes associated with each ED visit, collapsing the ICD-9 and ICD-10 codes into 258 clinically meaningful categories, which we label as CCS (Senathirajah et al. 2011).

Lastly, teaching status of hospitals and ED crowdedness at the time of patient's arrival were two other key variables in our analysis. To determine teaching status, we directly use the Teaching-hospital variable from the AHA data as described in HCUP data dictionary (HCUP 2014). For ED crowdedness, past studies used various measures such as waiting time, service time, the fraction of patients leaving without being seen, census in the waiting room, or census in the ED (Batt and Terwiesch 2015, 2012, Anderson et al. 2014, Song et al. 2014, Hwang and Concato 2004). Our approach approximates the relative census in the ED. In particular, we leverage our database by first creating a volume measure in the ED of a particular hospital across a given year and then identifying the volume for a given hour at any day. Combining the volumes for a given hour across the year, we obtain a distribution of volumes. We then use a year- and ED-specific volume threshold above which we label the ED as crowded and non-crowded otherwise. Specifically, we use the 0.75 quantile (ED- and year-specific) as the threshold to label an ED as crowded (Batt and Terwiesch 2012).

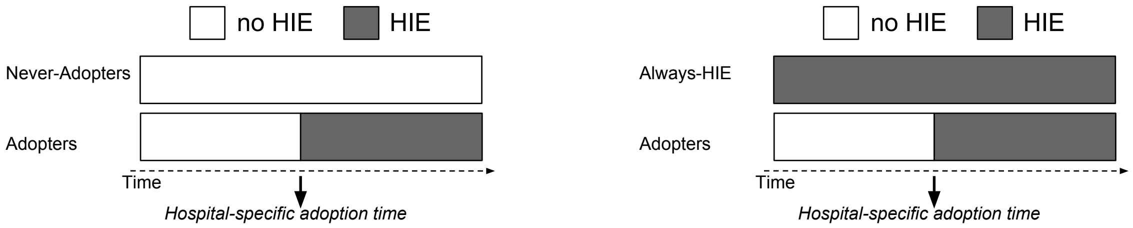
In addition to these key variables, our analyses require formulating additional variables, which we summarize next. In an alternative analysis ("the Index Visit Analysis"), we capture multiple ED visits using HCUP revisit variables. If the same patient visited the hospital multiple times in a given year, the HCUP would include separate records in the respective HCUP database for each visit. To facilitate analyses that focus on multiple hospital visits by the same person, AHRQ created a set of supplemental variables that can be linked to the HCUP state-level databases to track multiple (repeat) patient visits in the hospital setting, while adhering to strict privacy regulations. The encrypted patient identifiers together with the revisit variable allow tracking one person over multiple years and institutions, hence we are able to identify all the revisits (i.e., all subsequent visits by the same patient). In our Disease Analysis, we use the primary and secondary diagnosis codes of each ED visit record to associate a disease with a particular visit (Senathirajah et al. 2011).

We account for confounding effects that influence LOS by including factors that are related to patients and the hospitals. Consistent with the extant literature on LOS-influencing factors in the ED, we control for demographic factors (age group, gender, race/ethnicity), time indicators (year, day of week: Monday/ other weekday/ weekend), socioeconomic status (expected primary payers, the quartile of the patient's zip code median income), and hospital size.

1.4 Study Design and Econometric Specifications

We develop empirical models to assess the impact of HIE on ED LOS. As schematized in Figure 1.2, we group the hospitals appearing in our dataset into three subsets based on their adoption timing and status between the years 2009 and 2013 as i) hospitals that never adopted HIE during our study period (referred to as the *Never-Adopters group*), ii) hospitals that had HIE since the beginning of our study period (referred to as the *Always-HIE group*), iii) hospitals that adopted HIE during our study period (referred to as the *Adopters group*).

We employ a difference-in-differences (DID) approach to compare the LOS between the control and treatment groups. When interpreting our results, we do not highlight statistical significance according to conventional frequentist hypothesis testing. This is because, with large samples such as ours, interpretation of results should be based on the effect sizes



(a) Main Sample

(b) Adopters Sample

Figure 1.2: Depiction of Subsamples Used in the Study

and the uncertainty around the estimates instead of making conclusions based on p-values (Lin et al. 2013). Consistent with this approach, we draw on the most recent recommendations by the American Statistical Society: 1) capture the uncertainty in the estimates via standard errors and 2) highlight practical significance of the effects by providing easily-interpretable percentages (Wasserstein and Lazar 2016). That is, when presenting the results of regressions, we report the estimated effect (marginal percentage change in LOS), followed by the regression coefficient and and cluster-robust standard errors (accounting for clustering of visits within hospitals). This allows the reader to both see the effect size and calculate confidence intervals easily, a good practice recommended for empirical information systems research with large sample sizes (Lin et al. 2013).

1.4.1 Main Analysis

In our main analysis, we assess the LOS for an arbitrary visit to an HIE hospital as compared with an arbitrary visit to a non-HIE hospital. This perspective evaluates the overall value of health information sharing with respect to LOS in the ED setting by comparing Adopters to Never-Adopters subsets of hospitals (Figure 1.2a). Letting the indicator variable $HIE_{h,t}$ denote the HIE adoption status for hospital h in year t, we formulate a linear model, which we call *Overall* model, to estimate the association between HIE and ED LOS as follows:

$$\begin{aligned}\log(LOS_{v,h,t}) = \beta_0 + \beta_1 HIE_{h,t} + \beta_2 Crowded_{v,h,t} + \beta_4 Teaching_h + \beta_6 CCI_{v,h,t} \\ + \beta_8 Transport_{v,h,t} + \beta_{10} X_v + \beta_{11} X_h + \gamma_1 z_t + \gamma_2 Treat_h + \epsilon_{v,h,t},\end{aligned} \tag{Overall}$$

where $Crowded_{v,h,t}$ is the indicator for the hour of the visit v to hospital h during its crowded hours in year t, $Teaching_h$ is the teaching status of the hospital h, $CCI_{v,h,t}$ is the Charlson comorbidity index for visit v to hospital h in year t, $Transport_{v,h,t}$ is the indicator for the mode of transportation for visit v to hospital h in year t. X_v includes visit-level control variables (age group, gender, race/ethnicity, insurance, income) and X_h includes hospital-level control variables (bed size), and z_t is the fixed effect variable for year t. Lastly, $Treat_h$ is the dummy variable for treatment group, indicating whether hospital h is in the Adopters group and the variable $\epsilon_{v,h,t}$ captures the unobserved factors that change over visit v, hospital h, and time t.

To test the hypotheses on moderation effects, we formulate an analogous model to Overall model that captures interactions as follows:

$$\begin{aligned}\log(LOS_{v,h,t}) = {} & \beta_0 + \beta_1 HIE_{h,t} + \beta_2 Crowded_{v,h,t} + \beta_3 HIE_{h,t} Crowded_{v,h,t} \\ & + \beta_4 Teaching_h + \beta_5 HIE_{h,t} Teaching_h + \beta_6 CCI_{v,h,t} + \beta_7 HIE_{h,t} CCI_{v,h,t} \\ & + \beta_8 Transport_{v,h,t} + \beta_9 HIE_{h,t} Transport_{v,h,t} \\ & + \beta_{10} X_v + \beta_{11} X_h + \gamma_1 z_t + \gamma_2 Treat_h + \epsilon_{v,h,t}.\end{aligned} \tag{Interactions}$$

For alternative analyses, we create a sample based on visits to Always-HIE group (becomes the control) and those to Adopters group (becomes treatment). As for the model specifications, we retain the variable definitions as before, and in addition, introduce the interaction term $HIE\_Tr_{h,t} := HIE_{h,t} \times Treat_h$ as the variable of interest in our modeling setup, which will take a value of 1 for hospitals in the Adopters group after the time of HIE adoption, and be 0 otherwise. Then, based on this variable, we estimate the association between HIE adoption and LOS as follows:

$$\begin{aligned}\log(LOS_{v,h,t}) = \beta_0 &+ \beta_1 HIE\_Tr_{h,t} + \beta_2 Crowded_{v,h,t} + \beta_4 Teaching_h + \beta_6 CCI_{v,h,t} \\ &+ \beta_8 Transport_{v,h,t} + \beta_{10} X_v + \beta_{11} X_h + \gamma_1 z_t + \gamma_2 Treat_h + \epsilon_{v,h,t},\end{aligned} \tag{Alternative}$$

When analyzing the role of moderation effects on this subsample, we again utilize the Interactions model given above, where $HIE_{h,t}$ is replaced with $HIE\_Tr_{h,t}$.

While our quasi-experimental setup is a well-established approach to limit confounding effects, our estimation of the impact of HIE adoption on LOS based on *Overall* model could potentially be subject to endogeneity because 1) hospitals that perform poorly in terms of LOS may self-select into the HIE adoption, 2) more efficient hospitals may decide to adopt HIE earlier, and 3) hospitals may adopt concurrent LOS-improving initiatives together with HIE. We mitigate this potential endogeneity and demonstrate the validity of our results through i) conducting robustness checks as described in § 1.4.2 and ii) identifying relatively more homogeneous subsets of hospitals where comparability of treatment and control groups are less of a concern due to financial or operational incentives as described in § 1.4.3. Below we describe these analyses in more detail.

1.4.2 Robustness of the Main Analysis

Fixed Effects (F.E.) Analysis: We estimate our models with hospital fixed effects (alongside time fixed effects) to capture time-invariant differences across hospitals in any observable or unobservable predictors, such as differences in hospital quality metrics, managerial skills, or patient population (Litwin et al. 2012). The inclusion of hospital fixed effects robustifies our analysis if, for instance, the hospitals that adopt HIE have other time-invariant characteristics that also help them decrease LOS more effectively. We operationalize the

Fixed Effects model by estimating *Model-Overall* with hospital fixed effects, u_h, replacing the indicator $Treat_h$ and the hospital-level control variables X_h. Hospital and year fixed effects account for unobservable factors related to a hospital or time-based trends.

Disease Analysis: The effect of HIE on LOS is also likely to vary across disease groups. In particular, for some diseases, HIE can directly reduce the over-utilization of ED resources and hence reduce LOS (e.g., by reducing the rate of radiological imaging or laboratory testing, Lammers et al. 2014, Frisse et al. 2012), while for others, HIE may just provide more clinical background, leading to no change or even an increase in LOS. This line of reasoning is further supported by the observations that clinicians use HIE less frequently for certain conditions, presumably expecting that more information would change their behavior little (Vest et al. 2011). In addition to capturing the clinical complexity of a patient by CCI variable, we further conduct robustness on the main analysis by including diagnosis codes (i.e., the indicators for the CCS codes) associated with each visit. The robustness check with CCS help us identify possible differences in HIE's value in terms of the relationship with LOS in the context of different diseases/conditions.

Pre-trend Analysis: For a robust DID estimation, parallel trends in the control and treatment groups are preferable (Bertrand et al. 2004). When there is heterogeneity in the treatment and the control groups in regards to the efficiency before adoption, the observed post-adoption reduction in LOS in the Adopters group may be attributable to pre-adoption trends in hospitals belonging to this group. To address this concern, we estimate the *Pre-trend* model by adding pre-adoption year dummies to the *Overall* model. The dummy variable for focal hospital takes the value one in the year preceding the HIE adoption and zero otherwise.

Use Analysis: Our main analysis around adoption provides an aggregate effect based on average use. Per our discussion in § 1.2, the actual use rates could vary at the hospital level, and it is expected that any impact of HIE would scale up with increased level of use. To validate this intuition that the observed HIE effect would be even stronger under actual use, we proxy the rate of access to HIE based on the use data as reported in the CMS's Meaningful Use program. Specifically, we interact the HIE adoption variable with the reported use under the CMS Meaningful Use program (denoted by *MU*) and include it in the Overall model.

The criteria under the Meaningful Use incentive program were established by CMS to measure not just adoption of health information technology but actual "meaningful" use thereof (Blumenthal and Tavenner 2010). We create the MU variable by aggregating the variables from the Meaningful Use Program menu objectives that relate to HIE use. Namely, we measure the proportion of fulfilled HIE-related menu objectives that the hospitals attest to regarding the following: (1) "Provide summary of care record for patients referred or transitioned to another provider or setting," (2) "Submit electronic data on reportable laboratory results to public health agencies," (3) "Submit electronic syndromic surveillance data to public health agencies," (4) "Perform medication reconciliation between care settings," and (5) "Submit electronic immunization data to immunization registries or immunization information systems." For the attestation process, first, a random patient population is selected regarding a menu objective. Then, the hospital's attestation for performing the objective is deemed satisfactory based on a threshold as specified in the Meaningful Use Program. For example, the menu objective (1) is deemed as achieved when a summary of care record is provided for more than 60% of randomly chosen patients who are referred or transitioned to another provider or setting.

Instrumental Variable Analysis: In addition to Fixed Effects, and Use Analysis as well as several robustness checks that we introduce later, we further implemented an instrumental variable (IV) estimation approach to address any potential endogeneity problems. The IV should explain the variation in hospitals' HIE adoption decisions but not the variation in LOS except that through the IV. As such, we propose and use the number of security-related software applications that a hospital has, recorded in the HIMSS survey, as the IV in our analysis (HIMSS 2010). Using the "number of security-related software applications" as the IV, we apply two-stage least squares (2SLS) to the *Overall* model with hospital fixed effects. A variable indicating the number of the IT-security investments is a good IV because communication-related IT adoption, such as HIE adoption, exposes the hospital to additional security risks; therefore, HIE-adopting hospitals will likely enhance their IT security by investing in more security technologies.[1] However, an increased number of IT security investments is not expected to be correlated with the operational efficiency in the ED. We remark that our choice of IV is also consistent with the prior research that used non-clinical health IT adoption as an instrument to endogeneity due to time-varying unobserved factors (c.f., Hydari et al. 2014).

The second IV that we considered is the hospital-wide information strategy "to go paperless," as recorded in the HIMSS survey (HIMSS 2010). Past research has shown that hospitals' strategy to go paperless triggers the adoption of HIT across different departments in a hospital (Sands et al. 1997, Dykstra et al. 2009). As such, it is expected the strategy of going paperless would be correlated with HIE adoption decisions. On the other hand, the literature on going paperless does not suggest that hospitals with this objective would systematically make major operational changes in the ED other than IT adoption (Sands et al. 1997, Dykstra et al. 2009, Vezyridis et al. 2011). Hence, using the "going paperless" strategy as the IV, we apply two-stage least squares (2SLS) to the *Model-Overall* with hospital fixed effects.

It is possible that a hospital adopting HIE in the ED would change processes around its effective use, however, workflow redesign is a critical part of any technology adoption to realize its value (Goh et al. 2011), and hence such a change, if exists, should not bias our results. We also remark that concurrent adoption of EMRs with HIEs could potentially confound our estimates, because simultaneously adopted EMRs would also correlate with going paperless strategy. However, our analysis has shown that only two hospitals out of 63 adopted both advanced EMR and HIE during the study period, However, our analysis has

[1] The first stage of IV estimation corroborates our intuition with an effect size of 1.6%.

shown that only few hospitals adopted both advanced EMR and HIE during the study period (4% of all %visits in our sample), and none of them did so in the same year. Excluding the simultaneous adopters from our sample, and reestimating the coefficients accordingly provided similar insights. Hence, we believe concurrent adoption of EMR technologies in the ED setting does not affect our findings.

We have also considered as HIE the "HIE concentration in the vicinity". To define "vicinity", we consider areas called hospital referral regions (HRR) as defined in Wennberg and Cooper (1996). To define our IV, we counted for each year and each hospital, what the percentage of HIE adoption was among other hospitals in the same HRR. We reason that hospitals may be more induced to adopt HIE if many other hospitals nearby adopted because then HIE can provide more information. At the same time, we would not expect that HIE at other hospitals would directly affect LOS in the hospital in question. Unfortunately, the hospitals in our dataset fell under only five different HRR, which did not yield enough variation to leverage this IV.

1.4.3 Alternative Analyses and Robustness Checks

In addition to the robustness around the Overall model, to further strengthen our results, we conduct alternative robustness and falsification tests using systematically created subsamples. We select subsamples based on i) financial incentives of hospitals that makes information exchange and therefore actual use more likely (Systems Analysis), ii) choice of a more uniform set of hospitals so as to mitigate endogeneity concerns around hospitals' self-selection into HIE adoption (Adopters Only or Matched Analyses), iii) conditioning on a previous visit to an HIE hospital for assuring the usefulness of HIE in the subsequent visits (Index Analysis).

(Integrated) Systems Analysis: We acknowledge that some of the hospital characteristics that are unobservable to us may influence operational performance. To address this potential confounding effect, we conduct the (Integrated) Systems Analysis which focuses on a more homogeneous sample of hospitals. In particular, we restrict our focus to hospitals belonging to a “centralized health system”, which is defined as a “delivery system in which the system centrally organizes hospital service delivery, physician arrangements, and insurance product development,” representing the highest degree of integration in the American Hospital Association survey (Bazzoli et al. 1999). Therefore, if two hospitals are part of the same system, it is expected that their approach to care quality, cost-consciousness, and management would be similar. For example, the adoption and implementation of HIE for hospitals within the same hospital systems are more likely and incentives for using the shared information are stronger (De Brantes et al. 2006). In our Systems Analysis, because some of the hospitals from the same system in our sample varied in their HIE adoption timings, the confounding effect of differing control and treatment groups is mitigated in our DID setup. To conduct Systems Analysis, we use Overall model with hospital fixed effects and estimate it on the sub-sample of integrated health systems.

Furthermore, the ED physicians in these hospitals are motivated to use HIE at least for patients from within other system hospitals and physicians, and the factors influencing the HIE use may vary less across the hospitals within one system. Our results in Table A.2 show a decrease in ED LOS after HIE adoption for non-teaching hospitals but an increase for teaching hospitals. Other interactions remain largely consistent with the previous findings. Major barriers to information sharing are competitive concerns and interoperability Kuperman and McGowan (2013), McCarthy et al. (2009a), Shapiro et al. (2007). These concerns should be irrelevant or alleviated for hospitals within the same system. Indeed, Vest (2010) found that system hospitals are more likely to both adopt and actually implement HIE while hospitals in competitive environments are less likely to do so.

Adopters-only Analysis: As discussed earlier, organizational differences between HIE-adopting and non-adopting hospitals may confound our analyses (Robinson et al. 2009). Some of the previously presented analyses such as Hospital Fixed Effects account for time-invariant characteristics of hospitals, while the System Analysis creates a more uniform set of hospitals to be included in the control and treatment groups. To further alleviate any concerns about a potential self-selection bias, we also present Adopters-only Analysis, where we exclude the non-adopters and conduct the analysis based only on HIE-adopting hospitals and those that had HIE throughout the entire study period (recall that the corresponding subsample is called Adopters Sample). This subsample of HIE-adopter ensures a uniform occurrence of IT diffusion in our subsample and leads to more uniform set of hospitals with increased similarity of control and treatment groups with respect to HIE adoption behaviors/tendency. Because these characteristics may also influence operational performance, adopters-only analysis is expected to shed light on any confounding effects.

Matching Analysis: In addition to all of the above mentioned endogeneity-related analyses, we also conduct a matching analysis where we match hospitals into the two groups. A formal matching method such as propensity score matching is the preferable method in such cases. However, despite having a large sample in terms of number of visits, the limited number of hospitals in the Adopters or Never-Adopter groups hinders us from performing such an analysis. As an alternative, we match the hospitals based on the strongest reported predictors of HIE adoption: teaching status, hospital size (number of beds), and patient volume (e.g., Adler-Milstein and Jha 2014). More specifically, we create a subsample of Adopters and Never-adopters based on exact matching of the combination of these features, where seven hospitals are included in the final analysis in each group. In addition to Systems, Adopters-only, and Matching Analyses creating a more comparable control and treatment groups (albeit somewhat informally in the case of Matching Analysis), the fact that Massachusetts is a small state with similar hospitals in terms of infrastructure provides further support and strengthens the thought that treatment and control groups are mostly similar.

Index Visit Analysis: If the improvement in LOS is in fact attributable to HIE, one would expect that a patient should, on average, observe higher benefits in her subsequent visits following an initial visit. Building upon this idea, our Index Visit Analysis considers a subsample of patients where an initial visit to an HIE-carrying hospital is ensured. By comparing the subsequent ED LOS for visits to HIE-carrying hospitals against non-HIE hospitals, our Index Visit Analysis quantifies the value of HIE among patients who had their first visit to an ED with HIE. In other words, the Index Analysis serves as a falsification test where if there is indeed a causal link between HIE and ED LOS, then the observed ED LOS in the Index Analysis should be higher than that observed in the Overall Analysis. Otherwise, this Index analysis would falsify a causal relationship in the overall analysis, which estimates the average effect of HIE.[2]

To create the Index Sample, we initially select individuals with multiple ED visits from 2009 through 2013 from the Adopters Sample. Next, we subset this data to patients with at least one visit to any ED setting with HIE. Then, we identify the first visits to EDs with HIE for these patients and keep their entire subsequent visits (to EDs with and without HIEs) in our analytic files. Such patients, also termed as high utilizers or "frequent fliers" (Griswold et al. 2005), substantially burden EDs (LaCalle and Rabin 2010), and in line with our line of thoughts, clinicians generally expect HIE to help more substantially to this subpopulation (Thorn et al. 2013). In some sense, by controlling for the availability of information from an initial visit, we are increasing the chances of the index visit information being accessible to the caregivers in subsequent visits.

For the Index Analysis, we modify the *Alternative* model and replace the variable $HIE\_Tr_{h,t}$ with $HIE\_Tr(next)_{h,t}$ which takes value 1 for any visit to an ED with HIE where the patient's first visit is restricted to be an HIE hospital. Then, we estimate the

[2] We remark that even in the case when a patient does not have a prior visit to a hospital with HIE, there may still be information available through other sources that may be shared via HIE. Such information may come from labs, radiologists, pharmacies, and physician offices, etc.

relationship between HIE adoption and LOS in the Index Sample using:

$$\begin{aligned}\log(LOS_{v,h,t}) = \beta_0 + \beta_1 HIE\_Tr(next)_{h,t} + \beta_2 Crowded_{v,h,t} + \beta_3 Teaching_h + \beta_4 CCI_{v,h,t} \\ + \beta_5 Transport_{v,h,t} + \beta_6 X_v + \beta_7 X_h + \gamma_1 z_t + \gamma_T reat_h + \epsilon_{v,h,t}.\end{aligned} \tag{Index}$$

1.5 Results

1.5.1 Descriptive Results

We begin our analysis with a descriptive comparison of ED LOS based on HIE adoption statuses of hospitals. Table 1.2 presents the summary statistics stratified by the variables of interest and the groups of patients that we use in different analysis: 1) Never-adopters group, 2) Adopters group prior to HIE adoption, 3) Adopters group after HIE adoption, and 4) Always-HIE group. We observe a sizable difference in ED LOS distribution across cohorts. In particular, among the Adopters group, both mean and median LOS were lower in post HIE adoption period compared with pre-adoption period, while mean and median LOS were similar in the Always HIE and Never-adopters groups. The ratio of patients visiting a crowded ED to a non-crowded ED, severe to non-severe patients, and patients with CCI$<$2 to those with CCI$\geq$2 are similar in HIE and non-HIE hospitals. Ratio of visits happening to teaching in comparison to non-teaching hospitals is substantially higher in HIE-carrying hospitals for both subsets. This is expected because teaching hospitals are more likely to adopt HIE and therefore higher proportion of HIE visits would belong to the teaching hospitals. While treat-and-release visits are made by persons of all age, most of the visits belong to patients who are between 19 and 54 years old. This is not surprising because older patients are much more likely to be admitted to a hospital than being discharged. Lastly, while we do not present it in the table, we remark that most of the ED visits are paid by private insurance (38%), followed by Medicaid (27%) and Medicare (16%). We present the correlation matrix between all variables (except age) in the appendix, Table A.1.

Table 1.2: Treat-and-release emergency department visits' length of stay and total volume

	Never-adopters					Adopters, pre HIE adoption					Adopters, post HIE adoption					Always HIE				
	Total Visits	Mean LOS	Q1	Median	Q3	Total Visits	Mean LOS	Q1	Median	Q3	Total Visits	Mean LOS	Q1	Median	Q3	Total Visits	Mean LOS	Q1	Median	Q3
Total	4,881,612	202	84	146	237	1,337,002	231	92	155	249	1,185,993	217	97	162	255	2,319,153	215	87	149	244
Year																				
2009	978,644	200	83	142	228	483,769	255	96	159	254	38,094	161	52	104	206	456,457	203	83	145	239
2010	941,494	193	82	142	230	385,306	240	97	162	259	112,487	189	84	148	235	457,342	203	79	141	235
2011	938,648	196	84	146	238	240,819	197	86	146	237	273,375	226	107	173	266	459,319	210	84	146	239
2012	1,028,761	206	85	148	241	160,870	196	84	142	229	360,207	222	100	165	259	450,894	222	89	152	249
2013	994,065	213	87	152	247	66,238	216	82	142	241	401,830	219	96	160	253	495,141	236	97	161	258
Crowdedness																				
0(<75%)	3,762,784	201	83	145	236	1,019,542	228	90	153	249	923,278	216	96	162	256	1,783,399	214	86	148	244
1(>=75%)	1,118,828	206	88	150	238	317,460	242	97	158	249	262,715	219	101	164	251	535,754	221	90	153	245
Teaching Status																				
Non-Teaching	3,595,084	183	78	134	217	756,140	198	86	146	234	378,746	201	83	146	234	583,230	178	77	134	217
Teaching	1,286,528	255	110	185	293	580,862	274	100	166	268	807,247	224	104	169	264	1,735,923	228	90	155	254
Selected Modes of																				
Ambulance	735,614	308	139	215	331	221,890	307	136	211	325	222,766	309	144	216	326	361,735	330	135	216	347
Helicopter	346	296	115	229	400	41	380	102	198	362	63	233	108	202	307	342	506	157	319	601
Walk-in, including public or private transportation	3,765,222	181	79	135	219	1,080,043	217	87	144	233	930,448	196	90	151	239	1,648,868	159	81	139	227
Comorbidity																				
CCI<2	4,781,720	201	84	145	235	1,301,565	229	91	153	246	1,142,597	215	95	160	252	2,271,052	213	86	148	242
CCI>=2	99,892	253	134	206	300	35,437	327	153	227	329	42,396	262	153	221	308	48,101	337	152	235	354

Note: The data includes all treat-and-release emergency visits during 2009-2013 in Massachusetts. LOS is measured in minutes as the difference between admission time and discharge time for each visit. The data with cell size less than 10 are suppressed.

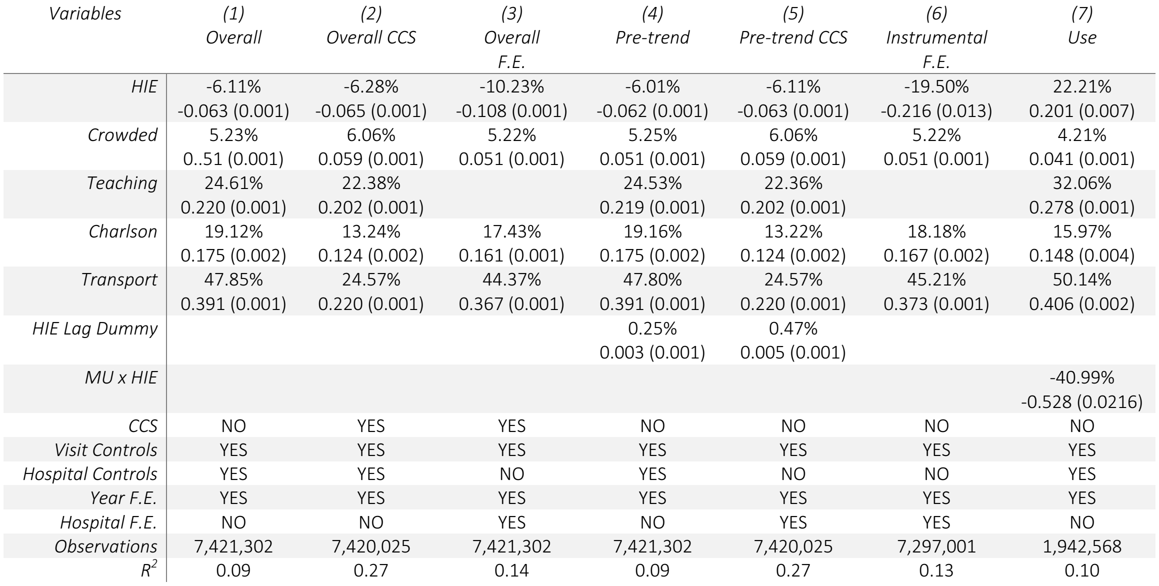
1.5.2 Results from the Overall Analyses

In Table 1.3, we present the main results based on *Overall* model with three main estimation methods: i) OLS estimation with time fixed effects, ii) estimation with hospital and time fixed effects, iii) estimation using IV along with the time and hospital fixed effects.

As we observe from this table, the impact of HIE adoption in treatment group (as captured by HIE in models (1)-(6)) on the average LOS per ED visit is negative. That is, HIE adoption decreases the LOS in all estimation methods, which is in line with our Hypothesis 1 (a). Based on model (1), the estimated average reduction for adopter hospitals is 6.1%. Given that the average LOS in these hospitals before adoption is 231 minutes, this practically corresponds to about 14 minutes average reduction in ED LOS, a highly significant reduction in the ED setting. As for the analyzed contextual variables (i.e., teaching status, patient severity and complexity), all three estimation methods lead to the expected signs for other relevant coefficients. In particular, teaching hospitals have on average longer LOS, presumably due to more complicated case mix of patients and the practice of more advanced care compared with non-teaching hospitals; patients who arrive to crowded EDs experience longer LOS as expected from basic queueing theory; and severe and complex patients stay longer, as their treatment may simply require more time.

The robustness check that controls for specific disease conditions captured by model (2) produces consistent estimates of HIE and LOS relationship with model (1). The results from the Fixed Effects Analysis in model (3) indicate a 10.2% decrease in LOS among Adopters, which translates to a 23 minutes reduction in LOS, following the HIE adoption. This result implies that while adopter hospitals may differ on their persistent characteristics (captured by the fixed effects), these differences do not confound the HIE and LOS relationship. The small and positive effect sizes and relatively larger standard errors for the pre-trend dummies in Models (4) and (5) indicate that adopting hospitals do not have a downward trend of LOS before adoption, and provide further support for causality between HIE adoption and reduced ED LOS.

Table 1.3: Regression results for full-set analyses



Variables	*(1) Overall*	*(2) Overall CCS*	*(3) Overall F.E.*	*(4) Pre-trend*	*(5) Pre-trend CCS*	*(6) Instrumental F.E.*	*(7) Use*
HIE	-6.11% -0.063 (0.001)	-6.28% -0.065 (0.001)	-10.23% -0.108 (0.001)	-6.01% -0.062 (0.001)	-6.11% -0.063 (0.001)	-19.50% -0.216 (0.013)	22.21% 0.201 (0.007)
Crowded	5.23% 0..51 (0.001)	6.06% 0.059 (0.001)	5.22% 0.051 (0.001)	5.25% 0.051 (0.001)	6.06% 0.059 (0.001)	5.22% 0.051 (0.001)	4.21% 0.041 (0.001)
Teaching	24.61% 0.220 (0.001)	22.38% 0.202 (0.001)		24.53% 0.219 (0.001)	22.36% 0.202 (0.001)		32.06% 0.278 (0.001)
Charlson	19.12% 0.175 (0.002)	13.24% 0.124 (0.002)	17.43% 0.161 (0.001)	19.16% 0.175 (0.002)	13.22% 0.124 (0.002)	18.18% 0.167 (0.002)	15.97% 0.148 (0.004)
Transport	47.85% 0.391 (0.001)	24.57% 0.220 (0.001)	44.37% 0.367 (0.001)	47.80% 0.391 (0.001)	24.57% 0.220 (0.001)	45.21% 0.373 (0.001)	50.14% 0.406 (0.002)
HIE Lag Dummy				0.25% 0.003 (0.001)	0.47% 0.005 (0.001)		
MU x HIE							-40.99% -0.528 (0.0216)
CCS	NO	YES	YES	NO	NO	NO	NO
Visit Controls	YES	YES	YES	YES	YES	YES	YES
Hospital Controls	YES	YES	NO	YES	NO	NO	YES
Year F.E.	YES	YES	YES	YES	YES	YES	YES
Hospital F.E.	NO	NO	YES	NO	YES	YES	NO
Observations	7,421,302	7,420,025	7,421,302	7,421,302	7,420,025	7,297,001	1,942,568
R^2	0.09	0.27	0.14	0.09	0.27	0.13	0.10

Note: The first number denotes the estimated percentage change, the second number the corresponding regression coefficient, and the third number the robust standard error. The symbol "x" indicates the presence of the specific controls or the fixed effects in the model while no entry suggests they are absent.

The Use Analysis in model (7) addresses the question of whether size of the effects is strengthened by increased use. The negative coefficient estimate for the $MU \times HIE$ suggests that the effect of HIE is higher as the hospital tends to use the HIE-related features more. This finding confirms our intuition in Hypothesis 1 (b) and supports the practitioners' expectation as argued in Halamka (2013).

In reference to IV analysis, we note that the first-stage results from the 2SLS confirm that our proposed IV—the number of security-related IT implementations—is correlated with HIE adoption (results presented in the Appendix). The second-stage results as captured by model (6) are in the same direction as the results from the prior analyses with an effect size of 19.5$\%$, even though the IV results exhibit a relatively higher standard error, as expected. We also note that the Cragg-Donald F statistic is significant and our instrument meets the Stock and Yogo (2002) strength test. The result based on IV analysis further corroborates our findings and is notable because estimation using an instrument, in principle overcomes many endogeneity threats, including simultaneity (e.g., hospitals that set on the journey of improving LOS also start implementing HIE) or omitted variable bias. On the other hand, despite the favorable and consistent findings based on the IV Analysis, we acknowledge that finding a strong instrument is not an easy task. In that regard, we note that other robustness checks and alternative analysis as presented next provide further evidence on our hypothesized relationship between HIE and LOS. We include the results from the "going paperless" analysis in the appendix.

1.5.3 Results from Alternative Analyses

Next, we turn our attention to the results on specific patient and hospital populations as presented in Table 1.4. In particular, we present the results for various estimations based on Systems, Adopters-Only, Matched, and Index Analyses. These analyses 1) serve as robustness checks because we restrict our data to study more homogeneous populations and 2) work as a generalization check to show that our main findings hold qualitatively in different settings, though the size of the HIE impact varies.

The Main Sample consists of eight centralized systems in Massachusetts, where hospitals within a system are typically more homogeneous compared with other hospitals and would have stronger incentives to use HIE when exchanging information with the affiliated system hospitals. The estimation uses fixed effect and robust standard errors clustered within the systems, and in line with our expectations, shows that the impact of HIE (-14.8%) is qualitatively similar and quantitatively stronger than the estimates from the Overall Analysis with fixed effects as in model (3) (-10.2%). Similar to the Systems Analysis, the results from Adopters-only and the Matched Analyses, which were also concerned with the possible non-homogeneity in the hospitals falling into the control and treatment groups, are consistent with the previous estimations.

Recall from Section 1.4.3 that in contrast to an arbitrary patient that we study in the Overall Analysis, Index Analysis studies those patients with multiple ED visits who had at

Table 1.4: Regression results for Alternative Analyses

Variables	*(8)*	*(9)*	*(10)*	*(11)*	*(12)*	*(13)*
	Overall Systems	*Adopters*	*Overall*	*Overall Matched*	*Index*	*Index F.E.*
	F.E. CCS	*Only*	*Matched*	*F.E. CCS*		
HIE (or HIE_Tr)	-14.84%	-6.91%	-15.31%	-6.43%	-10.24%	-21.03%
	-0.161 (0.004)	-0.072 (0.001)	-0.166 (0.002)	-0.067 (0.002)	-0.108 (0.002)	-0.236 (0.004)
Crowded	9.78%	4.83%	8.38%	8.92%	5.40%	5.59%
	0.093 (0.002)	0.047 (0.001)	0.081 (0.002)	0.085 (0.001)	0.053 (0.001)	0.054 (0.001)
Teaching		20.08%	17.89%		21.58%	
		0.183 (0.001)	0.165 (0.002)		0.195 (0.001)	
Charlson	9.66%	24.48%	17.87%	11.55%	24.53%	18.31%
	0.092 (0.004)	0.219 (0.002)	0.164 (0.003)	0.109 (0.003)	0.219 (0.004)	0.168 (0.004)
Transport	21.26%	45.85%	35.07%	16.03%	59.19%	56.69%
	-0.193 (0.002)	0.377 (0.001)	0.301 (0.002)	0.149 (0.002)	0.465 (0.002)	0.449 (0.002)
CCS	YES	NO	NO	YES	NO	NO
Visit Controls	YES	YES	YES	YES	YES	YES
Hospital Controls	NO	YES	YES	NO	YES	NO
Year F.E.	YES	YES	YES	YES	YES	YES
Hospital F.E.	YES	NO	NO	YES	NO	YES
Observations	965,281	4,860,829	1,306,700	1,306,700	1,840,425	1,840,425
R^2	0.32	0.09	0.13	0.30	0.09	0.14

Note: The first number denotes the estimated percentage change, the second number the corresponding regression coefficient, and the third number the robust standard error. The symbol "x" indicates the presence of the specific controls or the fixed effects in the model while no entry suggests they are absent.

least one previous ED visit to an HIE-enabled hospital. We expect such frequent utilizers of EDs to benefit more from HIE. In line with this intuition, we find that such patients on average experience a 10.2% reduction in LOS, compared with a 6.1% reduction for an arbitrary patient as captured by the Overall Analysis in model (1) of Table 1.3. Similarly, when estimated with hospital fixed effects, we observe a 21.5% reduction in this population, compared with a 10.2% reduction from the Overall Analysis in model (3) of Table 1.3.

1.5.4 Variation among HIE Networks

Depending on the year within our timespan, Massachusetts had about four HIE networks. Specifically, within our period, the following HIE networks were present: MaShare, Massachusetts Health Data Consortium, New England Healthcare EDI Network, CHAPS, Massachusetts eHealth Collaborative, and South Eastern Massachusetts Regional Health Information Organization. Since different networks may mean varying experience for adopting hospitals, we could explicitly test this variation. To do so, in our fixed-effect overall model, we could model the HIE coefficient as random, varying by the exchange network. This would allow us to estimate the variation across exchanges, and test whether this variation is statistically significant. This analysis would need to assume that there are no differential trends within hospitals in different networks. Unfortunately, many hospitals did not report their HIE network. There was a wide variation in the number of hospitals participating in different networks. In fact, a plurality of hospital visits do not even have HIE reported (we denote it as HIE Network unknown). Among the others, there is still substantial variation (see Table 1.5).

Table 1.5: Regression results for the HIE Networks analysis

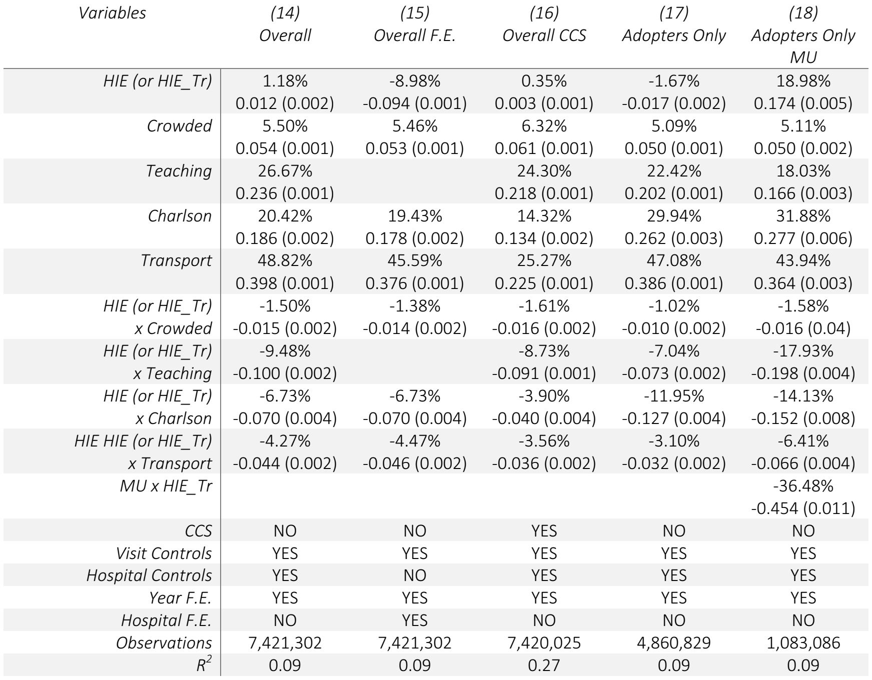
Coefficient	**HIE Networks**	
HIE Network unknown	0.6%	0.006 (0.001)
HIE Network 1	-13.7%	-0.148 (0.001)
HIE Network 2	-25.9%	-0.299 (0.002)
HIE Network 3	26.8%	0.237 (0.002)
HIE Network 4	0.0%	0.000 (0.002)
Controls:Visit	x	
Controls:Hospital	x	
FE: Year	x	
FE: Hospital		
N	4,687,266	
R^2	0.09	

Note: The first number denotes the estimated percentage change, the second number the corresponding regression coefficient, and the third number the robust standard error. The symbol "x" indicates the presence of the specific controls or the fixed effects in the model while no entry suggests they are absent.This analysis includes only the adopter hospitals. The baseline group are the visits that occurred before HIE was implemented.

1.5.5 Results for Interaction Analyses

In Table 1.6, we present the results with interactions based on variations of Overall and Adopters-only Analysis estimations to assess the effects of the moderating variables on the HIE-LOS relationship, as conjectured in Hypotheses 2-4. While the interaction results are consistent both in direction and standard errors across the five analyses, below, we discuss in detail the results from the Overall and Adopters-only estimations as these estimations, unlike the fixed effects counterparts, explicitly capture hospitals' teaching status.

Table 1.6: Regression results for interaction analyses



Variables	(14) Overall	(15) Overall F.E.	(16) Overall CCS	(17) Adopters Only	(18) Adopters Only MU
HIE (or HIE_Tr)	1.18% 0.012 (0.002)	-8.98% -0.094 (0.001)	0.35% 0.003 (0.001)	-1.67% -0.017 (0.002)	18.98% 0.174 (0.005)
Crowded	5.50% 0.054 (0.001)	5.46% 0.053 (0.001)	6.32% 0.061 (0.001)	5.09% 0.050 (0.001)	5.11% 0.050 (0.002)
Teaching	26.67% 0.236 (0.001)		24.30% 0.218 (0.001)	22.42% 0.202 (0.001)	18.03% 0.166 (0.003)
Charlson	20.42% 0.186 (0.002)	19.43% 0.178 (0.002)	14.32% 0.134 (0.002)	29.94% 0.262 (0.003)	31.88% 0.277 (0.006)
Transport	48.82% 0.398 (0.001)	45.59% 0.376 (0.001)	25.27% 0.225 (0.001)	47.08% 0.386 (0.001)	43.94% 0.364 (0.003)
HIE (or HIE_Tr) x Crowded	-1.50% -0.015 (0.002)	-1.38% -0.014 (0.002)	-1.61% -0.016 (0.002)	-1.02% -0.010 (0.002)	-1.58% -0.016 (0.04)
HIE (or HIE_Tr) x Teaching	-9.48% -0.100 (0.002)		-8.73% -0.091 (0.001)	-7.04% -0.073 (0.002)	-17.93% -0.198 (0.004)
HIE (or HIE_Tr) x Charlson	-6.73% -0.070 (0.004)	-6.73% -0.070 (0.004)	-3.90% -0.040 (0.004)	-11.95% -0.127 (0.004)	-14.13% -0.152 (0.008)
HIE HIE (or HIE_Tr) x Transport	-4.27% -0.044 (0.002)	-4.47% -0.046 (0.002)	-3.56% -0.036 (0.002)	-3.10% -0.032 (0.002)	-6.41% -0.066 (0.004)
MU x HIE_Tr					-36.48% -0.454 (0.011)
CCS	NO	NO	YES	NO	NO
Visit Controls	YES	YES	YES	YES	YES
Hospital Controls	YES	NO	YES	YES	YES
Year F.E.	YES	YES	YES	YES	YES
Hospital F.E.	NO	YES	NO	NO	NO
Observations	7,421,302	7,421,302	7,420,025	4,860,829	1,083,086
R^2	0.09	0.09	0.27	0.09	0.09

Note: The first number denotes the estimated percentage change, the second number the corresponding regression coefficient, and the third number the robust standard error. The symbol “x” indicates the presence of the specific controls or the fixed effects in the model while no entry suggests they are absent.

Similar to existing studies (Le and Hsia 2014, Karaca et al. 2012), the regression results for Overall Analysis (Adopters-only Analysis) show that the average LOS per ED visits to teaching hospitals is 26.7% (22.4%) higher than the one for non-teaching hospitals. However, HIE adoption decreases the ED LOS in teaching hospitals by additional 9.5 (7.0) percentage points compared with non-teaching hospitals for an arbitrary patient. These findings corroborate Hypothesis 2 that teaching hospitals can possibly leverage their HIE capabilities more, as compared with non-teaching hospitals.

Patients arriving at a crowded ED on average wait longer, which is expected and consistent with the existing studies (Batt and Terwiesch 2012, Kc 2014). Perhaps more interestingly, we observe that HIE adoption is associated with more reduction in LOS in patients arriving at a crowded setting as compared with a non-crowded setting, specifically an additional 1.5 percentage points relative reduction for an arbitrary patient. The result conflicts with the hypothesized effect of HIEs under crowded settings. However, the small percentage reduction in crowded settings suggests that for an average patient in a crowded ED, HIE still decreases LOS, but the differential effect as compared with non-crowded setting is very small, and is dominated by the main effect due to crowding. Hence, although negative, the small effect of HIE in crowded vs. non-crowded settings hinders us from making a definitive judgment on the moderation analysis for Hypothesis 2.

For clinically complex patients (measured by CCI) and severe patients (proxied by transportation), in line with the findings of the previous studies (Herring et al. 2009, Tanabe et al. 2004), we find that such patients on average stay longer in EDs; yet HIE adoption is associated with a reduction in LOS. In particular, under the Overall (Adopters-only) Analysis, as compared with non-complex patients, clinically complex patients spend on average 20.4% (29.9%) more time and HIE reduces the LOS by an additional 6.7% (12%) for such patients. Similarly, Overall (Adopters-only) Analysis suggests that compared with non-severe patients, severe patients spend substantially more time in the ED, an increase of 48.8% (47.1%), and HIE reduces the LOS on average by an additional 4.3.% (3.1%) for such patients. Therefore, our findings support Hypotheses 4a and 4b.

1.6 Discussion

In this study, we explore how the adoption of HIE affects ED LOS, an important measure of operational performance in EDs. Past research on HIE has been narrowly focused, considering the effect of HIE on reducing radiological imaging and laboratory testing, ED and hospital admissions, and aggregate cost of care (Rahurkar et al. 2015). This limited research attention was possibly due to limited data availability. By using LOS data for the entire state of Massachusetts, we were able to establish a stronger research design and overcome some of the data challenges. In particular, the Massachusetts Division of Health Care Finance and Policy granted a special permission for the sensitive data elements, including exact admission and discharge times from which the ED LOS can be calculated. The data fields are restricted to be used internally by AHRQ, and we were fortunate to utilize the strengths of this dataset in overcoming some of the challenges faced by the prior research. Moreover, the granular nature of our dataset allowed us to study HIE effect on LOS under various hospital and patient related factors.

Our findings indicate that overall, HIE adoption is associated with an improvement in LOS, a somewhat widely believed but empirically not well established result. We further find that the HIE-LOS relationship is subject to sizable moderation effects by several other factors including a hospital's teaching status, crowdedness of the ED at the hour of patient arrival, clinical severity/complexity of the patients, and the type of the index disease/condition. In particular, we find that teaching hospitals observe substantial LOS benefits with the adoption of HIE as compared with non-teaching hospitals, and patients arriving at a crowded ED benefit less from HIE in terms of LOS. For clinically severe and complex patients, we find that HIE adoption is associated with even lower LOS, possibly due to higher need for information coordination for such patients. Our additional analyses suggest that the LOS reductions due to HIE differ with respect to diseases/conditions and the presence of previous history of a visit to an ED with HIE.

We believe our results may help healthcare administrators decide whether or not to adopt HIE. In particular, many healthcare administrators and other stakeholders have questioned the business case for and the sustainability of HIEs (Adler-Milstein et al. 2013). In that regard, our findings based on large-scale data analysis could reliably inform the stakeholders regarding the adoption decisions, especially given the ED throughput being a significant concern for the managers (Helm et al. 2011a, Handel et al. 2010, Pallin et al. 2013). Furthermore, the contextual analysis may further help the managers in tailoring their HIE adoption decisions to the specific settings they operate in and the type of patients that their ED serves.

1.6.1 Moderators of HIE and Length of Stay Relationship

Teaching hospitals have historically been experiencing longer waiting times, partially due to serving more uninsured and Medicaid patients with clinically complex conditions (Grover et al. 2014). Also, teaching hospitals are among the first adopters of HIE and are well positioned to adopt HIE effectively due to their past experience in HIT as explained before. Our analysis suggests that teaching hospitals could further improve their longer LOS (compared with non-teaching hospitals) and mitigate some of the pressure due to the so-called crowdedness epidemic by adopting HIE.

With respect to crowdedness, we observe that LOS decreases due to HIE less. Operations management literature suggests that overloaded clinicians typically spend less time on non-critical activities. And indeed, our findings suggest that in HIE-carrying hospitals, ED physicians are apparently unable to utilize HIE as productively when the ED is crowded. Given that some of the resistance to HIE implementation comes from physicians who suspect that using HIE will disrupt their workflow and make them less productive, especially when hospital is crowded and physicians are under heavy workload (Rudin et al. 2011), we believe our finding confirms these worries and suggests that hospital administrators and HIE providers must focus on making HIE more user friendly and less disruptive.

Our findings regarding patient-related factors suggest that HIE adoption is associated with shorter LOS for clinically severe and complex patients. This is likely because of higher information coordination needs among these patients. Capturing how HIE affects the quality of care is beyond the scope of this study, but based on previous research, we are inclined to believe that the quality also improves. In particular, past research suggested that additional information collected via HIT improves the quality of care more significantly in severe and complex patients. For example, McCullough et al. (2013) studied the impact of EHR in the inpatient setting on different diagnoses and found that the use of EHR improved the quality of care only little in average cases but more significantly in severe cases. A complementary study by Haque (2014) focuses on the impact of EHR on inpatient LOS and finds some decrease in LOS for non-complex patients but no decrease in complex patients. While it is difficult to make a definitive statement on quality implications for the severe or complex patients, we believe that our research is a critical first step in showing the differential effect of HIE on LOS depending on patients' conditions, while highlighting the need for more focused research in this area.

1.6.2 HIE Use

We do not observe in our data the HIE use for individual ED visits, and we acknowledge that the actual rates of access may vary. Rudin et al. (2014) in their recent systematic review on HIE document that the HIE access rates mostly ranged from 2% to 10%, a low rate relative to the need, and highlight that the HIE use may be driven by local context and implementation factors including the patient's condition, past utilization, age, comorbidities, crowdedness, race/ethnicity. To our knowledge, the studies that employed HIE use data were restricted to typically only one exchange network or just a few hospitals (e.g., Overhage et al. 2002, Frisse et al. 2012, Yaraghi et al. 2015, Vest et al. 2014). By considering adoption, our study only captures the overall effect, similarly to other large scale HIE evaluation studies (e.g., Jones et al. 2011, Lammers et al. 2014, Vest and Miller 2011). We also remark that the use relates to the idiosyncrasies of an actual implementation at a micro level whereas we focus on the relation between the HIE adoption and LOS at a high level. Assuming that our results are valid and that HIE adoption indeed decreases LOS for a typical patient, the effect we find would possibly be larger should the use data be available. Indeed, in Systems Analysis, because we consider hospitals that are part of integrated health systems which are more likely to be financially incentivized to use the HIE, we observe higher improvement in LOS consistent with our intuition. (Vest et al. 2013, 2011, Rudin et al. 2011). While we do not explicitly model HIE use, could be in general interested in both the frequency of use and in the effect per use, but our analysis will confound these two quantities, only capturing the overall effect of HIE adoption. However, this is similar to other large scale evaluation studies of HIE that have been conducted to the date (e.g., Jones et al. 2011, Lammers et al. 2014, Vest and Miller 2011). Our research differs from and complement the HIE use studies in the following senses: 1) previous research reports fairly consistent association between HIE adoption and various forms of quality and efficiency gains, 2) access rates vary and they are low relative to the potential need (Rudin et al. 2014), 3) use relates to the idiosyncrasies of an actual implementation at a micro level whereas we focus on the association between the HIE adoption and LOS at a high level, and 4) assuming that our results are valid and that HIE adoption indeed decreases LOS for a typical patient, the effect we find would possibly be larger should the use data be available. By considering adoption, our study only captures the overall effect as similar to other large scale HIE evaluation studies. Since we are unaware of any large-scale dataset with HIE use data available, we see these two types of studies, large-scale on HIE adoption and small-scale on HIE use, as complementary. Finally, we note that a study with the same dataset that would have the HIE use information available would find the main HIE coefficient with the same sign but possibly larger (assuming that our results are valid and that HIE indeed decreases LOS for a typical patient).

1.6.3 Limitations and Future Work

Our study has limitations. First, although our quasi-experimental setup and use of longitudinal data for an entire state provide a strong empirical basis for analysis, and we have done our best to address endogeneity concerns by using appropriate estimation procedures, one may still worry if causality has been fully established. Fixed Effects Analysis addresses endogeneity that would arise due to secular time trends or time-invariant hospital-specific factors correlated with HIE's impact on LOS. Instrumental Variable Analysis with hospital fixed effects further addresses the endogeneity concerns and helps us obtain consistent estimates. Index Analysis uses strong sub-sample design where the proposed relationship should hold strongly. Systems Analysis provides a more uniform set of hospitals in the treatment and control groups. We believe that these estimation methods and subsample-based robustness checks help us alleviate any serious concerns regarding the endogeneity.

Second, our calculated LOS corresponded to the sum of the waiting time and the service time for a patient (Welch et al. 2011). While ideally one would separately consider waiting and service times to identify the HIE effect on each of these components of the total service duration, we were not able to separate out the waiting and the treatment durations because of the way the time logs are coded in our dataset.

Third, while we acknowledge that reduced LOS does not always mean simultaneous improvement in other metrics (e.g., quality metrics), we unfortunately cannot assess the impact of HIE on other metrics concurrently with the LOS due to data limitations. Finally, our definition of crowdedness could be improved by considering the patients who are eventually admitted. Despite this limitation, our volume estimate used in the crowdedness definition is a good proxy because treat-and-release patients comprise 81% of all ED visits (National Center for Health Statistics 2013).

Following Dranove et al. (2014)'s approach we define basic EMR when the hospital has adopted either of the clinical data repository, clinical decision support, or order entry/communication, and advanced EMR when the hospital has adopted one of the more advanced EMR capabilities such as computerized provider order entry or physician documentation. Using these definitions, we found that all hospitals in our sample already had basic EMR, and most of them also already had advanced EMR. Advanced EMR had been adopted by only few hospitals during our study period. Moreover, only two hospitals out of 63 adopted both advanced EMR and HIE during the study period and none of them did so in the same year. Excluding the simultaneous adopters from our sample, and reestimation provides similar insights. Hence, concurrent adoption of EMR technologies does not affect our findings.

Patient Choice of ED

Emergency departments might attract more patients by advertising a short waiting time. This practice seems to emerge around early 2010's along with national reporting initiatives (Emerman 2012, Members of the Emergency Medicine Practice Committee 2012). As we argued in Section 1.4.1 for the instrumental variable, the literature suggests that this was not the case in 2000's and this might not have yet been so common in Massachusetts because of the emergency care is so highly demanded. We could further check that HIE does not drive (directly or indirectly through changing LOS) patient choices of ED, which could in turn influence LOS. To do so, we might consider the following model:

$$\log(\text{Volume}_{h,t}) = \beta_0 + \beta_1 HIE_{h,t} + \beta_2 \log(\text{Volume}_{h,t-1}) + \gamma_1 z_t + \epsilon_{h,t}, \tag{1.1}$$

where $\text{Volume}_{h,t}$ is volume in hospital h in year t, and the rest is similar as in the Overall model. We would test whether $\beta_1 \neq 0$, in which case HIE might influence patient volumes and hence care choices which could be worrisome for our analysis.

Severity and Complexity

Our measures of severity and complexity are indirect and we would like to validate them using alternative measures. For severity, the ideal measure would be the triage score, but this one is not available. For complexity, a good measure would be the effort put into treating a patient. Medicare measures such effort using so-called "relative value units" for physician procedures. However, counting these for the visit in question might cause endogeneity. Therefore, we instead propose the following alternative measure for complexity: We restrict our attention to the patients with multiple ED visits whose latest visit was not preceded by another ED visit in the past month (to exclude frequent ED patients). We focus on their latest visits. For these, we translate their past procedure codes into relative value units and then average these. This average is then included as an alternative measure of complexity. We then compare for this same patient population the coefficient sign for the analysis using this measure with the sign for the analysis using CCI, we expect these will go in the same direction.

HIE Implementation and LOS Variability

We could further support the hypothesis of the effect of HIE by studying the variability of LOS after the HIE implementation. Specifically, we would expect that the variability of LOS will increase if HIE is being used. The variability would be the result of possible multiple scenarios when a patient is treated. In addition to care pathways available before HIE implementation, some patients will have HIE information available, which will lead to a different LOS (often lower and maybe sometimes higher). Hence, suddenly, even for patients that would previously had the same distribution of LOS, there are suddenly more possible distribution, depending on whether and how HIE would be used.

1.7 Conclusions

Our analysis fills an important gap in the literature as it is among the first to study HIE and operations relationship. While previous HIE research has focused predominantly on utilization measures, we are not aware of any prior large-scale study that considered the association between HIE and LOS in the ED setting. As HIE gains critical mass (Adler-Milstein et al. 2013) and richer sources of data become available, we believe that more work needs to be conducted at the intersection of HIT and healthcare operations. Our research may motivate future studies to assess the relationship between HIE and more granular LOS measures or other operational performance metrics.

1.8 Acknowledgments

We would like to acknowledge Massachusetts Division of Health Care Finance and Policy that contributed data to HCUP that used in this study. We also gratefully acknowledge

The Dorenfest Institute for H.I.T. Research and Education at the HIMSS Foundation for providing us access to the historical HIMSS Analytics database records.

This project used data from the Healthcare Cost and Utilization Project (HCUP), which is considered exempt by AHRQ IRB standards.

CHAPTER 2

PHYSICIAN INTEGRATION IN BUNDLED PAYMENTS

The U.S. healthcare system suffers from poor coordination, large inefficiencies, and misaligned incentives. As a result, the healthcare system in the U.S. incurs high costs and often delivers low quality of care compared with many other developed countries (Davis et al. 2014). The U.S. healthcare spending is estimated to range from 3 to 4 trillion dollars (Patel 2016), of which about one third was being wasted (Berwick and Hackbarth 2012). From an efficiency standpoint, new payment models are viewed as important tools in reducing the overall healthcare costs and improving quality. However, while promising in value, many of the new payment models are unproven and the true impact on providers and overall healthcare system is yet to be seen. Therefore, understanding the financial incentives of providers and characterizing conditions that are essential for achieving the aimed objectives of these new payment models is critical (Schoen 2016).

The U.S. healthcare system generally uses separate methods for reimbursing hospitals and physicians for the services they provide. Typically, hospitals receive a fixed payment per visit specific to a standardized grouping of diseases, called the *diagnosis-related group (DRG)-based prospective payment* model. In contrast, physician payments are made on a *fee-for-service (FFS)* basis where physicians are reimbursed separately for every service provided and procedures performed around the visit. In the remainder of this chapter, we will refer to the DRG-based prospective payments for hospitals and FFS-based payments for physicians services together as the FFS model, as widely done in the literature (Mechanic and Altman 2009, McClellan 2011). As a result of this disparate payment setup, incentives of hospitals and physicians are misaligned. The hospital can be financially better off by eliminating services to reduce costs, while the physicians can be financially better off by increasing the number of services. The overuse of medical services is associated with higher costs, while there continues to be a heated debate on whether the resulting quality is better or worse (Fisher et al. 2003, Grady and Redberg 2010). Under such misaligned incentives, physicians and hospitals often fail to coordinate care and subsequently forego opportunities to improve quality and decrease costs (Mehrotra and Hussey 2015).

To better align the incentives, the Center for Medicare and Medicaid Services (CMS), the largest payer of healthcare services in the U.S., introduced *bundled payments*, which aims to combine hospital and physician reimbursements for an entire episode of care into one payment. For example, for an episode of knee surgery, the payer pays a fixed, bundled amount to the hospital (or "convener") to cover all related services and procedures, including tests, treatments, as well as physician fees. Such a "bundle" typically includes fees of the surgeon, the anesthesiologist, the hospital, and the costs of rehabilitation, implants, and other medical devices. The hospital bears the financial risk due to uncertainty in the costs incurred during the episode and is also responsible for coordinating and reimbursing the physicians. In return, the hospital may keep any savings or share these with physicians using *gainsharing*, a mechanism that allows hospitals to induce physicians to make cost-conscious decisions in alignment with hospital's incentives, which is not allowed in FFS-based models (Froimson et al. 2013).

Bundled payments offer some opportunities in improving healthcare services and delivery. The bundling of payments is expected to realign the incentives for hospital and physicians, misalignment of which is currently plaguing the healthcare industry. Engaging physicians in containing hospital costs could mitigate the overuse problem, and therefore help in reducing the excessive healthcare costs. In addition, by eliminating one of the two separate billing systems, one for hospital payment and one for physician payment, the lump sum payments to involved parties could also decrease the high administrative costs, currently accounting for a quarter of all hospital spending (Mehrotra and Hussey 2015).

However, despite their promise of improved efficiency, bundled payments are often resisted by the physicians, which typically deters hospitals from more widely enrolling in bundled payments (Tsai et al. 2015). Physicians resist because they believe bundled payments may encroach on their autonomy, constrain how they practice medicine, and possibly reduce their profits, which is why physicians have been previously excluded from the hospital DRG-based payments in FFS models (Mehrotra and Hussey 2015).

Hospitals differ in their relationships with physicians in influencing care (typically referred to as *level of alignment* or *level of integration* in care coordination), and subsequently in their influence on care intensity. Some hospitals constrain physicians by developing care protocols thus substantially limiting physician autonomy while other hospitals leave most treatment decisions completely to physicians (Burns and Muller 2008). For example, integrated hospital systems and larger hospitals typically have more standardized care protocols as well as more aligned physicians (Bloom et al. 2013). On the other hand, stand-alone nonprofit hospitals are *de facto* physician-controlled, and therefore the management in such hospitals may find it difficult to restrain physicians' excessive treatment choices (Pauly and Redisch 1973, Sloan 2000). The hospital's capability to influence physician's care intensity, which we henceforth refer to as *physician alignment*, determines the cost and quality of care. While bundled payments are expected to promote cost reduction, it remains unclear whether they are suitable for all hospitals or only for certain types of hospitals in the spectrum of alignment levels.

Because the interest in bundled payments has reemerged only recently with a perceived commitment from the CMS for implementing them, there has been an increasing interest in the subject by the Operations Management (OM) researchers. Gupta and Mehrotra (2014) take the payer's perspective and examine how bundled payment contracts that providers propose should be selected by CMS, the major bundled payments contractor. On the other hand, Adida et al. (2016) consider how contending healthcare payment models (including bundled payments) appear from the perspective of a single integrated risk-averse provider and how the models impact patient selection, intensity of care, and the system payoff. A paper by Andritsos and Tang (2018) compares how readmissions occur under fee-for-service, pay for performance, and bundled payments. Finally, Han et al. (2017) consider the strategic interaction of hospitals when determining care quality under bundled payments. While we also consider bundled payments, our focus is very different than the above mentioned studies. In particular, we study the interaction between the hospital and physician and focus on the implications of this interaction on bundling decisions and corresponding outcomes. Considering the role that physicians play in determining the financial bottom line of hospitals and the idiosyncrasies of hospital-physician relations that we discussed earlier, our work makes the first attempt in healthcare operations literature to study the interdependency between hospitals and physicians, the resulting trade-offs, and the obstacles for setting up a bundled payment model for the payers.

Our findings suggest the following. First, in regards to who will benefit and who will lose in a bundled payment environment, we confirm the experts' expectation that hospitals with very loosely aligned physicians would not benefit from bundling and be better off under FFS. However, somewhat unexpectedly, we also find that the hospitals with highly-aligned physicians in general are less likely to benefit as well; and that those hospitals which lie in between these two cases in the spectrum of alignment levels will benefit the most (Theorem 2.2.1 and Proposition 2.2.2). Second, in regards to quality implications of bundling, we characterize hospital care contexts where the quality will deteriorate under bundled payments (Proposition 2.2.3). We demonstrate how the payer can employ quality constraints to prevent the quality from decreasing excessively and, we find that FFS can sometimes be the better choice when the payer worries about quality (Theorem 2.2.2). Further, we show that quality initiatives that motivate hospitals to safeguard quality, such as the ongoing Hospital Readmissions Reduction Program, might actually demotivate hospitals from adopting bundled payments (Corollary 2.4.1). Finally, we extend our model to capture a setting where physicians are hospital employees (salaried physicians), as opposed to being paid for care services independently from hospitals—a specific case of highly aligned physicians. We find that initiatives that further hospitals' accountability for care quality may dampen the incentives for bundling in hospitals with independent physicians, whereas they are likely to enhance incentives for bundling in hospitals with salaried physicians.(Proposition 2.4.2 vs. Corollary 2.4.1).

The rest of this chapter is organized as follows. In Section 2.1, we introduce some key concepts that will be present throughout the paper and also review the relevant literature. In Section 2.2, we introduce and analyze our Initial model. In Section 2.3, we discuss a "Coproduction" model suggested by one of our reviewers. The Coproduction model includes quality awareness from both the hospital and the physicians, and, in contrast to the Initial model, it includes "observable" transfer payments, not dependent on unobserved efforts. Next, we extend the Initial model in Section 2.4 to consider a Quality-aware model, a model with salaried physicians ("Salary model"), a model with explicit modeling for risk aversion, and finally a Physician-driven model where physicians control care. In Section 2.5, using real data, we present a data-driven approach to construct clinical pathways, a key concept in our analysis. Finally, we summarize and conclude in Section 2.6.

2.1 Background

In this section, we provide background information, summarize some of the key concepts related to bundled payments, and review related literature.

2.1.1 Current Bundling Initiatives by the CMS

Currently, there are several BP initiatives, which differ mainly with respect to how they are implemented. The oldest of these initiatives, the "Bundled Payments for Care Improvement" (BPCI) program, is the largest one among the existing BP initiatives and is based on voluntary participation. More specifically, the BPCI allows providers to apply voluntarily for bundling in any or all of 48 designated conditions (DRGs, such as major joint replacement of the lower extremity, acute myocardial infarction, congestive heart failure, or simple pneumonia and respiratory infections). The BPCI program involves four sub-models: BPCI Models 1–4 vary primarily with respect to being prospective or retrospective and inclusion or exclusion of post-acute care in addition to acute care services. Under retrospective models, providers operate financially in the same way as in FFS, but the budgets are reconciled against a virtual BP setup after each accounting period; while under prospective models, providers are paid directly through a bundled payment, not just virtually through a reconciliation process. Another relatively less common and mandatory BP program called "Comprehensive Care for Joint Replacement (CCJR)" builds upon BPCI Model 2 and only applies to joint replacement DRGs in select metropolitan areas. More recently, the CMS has introduced BPCI Advanced, a program mostly similar to BPCI with some minor differences. Most notably, in BPCI Advanced, program requirements have been further simplified with the elimination of alternative options, i.e. unlike regular BPCI, BPCI Advanced does not have sub-models. As of writing this manuscript, the CMS continues to actively experiment with alternative models with variations in implementation details, and it is possible to see new models forthcoming in the near future. However, despite all their differences in implementation details, it would be fair to say that all of these BP models share some common features that make it possible to analyze BPs at a broader level from an incentive alignment perspective. In the following, we summarize and discuss these general features that are critical from an incentive alignment perspective.

2.1.2 Common Features of Bundling Initiatives

The overarching feature of the existing bundled payment models is the move to a system where the payer provides a single fixed payment for an episode of care around a DRG (e.g., knee surgery); in the status quo, hospitals are paid a fixed amount on a prospective basis (i.e., DRG) while physicians are paid based on the intensity of care provided (i.e., all the services provided). Under BP, a "convener" such as the hospital is responsible for distributing the payments to involved parties. Typically, the convener continues to pay the physicians based on delivered intensity. However, unlike FFS, physicians may be paid extra in BP based on savings achieved, a mechanism called *gainsharing*, which is disallowed under the status quo. Indeed, gainsharing is a common and critical feature in all bundled payment initiatives, and hence, we further describe this mechanism in greater detail next.

Gainsharing has long been promoted by experts for aligning hospitals and physicians (Wilensky et al. 2007, Grandusky and Kronenberg 2006). Through gainsharing, hospitals reward physicians for the savings realized when physicians use more standardized supplies, choose cheaper devices when appropriate, and, more generally, help reduce unnecessary utilization. In contrast, under the standard FFS model, the physicians often lack incentives to decrease the utilization; indeed, a higher utilization means larger financial gains for them. Despite its promise, both the CMS and related legislation such as the Stark Laws and the Civil Monetary Penalties Law have prohibited or discouraged gainsharing in the past because of antitrust concerns. Recently, the CMS authorized limited use of gainsharing (limited for example by capping the payments and placing constraints on quality) in demonstration initiatives for BP (Froimson et al. 2013).

While there are many different ways to implement gainsharing between the hospital and physicians, these contracts are unfortunately often proprietary. However, the publicly available records from BPCI participants suggest that gainsharing is almost universally prevalent when physicians are involved (Dummit et al. 2015) and that the gainsharing mechanisms were setup to reward physicians for realized efficiencies with respect to care intensity. The conceptualization of gainsharing in our models was chosen to reflect these realities.

2.1.3 Motivating Example: Bundling Coronary Artery Bypass Grafting at the Maine Heart Center

We present an example that motivated our models, based on the presentation by the Maine Heart Center and their coronary artery bypass grafting (CABG) bundling (Cutler and Seekins 2014). Their bundled payments are based on the BPCI Model 2, which covers the acute hospital stay and post-acute care. Next, we describe some of the salient features in their bundled payment program. They describe four care pathways that encompaas different modes of post-acute care: 1) home health agency, 2) skilled nursing facility (SNF), 3) home (no further care), or 4) inpatient rehabilitation. Most of their CABG episodes costs are evenly distributed between \$20000 and \$50000, but several ”outlier” episodes involving a readmission or prolonged inpatient rehabiliation incur much higher costs. The BPCI contract is with CMS, so it outlines certain minimum quality requirements. The CMS also requires a 2% discount compared to the pre-bundled payments FFS period.

They also describe how they redesigned their care, including changes to care pathways. First, they standardized and streamlined their home health and SNF pathways, which reduced costs. Next, they focused on reducing readmissions, to improve quality and also decrease costs. Finally, they critically inspected the home health pathway and moved patients who did not needed to the home pathway. Notably, since BPCI Model 2 assumes ”retrospective bundled payment”, the center did not have to make substantial infrastructure or IT investments.

2.1.4 Retrospective vs. Prospective Bundling

To date, bundled contracts have been mostly *retrospective*, that is, all providers continue to receive individual FFS payments based on their standard reimbursement rates. Under this model, at the end of the year, CMS (payer) compares the total reimbursed amount with the pre-established, discounted bundle price (benchmark). If the providers succeeded in care redesign, and the amount spent is less than the benchmark, CMS pays the difference to the providers. Otherwise, the providers have to repay CMS. Since the retrospective payments allow providers to experiment with bundling without substantially changing their existing operations, retrospectively paying the FFS rate under bundled payments has been the default, and by far the most common among the nt bundled payment models (CMS 2014, Dummit et al. 2015). As such, we consider and focus on retrospective models in this study. On the other hand, under a prospective bundling model, the hospital receives a single, lump sum payment from CMS and then distributes the payment among all the providers involved in the episode of care. Clearly, implementing a prospective bundling model is more costly to set up; therefore compared with the retrospective models, prospective models have been less common so far.

2.1.5 Key Drivers

Healthcare administrators, experts, and analysts suggest several key drivers affecting provider's decision to engage in bundled payments for a given care episode (i.e. an ”extended DRG”). While we do not model all of them, we list them here for future reference and research. An oft-cited deterrent of bundling for hospitals and physicians are the initial investment costs. These include new information and accounting systems, advancements in cost measurement, and defining care episodes. However, retrospective bundled payments do not require much of these and be typically set up without large upfront costs. Additionally, hospitals also face a require 3-3.5% discount to CMS compared to the current practice.

Despite these drawbacks, some hospitals still seek out bundled payment opportunities. First, they see it as an opportunity to learn for the future of healthcare, the future with new payment models. Even CMS notes that "competencies learned in bundled payment position physicians for success in value-based contracting," and that bundled payments provide an "opportunity to work and learn from others nationally and receive data." Indeed, CMS offers the hospital claims information that they could not access otherwise. Furthermore, bundled payments need not be sudden: CMS offers a trial period ("Phase 1") when it shares claims with providers, so that they can learn the true cost of their operations, but they do not suffer from any cost implications if they exceed the budget.

However, the hospitals that eventually decide to participate in the bundled payments are challenged in other ways. First, in contrast to fee-for-service, the hospital actually bears the consequences of cost variation throughout the episode, having to hedge against outlier patients. In fact, American Hospital Association (2013) suggest that "episode types should be selected that have enough variation to provide opportunities for cost reduction, but not so much variation as to pose excessive risk to the organization." Related to the cost variation is the question of patient volume: the higher the volume, the lower variation in the mean hospital profit. Vice versa, small hospitals do not hazard entering bundled payment contracts because they face formidable episode cost variation. Furthermore, hospitals need high patient volume is important to amortize the startup and administrative costs of developing and implementing the bundles (Ridgely et al. 2014). To further motivate hospitals to bundle, some propose that hospitals should be rewarded by increased patient volumes (i.e., market share), either from payers or patients (if payers motivate them, for example by a promise of lower copays or better quality). In addition to cost variation, hospitals must be wary of quality implications. In fact, the contract with CMS stipulates minimal requirements on bundle quality. For instance, some participants in BPCI Models 2 and 3 are required to track B-CARE measures and, if they choose so, other measures as well.

For hospitals and physicians alike, bundling may also interact with other pay-for-performance initiatives. For instance, for readmissions (the HRRP program), high readmission rates for certain conditions lead to cuts in CMS reimbursements for all conditions. Furthermore, hospitals may worry about their reputation tarnished by being on the list of penalized hospitals. Notably, all the conditions currently subject to HRRP are high-volume, so potentially good bundling targets. As another example of a pay-for-performance initiative, the value-based payment (VBP) program monitors several different quality measures across all conditions and again, the adjustments apply to all CMS reimbursements.

Finally but very importantly, bundled payments impose on hospitals a coordination problem (Ridgely et al. 2014). Specifically, on the provider side, bundled payments require involvement of several different providers, including hospitals and physicians. Notably, about a half of the BPCI Model 4 initiatives are centered around a convening organization other than a hospital. And in fact, according to Hussey et al. (2012), most bundled-payment initiatives so far faced a resistance from providers.

2.1.6 Relevant Literature

Here, we describe the relevant literature, namely OM literature, health economics literature on care coordination among providers and medical and health policy literature on the expectations and organizational perspectives about bundled payments.

Payment Models in Operations Literature

Recent Operations Management (OM) literature explored various aspects of payment models in the healthcare context. Ata et al. (2013) investigate how existing CMS policies for hospice reimbursement tie to providers' patient selection decisions, and they propose an improvement over the current policy. Several studies investigate the role of performance-based payment models in various contexts. Taking a social planner's perspective, Lee and Zenios (2012) find the socially optimal decisions in regards to implementing payment policies based on process compliance vs. outcomes for the End-Stage Renal Disease (ESRD) patients. In a principal-agent framework, Jiang et al. (2012) propose a penalty-based contract for coordinating providers' capacity allocation decisions to ensure timely access to outpatient services while Zhang et al. (2016) study readmission penalties implemented under the Hospital Readmissions Reduction Program and show the unintended consequences of benchmarking when setting the penalty thresholds.

Because the interest in bundled payments has reemerged only recently with a perceived commitment from the CMS to implement them, there has been an increasing interest in the subject by the OM researchers. Gupta and Mehrotra (2014) take the payer's perspective and examine how bundled payment contracts that providers propose should be selected by CMS, the major bundled payments contractor. On the other hand, Adida et al. (2016) consider how contending healthcare payment models (including bundled payments) appear from the perspective of a single integrated risk-averse provider and how the models impact patient selection, intensity of care, and the system payoff. A working paper by Andritsos and Tang (2018) compares how readmissions occur under fee-for-service, pay for performance, and bundled payments. Finally, Han et al. (2017) consider the strategic interaction of hospitals when determining care quality under bundled payments. While we also consider bundled payments, our focus is very different than the above mentioned studies. In particular, we study the interaction between the hospital and physicians and focus on the implications of the level of integration on bundling decisions and corresponding outcomes. Considering the role that physicians play in determining the financial outcomes of hospitals and the idiosyncrasies of hospital-physician relations that we discussed earlier, our work makes the first attempt in healthcare operations literature to study the interdependence between hospitals and physicians, the resulting trade-offs, and the obstacles for setting up a bundled payment model for the payers.

Health Economics Literature

In the health economics literature, several papers study care coordination among providers. While these studies are relevant to ours from a modeling perspective, our work is distinct as we explore when bundling occurs and what is needed for care coordination, and discover how the hospital and the physicians are likely to react to the most recent bundled payment initiatives. Harris (1977) provides a modeling framework to study incentive relationships and interactions between hospital management and physicians. Ma (1994) wrote a seminal paper that theoretically models care delivery coordination by hospitals and physicians, which also partially motivates our model setup. Crainich et al. (2008) extend this work by incorporating features from international health systems. Custer et al. (1990) propose a model of how physicians react to the Prospective Payment System (implemented in the 80s), and how this affects the hospital under several different modes of hospital-physician coordination. Boadway et al. (2004) model the two-way contracts among doctors, hospitals, and a social planner where the purpose of the contracts is to address the inefficiencies due to information asymmetry around patient severity. Huang and McCarthy (2015) explore how coordination between the hospital and physicians changes as the insurance market shifts.

While relevant from a modeling perspective, all of these studies are concerned with prior payment mechanisms and are not directly applicable to current bundled payment models. In this regard, two studies published in the 1990s and 2000s in the wake of earlier bundled payment discussions are relevant to our work. The first one by Jelovac and Macho-Stadler (2002) focuses on the payer's problem in terms of the contract design with the hospital and the physicians. The second one, a short note by Dor and Watson (1995), examines the coordinating split of a bundled fee for an efficient outcome. The authors conclude that there is a clear need for investigating i) the role of varying hospital-physician alignment in bundling decisions, ii) the efficiency implications of quality contracts when offered alongside bundling, which, in some sense, motivate our work.

Medical and Health Policy Literature

Bundled payments as an alternative to FFS have for long interested health-policy researchers. While most of this literature is qualitative, we summarize the relevant findings concerning bundled payment models and physician-hospital relationships.

Bundled Payments and Expectations Bundled payments are expected to better control care intensity (utilization), encourage high quality, promote provider coordination and integration but can be readily implemented (Mechanic and Altman 2009). These expectations are largely speculative. For instance, Hussey et al. (2012) conduct a thorough literature review and find that almost all existing work were observational or descriptive. The study reports a consistent drop in intensity with ambiguous quality impact among the reviewed articles, however, the body of evidence was rated as low due to concerns about bias, confounding, a lack of design and contextual factors.

Another stream of research addresses questions around how to form bundles. For instance, Dobson et al. (2012) analyze historical Medicare claims data and provide widely-cited guidelines on how to define, price, and manage a bundle. Sood et al. (2011) review the existing evidence and focus on which conditions to choose for bundling and how to choose the bundle episode length. Ridgely et al. (2014) describe an unsuccessful bundled payments initiative, Prometheus in California, to inform future bundled payments initiatives. Finally, many studies from physician communities discuss how bundled payments will affect the physicians and how physicians should react (e.g., Bozic et al. 2014, Shih et al. 2015, Mukherji and Fockler 2014, Brill et al. 2014).

Bundled Payments and Organizational Perspectives Health policy studies also explore how new payment models relate to physician-hospital relationships. The new payment models are expected to further encourage ongoing provider integration (Mechanic and Altman 2009). Gaynor and Town (2012) review the effects of such consolidation under the employment model, while Casalino et al. (2008) and Berenson et al. (2007) highlight the rise of physician-owned facilities, particularly ambulatory surgery centers and physician-owned hospitals that directly compete with traditional hospitals. Friedberg et al. (2015) survey physician practices about how they perceive and worry about new payment models. They find that the new payment models may induce some physicians to more closely collaborate with hospitals, force them to face higher expectations, but possibly also provide them with opportunities to improve care quality.

Bundling in other industries The concept of “bundling” exists in industries outside of healthcare, but the concept has a somewhat different meaning. In healthcare, bundling embodies essentially two ideas: 1) price transparency and 2) a healthcare procedure as a well-defined product (e.g., a surgery with all related services included and with a warranty). It is uncommon for patients to want to buy the components of the bundle separately (e.g., spend three days in a nursing home without having the surgery first). In contrast, traditionally, bundling refers to deciding how to combine and sell several independent products as a larger “product bundle” (Venkatesh and Mahajan 2009), often with the purpose of charging customers more. However, the components of the bundle are in principal independent, and customers might legitimately want to buy them separately. Interestingly, there is a setting in healthcare that bundles in this more traditional sense, so-called group purchasing.

2.2 Initial Model

As discussed earlier, under the prevailing FFS model, physician and hospital incentives do not align well, which leads to inefficiencies in the care delivery. The potential success of bundling, on the other hand, is argued to depend on the cooperation between the hospital and physicians in cost reduction and quality improvement efforts (Mechanic and Altman 2009). Our Initial model studies the incentive alignment problem between the hospital and physicians, and analyzes, in the spectrum of level of alignment/integration, when bundling becomes attractive for both parties. The model captures main dynamics of the coordination problem by considering index admissions[1] in hospitals with non-salaried physicians, where the hospital is profit-driven and quality of care is predominantly determined by physician efforts. Later, we extend the Initial model to capture cases where a) the hospital is also cognizant of quality in addition to being profit-driven and the readmissions are also accounted for (Section 2.4.1), and b) physicians are salaried employees of the hospital (Section 2.4.2).

Without loss of generality, we build our model around payments made for a single medical condition such as knee-replacement, characterized by a DRG categorization. For example, a patient with “major joint replacement or reattachment of lower extremity without major complications” will be assigned DRG 470 and another patient with “simple pneumonia and pleurisy without complications” will be assigned DRG 195 for billing purposes. Under the FFS model, once the care is complete, the hospital will receive predetermined amounts for the respective DRGs regardless of the costs for treating the patient; whereas physician is paid separately for each service he provides. Bundled payments, on the other hand, brings the hospitals and phsycian payments together, by paying a single lump-sum

[1] Index admission corresponds to the initial admission in a care episode.

amount for a given DRG (e.g., 470-knee replacement), which is then shared between the hospital and physician.

Table 2.1 summarizes the variables that appear in the Initial model. Next, we describe the key model components.

Table 2.1: List of symbols in the Initial model

$r_{1,p},\ r_{2,p}$	Physician reimbursement for pathway 1, 2.
$c_1,\ c_2$	Per-patient costs for inpatient, acute pathway 1, 2.
Δc	Cost differential for hospital $\Delta c = c_1 - c_2$.
Δr_p	Revenue (reimbursement) differential for physicians: $\Delta r_p = r_{1,p} - r_{2,p}$.
$I = I(i_h, i_p) \in [0, 1]$	The intensity of care, capturing fraction of patients assigned to the more expensive pathway.
I_0	Optimal value of I that maximizes care quality outcomes.
$i_h \in [0, 1]$	Hospital influencing effort (toward the cheaper pathway).
$i_p \in [0, 1]$	Physician influencing effort (toward the costlier pathway); between 0 and 1.
$\Psi \in (0, 1)$	Physician alignment coefficient.
T	Gainsharing amount (relative to physician's effort).
r_h	Hospital reimbursement under FFS.
$F_h^{\text{FFS}}, F_p^{\text{FFS}}, F_h^{\text{BP}}, F_p^{\text{BP}}$	Per-patient objective functions for the hospital/physicians under FFS/BP.
w_b	Physician benevolence coefficient.
$x^*, x^\sharp$	A solution feature under FFS, respectively bundled payments (x can be any symbol).
r^{FFS}	Total payment from the payer under FFS.
r^{BP}	Total payment from the payer under bundled payments.
R_0	Baseline readmission rate.
c'_1, c'_2, c_Ω	Per-patient costs for post-acute pathway 1, 2, and for a readmission respectively.
ι_p, ι_h	Physicians', respectively the hospital's best response functions.

A key feature in our analysis is the concept of a clinical *pathway*, which represents the set of medical procedures a patient in a given DRG category follows, including diagnostic tests, medications, and consultations, conducted during care delivery (De Bleser et al. 2006). In practice, when providers treat a condition, they often vary in terms of the clinical pathways chosen, resulting in different ranges of costs and health outcomes. Under FFS, because each service is billed for separately, hospitals and physicians tend to not worry much about identifying and coordinating on a common clinical pathway. In contrast, because cost accounting is a bigger concern under bundled payments, hospitals and physicians need to better understand their common clinical pathways and identify the most cost-effective ones.

For any given condition, without loss of generality, we consider two pathways, one being more intensive (and costly), and the other one being less intensive (and cheaper). For the ease of interpretation, we assume that the costlier pathway is a superset of the cheaper pathway, with more procedures performed. The corresponding costs to the hospital for these pathways are respectively captured by c_1, and c_2, where $c_1 > c_2$, and the corresponding reimbursements to the physician are respectively captured by $r_{1,p}$, and $r_{2,p}$, where $r_{1,p} > r_{2,p}$. We denote the payments to the hospital under FFS using r_h. Lastly, we remark that physician costs are not directly related to care intensity (Weiss 2003), and hence are not included in the analysis.

For each DRG, there is a level of care intensity, $I_0 \in [0, 1]$ that corresponds to the "best" clinical outcome from the "patient perspective" without cost concerns. In what follows, we refer to I_0 simply as "quality-maximizing" intensity. I_0 then captures the fraction of patients that should be assigned to the more expensive pathway when financial considerations are put aside, and the only objective is to maximize outcomes from patients' perspective (e.g., $I_0 = 0.65$ for a specific DRG implies when the objective is to maximize care quality, 65% of patients should follow the more intense and costlier pathway). Typically, optimal patient outcomes are achieved at an interior point of the support $[0, 1]$ for care intensity, because both overtreatment and undertreatment lead to suboptimal patient outcomes (Fisher et al. 2003, Doyle et al. 2015). The practiced level of care intensity, I, however often deviates from the optimal level of care intensity. In particular, given that the hospital is paid a fixed DRG-based rate under the FFS, it may try to reduce the cost by influencing physician practices through various means such as care protocols to reduce the intensity.[2] In contrast to the hospitals, the physician aims to increase the intensity in order to provide appropriate quality of care and sometimes to also increase their profit margin. As such, similar to the established literature in health economics and policy (Dor and Watson 1995, Jelovac and Macho-Stadler 2002, Crainich et al. 2008), we model the practiced *care intensity*, I, to be jointly produced by the hospital and physicians as follows:

$$I(i_h, i_p) := (1 - i_h)\Psi + (1 - \Psi)i_p, \tag{2.1}$$

[2] Such hospital-driven intensity reduction was observed during the transition to the DRG-based system in the 80s from the cost-based reimbursement, where hospitals were paid based on average cost of a patient per-diem and the length of stay. At that time, hospitals were able to reduce the average length of stay for non-surgical patients from 9.4 days to 7.2 days within only a few years (Altman 2012).

where intensity I represents the practiced level of care intensity and captures the fraction of patients assigned to the more expensive pathway, $i_h \in [0,1]$ represents the physician influencing effort, $i_p \in [0,1]$ is the hospital influencing effort, and $\Psi \in (0,1)$ is the *physician alignment* coefficient. Physician alignment (also referred to as physician integration) is a well-established and widely studied concept in medical and health economics literature, and is defined as the degree to which physicians share the same mission, vision, and objectives with their hospital systems and work toward their success (Shortell et al. 2001, Ma 1994, Huang and McCarthy 2015). Empirical research has found that physicians within large group practices, those physicians that receive a stipend, and older physicians[3] in general have higher alignment level with the hospital (Shortell et al. 2001).

2.2.1 Fee-for-service (FFS) Payment Model and Analysis

In the Initial model, we assume the hospital's objective is to reduce costs and increase its profits while the physicians weigh both their monetary benefits and the patient's interest as captured by the quality of care. This dichotomy in physician's behavior between monetary benefits and some measure of benevolence, altruism, or professionalism is well established and is used widely in the health economics literature (e.g., Ellis and McGuire 1986), and is captured by the physician benevolence coefficient w_b in our analysis. We model FFS case, the status quo, as a Nash equilibrium of a single-stage game where hospital and physician simultaneously choose their effort levels. The utility functions for the hospital and physician are characterized as:

$$\begin{aligned} F_h^{\text{FFS}} &= r_h - c_1 I(i_h, i_p) - c_2(1 - I(i_h, i_p)) \\ F_p^{\text{FFS}} &= -w_b(I(i_h, i_p) - I_0)^2 + r_{1,p} I(i_h, i_p) + r_{2,p}(1 - I(i_h, i_p)). \end{aligned} \tag{2.2}$$

[3] The authors of the conducted empirical work argue that this is perhaps due to lower level of competitiveness and fewer available alternatives older physicians have, given that they are at the later stages of their career.

We remark that F_h^{FFS} and F_p^{FFS} respectively represent the average per-patient utility for the hospital and physician. Under FFS, the hospital receives a single DRG-based payment, r_h, and incurs costs due to patients receiving care via either the costly or cheaper pathway. In contrast, the physician receives pathway-dependent payments and incurs disutility due to any deviation from the optimal care intensity. The following lemma characterizes the physician preference if the care intensity is solely determined by her.

Lemma 2.2.1. *Under FFS, the physician's utility function is maximized when the level of care intensity I is equal to $I_0 + \frac{\Delta r_p}{2w_b}$.*

The above lemma shows that, under FFS, the physician prefers to increase the care intensity beyond the optimal care intensity, I_0, by the benevolence-adjusted financial motives, $\frac{\Delta r_p}{2w_b}$, when the hospital has no influence on the care intensity. As expected, the financial motives, measured by the physician payment/revenue differential between the costly and cheaper pathways, Δr_p, increases the extent of deviation while the benevolence factor, w_b, decreases the extent of deviation from the optimal care intensity. Next, we present an intuitive but helpful result, which states that the optimal intensity under FFS (i.e., status quo) is inversely related to the physician alignment. That is, the higher the alignment, the lower the intensity (i.e., more patients will follow the cheaper pathway). Let I^* denote the equilibrium intensity under FFS, then the following lemma provides an upper bound for I^*.

Lemma 2.2.2. *The equilibrium intensity under FFS is bounded from above by $1 - \Psi$, that is*

$$I^* \leq 1 - \Psi. \tag{2.3}$$

Lemma 2.2.1 and 2.2.2 together imply that, under FFS, physicians would be inclined to set the optimal intensity level to $I_0 + \frac{\Delta r_p}{2w_b}$, whenever they can; however, when physicians are highly aligned with the hospital, the level of alignment dominates the physicians' financial incentives to set the intensity at the benevolence-adjusted level, hence the equilibrium intensity would be bounded by $1 - \Psi$.

In the following result, we fully characterize the optimal intensity under FFS as a function of the physician alignment level, Ψ. In particular, we show that in hospitals with non-salaried physicians, where physician alignment is lower than a certain threshold, $\bar{\Psi}$, the equilibrium intensity under FFS is driven by the physicians only and is set to the maximizing value of the physician utility function, $I_0 + \frac{\Delta r_p}{2w_b}$. On the other hand, in hospitals where physician alignment is higher than this threshold $\bar{\Psi}$, physician utility maximizing intensity, $I_0 + \frac{\Delta r_p}{2w_b}$, becomes larger than the maximum intensity level, $1 - \Psi$ (by Lemma 2.2.2), and hence the intensity level is set to its maximum value, $1 - \Psi$. We define this threshold value $\bar{\Psi}$ in the lemma below, and throughout the remainder of this paper, we refer to those hospitals with $\Psi > \bar{\Psi}$ ($\Psi < \bar{\Psi}$) as hospitals with high (low) physician alignment.

Lemma 2.2.3 (Status-quo intensity). *The equilibrium intensity under FFS, I^*, is given by*

$$I^* = \begin{cases} I_0 + \frac{\Delta r_p}{2w_b} & \textit{if } \Psi \leq \bar{\Psi} \\ 1 - \Psi & \textit{otherwise} \end{cases}$$

where

$$\bar{\Psi} := 1 - I_0 - \frac{\Delta r_p}{2w_b} \tag{2.4}$$

2.2.2 Bundled Payment Model and Analysis

In line with the commonly practiced retrospective bundled payments models, we assume that the hospital is initially operating under FFS. The hospital is then offered bundled payments, under which the hospital is allowed to reward the physicians for cooperating with the hospital for cost reductions through a gainsharing contract. In studying the equilibrium under the bundled setting, we consider a two-stage game, where the hospital first offers a gainsharing contract to the physicians and sets its effort, and then, in response, the physicians choose their effort. We derive the subgame perfect Nash equilibrium using backward induction.

While under FFS, the hospital and physicians are paid separately, under bundled payments, the hospital (or its parent health system) is responsible as the "convener" to receive and distribute the payment. The total payment under BP, which is paid to the hospital and is shared with physicians, is captured by r^{BP}. We note that r^{BP} accounts for all the costs from the services included in the bundle, in addition to hospital payment. In line with the retrospective bundled payment practice, we assume that the hospital reimburses the physicians at the FFS rates, but the hospital can now also pay an additional *gainsharing* amount that is proportional to the physicians' influencing effort toward the cheaper pathway, i.e, $(1-i_p)T$, to the physicians to incentivize them to reduce costs by reducing intensity. Hence, the decision variables are i_h and T (gainsharing) for the hospital and i_p for the physicians. Then, the hospital's and physician's utility functions under the bundled payments become:

$$\begin{aligned} F_h^{\text{BP}} &= r^{\text{BP}} - (c_1 + r_{1,p})I(i_h, i_p) - (c_2 + r_{2,p})(1 - I(i_h, i_p)) - (1 - i_p)T \\ F_p^{\text{BP}} &= -w_b(I(i_h, i_p) - I_0)^2 + r_{1,p}I(i_h, i_p) + r_{2,p}(1 - I(i_h, i_p)) + (1 - i_p)T. \end{aligned} \tag{2.5}$$

Bundling will occur if each of the stakeholders—the physicians, the hospital, and the payer—has higher payoffs, compared with FFS. Clearly, in order for the hospital not to lose compared with the FFS, the total payments to the physicians under bundled payments should not be too high, and the overall payment to the hospital from the payer under bundled payments, r^{BP}, should not be too small. This means that in addition to r^{BP} being less than r^{FFS} (the total payment from the payer under FFS, including hospital and physician reimbursements), the difference should not be too high in absolute terms. As we study the effect of hospital-physician alignment on bundled payments in this paper, we state this payer-relevant condition[4] as an assumption below and characterize the hospital and

[4] Note that if this condition is not satisfied, that is, if the payment by the paper under BP is too small, then,

physician-related conditions for bundling in Theorem 2.2.1.

Assumption 2.2.1. $r^{FFS} - r^{BP}$ *is not too large; specifically*

$$0 < r^{FFS} - r^{BP} \leq \begin{cases} \begin{aligned} &(\Delta c + \Delta r_p)(1 - I_0 - \Psi) - \frac{1}{2w_b} \cdot \Bigg[\frac{T^2}{(1-\Psi)^2} \\ &\qquad - \frac{T}{(1-\Psi)}(\Delta c + 2\Delta r_p - 2w_b(1 - I_0 - \Psi)) + \Delta r_p(\Delta c + \Delta r_p) \Bigg], \end{aligned} \\ \textit{if } \Psi \geq \bar{\Psi} \\ \\ \dfrac{\Delta r_p^2}{2w_b} - \dfrac{1}{2w_b} \cdot \left[\dfrac{T^2}{(1-\Psi)^2} - \dfrac{T}{(1-\Psi)}(\Delta c + 2\Delta r_p - 2w_b(1 - I_0 - \Psi)) + \Delta r_p^2 \right], \\ \textit{if } \Psi \leq \bar{\Psi} \end{cases},$$

where T is quantified later in Proposition 2.2.1, Δc is the cost differential, and Δr_p is the revenue differential between the pathways.

The following theorem characterizes the incentives for bundling in the context of physicians' alignment with hospital.

Theorem 2.2.1 (When do they bundle?)**.** *Suppose the condition in Assumption 2.2.1 holds. Then, hospitals and physicians will bundle if and only if:*

$$|\Psi - \bar{\Psi}| < \frac{\Delta c + \Delta r_p}{2w_b} \tag{2.6}$$

Theorem 2.2.1 has several important implications. First, we observe that the following factors are critical for bundling to occur: i) hospital influence relative to physician in determining the care intensity (physician alignment), ii) the financial incentives for each party as determined by cost and revenue differentials, iii) optimal care intensity associated with the disease condition (I_0 as captured within $\bar{\Psi}$) relative to physician alignment ($\Psi - \bar{\Psi}$) and iv) physician's level of care for quality (as measured by the benevolence

clearly bundling will not be feasible.

factor w_b). Because bundling is a joint decision under conflicting incentives, it is intuitive that the relative influence matters in bundling decisions. The incentives of physicians and hospitals respectively manifest in the cost and revenue differentials, and therefore, the chance of bundling increases with increasing differentials. In some sense, the cost and revenue differentials represent the cost saving and revenue enhancement opportunities from hospital's and physicians' perspectives, respectively. Second, we observe that when the physician alignment level and the care intensity threshold ($\bar{\Psi}$) are within a range defined by the cost and revenue differentials, hospital and physician will bundle. Given the care intensity threshold, bundling will not occur in hospitals with too high physician alignment or with too low physician alignment. An intuitive explanation for this interesting finding is as follows. We know from Lemma 2.2.3 that as the physician alignment level increases, level of care intensity under FFS decreases. Thus, in hospitals where physician alignment is high, most of the potential cost savings would have been already realized under FFS, and there would be very little room for further cost reduction and hence savings through bundled payments. On the other hand, in hospitals with low physician alignment, level of care intensity under FFS would be higher than the preferred intensity from patient's perspective (I_0), and hence there will be more opportunities for cost-reduction via bundling. However, in the bundling scenario, the relative revenue loss for physicians outweighs the revenue gain from cost reduction. Physicians, therefore, will not have enough incentives to cooperate with the hospital and engage in bundling activities, and given the loose level of alignment, hospital lacks the power to influence physicians and integrate them to the bundling initiatives. As such, when physician alignment level is low, although there is much room for cost-reduction, bundling will not occur.

Next, in Proposition 2.2.1, we characterize the optimal solution when bundling occurs, and highlight the role of gainsharing, an incentive mechanism for physicians that is not allowed under FFS.

Proposition 2.2.1 (Optimal solution and the role of gainsharing)**.** *If bundling occurs as* *outlined in Theorem 2.2.1, then the optimal solution becomes the following:*

$$\begin{aligned}
&i_h = 1, \\
&T = 2w_b(1-\Psi)\min\{\frac{1}{2}(I_0 + \frac{\Delta c + 2\Delta r_p}{2w_b} - (1-\Psi)), I_0 + \frac{\Delta r_p}{2w_b}\}, \\
&i_p = \frac{1}{1-\Psi}(I_0 + \frac{\Delta r_p}{2w_b} - \frac{T}{2w_b(1-\Psi)})
\end{aligned} \tag{2.7}$$

In this case, it always holds that $T > 0$*.*

One notable finding in Proposition 2.2.1 is that the gainsharing amount, T, is always positive, which is in line with the expert opinions that gainsharing is critical in moving bundled payments forward (Froimson et al. 2013). Intuitively, this is because while the hospital would be inclined to minimize intensity, the physicians in return may resist and attempt to keep the intensity as high as it was in FFS. Therefore, in order for the hospital to incentivize physicians to reduce the level of care intensity, the hospital would have to compensate physicians through gainsharing.

Gainsharing aims to get physicians to cooperate in cost reduction efforts, which in turn creates value for hospitals. Under the bundled payments, physicians would also be interested in how much gainsharing, as a "value-based" part of their compensation, they actually receive. We characterize the gainsharing amount T as a function of the physician alignment factor Ψ. The amount that is gainshared with physicians is maximal at $\Psi = \bar{\Psi} + \frac{1}{2}(I_0 - \frac{\Delta c}{2w_b})$, first increasing with physician alignment but later decreasing, as showed in Figure 2.1. For lower physician alignment level values, $\Psi < \bar{\Psi} + \frac{1}{2}(I_0 - \frac{\Delta c}{2w_b})$, an increase in the level of physician alignment requires larger gainsharing to compensate physicians' higher financial loss from bundling. For higher physician alignment level values, $\Psi > \bar{\Psi} + \frac{1}{2}(I_0 - \frac{\Delta c}{2w_b})$, an increase in the physician alignment requires smaller gainsharing because reduced cost saving opportunities along with better cooperating physicians facilitate a reduction.

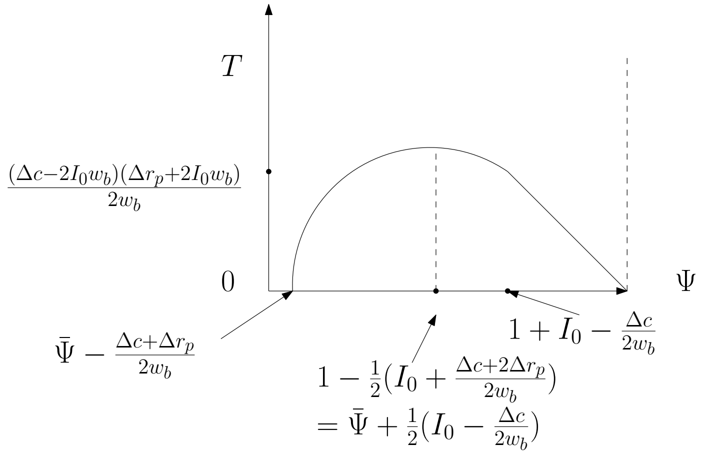


Figure 2.1: Gainsharing amount by physician alignment. Notice that the maximum is not attained at $\bar{\Psi}$ but rather at $\bar{\Psi} + \frac{1}{2}(I_0 - \frac{\Delta c}{2w_b})$, as explained in the main text.

Alternative parameterization

Here, we provide a version of Theorem 2.2.1 using an alternative parameterization, a parameterization that allows plotting the feasible region in three dimensions of key variables: physician integration, cost differential, and revenue differential. This allows us to infer some new insights and derive certain previous insights more easily.

Corollary 2.2.1 (When do they bundle? (Version with $\chi,\ \gamma,\ \rho$)). *Let* $\rho = \frac{\Delta r_p}{2w_b}$, $\gamma = \frac{\Delta c}{2w_b}$, *and* $\chi = I_0 - (1 - \Psi)$, *and suppose the condition in Assumption 2.2.1 holds. Then, the hospital and physicians will benefit from bundling if:*

$$\gamma > \chi \geq -\rho, \qquad \textit{or} \tag{2.8}$$

$$-\rho > \chi > -(\gamma + 2\rho) \tag{2.9}$$

We scale physician integration, the cost differential, and the revenue differential and define three corresponding variables in Corollary 2.2.1. Namely, we denote the optimal intensity-adjusted physician alignment using χ, the benevolence-adjusted cost differential using γ, and the benevolence-adjusted revenue differential using ρ. To better present the

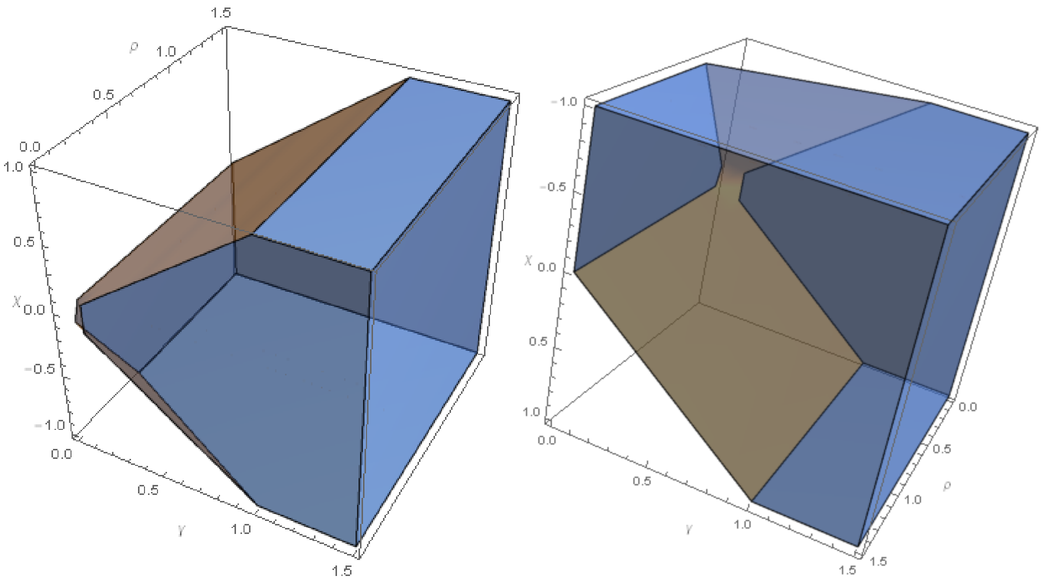


Figure 2.2: The shaded regions indicate when bundling is preferred, with views from different angles. The variables are $\rho = \frac{\Delta r_p}{2w_b}$, $\gamma = \frac{\Delta c}{2w_b}$, $\chi = I_0 - (1 - \Psi)$.

insights from Corollary 2.2.1, we visualize the trade-off between the three factors in Figure 2.2, where the shaded regions represent the instances in which bundling is preferred. From this figure, we observe the following. First, we confirm that, as expected by healthcare experts, bundling would be more appealing when opportunities for savings on hospital costs (i.e., Δc) are high. In particular, we observe that when the cost differential between the pathways is very high, hospital and physicians choose to bundle regardless of revenue differentials and the level of integration. Second, we observe that bundling is mostly likely when physician alignment level is about $1 - I_0$ (i.e. $\chi = 0$), and it becomes progressively more difficult when physician alignment deviates (in either direction) from this level.

One implication is that bundling is a good option when both optimal care intensity, I_0, and physician integration level are not simultaneously high or simultaneously low. This is because if the optimal required care intensity, I_0, is high and integration level is low, under FFS, then most patients would be treated through the costly pathway, and some would experience overtreatment. Hence, with bundling, there would be opportunity for cost reduction by reducing overtreatment. On the other hand, if I_0 is low, bundling would be profitable only if the physician integration level is high. This is because in this case, because of the hospital influence, physicians' would make at least some patients go through the costlier pathway under FFS, which implies that there would be a room for cost reduction under BP. This is because, in hospitals where physician integration is already too high, most of the potential cost savings would have been already realized under FFS, and there would be very little room for further cost reduction and hence savings through bundled payments. In such hospitals, bundling would be only desirable when the cost differential is very large. Third, we observe that bundling also becomes difficult as physician alignment level decreases further below from $1 - I_0$. However, unlike the case when alignment level is above $1 - I_0$, bundling is still preferable even when the hospital cost differential is low, as long as the physician revenue differential (Δr_p) is high enough. This is because under the bundled payments, unlike FFS, the hospital assumes the responsibility for reimbursing the physicians. Hence, if the physician reimbursement is high, the hospital may decrease intensity and save on physician reimbursement in addition to hospital costs, then gainshare part of the savings with the physicians to keep them involved while still keeping some savings for itself. Hence, we see that while bundling is very sensitive to the hospital savings opportunities (Δc), it is less sensitive to the saving opportunities on physician reimbursement (Δr_p). Overall, these findings suggest that, in addition to the hospital savings opportunities (Δc), saving opportunities based on physician reimbursement is also very critical.

Savings under bundled payments

In the previous section, we have characterized the conditions under which bundling is preferred and the corresponding optimal solution. In this subsection, we analyze the extent of savings achieved under bundled payments as the alignment level Ψ changes. Let $\Sigma := r^{\text{FFS}} - r^{\text{BP}}$, the difference between the total reimbursement under FFS and the minimal acceptable reimbursement under bundled payments, to represent the overall savings from bundled payments. Then, we have the following result characterizing the overall savings from bundled payments as a function of the physician alignment level, Ψ:

Proposition 2.2.2. *When bundling is feasible and preferred, in hospitals with relatively low physician alignment (i.e., $\Psi \leq \bar{\Psi}$), increasing alignment level leads to higher savings. In contrast, in hospitals with relatively higher physician alignment (i.e., $\Psi \geq \bar{\Psi}$), increasing alignment level further leads to lower savings. More specifically, for $i_p^\sharp > 0$, the savings are given by:*

$$\Sigma = \begin{cases} \dfrac{(\Delta c + 2\Delta r_p - 2w_b(1 - I_0 - \Psi))^2}{8w_b}, & \text{if } \Psi \leq \bar{\Psi} \\ \\ \dfrac{(\Delta c + 2w_b(1 - I_0 - \Psi))^2}{8w_b}, & \text{if } \Psi \geq \bar{\Psi} \end{cases},$$

and the savings Σ are maximized when $\Psi = \bar{\Psi}$, as also illustrated in Figure 2.3.

Figure 2.3 visualizes Proposition 2.2.2. As seen from this figure, savings initially increase, with the highest savings occurring when $\Psi = \bar{\Psi}$, and then decrease as the alignment level further increases. Interestingly, this result implies that hospitals that have the opportunity for highest savings from bundling are not the ones with very high or low physician level alignment, but instead are the ones with moderately high level of physician alignment. A conclusive matching of alignment levels and specific hospital types is a difficult task. However, some good examples to hospitals with high, low, and moderate physician alignment levels could be integrated healthcare systems, stand-alone hospitals in competitive markets, and stand-alone community hospitals in less competitive markets, respectively. Based on these examples, the result in Proposition 2.2.2 implies that when bundling is feasible, integrated networks or stand-alone hospitals in competitive markets are expected to achieve relatively lower savings, compared with stand-alone community hospitals in less competitive markets. Although such a finding may appear counterintuitive at first, it has an intuitive explanation: inefficiencies, and hence the potential for savings, are highest in hospitals with low physician alignment. As the alignment level increases, the proportion of these savings that is realized, increases. Then, at a certain threshold alignment level, all potential savings are realized, and as the alignment level further increases, the room for savings tends to decrease.

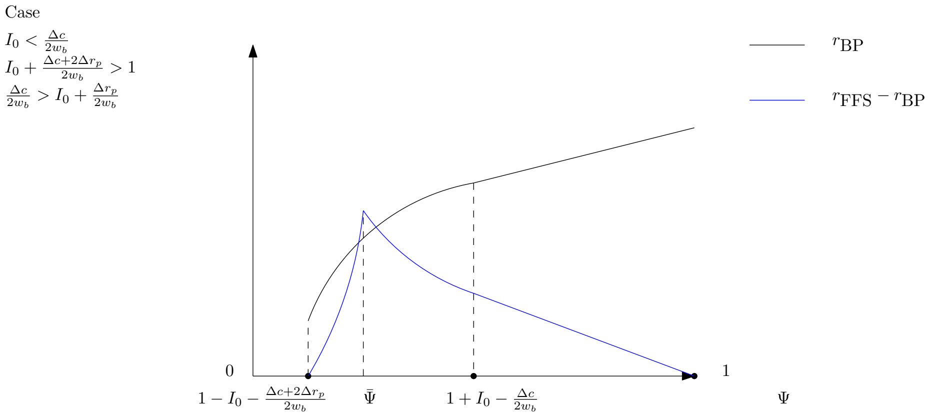


Figure 2.3: Savings as a function of Ψ under bundled payments. The y-scales of r^{BP} and $r^{\mathrm{FFS}} - r^{\mathrm{BP}}$ are not comparable. For other cases of parameters, the figure does not differ dramatically, even though the bounds of Ψ where bundling occurs may vary (even down to 0 or up to 1). Furthermore, the right, linear part of r^{BP} may be decreasing rather than increasing. Details are given in Appendix B.1.1.

Intensity and quality

In this subsection, we analyze how care intensity and quality will be influenced by bundled payments.

Corollary 2.2.2. *The optimal care intensity under bundled payments, $I^\sharp$, is less than that under the FFS, where $I^\sharp$ is given by*

$$I^\sharp = I_0 + \frac{\Delta r_p}{2w_b} - \frac{T}{2w_b(1-\Psi)} \leq I^*. \tag{2.10}$$

Corollary 2.2.2 corroborates experts' intuition that, compared with FFS, bundled payments are expected to decrease intensity, and hence utilization and costs, which underlies the motivation of CMS to implement bundled payments (Mechanic and Altman 2009). However, it is unclear whether this decreased intensity would lead to a decrease or an increase in quality, which we investigate next.

Let ΔQ represent the extent of deviation from the optimal care intensity under FFS and bundled payments, i.e., $\Delta Q := |I^* - I_0| - |I^\sharp - I_0|$, where $I^* > I^\sharp > 0$. $\Delta Q > 0$ implies less deviation from the optimal care intensity, I_0, under bundled payments, which may be interpreted as quality improvement under bundled payments as compared with the FFS. On the other hand, the higher the deviation from the optimal care intensity under bundled payments is (i.e., $|I^\sharp - I_0|$), the lower ΔQ becomes, which may eventually become negative, implying a reduction in quality.

Proposition 2.2.3. *Compared with FFS, quality of care under bundled payments may decrease or increase, depending on the physician alignment level, Ψ. In particular:*

1. *For $\Psi \geq \bar{\Psi}$, we have*

$$\Delta Q = \begin{cases} \frac{1}{2}(1 - \Psi - I_0 - \frac{\Delta c}{2w_b}) > 0 & \textit{if } I_0 + \frac{\Delta c}{2w_b} < 1 - \Psi \\ \frac{1}{2}(3(1 - \Psi - I_0) - \frac{\Delta c}{2w_b}) \lesseqgtr 0 & \textit{if } I_0 < 1 - \Psi < I_0 + \frac{\Delta c}{2w_b} \\ \frac{1}{2}(I_0 + \Psi - 1 - \frac{\Delta c}{2w_b}) < 0 & \textit{if } 1 - \Psi < I_0. \end{cases} \tag{2.11}$$

2. *For $\Psi < \bar{\Psi}$,*

$$\Delta Q = \begin{cases} -\frac{1}{2}(I_0 + \frac{\Delta c - 2\Delta r_p}{2w_b} - (1 - \Psi)) \textit{ can be } < \textit{ or } > 0 & \textit{if } I_0 + \frac{\Delta c}{2w_b} \geq 1 - \Psi \\ \frac{1}{2}(I_0 + \frac{\Delta c + 2\Delta r_p}{2w_b} - (1 - \Psi)) = 2w_b(1 - \Psi)T > 0 & \textit{if } I_0 + \frac{\Delta c}{2w_b} < 1 - \Psi \end{cases} \tag{2.12}$$

Proposition 2.2.3 presents a somewhat surprising result, which suggests that the care quality under bundled payment may increase or decrease, depending on the hospital and physician alignment level.

When the Payer Adjusts Payments to Achieve Certain Quality

Our analysis in Section 2.2.2 suggests that under bundled payments, intensity decreases and the associated quality may decrease or increase when a hospital is allowed to operate as an unconstrained profit maximizer. Expecting the intensity reduction with a concern for quality, it is plausible that under BP, the payer may set quality guarantees. Namely, he may set a lower bound on $I^\sharp$ by paying just enough for the hospital to raise the quality at or above $I^\sharp$. The analysis then corresponds to the characterization of the efficiency frontier for payments vs. the achievable quality. Our main findings are summarized in the following theorem.

Theorem 2.2.2. *When the payer sets minimum quality requirements, three scenarios are possible:*

(1.) If $I^* > I_0$, when bundling occurs, the resulting quality under bundled payments may be lower or higher than that under the FFS.

(2.) If $I^* < I_0$ and $\Delta r_p > \Delta c$, then a higher quality level can always be achieved for cheaper with bundled payments than FFS.

(3.) If $I^* < I_0$ and $\Delta r_p < \Delta c$, then a higher quality level may be achieved for cheaper with FFS, compared with bundled payments.

Figures 2.4 and 2.5 illustrate Cases 2 and 3 in Theorem 2.2.2, respectively. Perhaps the more interesting case is Case 3 presented in Figure 2.5, which implies that the payer may reach high quality more easily under FFS than under bundled payments. In other words, when the physician reimbursement differential between the two pathways is lower than the hospital cost differential, FFS may offer higher quality for lower cost. While this is a somewhat counterintuitive result, there is a reasonable explanation: under FFS, increasing intensity (hence in this case quality) increases physician reimbursement but not hospital reimbursement. Hence, if Δr_p is relatively low compared to Δc, the payer can easily and cheaply motivate physicians, and FFS is then more desirable if high quality is required.

A typical example for $\Delta r_p < \Delta c$ may be the common problem faced by the hospitals when physicians' preference of implants determine the cost of knee replacement to the hospital. When the choice is between a cheaper and an expensive implant, the two pathways is determined by physician's implant choice. The procedure conducted by the physician may be similar in both cases (i.e., Δr_p is small), but the hospital incurs a higher cost when the expensive implant is used (Δc is large). In contrast, $\Delta r_p > \Delta c$ may hold, for instance, if the physician can order additional laboratory or radiological testing in the costlier pathway which generates relatively higher revenues for physicians, even though the cost differential for the hospital may be lower as compared with revenue differential for the physician.

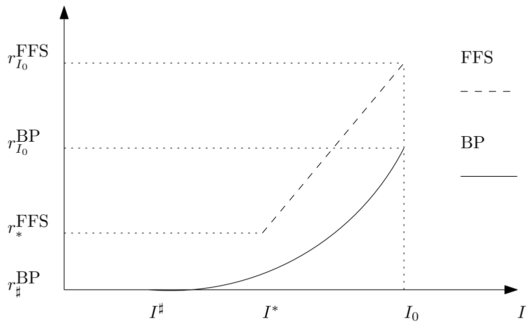


Figure 2.4: Comparison of efficiency frontiers for $\Delta c < \Delta r_p$.

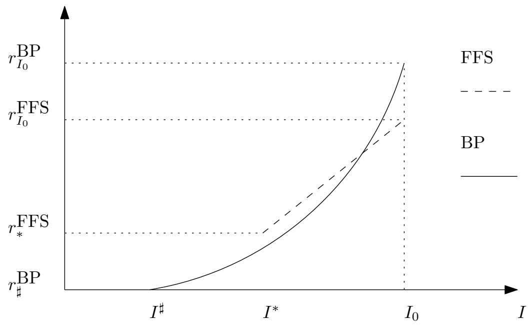


Figure 2.5: Comparison of efficiency frontiers for $\Delta c > \Delta r_p$.

2.3 Observable Coproduction Model

Based on the referee comments, we propose another model, where the gainsharing is depending on an observable quantity, namely I instead of i_p. Per referees' request, we also immediately include the hospital quality concern, something that only constituted an extension for the Initial model (see Section 2.4.1). In order to make a model with these two changes solvable, we also need to change the game structure of the bundled payments: In the first stage, the hospital offers the gainsharing amount T, and in the second stage, the hospital and physicians jointly optimize for i_h and i_p. With this perspective in mind, we are ready to describe the model. In this section, we repeat some background information from the Initial model in order to make this section self-contained.

In this "Coproduction" model, thee care delivery is co-produced by physicians and the hospital. This proposed model considers quality and financial concerns by both the hospital and physicians. In Sections 2.3.1 and 2.3.1, we present the key components of the general modeling framework. Then, in Section, 2.3.2, we analyze a special case where only the hospital is quality concerned, which provides a simple baseline for understanding key dynamics. Later, in Section 2.3.3, we analyze the more general case introduced in Section 2.3.1 for a complete analysis. Finally, in Section 2.4.4, we consider an alternative "Physician-driven" model where physicians are the sole drivers of the care delivery, in contrast to the Coproduction model. While the Coproduction modeling setup would be representative of most hospitals in the US system, this Physician-driven model would be especially relevant to the physician-controlled hospitals.

2.3.1 Setup

Intensity of Care under the Coproduction Model

As discussed earlier, under the prevailing FFS model, physician and hospital incentives are not well-aligned, leading to increased costs and inefficiencies in care delivery. While introduced as a potential remedy for this incentive alignment problem, the potential success of BP depends on the effectiveness of coordination between the physician and the hospital and hence the level of integration between the these two entities (Mechanic and Altman 2009). Our Coproduction model is developed to analyze the effectiveness of bundling in addressing the incentive alignment problem and the resulting cost and quality outcomes in the spectrum of physician integration levels. The model captures the main dynamics of the coordination problem by considering index admissions[5] in hospitals with non-salaried

[5] Index admission corresponds to the initial admission in a care episode.

physicians. Unless included in the main text, all proofs are included in the online appendix.

Without loss of generality, we build our model around payments for a single medical condition such as knee-replacement, characterized by a DRG categorization. For example, a patient with “major joint replacement or reattachment of lower extremity without major complications” will be assigned DRG 470 and another patient with “simple pneumonia and pleurisy without complications” will be assigned DRG 195 for billing purposes. Under the FFS model, once the care is complete, the hospital receives a predetermined amount of payment for each DRG regardless of the costs for treating the patient; whereas physician is paid separately for each service she provides. BP, on the other hand, bring the hospital and physician payments together, by paying a single lump-sum amount for a given DRG (e.g., DRG 470 – knee replacement), which is then shared between the hospital and physicians. Table 2.2 summarizes the variables that appear in our analyses, followed by a description of the key model components.

Table 2.2: Summary of notation used.

$r_{1,p},\ r_{2,p}$	Physician reimbursement for pathway 1, 2.
$c_1,\ c_2$	Per-patient costs for pathway 1, 2.
Δc	Cost differential for hospital; $\Delta c = c_1 - c_2$.
Δr_p	Revenue (reimbursement) differential for physicians; $\Delta r_p = r_{1,p} - r_{2,p}$.
$I = I(i_h, i_p) \in [0, 1]$	Care intensity, interpreted as the fraction of patients assigned to the more expensive pathway.
I_0	Quality-maximizing care intensity from the patient’s perspective.
$i_h \in [0, 1]$	Hospital influencing effort (toward the cheaper pathway).
$i_p \in [0, 1]$	Physician influencing effort (toward the costlier pathway).
$\Psi \in (0, 1)$	Level of physician integration.
T	Gainsharing amount (relative to care intensity).
r_h	Hospital reimbursement under FFS.
$F_h^{\text{FFS}}, F_p^{\text{FFS}}, F_h^{\text{BP}}, F_p^{\text{BP}}$	Respective objective functions for the hospital/physicians under FFS/BP (per patient).
w_b	Physician quality concern.
w_q	Hospital quality concern.
$x^*, x^\sharp$	Respective solution feature under FFS and BP (x can be different variables).
r^{FFS}	Total payment from the payer under FFS.
r^{BP}	Total payment from the payer under BP.
ι_p, ι_h	Respective best response functions of the hospital and physicians.
Σ	Total surplus under BP, to be distributed between the payer and the hospital.
π_p	Physician surplus under BP.
ΔQ	Quality difference between FFS and BP (> 0 means worse quality under bundling).

A key feature in our analysis is the concept of a clinical *pathway*, which represents the set of medical services and procedures a patient follows for a given DRG, including diagnostic tests, medications, and consultations, conducted during care delivery (De Bleser et al. 2006). In practice, when providers treat a condition, they often vary in terms of the clinical pathways chosen, resulting in different ranges of costs and health outcomes. Under FFS, because each service is billed for separately, hospitals and physicians tend to not worry much about care coordination through cost-effective clinical pathways. In contrast, because cost accounting is a bigger concern under BP, hospitals and physicians need to better understand their common clinical pathways and standardize the choice of the most cost-effective ones.

For any given condition, we consider two pathways, one being more intensive (and costly), and the other one being less intensive (and cheaper). For example, for DRG 470 (joint replacement), compared with the less intensive (and cheaper) pathway, the more intensive (and costlier) pathway may include an additional day of hospital stay, additional X-rays and MRIs, and use of a more advanced robotic technology during surgery. For the ease of interpretation, we assume that the costlier pathway is a superset of the cheaper pathway, with a higher volume or intensity of the procedures performed. The corresponding costs to the hospital for these pathways are respectively captured by c_1 and c_2 ($c_1 > c_2$) while the corresponding reimbursements to the physician are respectively captured by $r_{1,p}$ and $r_{2,p}$ ($r_{1,p} > r_{2,p}$).[6] Lastly, we denote the payments to the hospital under FFS using r_h.

For each DRG, there is a level of care intensity, $I_0 \in [0, 1]$ that corresponds to the "best" clinical outcome from the "patient perspective" without cost concerns. In what follows, we refer to I_0 simply as "quality-maximizing" intensity. The practiced level of care intensity, I, however, may deviate from the quality-maximizing intensity. In particular, depending on the payment model adopted, hospitals and physicians may have different preferences for the practiced level of intensity. For example, under FFS-based payments, while physicians, who are paid separately for each service, are incentivized to operate above I_0 due to financial motivations, hospitals, being paid a fixed flat rate for a given DRG, would prefer

[6] We remark that we do not explicitly model physician costs as a function of care intensity. However, such a relationship can be easily captured by interpreting physician reimbursements,$r_{1,p}$ and $r_{2,p}$, as cost-adjusted reimbursements.

to reduce costs and hence lower intensity. In this case, the hospital may choose to influence physician practices through various means, such as care protocols.[7] In this case, intensity is "*coproduced*" by the physician and the hospital. To model the coproduction of care, we follow the established literature in health economics (Dor and Watson 1995, Jelovac and Macho-Stadler 2002, Crainich et al. 2008) and model the practiced *care intensity*, I, to be jointly produced by the hospital and physicians as follows:

$$I \equiv I(i_h, i_p) := (1 - i_h)\Psi + (1 - \Psi)i_p, \tag{2.13}$$

where $i_h \in [0, 1]$ represents the hospital influencing effort, $i_p \in [0, 1]$ is the physician influencing effort, and $\Psi \in (0, 1)$ is the *physician integration* coefficient. Physician integration (sometimes also referred to as hospital and physician alignment) is a well-established and widely studied concept in medical and health economics literature, and is defined as the degree to which physicians share the same mission, vision, and objectives with their hospital systems and work toward their success (Shortell et al. 2001, Ma 1994, Huang and McCarthy 2015).

The co-produced intensity I captures the fraction of patients assigned to the more expensive pathway. Note that the two extreme points, $I = 0$ and $I = 1$, correspond to the cases where all patients go through the less intensive pathway and more expensive pathway, respectively; and the intermediate values of I (e.g., $I = 60\%$) are interpreted as fraction I of patients going through the more intensive pathway and $(1 - I)$ fraction going through the less intensive pathway. We finally remark that when $I < I_0$, underprovision of care is prevalent; and when $I > I_0$, overprovision of care is prevalent.

[7] Such hospital-driven intensity reduction was observed during the transition to the DRG-based system in the 80s from the cost-based reimbursement, where hospitals were paid based on average cost of a patient per-diem and the length of stay. At that time, hospitals were able to reduce the average length of stay for non-surgical patients from 9.4 days to 7.2 days within only a few years (Altman 2012).

Hospital and Physician Objectives

In this section, we model hospital and physician objectives and the corresponding utility functions. Specifically, we first model the FFS case, the status quo. In line with the commonly practiced BP models, we assume that the hospital is initially operating under FFS and is next offered BP. The hospital's objective is to reduce costs and therefore increase its profits, while maintaining high quality of care. The physicians also weigh both patients' health outcomes and satisfaction as captured by the quality of care, and their own monetary benefits. We remark that this dichotomy between quality concern (or benevolence or altruism) and financial benefits in providers' objectives is well established empirically (Chandra et al. 2012) and widely adopted in the literature (e.g. Ellis and McGuire 1986, Kolstad 2013). Given this setup, the utility functions for the hospital and the physician under FFS are defined as follows:

$$\begin{aligned} F_h^{\text{FFS}} &= r_h - c_1 I - c_2(1-I) - w_q(I-I_0)^2, \\ F_p^{\text{FFS}} &= -w_b(I-I_0)^2 + r_{1,p}I + r_{2,p}(1-I), \end{aligned} \tag{2.14}$$

where F_h^{FFS} and F_p^{FFS} respectively represent the average per-patient utility for the hospital and the physician. Under FFS, the hospital receives a single DRG-based payment, r_h, and incurs costs due to patients receiving care via either the costlier or cheaper pathway. In addition, the hospital incurs disutility (e.g., reduction in reputation, modulated by the hospital quality concern, w_q) in case of any deviation from the quality-maximizing intensity. Similarly, the physician receives pathway-dependent payments and incurs disutility due to any deviation from the quality-maximizing intensity (modulated by the physician quality concern, w_b).

While under FFS the hospital and physicians are paid separately, under BP, their payments are "bundled." Under bundling, the hospital (or its parent health system) is responsible as the "convener" to receive and distribute the payment. The total payment under BP, which is paid to the hospital and is shared with physicians, is captured by r^{BP}. The amount r^{BP} accounts for all the costs of services included in the bundle, in addition to the hospital payment.

In line with the prevailing bundled payment practice (Dummit et al. 2015), the hospital continues to reimburse the physicians at the FFS rates. However, unlike FFS practice, under BP, the hospital can also reward the physicians for cooperating with the hospital in cost reduction efforts, a practice referred to as *gainsharing*, which is not allowed under FFS models. That is, in addition to the exerted effort level, i_h, the hospital has a second decision variable under BP, denoted by T, which represents the maximum amount it can share with the physician to incentivize cost reduction. The size of this gainsharing amount decreases as the physicians are less cooperative with the hospital in reducing costs, and is inversely proportional to the realized intensity (equivalently, proportional to $(1-I)$). That is, the lower the realized intensity, the higher the fraction of patients undergoing cheaper pathway is; hence, as the realized intensity decreases, savings increase and therefore gainsharing amount also increases. Then, the hospital's and the physician's utility functions under the BP are as follows:

$$\begin{aligned} F_h^{\text{BP}} &= r^{\text{BP}} - (c_1 + r_{1,p})I - (c_2 + r_{2,p})(1-I) - w_q(I - I_0)^2 - (1-I)T, \\ F_p^{\text{BP}} &= -w_b(I - I_0)^2 + r_{1,p}I + r_{2,p}(1-I) + (1-I)T. \end{aligned} \tag{2.15}$$

In studying the BP case, we consider a two-stage game where the hospital first announces T, the gainsharing contract, and then the hospital and the physicians simultaneously choose their respective efforts i_h and i_p. We derive the subgame perfect Nash equilibrium using backward induction.

In our base case analysis presented in 2.3.2, for simplicity of the analysis and ease of presentation, we omit the quality concerns by the hospital w_q by setting it to zero and only consider the quality concerns by the physicians, i.e. $w_b > 0$. Practically, this setting corresponds to the case where the physician is the main driver of the quality. While this is a simplified setting, as we show later in the analysis of the general case in 2.3.3, the findings that we obtain through this simpler case are parallel to those for the general case with $w_q > 0$.

In the next subsection, we analyze the base case in which care quality is solely driven by physicians (we henceforth refer to it as the Base model). Next, in Section 2.3.3 , we analyze the more general case in which hospital is also quality concerned (called the General model), and draw parallels between the two sets of results. In our analyses, we consider that bundling will occur if both the hospital and physicians have strictly higher payoffs compared to the FFS setting, and the total amount of the payments to these two stakeholders are lower (so that the payer will save). We also remark that in practice, payer decides on r^{BP} for a given r^{FFS}, which is estimated based on historical reimbursements.

2.3.2 Analysis of the Base Coproduction Model

We start with characterizing the equilibrium intensity under FFS as a function of the physician integration level, Ψ. In particular, we show that in hospitals where physician integration is lower than a certain threshold, $\bar{\Psi}$, the equilibrium intensity under FFS is driven by the physicians only and is set to the maximizing value of the physician utility function, $I_0 + \frac{\Delta r_p}{2w_b}$. On the other hand, in hospitals where physician integration is higher than this threshold $\bar{\Psi}$, FFS equilibrium intensity is co-produced by the hospital and physicians and is set to its maximum value, $1 - \Psi$. We define this threshold value $\bar{\Psi}$ above which the care is co-produced in the lemma below, and throughout the remainder of this paper, we refer to those hospitals with $\Psi > \bar{\Psi}$ ($\Psi < \bar{\Psi}$) as hospitals with high (low) physician integration.

Lemma 2.3.1 (Status-quo intensity)**.** *The equilibrium intensity under FFS, I^*, is given by*

$$I^* = \begin{cases} I_0 + \frac{\Delta r_p}{2w_b} & \textit{if } \Psi \leq \bar{\Psi} \\ 1 - \Psi & \textit{otherwise} \end{cases}$$

where

$$\bar{\Psi} := 1 - I_0 - \frac{\Delta r_p}{2w_b} \tag{2.16}$$

Two observations follow from Lemma 2.3.1. First, under FFS, the physician operating under low level of integration (i.e., $\Psi \leq \bar{\Psi}$) prefers to increase the care intensity beyond the quality-maximizing intensity, I_0, by the quality-adjusted financial motives, $\frac{\Delta r_p}{2w_b}$. As expected, the financial motives, measured by the physician payment/revenue differential between the costly and cheaper pathways, Δr_p, increase the extent of deviation while the physician quality concern, w_b, decreases the extent of deviation from the quality-maximizing care intensity.

Second, when the level of integration is high (i.e. $\Psi \leq \bar{\Psi}$), the equilibrium intensity under FFS (i.e., status quo) is inversely related to the physician integration. That is, higher the level of integration, the lower the intensity, therefore more patients following the cheaper pathway (to see this, note that when $\Psi = \bar{\Psi}$, $I^* = I_0 + \frac{\Delta r_p}{2w_b}$, and gradually decreases afterwards as Ψ is increased further).

We remark that when evaluated together, these two observations based on Lemma 2.3.1 suggest that under FFS, physicians would be inclined to set the intensity level to $I_0 + \frac{\Delta r_p}{2w_b}$, whenever they can; however, when physicians are highly integrated with the hospital, the level of integration dominates the physicians' financial incentives to set the intensity at the quality-adjusted level, hence the equilibrium intensity would be bounded by $1 - \Psi$.[8] Key

[8] In this Base model in which the hospital is not quality concerned and a pure profit maximizer, i_h would be set to 1 so as to minimize costs and maximize profit. On the other hand, when the model is extended as in Section 2.3.3 to capture the quality concern of the hospital in addition to that of physicians, then i_h could be

qualitative results continue to hold in the general case presented in Section 2.3.3. Although this finding is somewhat intuitive, it reassures the validity of our model. Next, we examine when BP is beneficial for all parties, i.e., hospital and physicians (as well as the payer), as captured by their utility functions.

Theorem 2.3.1 (When does bundling benefits all parties?)**.** *Payers, hospitals, and physicians will all benefit from bundling if and only if the following conditions hold:*

(Hospital/physician coordination condition):

$$\Delta c + 2\Delta r_p > 2w_b(1 - I_0) \tag{2.17}$$

(Physician incentives condition):

$$\Psi \le \Psi_+ =: \frac{2w_b(1 - I_0) + \Delta c}{4w_b} \tag{2.18}$$

(Payer savings and alignment condition):

$$\Psi \le \Psi_0 =: \frac{(\Delta c + 2w_b(1 - I_0))^2}{8w_b(\Delta c + \Delta r_p)} \tag{2.19}$$

We have the following insights from Theorem 2.3.1. Inequality (2.17) is the coordination condition for bundling to occur. Note that under BP, hospital and physicians operate as a single entity and share savings. Condition (2.17) says that the potential size of the total amount of financial gain through bundling (as measured by revenue differentials plus cost differentials) should be higher than the potential incentives for overtreatment, as measured by $2w_b(1 - I_0)$. To see this, note that $(1 - I_0)$ corresponds to the potential room for overtreatment, and when multiplied by w_b, the quantity corresponds to the minimum financial gain that the physician is willing to accept in exchange of overtreatment. Then,

set to a value < 1. We remark that, as compared with the General model, this Base model may overestimate the practiced care intensity I; yet, the Base model is general enough to capture the main dynamics and better reflects the key trade-offs when studying hospital-physician interaction the condition says that the potential amount of financial gain under bundling should be at least as high as the financial gain that could be achieved through overtreatment under FFS, as otherwise it would be difficult to convince physicians to coordinate care under BP.

Condition (2.18) says that there are two critical factors to incentivize physicians for bundling. First, the size of potential incentives for overtreatment, captured by $(1 - I_0)$ should be large enough. This is because, otherwise, there will not be enough room for intensity reduction, resulting in savings through bundling and hence gainsharing with physicians. Second, cost differential between expensive and cheaper pathway, Δc, should be high enough, as otherwise, physician financial gain due to cost-savings won't be high even if the size of overtreatment in the system is reduced significantly. Finally, the inequality suggests that it is easier to incentivize physicians for bundling in hospital systems with lower physician integration, as measured by Ψ (see more on this in the following paragraph).

Inequality (2.19) is payer's savings condition around physician integration. The inequality suggests that bundling may be easier, and hence savings are expected to be larger, in hospital systems with lower physician integration, and becomes progressively more difficult as the initial level of physician integration increases. The intuition behind this finding is that in hospital systems with low level of physician integration (prior to bundling), there is more inefficiency and hence potential for cost reduction. This is because when physician integration is low, physicians are more powerful and can more easily practice overtreatment leading to inefficiencies in the system. Whereas in hospital systems with initially high level of physician integration, there is less inefficiency under FFS and hence smaller room for savings through bundling.

Lemma 2.3.2 characterizes the optimal amount of gainsharing under bundling, an incentive mechanism for physicians that is not allowed under FFS.[9]

Lemma 2.3.2 (Role of gainsharing)**.** *When bundling occurs as outlined in Theorem 2.3.1,*

[9] The full characterization of the optimal solution is given in Lemma B.1.2 in Appendix B.1.5.

the gainsharing amount, T, is always positive and is given by:

$$T = \min(\Delta r_p + 2I_0 w_b, \frac{1}{2}(-2w_b(1 - I_0) + \Delta c + 2\Delta r_p)) > 0. \tag{2.20}$$

The gainsharing amount T being always positive under bundling is consistent with the expert opinion that gainsharing is critical in advancing BP (e.g., see Froimson et al. 2013). Intuitively, this is because in order for the hospital to incentivize physicians to reduce the level of care intensity, the hospital needs to compensate physicians through gainsharing.

Assessment of Outcomes Under Bundled Payments

In the following set of results, we analyze how BP will influence outcomes, including care intensity, costs/savings, and quality of care under the setup that physicians care about quality and hence exert some positive efforts, i.e. $i_p^\sharp > 0$. We start with intensity as follows.

Corollary 2.3.1 (Intensity under BP vs. FFS)**.** *The equilibrium intensity under BP, $I^\sharp$, is less than that under the FFS, I^*, where $I^\sharp$ is given by*

$$I^\sharp = I_0 + \frac{\Delta r_p}{2w_b} - \frac{T}{2w_b} \leq I^*. \tag{2.21}$$

Corollary 2.3.1 corroborates experts' intuition that, compared with FFS, BP is expected to decrease intensity, and hence utilization and costs, which underlies the motivation of CMS to implement BP (Mechanic and Altman 2009). However, it is unclear whether this decreased intensity would lead to a reduction or an increase in quality, which we investigate next.

Let ΔQ be the difference in the extent of deviation from the quality-maximizing care intensity under FFS and BP, representing the quality difference between the two regimes. Specifically, let $\Delta Q := |I^* - I_0| - |I^\sharp - I_0|$, where $I^* > I^\sharp > 0$. Then, $\Delta Q > 0$ implies smaller deviation from the quality-maximizing care intensity, I_0, under BP, which can be interpreted as quality improvement under BP as compared with the FFS. Conversely,

$\Delta Q < 0$ implies higher deviation from the quality-maximizing care intensity under BP as compared with FFS, and hence indicates a quality reduction due to bundling.

Proposition 2.3.1 (Quality under BP vs. FFS)**.** *Compared with FFS, quality of care under BP may decrease or increase, depending on the physician integration level,* Ψ*. In particular:*

I. *If integration is low, i.e.,* $\Psi < \bar{\Psi}$*, overprovision of services characterizes FFS. In such a case, BP will improve quality (i.e.,* $\Delta Q > 0$*) if*

$$\frac{\Delta c}{2w_b} < \Delta r_p(1 - I_0) \tag{2.22}$$

and worsen it otherwise.

II. *If integration is high, i.e.,* $\Psi > \bar{\Psi}$*, then both underprovision and overprovision of services can characterize FFS. If the former was the case (i.e. if* $1 - \Psi < I_0$*), then the quality under BP is expected to be even worse. If it was the latter (i.e. overprovision,* $1 - \Psi > I_0$*), then BP will improve quality (*$\Delta Q > 0$*) if*

$$\frac{\Delta c}{2w_b} + 2\Psi < 3(1 - I_0). \tag{2.23}$$

Proposition 2.3.1 suggests that the care quality under BP may decrease or increase, depending on the level of physician integration along with other factors. It is especially worth noting that in hospital systems with low Ψ (and hence higher inefficiency due to overtreatment) as in Part I, we find that bundling payments will improve quality, as long as Condition (2.22) is satisfied, which ensures that the intensity is not reduced too much under BP. The intuition behind this finding is as follows: recall that from Corollary 2.3.1, we found that the intensity under BP would be reduced. In hospital systems with low level of physician integration (where overprovision of care is prevalent), some decrease in intensity through BP would result in reduction in overprovisioning and hence quality improvement.

However, too much decrease in intensity would ultimately result in underprovision of care (and hence quality reduction), which is characterized by Condition (2.22). This finding is important because it says bundling payments will work in the intended direction both in terms of quality and costs especially in hospital systems with low level of physician integration and hence high level of inefficiencies.

On the other hand, in hospital systems with higher level of integration as in Part II, quality may go in either direction after bundling, depending on the size of over/under provision of care prior to bundling. This is because when the level of physician integration is high, the hospital has stronger influence on the physician's care provision behavior. As a result of this interaction, both overprovision or underprovision of services under FFS are possible in this case. If overprovision was persistent under FFS, then the interpretation would be similar to that of the first case and Condition (2.23) plays a similar role as in Condition (2.22). On the other hand, if underprovision was persistent, a further reduction in intensity as a result of BP would lead to a quality reduction.

Finally, we analyze the extent of cost reduction (realized by the payer) under BP as the integration level Ψ changes. Let $\Sigma := r^{\text{FFS}} - r^{\text{BP}}$, the difference between the total reimbursement under FFS and the minimal acceptable reimbursement under BP, represent the overall savings from BP. Then, we have the following result characterizing the overall cost reduction under BP as a function of the physician integration level, Ψ:

Proposition 2.3.2 (Cost Savings in BP vs. FFS)**.** *When bundling is feasible and preferred, in hospitals with relatively low physician integration (i.e., $\Psi \leq \bar{\Psi}$), savings are highest. As the physician integration increases beyond $\bar{\Psi}$, the savings start to decrease and ultimately disappear altogether when physician integration is beyond $\tilde{\Psi} := min(\Psi_0, \Psi_+)$, where Ψ_0, and Ψ_+ are the bounds for the feasibility of bundling from Theorem 2.3.1. More specifically, the savings are characterized as follows:*

$$\Sigma = \begin{cases} \dfrac{(\Delta c + 2\Delta r_p - 2w_b(1-I_0))^2}{8w_b}, & \textit{if}\, \Psi \leq \bar{\Psi} \\ \\ \dfrac{(\Delta c + 2w_b(1-I_0))^2}{8w_b} - (\Delta c + \Delta r_p)\Psi, & \textit{if}\, \bar{\Psi} \leq \Psi \leq \tilde{\Psi} := min(\Psi_0, \Psi_+). \end{cases}$$

Proposition 2 provides the following insights: we find that when alignment level is below $\bar{\Psi}$, savings are highest. This is because, as we have shown earlier, care intensity and hence also inefficiencies are highest in hospitals with low physician integration. As a result, those inefficient hospitals under FFS with low physician integration have the highest potential for savings.[10] As the integration level increases further beyond $\bar{\Psi}$, hospitals have increasingly more influence on physician's choice of care intensity under FFS. As a result, the higher Ψ, the less inefficiencies under FFS are, and hence smaller room for savings under BP.

Given these results, it is pertinent to discuss what type of hospitals would have the highest potential for savings through bundling in real world. We remark that conclusive matching of alignment levels with specific hospital types is not an easy task as there will be many factors at play in determining alignment level. However, a case for high hospital and physician alignment could be large integrated healthcare systems (Budetti et al. 2002). In contrast, a case for low hospital and physician alignment could be stand-alone specialty (e.g., surgical) hospitals. Based on these examples, the result in Proposition 2.3.2 imply that when bundling is feasible, integrated systems are expected to achieve relatively lower savings compared with stand-alone specialty hospitals.

[10] Under low physician integration (i.e. $\Psi < \bar{\Psi}$), because physicians are the sole drivers of the intensity under FFS and hospitals have no influence on the intensity (and hence costs), savings are constant and do not depend on Ψ.

2.3.3 Analysis of the General Coproduction Model

In this section, we generalize the results from the Base model, by introducing back the quality concern of the hospital. That is, we now focus on analyzing the General model given by (2.14) and (2.15).

Each result presented in this section mirrors a corresponding result presented in the base case analysis presented in Section 2.3.2. After presenting each result, we compare and contrast it to its base case analogue and discuss implications. As before, we start with the FFS analysis and characterize the equilibrium intensity under FFS as a function of the physician integration level, Ψ, as follows:

Lemma 2.3.3 (Status-quo intensity)**.** *The equilibrium intensity under FFS, I^*, is given by*

$$I^* = \begin{cases} I_0 + \frac{\Delta r_p}{2w_b} & \textit{if } \Psi \leq \bar{\Psi} \\ 1 - \Psi & \textit{if } \bar{\Psi} \leq \Psi \leq \hat{\Psi} \\ I_0 - \frac{\Delta c}{2w_q} & \textit{if } \Psi \geq \hat{\Psi} \end{cases}$$

$$\textit{where } \bar{\Psi} := 1 - I_0 - \frac{\Delta r_p}{2w_b}, \;\; \hat{\Psi} := 1 - I_0 + \frac{\Delta c}{2w_q}.$$

This result is the analogue of Lemma 2.3.1 in Section 2.3.2. As we see from a comparison of the two results, the results are almost identical with the exception that in Lemma 2.3.3, an additional third scenario appears for the value the realized intensity can attain. Specifically, we see an additional case with very high physician integration, $\Psi > \hat{\Psi}$, where the hospital chooses to keep the intensity higher so as to maintain quality. In contrast to the Base case, the equilibrium intensity under FFS is also bounded from below, a bound that becomes binding for $\Psi > \hat{\Psi}$. This is because the hospital is now also quality concerned.

Similar to the base case, we proceed with analyzing when bundling payments would be preferred by all parties and present the results in Theorem 2.3.2 below. As before, we focus on the setup that physicians care about quality and hence exert some positive efforts, i.e. $i_p^\sharp > 0$.

Theorem 2.3.2 (When bundling benefits all parties?)**.** *Payers, hospitals, and physicians will all benefit from bundling if and only if the following conditions hold:*

(Hospital/physician coordination condition):

$$\Delta c + \Delta r_p(2 + w_q/w_b) > 2w_b(1 - I_0) \tag{2.24}$$

(Physician incentives condition):

$$\Psi < \Psi_+^\sharp =: \frac{2(w_b + w_q)(1 - I_0) + \Delta c}{(4w_b + 2w_q)} \tag{2.25}$$

(Payer savings and alignment condition):

$$\Psi < \Psi_0^\sharp =: (1 - I_0) + \frac{\Delta c + \Delta r_p}{2w_q} - \frac{1}{2w_q}\sqrt{\Delta r_p^2 + 2(\Delta c - 2(1 - I_0)w_q)\Delta r_p + \frac{2}{2w_b + w_q}(\Delta c^2 + 2w_q(1 - I_0)(\Delta c - (1 - I_0)))} \tag{2.26}$$

Findings in Theorem 2.3.2 are parallel to that of Theorem 2.3.1. The primary difference of the theorem is the additional impact of w_q, the hospital quality concern variable, on incentive dynamics for bundling. First, the inequality in (2.24) characterizes incentives for coordination against overtreatment. In comparison to the base case, hospital's quality concern strengthens the left hand side of the inequality, namely the expected benefits of coordination. An increase in hospital's quality concern will make this coordination condition easier to satisfy. Second, the addition of hospital quality concern reinforces the incentives for avoiding overtreatment as it increases the right hand side of the inequality in (2.25). In comparison to base case, physicians have more incentives to bundle as $\Psi_+^\sharp > \Psi_+$. Finally, although a one-on-one comparison is not easy to make, the inequality in (2.26) reflects the spirit of its analogue in (2.19) and suggests that bundling becomes progressively more difficult as physician integration increases. Overall, the theorem also suggests that efforts for making hospitals more quality concerned could facilitate bundling. Therefore, healthcare reform efforts focusing on coupling quality with payments (e.g., value-based adjustments to hospital payments) could further align hospitals and physicians to bundle.

Lastly, before closing this subsection, we note that we characterize the optimal gainsharing amount under bundling when hospital is quality concerned in Equations (B.35) and (B.37) in the Appendix. Similar to the base case, we observe that gainsharing is an essential element of bundling and is always positive in successful bundling arrangements (cf. Lemma 2.3.2).

Assessment of Bundling Outcomes Under the General Model

In the remainder of this section, similar to the base case analyses, we focus on how BP is expected to change outcomes; i.e., care intensity, costs/savings, and quality of care when both the hospital and physicians are quality concerned.

As for the intensity, we show that the the corresponding result is quite similar to that in Corollary 2.3.1, and that intensity is expected to decrease under BP (the results are formally presented in Corollary B.1.1 in the Appendix). However, it is unclear whether this decreased intensity would lead to a decrease or an increase in quality, which we investigate next. As before, let ΔQ be the difference in the extent of deviation from the quality-maximizing care intensity under FFS and BP, representing the quality difference between the two regimes. Then, we have the following result, which is analogous to Proposition 2.3.1.

Proposition 2.3.3 (Quality under BP vs. FFS). *Compared with FFS, quality of care under BP may decrease or increase, depending on the physician integration level, Ψ. In particular:*

I. If integration is low, i.e., $\Psi < \bar{\Psi}$, then overprovision of services characterizes FFS;

and in that case BP will improve quality (i.e., $\Delta Q > 0$) if

$$\frac{\Delta c}{2w_b} < (1 - I_0) + \Delta r_p \frac{2w_b + w_q}{2w_b} \tag{2.27}$$

and worsen it otherwise.

II. *If integration is moderate, i.e., $\hat{\Psi} > \Psi > \bar{\Psi}$, then both underprovision and overprovision of services are possible under FFS. If the former is the case (i.e. if $1 - \Psi < I_0$), then the quality under BP is expected to be even worse. If the latter is the case (i.e., overprovision of services, $1 - \Psi > I_0$), then BP will improve quality ($\Delta Q > 0$) if*

$$\frac{\Delta c}{2w_b} + \Psi(2w_b + w_q) < (1 - I_0) + (1 - I_0)(2w_b + w_q) \tag{2.28}$$

and worsen it otherwise.

III. *If integration is high, i.e., $\Psi > \hat{\Psi}$, then underprovision of services characterizes FFS; and in that case BP will worsen the quality by magnifying the level of underprovision.*

As we see above, results in Proposition 2.3.3 are quite similar and analogous to the results in Section 2.3.2. The intuition for Parts I and II parallels those of base case as in Proposition 2.3.1. In particular, the quality improvement under bundling will be a function of (1) hospital and physician integration and (2) trade-offs in terms of over- or underprovision of services under FFS as a result of the level of alignment. Namely, when overprovision of services characterize FFS, bundling may improve quality when incentives for cost reduction is relatively less as compared with incentives for achieving quality-maximizing care intensity. One difference in the general case is that conditions for quality improvement (Conditions 2.27 and 2.28) are easier to satisfy in comparison to the base case (i.e., larger bounds). This is intuitive because, under general case, both the hospital and the physicians are quality-concerned while under the base case only the physician is quality concerned. Another difference from the base case is Part III of the proposition in which quality always deteriorates when hospital and physicians are highly aligned. The intuition of the finding is similar in that, under high alignment, underprovisioning characterizes FFS and bundling will further reduce intensity. Such a reduction will only worsen quality.

Next, we analyze the extent of cost reduction/savings realized by the payer under BP as the integration level Ψ changes. As before, let $\Sigma := r^{\text{FFS}} - r^{\text{BP}}$, the difference between the total reimbursement under FFS and the minimal acceptable reimbursement under BP, represent the overall savings from BP. Then, under general case, we have the following result characterizing the overall savings from BP, which is the analogue of Proposition 2.3.2 in the base case.

Proposition 2.3.4 (Cost Savings in BP vs. FFS)**.** *When bundling is feasible and preferred, hospitals with relatively lower physician integration (i.e., $\Psi \leq \bar{\Psi}$) enjoy the highest savings. As the physician integration increases beyond $\bar{\Psi}$, the savings start to decrease and ultimately disappear altogether when physician integration is beyond $\tilde{\Psi}^\sharp := min(\Psi_0^\sharp, \Psi_+^\sharp)$, bounds for the feasibility of bundling from Theorem 2.3.2. More specifically, we have:*

$$\Sigma = \begin{cases} \dfrac{(\Delta r_p(2w_b + w_q) + (\Delta c - 2(1-I_0)w_b)w_b)^2}{4w_b^2(2w_b + w_q)} & \textit{if } \Psi \leq \bar{\Psi} \\ \\ \dfrac{(\Delta c + 2(1-I_0)(w_q + w_b))^2}{4(2w_b + w_q)} - (\Delta c + \Delta r_p + 2(1-I_0)w_q)\Psi + w_q\Psi^2 & \textit{if } \bar{\Psi} \leq \Psi \leq \tilde{\Psi}^\sharp \end{cases} \tag{2.29}$$

Figure 2.6 visualizes Proposition 2.3.4 for a certain set of parameter values. We observe that the results from Proposition 2.3.4 are parallel to those following from its analogue, Proposition 2.3.2. That is, when alignment level is below $\bar{\Psi}$, savings are highest. As the alignment level increases further beyond $\bar{\Psi}$, hospitals have increasingly more influence on physician's choice of care intensity under FFS, and as a result, the room for savings under BP becomes smaller.

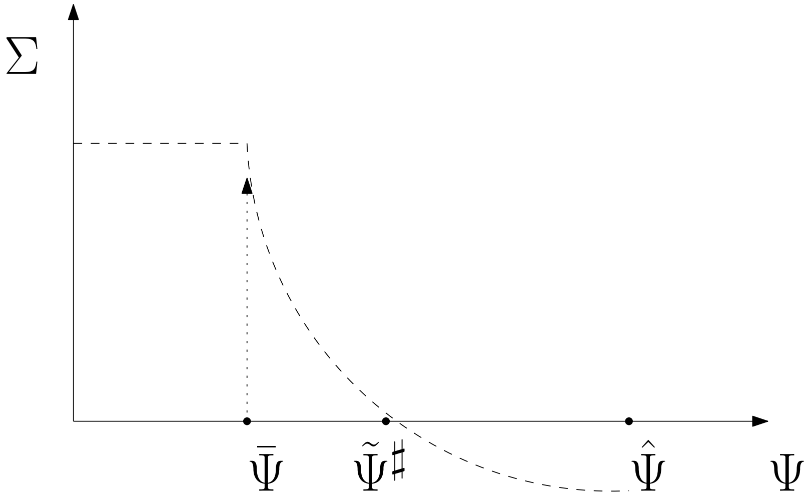


Figure 2.6: Cost Savings in BP vs. FFS.

2.4 Extensions

Hospitals are businesses that provide health services to make profits and rely on the physicians to provide high quality. However, some hospitals may value high-quality care in addition to the quality arising from physicians' altruism. Indeed, health economics literature suggests that non-profit hospitals tend to value quality more than for-profit hospitals (Chang and Jacobson 2012). In addition, reimbursement mechanisms that tie the hospital payments to value (i.e., quality) is becoming more common, as in the case of *performance-based payment models*. Under a performance-based payment model, hospital payments are adjusted based on quality performance (Rosenthal and Dudley 2007), as in the readmission penalties discussed earlier (Andritsos and Tang 2018, Zhang et al. 2016). Consistent with this theme, Section 2.4.1 extends the Initial model to capture the bundling decisions of a hospital whose utility function incorporates the resulting quality into their decision making beyond what physicians decide on. Then, in Section 2.4.2, we extend the Quality-aware model to analyze a setting where physicians are salaried employees of the hospital, physicians and hospitals contractually agree on a compensation with a pre-established performance requirements from the physicians. This latter analysis is relevant and important because several qualitative studies have discussed that different physician compensation models lead to differences in incentives and that hospital-physician integration through physician employment is a promising direction for success under new payment models, including bundled payments (Lee et al. 2012, OMalley et al. 2011). In Section 2.4.3, we extend the Initial model to explore the specification under which the hospital is risk-averse, possibly because a serious downside from the bundled payments might endanger the entire viability of the hospital operations, which was raised as a potential hospital motivation for instance in Dobson et al. (2012). Finally, in Section 2.4.4, we study the "Physician-driven" model where the physicians are the main determinant of the quality of care, a setup that may correspond for instance to physician-driven hospitals. This model can be also seen can be viewed as an opposite to the Salary model from Section 2.4.2 in the continuum of arrangements in terms of hospital power.

2.4.1 Quality-aware Hospital

As we discussed earlier, hospitals differ with respect to quality valuations. In this section, we consider a Quality-aware hospital model, which extends the Initial model by considering utility reduction that quality-conscious hospitals experience as the practiced intensity deviates from the optimal care intensity. We scale the deviation from the optimal intensity with a factor of w_q^{FFS} (respectively w_q^{BP}) under FFS (respectively under BP) which denotes the dollar weight that a hospital places on per unit quality. In the Quality-aware model, the hospital's and physician's utility functions under FFS and BP are given by:

$$\begin{aligned}
F_h^{\text{FFS}} &= r_h - c_1 I(i_h, i_p) - c_2(1 - I(i_h, i_p)) - w_q^{\text{FFS}}(I(i_h, i_p) - I_0)^2, \\
F_p^{\text{FFS}} &= -w_b(I(i_h, i_p) - I_0)^2 + r_{1,p} I(i_h, i_p) + r_{2,p}(1 - I(i_h, i_p)), \\
F_h^{\text{BP}} &= r^{\text{BP}} - (c_1 + r_{1,p}) I(i_h, i_p) - (c_2 + r_{2,p})(1 - I(i_h, i_p)) - w_q^{\text{BP}}(I(i_h, i_p) - I_0)^2 - (1 - i_p)T, \\
F_p^{\text{BP}} &= -w_b(I(i_h, i_p) - I_0)^2 + r_{1,p} I(i_h, i_p) + r_{2,p}(1 - I(i_h, i_p)) + (1 - i_p)T.
\end{aligned} \tag{2.30}$$

Our analysis in this section considers the case where $w_q^{\text{BP}} = w_q^{\text{FFS}} =: w_q$, namely the hospital's valuation of quality does not change in a bundled payment setup as compared with the FFS. Alternatively, in the Appendix, we also consider another practical case where $w_q^{\text{BP}} > w_q^{\text{FFS}}$, which indicates a lower disutility from lower quality under the FFS. This second case captures the possibility that hospitals could make a profit from the readmission stay under the FFS, resulting in a lower care provision during the index admission (while however patients are typically seen by different providers in readmissions, compared with index admissions). In contrast, under a bundled payment setup, the initial admission and readmissions will be bundled into one as both are part of a single episode.

Overall, our results resemble those of the Initial model with the exception that under the Quality-aware model, bundling is more difficult and the implied quality is no lower than that of the Initial model. In particular, under the quality-aware hospital setup, some FFS arrangements are a priori strictly superior to any bundling scenario (Corollary 2.4.1) where physicians' and payer's incentives are no longer fully aligned (Proposition 2.4.1). Although bundling is less likely under the Quality-aware model, when realized, the quality may be higher than that under the Initial model (Corollary 2.4.3). We proceed with presenting our main findings and start with a result which is the analogue of Lemma 2.2.3 in the Initial model.

Lemma 2.4.1 (Status-quo intensity under FFS)**.** *The equilibrium intensity under FFS I^* is given by*

$$I^* = \begin{cases} I_0 + \frac{\Delta r_p}{2w_b} & \textit{if } \Psi \leq \bar{\Psi} \\ 1 - \Psi & \textit{if } \bar{\Psi} \leq \Psi \leq \hat{\Psi} \\ I_0 - \frac{\Delta c}{2w_q} & \textit{if } \Psi \geq \hat{\Psi} \end{cases}$$

where

$$\hat{\Psi} := 1 - I_0 - \frac{\Delta c}{2w_q}$$

Note that the first two cases correspond to the cases in Lemma 2.2.3. The third case occurs in hospitals with highly aligned physicians, where setting intensity as $1-\Psi$ as in Lemma 2.2.3 would imply much deviation from the optimal care intensity I_0, which is not preferred when both physician and hospital are quality-conscious. As such, in this case, the intensity is bounded below by $I_0 - \frac{\Delta c}{2w_q}$. Next we present a result analogous to Theorem 2.3.1, assuming that the difference between r^{FFS} and r^{BP} will be small enough to allow for hospital profitability, as in Assumption 2.4.1.

Assumption 2.4.1. *$r^{FFS} - r^{BP}$ is not too large; specifically:*

$$0 < r^{FFS} - r^{BP} \leq \begin{cases} \frac{1}{4w_b^2(2w_q+w_b)}(w_q\Delta r_p + w_q(\Delta c + 2\Delta r_p) - 2w_b^2(1-I_0-\Psi))^2 \\ \quad \textit{if}\, T \leq T_{\max},\ \Psi \leq \bar{\Psi} \\ \frac{1}{4w_b^2}(2I_0 w_b + \Delta r_p)(w_q\Delta r_p + 2w_b(-I_0 w_q + \Delta c + \Delta r_p) - 4w_b^2(1-\Psi)) \\ \quad \textit{if}\, T > T_{\max},\ \Psi \leq \bar{\Psi} \\ \frac{1}{4(2w_b+w_q)}(\Delta c + 2w_b(1-I_0-\Psi) + 2w_q(1-I_0-\Psi))^2 \\ \quad \textit{if}\, T \leq T_{\max},\ \Psi \geq \bar{\Psi} \\ (-2I_0(w_b+w_q) + \Delta c + w_q(1-\Psi))(1-\Psi) \\ \quad \textit{if}\, T > T_{\max},\ \Psi \geq \bar{\Psi} \end{cases},$$

Under the Quality-aware model, hospital's and physicians' simultaneous interest in bundling does not necessarily translate into payer benefiting from bundling, which in turn results in many special cases in the solution space. For clarity of the presentation, in the following result, we present only one case, which is analogous to Theorem 2.2.1. The interpretation is also very similar to that of Theorem 2.2.1, but is applicable in a more limited setting. The full set of solutions is included in the Appendix, in Proposition B.1.2.

Proposition 2.4.1 (When do they bundle?)**.** *Suppose the condition in Assumption 2.4.1 holds. Then, hospitals and physicians will bundle if:*

$$\Delta c \geq 2I_0 w_q + 2w_b(1 + I_0 - \Psi),\ \Psi \leq 1 - I_0 - \frac{\Delta r_p}{2w_b}, \tag{2.31}$$

$$\begin{aligned} \frac{\Delta r_p}{2w_b} \cdot (\frac{w_q}{w_b} + 1) + \frac{\Delta c}{2w_b} &> \bar{\Psi} - \Psi, \textit{ and} \\ \frac{\Delta r_p}{2w_b} + \frac{\Delta c}{2(w_b + w_q)} &> -(\bar{\Psi} - \Psi). \end{aligned} \tag{2.32}$$

In the Initial model, getting physicians on board was sufficient to align hospital-physician-payer trio. Under the Quality-aware model, the payer will bundle only in certain cases, of which one is presented in Proposition 2.4.1: when the opportunities for savings are high (implied by $T = T_{\max}$) and the physicians are not aligned well with the hospitals (implied by $\Psi \leq 1 - I_0 - \frac{\Delta r_p}{2w_b}$), bundling would occur as long as Condition (2.32) holds. This condition is an almost immediate equivalent of Condition (2.6) in Theorem 2.2.1 except i) the lower and upper bounds for the distance from the critical threshold for the physician alignment level, $|\bar{\Psi} - \Psi|$, are shifted, ii) the bounds are no longer symmetric around the critical threshold $\bar{\Psi}$, iii) the alignment range where bundling is feasible may be smaller or larger as compared with the Initial model depending on the parameter values.

Corollary 2.4.1. *When $\Psi \geq 1 - I_0 + \frac{\Delta c}{2w_q}$, bundling is not attractive for hospitals and physicians.*

Hospitals can gain from bundled payments by decreasing intensity. However, this will often also decrease quality as well. When a hospital is concerned about the quality (as in the Quality-aware model), an increase in cost savings due to decreasing intensity may or may not outweigh the utility losses from decreasing quality. Many new payment mechanisms tie payments to quality, which in our framework essentially is the value of the hospital quality objective (specifically, it would increase w_q). Corollary 2.4.1 characterizes the region where bundling is not a profitable proposition for the hospitals and the physicians where an increase in w_q suggests a wider range of Ψ values. Hence, increasing w_q makes bundling more difficult. This observation warns CMS and other payers that they must be cautious when combining bundled payments with other quality-improving payment mechanisms (which would make the hospitals value the quality more). Indeed, we already see that hospitals participating in bundled payments complain about such conflicting payment mechanisms (Dummit et al. 2015). An uncareful launch of bundling and performance-based programs can discourage hospitals from bundling or it can lead them to incur losses, if bundling was mandatory. In the following result, we compare how a quality-aware hospital achieves different quality outcomes under bundled payments than a quality-blind hospital as in the Initial model. The Quality-aware model can be considered as the FFS vs. the bundled payment mechanism where the payer simultaneously offers performance-based payment programs to hospitals.

Corollary 2.4.2. *Under bundling, the intensity decreases as compared with the FFS.*

Corollary 2.4.2 compares care intensity under the FFS vs. the bundled payments for quality-aware hospitals, which is in line with the findings from the Initial model (see Corollary 2.3.1).

Corollary 2.4.3.

(1.) If $\Psi \geq \bar{\Psi}$, then the quality under bundling will be the same under the Initial model and the Quality-aware model.

(2.) If $\Psi < \bar{\Psi}$, then the quality under bundling will be higher in the Quality-aware model if

$$\frac{\Delta c - \Delta r_p}{2w_b} < \bar{\Psi} - \Psi. \tag{2.33}$$

and higher in the Initial model otherwise.

Corollary 2.4.3 has important implications for the payer in designing the payment mechanism and setting the quality expectations from it. When combining the bundled payments with performance-based programs (e.g., $w_q > 0$), a payer should be careful in the implementation and should account for i) the hospital-physician alignment levels, and ii) cost and revenue differentials for the disease condition. For hospitals with highly aligned physicians ($\Psi \geq \bar{\Psi}$), the quality will remain same when performance-based programs are jointly offered with bundling. However, for hospitals that work with less-aligned physicians ($\Psi < \bar{\Psi}$), combining performance-based programs with bundling may increase the quality as intended, or may lead to an unintended reduction of quality. Corollary 2.4.1 and 2.4.3 together imply that a payer should carefully balance the expected savings, incentives for bundling, and the quality implications when deciding to jointly offer the bundled and performance-based payment programs.

2.4.2 Physicians as Salaried Employees (Salary model)

In this section, we consider a salaried model setup, which represents a smaller but sizable portion of hospitals. Under the salaried setup, we find that the equilibrium intensity under FFS may be lower or higher in comparison to non-salaried setups (Lemma 2.4.2). However, even when the equilibrium FFS intensity is lower in comparison to non-salaried setups, it is still possible to bundle under some conditions. Further, among hospitals with salaried physicians and similar cost and revenue differentials, those that value quality highly are more likely to bundle; which is in contrast to the finding that quality-aware hospitals that do not employ physicians are less likely to bundle if they value quality highly (Proposition 2.4.2). Comparing the Salary model with the corresponding non-salaried analog based on the Quality-aware model, we find that bundling will be more feasible when cost savings between the pathways are low to moderate (Theorem 2.4.1).

When physicians are salaried employees, they are not active decision makers for bundling; instead the hospital pays physicians a fixed salary and collects the physician's part of the reimbursement itself. Therefore, the modeling is decision-theoretic (henceforth referred to as the Salary model), rather than a game-theoretical model, where $\Psi = 1$. Similar to Section 2.4.1, we consider a setting where the hospital is quality-conscious (i.e. $w_q > 0$), and at the end of this section, we compare the Quality-aware and the Salary models (Theorem 2.4.1). Given this setup, the hospital's utility function under FFS is given by

$$F_h^{\text{FFS}} = r_h - r_S - (c_1 - r_{1,p})I(i_h) - (c_2 - r_{2,p})(1 - I(i_h)) - w_q(I(i_h) - I_0)^2, \quad (2.34)$$

where r_S is the salary paid to physicians, prorated to a single patient visit. On the other hand, under bundled payments, the utility function is given by

$$F_h^{\text{BP}} = r^{\text{BP}} - r_S - c_1 I(i_h) - c_2(1 - I(i_h)) - w_q(I(i_h) - I_0)^2. \quad (2.35)$$

We also assume that the hospital's concern for quality, w_q, and also the salary to the physicians, r_S, is the same under FFS and bundled payments. This is reasonable because hospitals generally do not provide additional financial motivation to physicians, beyond possible gainsharing (Dummit et al. 2015).

Lemma 2.4.2 (Status-quo intensity)**.** *The status quo FFS intensity I^* is given by*

$$I^* = \min[I_0 - \frac{\Delta c - \Delta r_p}{2w_q}, 1]^+. \quad (2.36)$$

Lemma 2.4.2 suggests that the optimal care intensity under the Salary model is lower than the optimal care intensity under the Initial model and it can be lower or higher than the same under the Quality-aware model. Next, we characterize when hospitals with salaried physicians would be better off with bundling.[11]

[11] When bundled payments and FFS lead to same payments to hospitals and cost to the payer, we break the tie in favor of the bundled payments as coordination and less administrative costs are more desirable for the payer.

Proposition 2.4.2 (When do they bundle?). *The hospital will bundle iff*

$$\Delta c < \Delta r_p + 2w_q I_0 \tag{2.37}$$

and r^{BP} is high enough.[12]

Proposition 2.4.2 is a simple yet insightful finding. Based on the inequality (2.37), under the Salary model, the decision to bundle depends on the cost differential between pathways relative to potential gains from bundling. Specifically, the left-hand side of the inequality in (2.37) represents the opportunities for cost savings, and the right-hand side represents potential gains from i) payer's reimbursement of physician services that are paid to the hospital and ii) the value derived from the optimal quality. One important implication of this result is that among hospitals with salaried physicians and similar cost and revenue differentials, those that value quality highly are more likely to bundle. Note that this result is in contrast to the finding that quality-aware hospitals that do not employ physicians are less likely to bundle if they value quality highly. The key intuition for this difference in these two results is the following. In the case of salaried physicians, since the hospital receives a higher physician payment for the more expensive pathway from the payer but pays the physicians a fixed salary, the hospital has an incentive to increase the intensity of care under FFS. However, under bundled payments, the hospital does not have such an incentive and hence tends to reduce the intensity. The increase in hospital utility through this reduction in intensity is higher if the hospital values quality highly, and hence a hospital that values quality highly is more likely to bundle if it has salaried physicians. On the other hand, when the physicians are not salaried employees of the hospital, since the physicians have an incentive to choose a high intensity and the hospital will have to compensate the physicians to reduce the intensity, a hospital that values quality highly will have to compensate the physicians more (or equivalently limit the reduction in intensity consistent with

[12] The specifics for the conditions on r^{BP} are provided in the appendix

the valuation of the quality). This causes bundling to be less profitable compared to FFS for a hospital that values quality highly in case of non-salaried physicians. One example of such hospitals are academic medical centers, which typically salary the physicians and at the same time value quality highly. Indeed, in line with our findings, observational studies show that academic medical centers are more eager to bundle (Tsai et al. 2015).

Proposition 2.4.3 (Optimal solution)**.** *If bundling occurs as outlined in Proposition 2.4.2, then the optimal solution is as follows:*

$$I^{\sharp} = [I_0 - \frac{\Delta c}{2w_q}]^{+}. \tag{2.38}$$

Moreover, the care intensity decreases under bundled payments.

The optimal care intensity under bundling of the Salary model is lower than that of the FFS, and not higher than the optimal care intensity under bundling of the Quality-aware model (Lemma 2.4.1). The comparatively lower care intensity in settings where hospitals and physicians are highly aligned, as in the Salary model, emphasizes the importance of performance-based payment models. To achieve good quality, payers should offer bundling and performance-based payment models simultaneously (i.e., $w_q \gg 0$) to ensure good quality (i.e., $\lim_{w_q \to \infty}[I_0 - \frac{\Delta c}{2w_q}] = I_0$) in salaried settings or settings where hospitals and physicians are highly aligned.

The next theorem compares the Salary model with the Quality-aware model. Although physicians do not manage costs and do not assume any risk for low-quality care in salaried settings, they are highly aligned with hospitals.

Theorem 2.4.1. *Consider a hospital with highly-aligned physicians ($\Psi \to 1$). Then the following provides a characterization of bundling under the quality and the salary models:*

*(1.) There are cases when the hospital would bundle under the Salary model but not under the Quality-aware model, namely, when $\Delta c < 2w_q I_0$ (*relatively low savings*).*

*(2.) There are cases when the hospital would bundle under the Quality-aware model but not under the Salary model, namely, when $\Delta c - \Delta r_p > 2w_q I_0$ (*relatively high savings*).*

*(3.) Bundling may occur under both models when $2w_b I_0 + 2w_q I_0 < \Delta c < \Delta r_p + 2w_q I_0$ (*relatively moderate savings*).*

Theorem 2.4.1 emphasizes the role of salaried physicians and the resulting bundling incentives for hospitals, physicians, and the payers when physicians are highly aligned with the hospitals. The race between the hospital's valuation of quality and the savings opportunities with intensity reduction due to bundling determine the difference between the salaried and non-salaried settings. When hospital's valuation of quality is higher than the cost savings as in part *(1.)* of the theorem, the Salary model indicates presence of sufficient incentives for bundling whereas the independence of physicians under the non-salaried setting suggests FFS as a better option for hospitals and physicians. When the hospital's cost savings relative to physicians' revenue reduction is larger than hospital's valuation of quality as in part *(2.)* of the theorem, the non-salaried setting allows for hospital and physician alignment in bundling decisions whereas the hospitals operating in the salaried setting will not have the incentives to bundle because they would be already operating efficiently under the FFS. When savings are moderate as in part *(3.)* of the theorem, bundling occurs under both models with highly aligned physicians. Neither high nor low savings opportunities provide a balanced incentive environment for both the physicians and the hospital, thus facilitating bundling.

2.4.3 Risk-averse Model

Prior studies on bundled payments have highlighted the role of risk. For instance, an influential analysis (Dobson et al. 2012) commissioned by the American Hospital Association postulated four conditions that a hospital should consider when selecting DRGs suitable for bundling. One of the conditions is the "appropriate amount of variation in Medicare payment to achieve efficiency gains, but not so much that the risk of multiple outlier cases outweighs the reward." The risk enters under the more expansive definitions of bundled payments that also include post-acute care. Indeed, most hospitals do not control post-acute care providers, so these hospitals cannot predict the overall episode costs so well. At the same time, hospitals, especially the smaller ones, tend to be risk averse. This risk aversion then hinders the adoption of bundled payments.

In this risk-averse model, the FFS problem remains unchanged. Indeed, under the FFS, the hospital does not care about post-acute care costs. However, under bundled payments, the hospital suddenly becomes responsible for the these costs, $\tilde{c}_i'$ ($i \in \{1,2\}$), which are highly uncertain. In what follows, we assume that each acute pathway results in different post-acute care costs, and we formulate the hospital's and physician's problem as follows:

$$\begin{aligned} F_h^{\text{BP}} &= r^{\text{BP}} - (c_1 + \tilde{c}_1' + r_{1,p})I(i_h, i_p) - (c_2 + \tilde{c}_2' + r_{2,p})(1 - I(i_h, i_p)) - (1 - i_p)T, \\ F_p^{\text{BP}} &= -w_b(I(i_h, i_p) - I_0)^2 + r_{1,p}I(i_h, i_p) + r_{2,p}(1 - I(i_h, i_p)) + (1 - i_p)T, \end{aligned} \tag{2.39}$$

where $\tilde{c}_i'$ ($i \in \{1,2\}$) are random variables. Therefore, in contrast to the previous models, the hospital is now trying to maximize a stochastic utility function. For instance, if the risk aversion is expressed through the exponential utility function and normally distributed post-acute care costs, model (2.39) results in a hospital-driven mean-variance optimization problem. This simplified problem is what we will focus on. Specifically, this assumption means that $\tilde{c}_i' \sim \mathcal{N}(c_i', \sigma_i^2)$. Then, equivalently, the hospital can optimize the following:

$$\begin{aligned} F_h^{\text{BP}} = r^{\text{BP}} &- (c_1 + c_1' + r_{1,p})I(i_h, i_p) - (c_2 + c_2' + r_{2,p})(1 - I(i_h, i_p)) - (1 - i_p)T \\ &- \frac{\alpha}{2}(\sigma_1^2 I(i_h, i_p)^2 + \sigma_2^2(1 - I(i_h, i_p))^2) \end{aligned} \tag{2.40}$$

where α is the coefficient of risk aversion, and we defined $\zeta_i := \frac{\alpha}{2}\sigma_i^2$. Now, we are ready to state an analog of Proposition 2.3.1:

Proposition 2.4.4 (When do they bundle?). *(1.) There will be a case similar to Proposition 2.3.1 where $T > 0$, with the conditions to satisfy being*

$$\frac{\Delta r_p}{w_b}(\zeta_1 + \zeta_2) + \Delta c + \Delta c' + 2(\Delta r_p + I_0\zeta_1 - (1 - I_0)\zeta_2) > 2w_b(1 - I_0 - \Psi) \tag{2.41}$$

$$\Delta c + \Delta c' > 2((\zeta_1 + \zeta_2)\Psi - \zeta_1 - w_b(1 - I_0 - \Psi)). \tag{2.42}$$

This case is profitable for physicians, and often, it will also be profitable for the payer (paying r^{BP} high enough), as long as ζ_i are not too high.

(2.) There will be cases with $T = 0$ that resemble the cases from FFS, when none of the parties benefits, and the hospital loses from having to bear the risk from post-acute care.

(3.) There will also be a case with $T = 0$ that is different from FFS cases, where the hospital is trying to increase intensity because the cheaper pathway has much uncertainty (ζ_2 high and $\frac{\Delta r_p}{2w_b} < 1 - I_0$). This may be profitable for the physicians and may or may not be profitable for the payer.

2.4.4 Physician-Driven Model

In this section, we analyze an alternative model in which physicians are the sole drivers of the care delivery and are only indirectly influenced by the hospital. This alternative model is in contrast to the model we considered earlier where the care delivery and hence intensity were co-produced by the hospital and physicians (the Coproduction model). We remark that while coproduction of care is applicable to a wide range of hospitals, modeling physicians as the sole driver of care may be appropriate in certain settings where physicians are highly influential. We show throughout the remainder of this section that the results from this alternative model are mostly qualitatively similar to those from the co-production models. For brevity, we present results relating to "when bundling benefits all parties?" and "quality" results, which are slightly different than their analogues in the co-production model case. We present the full set of results from the physician-driven model in Appendix B.2.6.

In physician-driven model, physicians are the sole drivers of the intensity; i.e. we have $I \equiv i_p$. While the hospital has no influence on the intensity directly, it can indirectly influence the physician's behavior. More specifically, when making the care delivery decisions, the physician considers hospital's expectation of cost reduction, which deters the physician from practicing overprovision of care (captured by care intensity in the model). We model this indirect influence of hospital on the physician behavior by $-I\Phi$ in the physician's payoff function, where the intensity I is set solely by the physician and Φ captures the level of hospital's influence. Note that $\Phi = 0$ means physicians are fully autonomous in their decisions with no hospital influence and hospital influence increases with increasing Φ. We then have

$$\begin{aligned} F_p^{\text{FFS}}(I) &= -w_b(I - I_0)^2 + r_{1,p}I + r_{2,p}(1 - I) - I\Phi, \\ F_h^{\text{FFS}}(I) &= r_h - c_1 I - c_2(1 - I), \end{aligned} \tag{2.43}$$

and

$$\begin{aligned} F_p^{\text{BP}} &= -w_b(I - I_0)^2 + r_{1,p}I + r_{2,p}(1 - I) + (1 - I)T - I\Phi, \\ F_h^{\text{BP}} &= r^{\text{BP}} - (c_1 + r_{1,p})I - (c_2 + r_{2,p})(1 - I) - (1 - I)T. \end{aligned} \tag{2.44}$$

Given the payoff functions (2.43) and (2.44) for this alternative model, we assess when bundling is desirable for all parties in the next theorem.

Theorem 2.4.2 (When bundling benefits all parties?)**.** *Hospitals and physicians will bundle if the following condition holds:*

(Hospital/physician coordination condition):

$$\Delta c + 2\Delta r_p > 2w_b(1 - I_0) + \Phi \tag{2.45}$$

We observe that overall, Theorem 2.4.2 is analogous to Theorem 2.3.1 but simpler, with only one condition to satisfy. Specifically, the interpretation for Condition (2.45) is similar to that of Condition (2.17), but we no longer need analogues for Conditions (2.18) and (2.19). This is because, Conditions (2.18) and (2.19) implied low physician integration as necessary conditions, and in physician-driven scenario where physicians are the sole drivers of care delivery, obviously physicians have higher level of autonomy and hospitals have low influence on physicians' care delivery decisions. Hence such conditions are automatically satisfied and no longer needed. Similarly, the results for quality, presented in the next proposition, are analogous to that of Proposition 2.3.1.

Proposition 2.4.5 (Quality under BP vs. FFS)**.** *Compared with FFS, quality of care under BP may increase or decrease, depending on the physician integration level Φ:*

I. *If integration is low, $\Phi < \Delta r_p$, then overprovision of services characterizes FFS; and in that case BP will improve if either*

$$2(1 - I_0)w_b > \Delta c + \Phi \tag{2.46}$$

or

$$2w_b(1 - I_0 + 2\Delta r_p - 2\Phi) > \Delta c + \Phi > 2(1 - I_0)w_b. \tag{2.47}$$

II. *If integration is high, $\Phi > \Delta r_p$, then underprovision of services characterizes FFS; and in that case bundled payments will worsen the quality by magnifying the level of underprovisioning.*

Similar to the Coproduction models, both the level of hospital influence as well as the status quo care intensity under FFS (under- or overprovision of services) determine quality outcomes. Particularly, if overprovisioning characterizes the FFS quality can improve under bundling when hospital influence (and/or potential cost savings) are low enough. If underprovisioning characterizes FFS, however, the quality is expected to deteriorate under bundling.

2.5 A Machine-learning Approach to Identify Common Service Bundles and Clinical Pathways

We demonstrated in Sections 2.2, 2.3, and 2.4 that the costs associated with different clinical pathways play a key role in determining whether the current efforts related to bundling are likely to be successful. Consequently, it is imperative for hospitals and physicians to better understand and manage their combined costs resulting from different clinical pathways before proposing bundled payments to a payer or accepting a payer's proposal for bundled payments. However, identifying the common clinical pathways may be quite complicated (Curran et al. 2005). In particular, although specific pathways are well established and understood by some hospitals, especially the ones with strong information technology capabilities, they are less obvious to most hospitals. Therefore, an important issue to address when considering bundling decisions is how to identify these "naturally-occurring" pathways with different costs via a data-driven approach.

Our purpose in this section is to illustrate a practical machine learning method to identify such pathways using historical data. We remark that the "pathways" that inform bundling decisions may not necessarily represent actual "physical" pathways, but rather different ways of delivering care, resulting in different costs. For instance, a more expensive pathway may mean using a more expensive implant, ordering several unnecessary tests, or prescribing special drugs.

The machine learning method we propose uses an Institutional Review Board-exempted dataset obtained from a hospital specializing in orthopedic surgery. Standard and reproducible orthopedic surgeries such as knee/hip replacement are considered as ideal for bundling. Therefore, to illustrate our proposed approach, we use cost data for DRG 470, corresponding to knee replacement. Our dataset is obtained from a hospital specializing in orthopedic surgery and includes 364 visits of DRG 470 made over a span of about two years, which is a volume comparable with the hospitals that are actually engaged in bundled payments for this DRG under the BPCI program. In addition to detailed information on costs and a breakdown of charges, our dataset included demographic information, insurance type, additional diagnoses, BMI, and a code for attending and assisting physicians. The machine learning method we propose consists of two stages.

(Stage 1): In this stage, we characterize "service bundles" from patient charges using Latent Dirichlet Allocation (LDA) model. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics, namely, a set of tests and services provided (Blei et al. 2003). LDA is commonly used in text mining, where a number of documents is available, and each document comprises several (unobserved) topics with different frequencies. The words in the documents are randomly drawn according to the topics. The goal is to determine how words relate to different topics and how different topics are represented in each document. Analogously, in our problem, each word corresponds to a charged service, each topic corresponds to bundles of services ("service bundles", e.g., a blood draw and laboratory tests) that typically occur together, and each document corresponds to a patient.

(Stage 2): In this stage, we identify the cost clusters using Gaussian mixture regression for a given DRG. The mixing probabilities depend on the service bundles (or "topics") identified in Stage 1 and the costs are controlled for other relevant factors such as patient demographics. These final cost clusters then represent different pathways resulting in different costs for the same DRG.

Results. In our analysis, we assume normally distributed costs within each cluster. We fitted the clusters and costs jointly using the expectation-maximization algorithm and determined the number of clusters using the Bayesian Information Criterion (Leisch 2004).

The analysis of actual data yielded two cost clusters with the following statistics:

- $c_1 \approx \$19400, \sigma_1 \approx \$5000, I \approx 0.18,$

- $c_2 \approx \$8700, \sigma_2 \approx \$260, (1 - I) \approx 0.82,$

Table 2.3: Logistic regression estimates for a model predicting the probability of the expensive pathway using care topics.

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-2.35	0.67	0.0004
topic 1	1.24	0.86	0.15
topic 2	1.94	0.85	0.02
topic 3	-1.10	1.42	0.44
topic 4	0.10	0.96	0.92
topic 5	1.14	0.85	0.18
topic 6	2.62	0.84	0.002
topic 7	3.19	1.03	0.002
topic 8	-1.20	1.36	0.38
topic 9	-0.38	1.11	0.73

where c_i is the estimated average cost for pathway i and σ_i represents the estimated standard deviation of cost in pathway i. Similarly as in our theoretical models, intensity I represents the fraction of patients following the more expensive pathways and corresponds to the mixing probability of the more expensive pathway.

Our LDA analysis revealed that these cost clusters are partially explained by a set of service bundles—topics—as shown in Table 2.3. In the logistic regression estimates, a more positive coefficient for a topic indicates that the patients receiving the related service bundle are more likely to end up in the more expensive pathway. Similarly, a more negative coefficient for the topic indicates that the patients receiving the related service bundle are more likely to end up in the cheaper pathway. Altogether, the degree of presence of service bundles corresponding to the topics determines the overall probability of patient care being in the expensive.

Discussing these results with the executive team of the collaborating hospital, we found out that some of these service bundles do play a substantial role in total cost, and that some of them could be avoidable.[13] For instance, one of the topics with a positive coefficient corresponds to a particular brand of an implant, presumably an expensive one. Another topic (not significantly associated with the expensive pathway) corresponds to lab testing, including services such as sodium testing, potassium testing, chloride sera, hemoglobin testing, hematocrit testing, and a charge for a blood analyzer that the hospital was using. Other topics include drug combinations that are often administered together. The topics in the table that are not significant correspond to treatment patterns that are not correlated with either of the pathways; they instead correspond to treatment variations that on average incur similar costs.

The numerical exercise using real data demonstrated that the machine-learning approach could identify different inpatient clinical pathways. We further emphasize that such identification can be a starting point for hospital management and does not necessarily validate the insights we obtained from the analytical modeling. Our demonstration using a small number of observations suggests that the proposed method can $i)$ help in better understanding of the costs associated with various pathways, $ii)$ can encourage wider implementation of the optimal pathways once hospitals identify proper patient subgroups for each pathway, and $iii)$ can be a primer for an empirical approach when evaluating whether switching to bundled payments is profitable based on the existing care patterns as defined by clinical pathways. Future research can extend our analysis to larger and richer datasets to provide a deeper understanding of cost clusters.

2.6 Discussion

The emerging payment models offer a new business model for healthcare organizations and is set to change the way healthcare is delivered. In this work, we studied alignment

[13] We disguise the service bundles and refer to them as topics in order not to reveal the cost structure of the collaborating hospital.

of hospitals and physicians in the face of bundled payments. While a few previous studies have also studied bundled payments, no prior operations management study has considered care coordination and physician-hospital power struggle, which is acknowledged to be a key factor in this space (Goldsmith et al. 2016). Our study also adds to the understanding of the characteristics of medical conditions that are ideal for bundling. Different medical conditions are characterized by different quality requirements, treatment intensities, and coordination needs (Sood et al. 2011, Tsai et al. 2015); for example, while bundled payments are widely touted for use in knee replacement or hip replacement, they are not brought up as much in the context of stroke, where appropriate quality measures are critical but not yet well developed, costs along the entire episode of care are not well understood, and new payment models and gainsharing are less familiar to physicians-neurologists (McClellan et al. 2014). Our analysis could be a starting point for hospitals in determining required features of a medical condition for structuring the discussions around bundling.

We found that: i) hospitals with very low or very high levels of physician alignment are not ideal for bundling, and they may be worse off under bundled payments compared with FFS; ii) to engage physicians, hospitals need to gainshare, a mechanism that was not available in traditional FFS-based payment models; iii) bundled payments will decrease care intensity and, unless carefully regulated, bundling may also lead to a reduction in care quality, and iv) in an environment where hospitals are also held accountable for quality, the incentives for bundling will differ in hospitals employing salaried physicians than those where physicians are independent contractors.

2.6.1 Managerial and Policy Implications

While the finding that bundling is not ideal for hospitals with very low physician alignment is more intuitive, the finding that hospitals with very high physician alignment are not ideal for bundling is counter-intuitive. Physicians that are not well aligned with hospitals are unlikely to give up on their power, and hence it is more difficulty to coordinate care and reduce system inefficiencies (such as redundant medical exams). On the other hand, while highly aligned hospital and physicians are able to further coordinate care, the extent of inefficiencies, and hence the room for improvement is small in such hospitals. Because the total historical cost under FFS is taken as the benchmark in determining the rates under bundled payments, highly-aligned hospital and physicians may tend not to lower their margins by engaging in bundling initiatives.

There are important implications of our findings around the impact of the extent of physician alignment on hospitals' tendency to bundling. In general, physicians tend to be less aligned in competitive hospital marketplaces with access to alternatives (Wholey and Burns 1991, Burns et al. 2001). In contrast, physicians would expected to be more aligned in markets where health systems dominate, e.g. Intermountain Healthcare dominates in the state of Utah. Because the physician labor market is currently mostly undersupplied (Daly 2016), we anticipate physician presence or absence in competitive or monopolistic markets will determine the alignment spectrum. First, under the voluntary bundled payment models, the predominant form which is likely to grow further, hospitals with highly aligned physicians are unlikely to adopt bundled payments. However, this may not be a concern from a societal perspective, as such hospitals are already performing cost-efficiently and less of a burden in the overall healthcare costs. Second, as hospital with lowly aligned physicians lack the power to coordinate care and induce efficient care, the voluntary bundled payments will be ineffective in reducing inefficiencies and hence unnecessary costs in such hospitals. Third, because hospitals with a moderate-level physician alignment are more likely to embrace bundled payments, such hospital and physician groups are set to gain from the voluntary model. This would further hurt the low-alignment, inefficient hospitals that compete with the moderate-alignment hospitals. This could lead to mergers (which we expect to increase hospital power and thus alignment) or service-line closures, thereby further concentrating the market.

There have been opposing views about the role of gainsharing in bundled payments, which is used by hospitals as a device to coordinate with physicians. On the one hand, gainsharing is recognized to play a crucial role in bundled payments (Froimson et al. 2013), but on the other hand, the importance of gainsharing is underplayed (American Hospital Association 2013). Our findings reinforces the former view that the hospitals and other conveners should take gainsharing into account from the outset, when considering bundled payments. Our study is a first attempt on paving the way for determining how gainsharing proceeds can be quantified (as represented by T in our modeling) and in doing so, how the role of care intensity, quality, and pathways can be assessed.

Third, we showed that the intensity will decrease under bundled payments, but this may sometimes lead to a lower quality. The payer can use two standard approaches to avoid this. First, it may set quality thresholds, and second, it may set stop-gain limits to avoid excessive cost savings. Imposing a quality threshold may be challenging because quality is often multidimensional and defining the right quality measures and risk adjusting based on patient severity may be difficult to achieve (Dranove and Jin 2010). As for the second approach, we have showed that the quality decreases, particularly the case when the opportunities for savings are large. The second approach of stop-gain provisions seem more promising through which hospital and physicians may choose to reduce the truly unneeded care that are also a source of cost. However, stop-gain provisions may further shift the risk to providers and, in return, providers' incentives for bundling may reduce.

Finally, an organizational choice made by hospitals—whether or not physicians are employees—could have an impact on incentives for bundling. Our analysis suggests that, when hospitals value quality highly, a salaried physician setup may be more accommodating for bundling in contrast to a setup where physicians are independent contractors to hospitals. The implications for the hospital management is that, when engaging in bundled payments, the existing physician employment structure should be considered. From a policy-making perspective, organizational structures should be taken into account when offering bundled payments, especially in an era of value-based payment incentives.

2.6.2 Limitations and Future Work

Our model has numerous limitations and could be variously extended. We outline some of these extensions in the upcoming paragraphs. One of the main limitations is that we do not study in depth the post-acute care and associated hospital risk aversion. While we have outlined a simplified model in Section 2.4.3, the risk aversion is probably one of they key drivers of bundled payments non-adoption and hence worthwhile to study in depth. In fact, many hospitals derive profits from their bundles largely because of savings on post-acute care, and for such hospitals, our findings may not be accurate. Although the topic of risk aversion has been touched on by other researchers (Adida et al. 2016), this is an area with much research potential. Relatedly, while we studied the coordination between the hospital and its physicians, researchers will encounter another degree of complexity when also modeling the post-acute coordination. There will often be multiple post-acute providers and decreasing costs for the hospital may mean pitting them against each other, stopping contracting with some and requiring lower costs from others.

Related to care coordination is the problem of convening the bundle. This term refers to the need to administer the money collected from the care episode (bundle) and distribute them to the various care providers, including the hospital, physicians, and post-acute care providers. Currently, most bundled payment models assumes that the administrator, also termed the "convener", is the hospital. However, this arrangement materialized more by convenience and coincidence than contemplation. However, in general, policymakers and healthcare administrators would be interested to know who is the best suited to assume the role of the convener and how to distribute the bundled payment among providers.

While payers can use prevailing FFS prices to initially price the bundle (Dong et al. 2011), it is unclear how bundles would be repriced in future (Rosen et al. 2013). Repricing the bundled payment would be needed because of technological innovation and other input changes. This issue replicates the same problem with original hospital DRG-based payments, which replaced cost-based reimbursement. In the flow of the time, DRG-based payments disconnected from the reality as providers embraced new technology. Therefore nowadays, many DRG are overcompensated and many are too low to cover the costs.

While our model assumes that hospitals and physicians focus on the present, real-world hospital administrators also care for the long term. They need a long-term planning model, particularly with the healthcare changes under way and ahead. For instance, under the recent MACRA act, physicians will have to choose in the upcoming years either to stay under fee-for-service with stringent pay-for-performance incentives or transition to participating in new payment models, including bundled payments. On the other hand, hospitals need to choose which conditions to bundle and possibly which other new payment initiatives to adopt, with the view of learning for the future when new payment initiatives become mandatory.

Although we have touched the question of conflicting incentives in our Quality-aware model, the interaction of bundled payments with other new payment initiatives remains a widely open topic. For instance, the Prometheus bundled payments pilot in California failed partially because it collided with existing capitation arrangements (Ridgely et al. 2014). Also, some providers participate both in bundled payments and accountable care organizations (Dummit et al. 2015). Finally, many providers must participate in "pay for performance" schemes that modify their payoff under FFS. Therefore, it seems that the topic of blended payment models is worth further exploration.

Our model considers only a simple, linear gainsharing arrangement with physicians, but more complicated arrangements exist in practice. One can ask whether hospitals could offer more sophisticated gainsharing contracts to induce physician integration more effective. Furthermore, what if multiple physicians are involved in a bundle, are game-theoretic approaches to gainsharing needed? This endeavor to capture more complex gainsharing arrangement could draw on the extensive literature on optimal contract design (Bolton and Dewatripont 2005). Related to the concerns studied by the contract theory is the common knowledge of the optimal intensity I_0. In our models, we assume that both the physicians and the hospital know the quality-optimal intensity I_0. However, this is often probably not the case. In fact, it seems that physicians vary in intensity of care and rely on incentives and local customs exactly when they are uncertain about what the optimal intensity of care should be (Sirovich et al. 2008).

Another limitation is that we only model two pathways. However, the case of multiple pathways can be partially reduced to the two-pathways case and furthermore the two-pathways case embodies a natural dilemma of the hospital administrator. Specifically, let's assume first that there are multiple pathways. Then when we start bundled payments preparations, we can initially merge the cheaper pathways together and the costlier pathways together, ideally in a way that the hospital saves the most from moving a patient from the costly "superpathway" to the cheap superpathway. Furthermore, if the hospital's process is very complex and there would be many pathways capturing many combinations of care, the hospital could consider as the cheap pathway the "most efficient care" case while as the expensive pathway the most common care or the "least efficient care encountered." Hence, we believe that our decision to consider two pathways does not lose too much generality. Finally, we argue that the two-pathways case capture the most common decision-making of administrators: Indeed, questions asked will typically be binary, such as "should I provide this treatment or not," "should I use the existing implant or search for a cheaper one," "should the patient receive and X-ray or not." More generally, the expensive pathway will often be the status quo while the cheaper pathway will be the standard of care after a potential care redesign.

It is unclear how long the duration of the coverage by bundled payments should be, in other words, how many days from patient hospital admission are covered. Currently, some BPCI models offered 30, 60, or 90 days "episode length", mostly focusing on 90 days. That is, within 90 days, all more-or-less related services must be covered by the providers and are included in the bundle. However, how long and how flexible the episode length or other episode definition should be is a question. This relates to a separate question of defining the bundle: Which services should be included and are related to the bundle? Too narrow definitions stifle provider innovation but too extensive definitions put providers at risk for unrelated costs. This but bundled payments question is closely related to the "warranty" stream of research within the operations management community. Therefore, researchers may begin to explore this question from several extensive warranty literature reviews (Blischke and Murthy 1992, Murthy and Blischke 1992a,b, Murthy and Djamaludin 2002). Particularly relevant in the healthcare context may be the design of bundle episodes with low "failure" rate, episodes that pass the required quality guarantees smoothly (Chen et al. 1998).

In the current model, we also do not consider *fixed implementation costs* of bundling as they would not influence the qualitative outcomes given FFS or bundled payments. However, fixed costs may influence the hospital's decision to actually switch from FFS to bundled payments or the hospital's outcomes if it is forced to bundle. This perspective may justify why CMS has first introduced retrospective bundling schemes, in spite of being critized by some healthcare experts who may view retrospective bundling only as "just a new pay-for-performance system" (Miller 2015). However, fixed implementation costs for a full-fledged prospective bundling scheme from the outset may be prohibitive for most hospitals, as demonstrated for instance by the large-scale failed Prometheus bundling project in California (Ridgely et al. 2014). Therefore, we speculate that CMS only uses retrospective bundling to help hospitals overcome prohibitive initial fixed costs and this is also suggested by how CMS positions its differeent BPCI models and advertises that the hospitals might need to

Given the foundations laid by our and related studies of bundled payments, authors in the coming years need to explore emerging data sources related to bundled payments and validate some of the findings shown here. Furthermore, data will open for investigation many assumptions that we put forward. For example, how does the physician concern for quality compare to the one of the hospital. Or, how to assess the physician level of alignment using modern data sources that were not available 25 years ago when the first physician alignment studies were conducted. Answers to these questions would be of interest not only to academicians but also to hospital administrators and healthcare policymakers interested in new payment models to make the healthcare system more effective, efficient, and affordable.

CHAPTER 3
FLEXIBLE BED MANAGEMENT

3.1 Introduction

About 12.6 million emergency department (ED) visits every year in the U.S. result in a hospital admission (National Center for Health Statistics 2016). Typically, there is a preferred (primary) hospital unit for each admitted patient, depending on the patient's condition. For example, a patient with congestive heart failure might be best served by the cardiology unit, whereas a patient with pneumonia might be better placed in the pulmonary unit. When a bed in a patient's the primary unit is not available, the patient may have to be assigned to a non-primary unit or wait ("board") in the ED. Treatment in a non-primary unit is associated with higher risk of inpatient mortality and increased rate of medical errors (Song et al. 2019, Komajda et al. 2003, Hodgetts et al. 2002, Goulding et al. 2015). On the other hand, extended ED boarding times are also associated with poor health and operational outcomes, such as impaired clinical coordination, increased rate of medical errors, delays in receiving necessary care, ED overcrowding, and ambulance diversion (Moskop et al. 2009b, Institute of Medicine 2006c). Therefore, there is a trade-off between assigning patients to a non-primary unit versus boarding them in the ED, and a lack of clear guidelines regarding these decisions (Proudlove et al. 2007), which we refer to as the "flexible bed management problem."

In this study, we develop data-driven solution approaches to balance the key trade-off faced in bed management: whether to assign an admitted patient to a secondary unit or board in the ED, when a bed in the primary unit is not available. Specifically, we propose a novel "Generalized Reservation and Threshold Reinforcement Learning" (GREAT-RL) policy, which generalizes two commonly used policies in the literature, reservation and threshold policies, and outperforms them in extensive numerical analyses. Furthermore, the GREAT-RL policy can be practically implemented, as it can be parameterized through reinforcement learning.

The remainder of this paper is organized as follows. In Section 3.2, we review the relevant literature. In Section 3.3, we present the model formulation for the flexible bed management problem. In section 3.4, we discuss reservation and threshold policies, their favorable structural properties under special conditions, and their limitations in more general settings. In Section 3.5, we introduce the GREAT-RL policy framework, demonstrate how it generalizes threshold and reservation policies, and describe our reinforcement learning implementation. In Section 3.6, we describe the numerical simulation study and discuss the results. Finally, in Section 3.7, we summarize our findings and provide recommendations to hospital operations managers.

3.2 Literature Review

This research contributes to several streams in the literature, including: (a) queueing systems in healthcare (see Gupta (2013) for a review), and (b) revenue management.

Prior research on improving ED boarding times includes analysing the effect of surge occupancy through econometric analysis (Long and Mathews 2018), studying the effects of bed pooling through discrete-event simulation (Thomas Schneider et al. 2018), identifying novel nurse staffing policies using many-server asymptotics (Véricourt and Jennings 2011), and designing early discharge strategies (Dobson et al. 2010). Other relevant work include scheduling for overflows and secondary assignments between internal units (Thompson et al. 2009) or for stepdown units in intensive care units (ICU) (Armony et al. 2018). Our study differs from these by focusing on the assignments of patients from the ED to the internal units (beds) and primary vs. secondary assignment options during this transition.

Another stream of research focuses on admission decisions to the units, with policies employing future information (Xu and Chan 2016), measurement on different time-scales

(Dai and Shi 2017), coordinated admissions (Helm et al. 2011b), and urgency-based prioritization (Deglise-Hawkinson et al. 2018). While relevant, studies in this stream of literature do not consider primary and non-primary units for patients.

Several papers consider trade-offs between boarding and misallocation (i.e., allocation to secondary units). Kilinc et al. (2019) consider this trade-off in a setup with two patient classes and two units. Griffin et al. (2012) also investigate the trade-off but do not propose an algorithm that would perform effectively across a diverse set of scenarios. Ouyang et al. (2020) capture a similar trade-off between ICU and general ward admissions through a patient-state model where the patient condition deteriorates faster in the general ward. Dai and Shi (2019) tackle the problem using the value function approximation methodology assuming the same treatment time distribution across patient classes.

Revenue management models consider similar demand-supply assignment decisions (Talluri and Van Ryzin 2006). When item substitution is allowed (similar to primary vs. secondary units), customers can choose among several items, with applications in airline overbooking (Karaesmen and Van Ryzin 2004), retail inventory problems (Smith and Agrawal 2000), and assemble-to-order manufacturing (Shumsky and Zhang 2009). These models parallel our notions of assignment to primary and secondary units but do not consider the possibility of waiting for a primary assignment, i.e., boarding is not allowed. Other researchers explore algorithms for dynamic supply-demand matching problems, e.g., allocating inventories to sequentially-arriving demand (Ma and Simchi-Levi 2017), allocating advertisements through online stochastic matching (Bahmani and Kapralov 2010, Manshadi et al. 2012), or assigning vehicles for dynamic routing (Spivey and Powell 2004). These models feature a finite inventory that is assigned whereas in our setting, the inventory (beds) is renewable. Our work also relates to due date management literature in make-to-order manufacturing settings (Savaşaneril et al. 2010, Hafızoğlu et al. 2016, Keskinocak and Tayur 2004) where arriving customers are quoted due dates and then decide whether to place an order; the flexible bed management problem has a similar renewable capacity as in due date management, but all the demand must be met and there are different resource types.

In various contexts, researchers studied reservation and threshold policies (e.g., see Talluri and Van Ryzin (2006)), which we show to be special cases of our proposed GREAT-RL policy (Proposition 3.5.1). More complex policies include simulation-based reservation (Bertsimas and De Boer 2005), limited-information (Lan et al. 2008), and nested reservation policies (Brumelle and McGill 1993). Threshold policies are effective, e.g., in (classical) inventory management (Arrow et al. 1951), rental systems (Papier and Thonemann 2010), and call-center routing (Zhan and Ward 2013). The GREAT-RL policy generalizes these reservation and threshold ideas while also being amenable to reinforcement learning, which allows wider applicability.

3.3 Problem Formulation

To analyze the flexible bed management problem, we employ a queueing framework where patients with different clinical conditions $i \in \{1, \ldots, I\}$ ("patient classes") are assigned to units $j \in \{1, \ldots, J\}$. The notation used throughout the manuscript is summarized in Table 3.1.

In unit j, there are κ_j beds, of which $o_j \leq \kappa_j$ are occupied and $\epsilon_j = \kappa_j - o_j$ are empty at a given time. There is a *misallocation penalty* $\pi_{i,j}$ for assigning a patient of class i to unit j, where $\pi_{i,j} = \infty$ if patient class i cannot be served in unit j. The misallocation penalty would depend on the specific setting (e.g., determined by a team of medical professionals and operations managers in the hospital).

For patient class i: (i) If $j = \arg\min_{j' \in J} \pi_{i,j'}$, then j is a *primary unit* and (i, j) is a *primary pair*, denoted by indicator function $\chi_{i,j} = 1$. (ii) Otherwise, if $\min_{j \in J'} \pi_{i,j'} < \pi_{i,j} < \infty$, unit j is a *secondary unit* and (i, j) is a *secondary pair*. When $\pi_{i,j} < \infty$, the pair (i, j) is an *eligible pair*, denoted by the indicator function $\eta_{i,j} = 1$. The sets of all primary and eligible pairs are denoted by X and E, respectively.

A patient who waits in the ED before being assigned to a bed is a "boarding patient" and incurs a *boarding penalty* b_i per unit time for patient class i. The boarding penalty captures the dynamics that a boarding patient may not receive the most appropriate treatment, while consuming scarce ED resources and potentially preventing other patients from receiving timely care in the ED.

We next describe the system dynamics. At any point in time, the system is in state $S := (\{Q_i\}, \{o_{ij}\})$, where Q_i denotes the number of patients of class i in the boarding queue and o_{ij} is the number of patients of class i served in unit j (hence, $O_j = \sum_i o_{ij}$, the total number of patients in unit j, i.e., occupancy of unit j). When a bed becomes available, it can be assigned to a boarding patient or "reserved" for a future patient. Patients of class i arrive for a bed assignment (i.e., after their assessment in the ED, they are admitted for inpatient treatment and become ready for bed assignment) according to a Poisson process with arrival rate λ_i. When a patient is ready for assignment, depending on the availability of beds and the condition of the patient, she can be either assigned to an eligible unit right away or wait (i.e., join the boarding queue). When a patient from class i is assigned to a bed in unit j, the service time is exponentially distributed with rate μ_{ij}.

The objective is to minimize a weighted sum of boarding and misallocation penalties. We identify effective policies under a variety of parameters and discuss the practical implications of the proposed policies. For the special case where all misallocation penalties are equal, we normalize the misallocation penalty to 1 and set the boarding penalty as $b > 0$ (boarding-to-misallocation ratio).

3.4 Threshold and Reservation Policies

In this section, we structurally analyze threshold and reservation policies, which are commonly used in the literature and constitute the backbone for our proposed GREAT-RL policy in Section 3.5. We start by formally defining threshold and reservation policies in Section 3.4.1 and in Section 3.4.2 we establish the optimality properties of these policies

Table 3.1: Notation for indices, sets, and parameters

$i \in \{1, \ldots, I\}$	Patient classes
λ_i	Arrival rate for patient class i
b_i	Boarding penalty for patient class i
Q_i	Number of patients of class i boarding
$j \in \{1, \ldots, J\}$	Hospital units
κ_j	Capacity (number of beds) of hospital unit j
ϵ_j	Number of empty beds in hospital unit j
χ_{ij}, X	Primary pair indicator function and set of primary pairs
η_{ij}, E	Eligibility indicator function and set of eligible pairs
π_{ij}	Misallocation penalty for pair (i, j)
b	Unified boarding penalty if $b_1 = b_2 = \ldots = b_I$, also referred to as boarding-to-misallocation ratio when $\pi_{ij} = 1\ \forall i, j$
μ_{ij}	Service rate for pair (i, j)
μ_i	Unified service rate for patient class i when μ_{ij} are equal $\forall j$
$O_j,\ o_{ij}$	Occupancy of hospital unit j, total and by class i, respectively
S	System state, $S = (\{Q_i\}, \{o_{ij}\})$

Table 3.2: Notation for calculated metrics used in the policies

W_{ij}	Waiting time of class i for unit j (Section 3.4)
$\tau_{i,j}$	Threshold in the threshold policy (Section 3.4)
Ω_j	Opportunity cost of assigning a patient to unit j (Section 3.4)
B_i	Expected total boarding time for a patient of class i (Section 3.4)
w_i	Maximum elapsed boarding time of class i patient (Section 3.5)
α	Admission function (Section 3.5)
σ	Scheduling function (Section 3.5)
ρ	Primary pair utilization (Section 3.6)
ζ	Standard deviation of service times (Section 3.6)
β	Variation of utilization (Section 3.6)
y_{ij}	Steady state occupancy probabilities for class i in unit j (Section 3.6)

under certain conditions. In Section 3.4.3, we demonstrate scenarios where these policies may not perform well, which motivate our proposed GREAT-RL policy. Please refer to Tables 3.1 and 3.2 for notation.

3.4.1 Policy Definitions

Threshold Policy

In a *threshold policy* (see Algorithm 1), if primary pair (i, j) becomes available for assignment, it is assigned immediately; otherwise, secondary pair (i, j) (if available) is assigned if $Q_i > \tau_{i,j}$, i.e., if Q_i is higher than a certain threshold $\tau_{i,j}$, defined as follows:

```
Algorithm 1: Threshold Policy
Input: Set of thresholds $\tau_{i,j}$
if patient $p$ arrives then
    if primary unit $j$ for $p$ has free beds then
        Assign $p$ to $j$
    else if $J' := \{j : \tau_{i_p,j'} < Q_{i_p}\} \neq \emptyset$ then
        Assign $i_p$ into $j' \in J'$ where $j' = \arg\min_{j \in J} \pi_{i_p,j}$
else if bed becomes available in unit $j$ then
    if primary patient $p$ for $j$ is waiting then
        Assign $p$ to $j$
    else if $I' := \{\iota : \tau_{\iota,j} < Q_\iota\} \neq \emptyset$ then
        Assign $i' \in I'$ into $j$ for $i' = \arg\min_{i \in I} \pi_{i,j}$
```

$$\tau_{i,j} = \left\lfloor \frac{\pi_{ij} + \Omega_j}{b_i \cdot B_i} \right\rfloor, \tag{3.1}$$

Recall that π_{ij} is the misallocation penalty for assigning patient class i to unit j and Ω_j is the opportunity cost of assigning a patient to unit j at this point in time (considering the potential arrival of future patients who may be primary for unit j). B_i is the expected boarding time for a patient of class i at this point in time (until their assignment). Hence, the nominator and the denominator in $\tau_{i,j}$ are the estimated cost and benefit of the assignment (i, j), respectively.

We approximate the opportunity cost Ω_j by considering the expected time until a newly assigned patient departs and accounting for the boarding penalty for primary patients, arriving for unit j in the interim:

$$\Omega_{i,j} = \sum_{i' \in I_j} \frac{\lambda_{i'}}{\mu_i + \sum_{i'' \in I_j} \lambda_{i''}} \cdot \left(\frac{1}{\mu_i} - \frac{1}{\mu_i + \sum_{i'' \in I_j} \lambda_{i''}}\right) \cdot b_{i'}, \text{ where } I_j = \{i' : \chi_{i',j} = 1,\ i' \neq i\}.$$

We estimate the expected boarding time as, $B_i = 1/(\mu_i \sum_j \chi_{ij} \kappa_j)$.

Reservation Policy (Reserve k Beds Policy)

A *reservation policy* allows for secondary assignments but "reserves" a certain number of beds to be used solely by primary patient classes. Formally, if a primary pair becomes available for assignment, it is assigned immediately. Otherwise, secondary pair (i, j) is assigned if o_{ij} and O_j are not "too high." *Reserve k Beds* policy reserves $k > 0$ beds in each unit for the primary class, and the remaining beds can be occupied by secondary classes, as formalized in Algorithm 2.

Algorithm 2: Reserve k Beds policy

```
Input: Reservation level $k$
if patient $p$ arrives then
    if primary unit $j$ for $p$ has free beds then
        Assign $p$ to $j$
    else if secondary unit $j'$ for $p$ has $> k$ empty beds then
        Assign $p$ to $j' = \arg\min_{j'' \in J} \pi_{i_p, j''}$
else if bed becomes available in unit $j$ then
    if primary patient $p$ for $j$ is waiting then
        Assign $p$ to $j$
    else if secondary patients for $j$ are waiting from classes $I'$ and $j$ has $> k$
      empty beds then
        Assign to $j$ a waiting patient from class $i = \arg\min_{i \in I'} \pi_{i,j}$
```

3.4.2 Structural Properties of Threshold and Reserve-k-Beds Policies

In this section, we study the conditions under which reservation or threshold policies are optimal. The insights from this analysis shape our proposed GREAT-RL policy in Section 3.5, which builds upon reservation and threshold policies while addressing their key limitations.

Throughout this section, for analytical tractability, we consider systems with two patient classes and two units and assume that pairs $(1,1)$ and $(2,2)$ are primary. For ease of interpretation, we consider a constant boarding-to-misallocation ratio b. The proofs are presented in Appendices C.1 and C.2.

Theorem 3.4.1 (Optimality of the Threshold Policy)**.** *Suppose that each unit has one bed and that $\mu_{ij} \equiv \mu$ and $\frac{\pi_{ij}}{b} > 1$ for $i \neq j$. Then, a threshold policy, with (class-dependent) thresholds N_i is optimal.*

Next, we proceed with showing the optimality of the reservation policy under special conditions on stationary assignment.

Theorem 3.4.2 (Optimality of the Reservation Policy)**.** *Consider the special case where the first unit has only one bed and $\pi_{1,2} = \infty$ and class 2 patients can be assigned to unit 1 at a stationary rate $\ell_{21} \geq 0$. Then the optimal policy features a nontrivial secondary assignment rate $\ell_{21} > 0$ and reserves the bed in unit 1 for the primary class under the following condition:*

$$0 > \frac{\pi_{21}}{\lambda_2} + b\frac{\partial}{\partial \ell_{21}}[(1-\ell_{21})\,\mathbf{E}\,W_{22}(\ell_{21}) + \ell_{21}\,\mathbf{E}\,W_{21}(\ell_{21})]_{\ell_{21}=0} > -\Xi, \tag{3.2}$$

where $\ell_{22} := \lambda_2 - \ell_{21}$ and Ξ is the marginal boarding cost incurred by patient class 1 due to the assignment of some class 2 patients to unit 1.

We note that Ξ and the waiting times $\mathbf{E}\,W_{22}$, $\mathbf{E}\,W_{21}$ can be computed analytically, as shown in the proof, in the Appendix C.2.

Theorem 3.4.2 shows that reserving beds for primary class patients may be preferable to unrestricted secondary patient assignments. That is, even when a secondary assignment improves the objective function for the patients from the secondary class (the first inequality in (3.2)), that benefit may be outweighed by the longer boarding times (and penalties) for the primary class (the second inequality in (3.2)). Hence, the theorem outlines how the reservation policy improves the objective function by balancing boarding and misallocation penalties among multiple classes, which is an insight motivating the GREAT-RL policy in Section 3.5.

3.4.3 Suboptimality of Threshold and Reserve-k-Beds Policies in Realistic Bed Management Settings

Prior literature has shown that threshold and reservation policies are promising in many applications (e.g., Talluri and Van Ryzin (2006), Papier and Thonemann (2010), Bertsimas and De Boer (2005)). While these policies may be optimal in simplified settings (as shown in Theorems 3.4.1 and 3.4.2), they are likely to be suboptimal in more general and practical settings, particularly in systems with multiple units and patient classes, as we illustrate in the following two examples.

Example 3.4.1. *The Reserve-k-Beds policy is not optimal in a system with three patient classes and three units. In particular, it can be outperformed by a simple dynamic policy.*

We sketch the example here and defer parameter details to the Appendix C.8. Consider a system with three patient classes (C_1, C_2, C_3), three units (U_1, U_2, U_3), each with 1 bed, and the parameters summarized in Table 3.3. Then the Reserve-k-Beds policy will assign C_1 to U_1 or U_2, assign C_3 to U_3, and assign C_2 to U_2 or U_3. However, this policy will be outperformed by a dynamic policy where assignments of C_2 to U_2 are allowed if patients from C_3 are present but not allowed if they are not.

Table 3.3: Parameters for Example 3.4.1.

(a) Misallocation penalties.

	U_1	U_2	U_3
C_1	0	0.1	∞
C_2	∞	0.1	0
C_3	∞	∞	0

(b) Patient class characteristics.

	λ	μ	b
C_1	high	high	high
C_2	high	high	low
C_3	very low	very low	high

Example 3.4.2. *The Threshold policy may not be optimal in a system with three patient classes and three units.*

For the exact parameterization, please again refer to the Appendix C.8. There are again three patient classes and three units, with one bed each. The parameters are described in Table 3.4. A static threshold policy will assign C_2 to U_2, C_3 to U_3, and C_1 to either U_1 or to U_2 if the number of patients from C_1 waiting exceed a threshold. However, this policy, regardless of the threshold selected, can be outperformed by a dynamic policy that assigns C_1 to the secondary unit U_2 if and only if U_3 is empty.

Table 3.4: Parameters for Example 3.4.2.

(a) Misallocation penalties.

	U_1	U_2	U_3
C_1	0	1	∞
C_2	∞	0	0
C_3	∞	∞	0

(b) Patient class characteristics.

	λ	μ	b
C_1	low	low	low
C_2	moderate	moderate	high
C_3	long	long	very high

In summary, while reservation and threshold policies gained much attention in the literature and are promising for the bed management problem, they may perform suboptimally, and even sometimes poorly, due to the aforementioned inherent limitations. In the next section, we propose a generalized policy that builds upon threshold and Reserve-k-Beds policies, while addressing their key limitations, and hence improving their performance.

3.5 Generalized Reservation and Threshold Reinforcement Learning (GREAT-RL) Policy

In this section we propose in Section 3.5.1 the Generalized Reservation and Threshold Reinforcement Learning (GREAT-RL) policy, which introduces flexibility to overcome the limitations of reservation and threshold policies (as discussed in Section 3.4), especially in complex systems with multiple units and patient classes, while maintaining their success under special settings. Furthermore, the GREAT-RL policy can be easily parameterized through reinforcement learning, as illustrated in Section 3.5.2.

3.5.1 Generalized Reservation And Threshold (GREAT) Policy: Definition

GREAT-RL policy, outlined in Algorithm 3, goes through two steps at each iteration: *first*, the policy determines a subset of eligible pairs to allow for assignment, and *second*, it scores each pair in the subset and assigns the pair with the highest score. Formally, the first step is implemented through an *admission function* $\alpha(.)$, where $\alpha(i, j, S) = 1$ if pair (i, j) is recommended for assignment under state S, and $\alpha(i, j, S) = 0$ otherwise. The second step is implemented through a *scheduling function* $\sigma(i, j, S)$, which assigns a real-valued score to each pair. While a well-performing scheduling function $\sigma(.)$ can be derived from the queueing literature, the admission function $\alpha(.)$ is specific to the flexible bed management problem and can be fine-tuned through reinforcement learning as demonstrated in Section 3.5.2.

Algorithm 3: GREAT-RL Policy

```
Input: Admission function $\alpha$; scheduling function $\sigma$
if patient $p$ arrives then
    if primary unit $j$ for $p$ has free beds then
        Assign $p$ to $j$
    else if $A_p = \{j : \alpha(i_p, j, S) = 1\} \neq \emptyset$ then
        Assign $p$ to $j' = \arg\max_{j'' \in A_p} \sigma(i_p, j'', S)$
else if bed becomes available in unit $j$ then
    if primary patient $p$ for $j$ is waiting then
        Assign $p$ to $j$
    else if $A_j = \{i : \alpha(i, j, S) = 1\} \neq \emptyset$ then
        Assign to $j$ a waiting patient from class $i = \arg\max_{i' \in A_j} \sigma(i', j, S)$
```

Next, we show that GREAT-RL policy generalizes Reserve-k-Beds and Threshold policies:

Proposition 3.5.1. *The GREAT-RL policy framework generalizes the Reserve-k-beds policy and Threshold policy. In particular, these two policies can be emulated by the GREAT-RL policy under a particular choice of α and σ.*

The proof is included in the Appendix C.3. In the following section, we demonstrate an RL implementation of our proposed GREAT-RL policy.

3.5.2 A Reinforcement Learning Implementation of GREAT-RL Policy

For Step 1 of the GREAT-RL policy, we consider the Gcμ scheduling function ($\sigma(i, j, S) = b_i \mu_{ij} w_i(S)$) (Van Mieghem 1995).

We estimate admission function $\alpha(.)$ through the one-step actor-critic reinforcement learning algorithm (Sutton and Barto 2018). Two key ingredients of the actor-critic-type algorithms are the *policy function* and the *value function*, both estimated in the course of the algorithm (Sutton and Barto 2018). We estimate the policy function for each eligible pair by inverse logit with the following features: the number of waiting patients (queue) for each patient class (Q_i for $i \in 1, \ldots, I$), and the number of patients from each class assigned (occupancy) in each unit (o_{ij} for $i \in 1, \ldots, I$, $j \in 1, \ldots, J$). We determine the value function *learning rate* using the REINFORCE algorithm in (Sutton and Barto 2018) and the policy function learning rate by a fixed learning rate and an adaptive learning rate (c.f. (Sutton and Barto 2018), Section 9.6).[1]

3.6 Computational Study

We compare the performances of the GREAT-RL and benchmark policies via a computational study. Section 3.6.1 describes the computational setup, Section 3.6.2 lists the benchmark policies, and Section 3.6.3 presents the results.

3.6.1 Experimental Design

Parameterization

In the computational study, we vary four parameters:

[1] The later numerical study includes for each scenario the learning rate that performed better for that scenario.

- *Utilization of primary pairs* $\rho = \sum_i \lambda_i / \sum_i \mu_i$
- *Boarding-to-misallocation ratio* b
- *Standard deviation of service time distribution* ζ, such that $\mu_i = 1 + \zeta\Delta\mu_i$ where $\mu_i \sim N(0,1)$ for every $i \in I$
- *Variation in utilization of primary pairs* β, such that $\rho_i = \lambda_i/\mu_i \sim \text{Beta}(\alpha, \beta)$, where α is selected such that $\mathbf{E}[\sum_i \lambda_i / \sum_i \mu_i] = \rho$

The considered intervals for these parameters and for constants are listed in Table 3.5. For a single set of parameters, we use the term *scenario*. For each scenario, we generate $\mu_i,\ \lambda_i$ and the interarrival and service time random variables multiple times, calling each one of these instantiations an *instance*. Each instance is used to evaluate each policy, using common random numbers (Nelson 2013). The GREAT-RL admission function is estimated on separate instances before the main evaluation. We generate 24 scenarios based on the Latin hypercube design and generate 10 instances for each scenario, for a total of 240 instances, which is a sufficient size to statistically test the mean difference in performance metrics between policies. Each instance consists of 50,000 arrival events, selected empirically based on the variance of the per-patient reward. There is a warm-up period of 2,500 arrivals selected empirically based on average waiting times.

Table 3.5: List of Parameters

Parameter	**Interval**
Number of patient classes	4
Number of units	4
Number of beds per unit	10
Fraction of primary pairs	0.3
Utilization of primary pairs ρ	(0.5, 0.99)
Boarding-to-misallocation penalty ratio b	(0.1, 1.0)
Standard deviation of service times ζ	(0.0, 0.5)
Variation of utilization β	(1.0, 5.0)

Details of parameter estimations and scenario generation are described in the Appendix C.4. The details of the implementation are provided in the Appendix C.10.

3.6.2 Benchmark Policies

We consider four benchmark policies: Strict, Greedy, Generalized $c\mu$ ("Gcμ"), and LEWC-p. The Strict and Greedy policies are simple baselines while Gcμ and LEWC-p had been considered in the literature. The *Strict policy* assigns a patient to a bed in their primary unit if such a bed is available and boards the patient otherwise. The *Greedy policy* assigns patient to the unit with the lowest misallocation penalty available on patient arrival. When a bed becomes available, the policy assigns the patient with the lowest misallocation penalty. The *Gcμ policy* (Van Mieghem 1995), assigns patient p from patient class i_p such that the following product is maximal:

$$w_p \cdot b_{i_p} \cdot \mu_{i_p}, \tag{3.3}$$

where w_p is the longest waiting time for a currently waiting patient from patient class i_p (Van Mieghem 1995). The *LEWC-p policy* extends the GCμ policy by adding to (3.3) a term proportional to π_{ij} (Kilinc et al. 2019).

Observe that the GREAT-RL policy generalizes all these four policies because the admission and scheduling functions in GREAT-RL can be chosen as follows:

- For Greedy policy: $\alpha(i,j) = 1,\ \sigma(i,j) = -\pi_{ij}$ (Section 3.6.2).
- For Strict policy: $\alpha(i,j) = 0,\ \sigma(i,j) = -\pi_{ij}$ (Section 3.6.2).
- For Gcμ policy: $\alpha(i,j) = 1,\ \sigma(i,j) = b_i \mu_{ij} w_i(t)$ (Van Mieghem 1995).
- For LEWC-p policy: $\alpha(i,j) = 1\{b_i/\mu_i > \pi_{ij} y_{ij}\}$ and then

$$\sigma(i,j) = \frac{b_i Q_i}{\sum_{j \in J} y_{ij} \mu_i} - \pi_{ij} \frac{y_{ij} Q_i}{\sum_{j \in J} y_{ij}}, \tag{3.4}$$

  where y_{ij} are estimated optimal steady state occupancy probabilities (Kilinc et al. 2019).

3.6.3 Results

Overall Performance

In Figure 3.1, each point corresponds to the *relative normalized reward* (relative to the Greedy policy) for one policy evaluated on one instance. The relative normalized reward for policy P on an instance is defined as $\frac{-\sum_k(\pi^P_{p_k}+b^P_{p_k})}{|\{p_k\}|} - \frac{-\sum_k(\pi^G_{p_k}+b^G_{p_k})}{|\{p_k\}|}$, where $\{p_k\}$ is the set of patients in the instance, π_{p_k} and b_{p_k} denote the actual realized values of misallocation and boarding penalties for the patient and superscripts P and G denotes whether the realized value was under the evaluated policy or the Greedy policy. The boxplots visualize the distribution of the relative normalized reward for a policy across all instances. The differences in means are statistically significant at level 0.05.

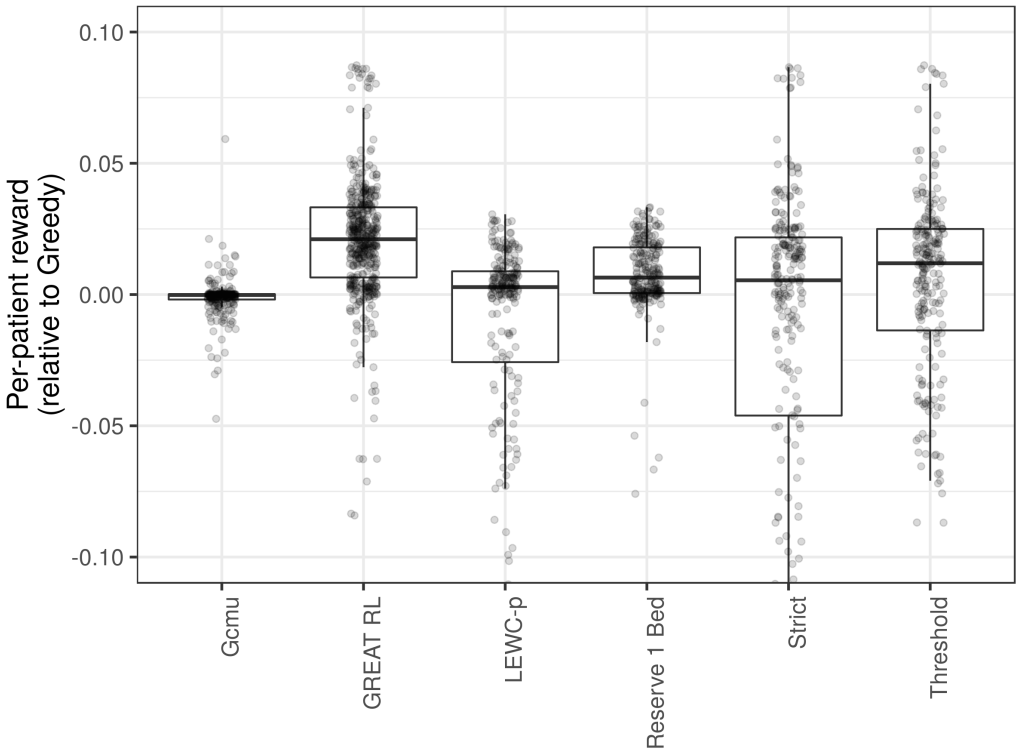


Figure 3.1: Distribution of relative normalized rewards (see the text for the definition) by policy

Figure 3.1 illustrates that the GREAT-RL policy outperforms the benchmarking policies. The Strict policy performs well when the boarding-to-misallocation ratio and utilization are low but performs poorly in other cases. Gcμ performs or slightly better than the Greedy policy. LEWC-p performs slightly better than GCμ in some cases but poorly in other cases.

Performance in Different Parameter Regions

We assess the sensitivity of policy performance to the parameters specifically looking at a) the utilization and b) the boarding-to-misallocation ratio. The results are summarized in Figure 3.2: The GREAT-RL policy performs strongly in regions with moderate utilization and boarding-to-misallocation ratio (the middle region in Figure 3.2). When either the utilization or the boarding-to-misallocation ratio are high (the upper right corner in Figure 3.2), the boarding penalty dominates, and the performance of GREAT-RL compares to that of Gcμ. If either the utilization or the boarding-to-misallocation ratio are very low (the bottom left corner of Figure 3.2), the Strict policy performs strongly, and the assign-

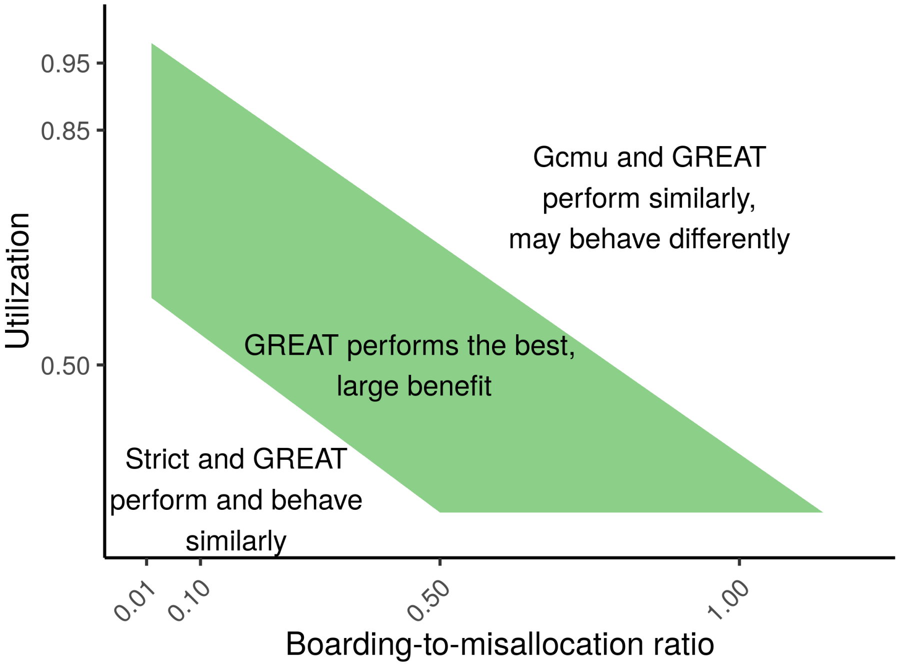


Figure 3.2: Performance in Different Scenario Regions: Illustrative Summary.
Note 1: “Behavior” refers to the structure of the policy, “performance” refers to the value of the objective function.
Note 2: Specific values of utilization and boarding-to-misallocation ratio are illustrative and will depend on the hospital layout and other parameters.

ment decisions taken by GREAT-RL policy tend to mimic those of the Strict policy. More detailed discussion is provided in the Appendix C.11.

Behavior of the GREAT-RL policy

This section interprets the *behavior* of the GREAT-RL policy, i.e. how the policy actually assigns patients to units:

1. When the utilization is low or the boarding-to-misallocation ratio is low (as shown in the lower left region of Figure 3.2), then the GREAT-RL policy tends to only assign primary pairs, behaving similarly to the Strict policy.

2. When the utilization is very high or the boarding-to-misallocation ratio is very high (as shown in the upper right region of Figure 3.2), then the GREAT-RL policy tends to treat primary and secondary assignments equivalently, behaving similarly to the Gcμ policy.

3. In the middle region of Figure 2, the GREAT-RL policy tends to only assign a certain subset of secondary pairs and treat them equivalently with primary pairs, hence behaving similarly to a static GCμ policy with a smaller set of eligible pairs.

4. Finally, there are cases in the middle region where the GREAT-RL policy tends to be a more complex dynamic policy.

3.7 Conclusions

In this paper, we considered the flexible bed management problem: how to assign beds in hospital internal units in the presence of boarding and non-preferred, secondary units. This is a critical problem because patients suffer both when they are assigned to their non-preferred unit as well as when they board for too long. The problem is further complicated by the patient competition for a limited number of beds. To address the problem, we first prove two results on optimality properties of threshold and reservation policies under special conditions with two units and two patient classes. However, these policies are not optimal in general scenarios with multiple units and patient classes. Thus, we propose a novel Generalized Reservation and Threshold Reinforcement Learning policy (GREAT-RL), which generalizes the threshold and reservation policies, performs well in a wider range of settings, and can be easily parameterized using reinforcement learning. In an extensive simulation study, the GREAT-RL policy outperforms both naive and state-of-the-art policies on a set of diverse and complex scenarios.

3.7.1 Managerial Implications

Our results offer intuitive bed management strategies for hospital operations managers. The choice of the strategy depends on the utilization and boarding-to-misallocation ratio in the hospital. The utilization figures are usually available to the operations managers through the midnight census data. The boarding-to-misallocation ratio is less commonly known, but we estimate this ratio to range between 0.05 and 0.30 when patients are assigned to clinically inappropriate units (e.g., surgery unit instead of cardiology unit) and higher when they are assigned to acceptable but still non-primary units (e.g., internal medicine instead of cardiology).[2]

Equipped with their estimates of utilization and boarding-to-misallocation ratio, the operations managers can adopt the following heuristic strategies derived from Figure 3.2 and Section 3.6.3:

1. When utilization is low (e.g., below 50%) *and* secondary units are medically inappropriate, then operations managers should only assign to primary units.

2. When utilization is very high (e.g., above 90%) or secondary units offer a quality of care comparable to the primary units, then operations managers may ignore the dis-

[2] We derive this range by combining the estimates from Song et al. (2019), Sun et al. (2013), and McCarthy et al. (2009b).

tinction between the primary and secondary units and purely maximize throughput, adopting for instance the G$c\mu$ policy.

3. In other cases, the operations manager should allow a limited number of secondary pairs to be assigned without limits while disallowing all other secondary assignments. To determine the exact number of allowed secondary pairs, the operations manager can gradually allow additional secondary pairs until the emergency department is no longer crowded. When selecting secondary pairs, the operations manager should first focus on patient classes that suffer from the longest boarding times and the secondary pairs that are the most medically appropriate.

Appendices

APPENDIX A

HEALTH INFORMATION EXCHANGES: SUPPLEMENTAL CONTENT

A.1 Correlation Matrix

Table A.1: Correlation matrix between study variables

	Duration	HIE	Crowded	Teaching	Charlson	Transport	Large Bedsize	Weekend	Monday	Female	White	Black	Asian	Hispanic	Medicare	Medicaid	Private	Unins.	High Income
Duration	1.00																		
HIE	0.00	1.00																	
Crowded	0.00	0.00	1.00																
Teaching	0.01	0.39	0.00	1.00															
Charlson	0.00	0.00	-0.01	0.01	1.00														
Transport	0.01	0.01	-0.02	0.05	0.09	1.00													
Large Bedsize	0.01	-0.11	0.00	0.00	0.00	0.03	1.00												
Weekend	0.00	-0.01	0.00	-0.01	0.00	-0.02	-0.01	1.00											
Monday	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	-0.27	1.00										
Female	0.00	-0.01	0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	1.00									
White	0.00	-0.09	-0.01	-0.20	0.02	0.04	-0.12	0.02	-0.01	-0.01	1.00								
Black	0.00	0.15	0.00	0.19	-0.01	0.00	0.10	-0.01	0.00	0.00	-0.51	1.00							
Asian	0.00	-0.03	0.01	0.05	-0.02	-0.04	0.07	-0.01	0.01	0.01	-0.64	-0.15	1.00						
Hispanic	0.00	0.01	0.00	0.04	0.00	-0.01	-0.03	0.00	0.00	0.00	-0.19	-0.04	-0.05	1.00					
Medicare	0.00	-0.01	-0.01	-0.03	0.15	0.19	-0.01	0.00	0.00	0.03	0.13	-0.06	-0.09	-0.03	1.00				
Medicaid	0.00	0.00	0.02	0.00	-0.04	-0.05	0.06	-0.01	0.00	0.05	-0.20	0.07	0.18	0.00	-0.31	1.00			
Private	0.00	0.01	0.00	0.03	-0.04	-0.06	-0.04	0.03	0.00	0.01	0.10	-0.04	-0.09	0.02	-0.34	-0.44	1.00		
Uninsured	0.00	-0.03	0.00	-0.01	-0.03	-0.04	0.02	-0.01	0.00	-0.06	-0.03	0.03	0.01	-0.01	-0.14	-0.18	-0.20	1.00	
High Income	0.00	0.04	0.00	-0.04	0.02	0.00	-0.18	0.01	0.00	0.00	0.32	-0.15	-0.28	0.02	0.04	-0.20	0.15	-0.02	1.00

A.2 Disease Analysis

Here, we provide supplementary results for the subsets of patients corresponding to major diseases. This analysis can be seen as a supplement of the severity and complexity analysis.

The effect of HIE on LOS is also likely to vary across disease groups. In particular, for some diseases, HIE can directly reduce the over-utilization of ED resources and hence reduce LOS (e.g., by reducing the rate of radiological imaging or laboratory testing (Lammers et al. 2014, Frisse et al. 2012)), while for others, HIE may just provide more clinical background, leading to no change or even an increase in LOS. This line of reasoning is further supported by the observations that clinicians use HIE less frequently for certain conditions, presumably expecting that more information would change their behavior little (Vest et al. 2011).

To analyze possible differences in HIE's value in terms of the relationship with LOS in the context of different diseases/conditions, we build upon the setup for the Index Visit Analysis and separately run regression models for the visit subsets belonging to select conditions. In particular, consulting with an ED physician as well as screening the HIE literature in the ED setting, we choose two diseases/conditions to analyze the HIE and LOS relationship in the specific context of revisits to the ED for 1) strains-and-sprains, and 2) headache.

When we consider the effect of HIE in the context of patients visiting the ED for specific diseases/conditions, we find that the magnitude of the HIE effect differs by the disease/condition considered. In particular, the results presented in Table A.2 show that the average ED LOS decreases substantially for headache (by 11.4%) and less so for strains-and-sprains (by only 3.6%). This confirms our initial expectations based on our discussions with ED physicians that there is significant variation in headache cases, which can originate from several underlying problems (e.g., inflammation of cranial and spinal nerves, migraine, blood clots), and "in deciding which test to perform, emergency physicians must assess pretest risk for the condition” (Edlow et al. 2009). Hence, additional information obtained through HIE from the previous headache-related visits may allow the physicians to better assess the risk and identify the needed tests more effectively. This in turn may lead to lower LOS. A typical case would be drug-seeking patients presenting in the ED with headache. Without prior information, a physician needs to perform substantial testing to discover that patients’ concerns are not genuine. However, if HIE is present, the physician sees that the patient had multiple prior visits to other providers, and possibly these even include a note that the patient sought drugs. In contrast to headache, strains-and-sprains typically require a diagnosis by an x-ray or a workup with little patient-to-patient variation (Marx et al. 2010). Hence, the physicians will usually not change the treatment choices much regardless what they see in a patient’s prior history. Hence, HIE is not expected to improve LOS much, which is also what we observe.

Table A.2: Regression results for disease analyses

Coefficient	**Headache Index**		**Sprains Index**	
HIE_Tr	-11.4%	-0.12 (0.021)	-3.6%	-0.037 (0.017)
Crowded	6.2%	0.061 (0.008)	9.0%	0.086 (0.006)
Teaching	14.5%	0.135 (0.008)	25.1%	0.224 (0.006)
Charlson	29.2%	0.256 (0.026)	21.4%	0.194 (0.036)
Transport	17.0%	0.157 (0.01)	38.5%	0.326 (0.008)
Controls:Visit		x		x
Controls:Hospital		x		x
FE: Year		x		x
FE: Hospital				
N		47,914		75,366
R^2		0.06		0.08

Note: The first number denotes the estimated percentage change, the second number the corresponding regression coefficient, and the third number the robust standard error. The symbol “x” indicates the presence of the specific controls or the fixed effects in the model while no entry suggests they are absent.

A.3 Going Paperless

Here, in Table A.3, we present the second-stage results for the “going paperless” IV. These results are based on the restricted dataset, 2009-2011.

Table A.3: Regression results for Going-paperless analysis

Coefficient	**Overall IV**	
HIE_Tr	-12.7%	-0.136 (0.012)
Crowded	6.5%	0.063 (0.001)
Charlson	19.4%	0.177 (0.002)
Transport	44.5%	0.368 (0.001)
Controls:Visit	x	
Controls:Hospital		
FE: Year	x	
FE: Hospital	x	
N	5,795,033	

Note: The first number denotes the estimated percentage change, the second number the corresponding regression coefficient, and the third number the robust standard error. The symbol “x” indicates the presence of the specific controls or the fixed effects in the model while no entry suggests they are absent.

APPENDIX B

BUNDLED PAYMENTS: SUPPLEMENTAL CONTENT

B.1 Supplemental Results

B.1.1 Cases for Figure 2.6

Here, we discuss the alternative cases for Figure 2.6 in the main text. Consider Ψ-vs-r^{BP} plots. Consider the following alignment thresholds:

- $\Psi^{-} = 1 - I_0 - \frac{\Delta c + 2\Delta r_p}{2w_b}$: No bundling below this point
- $\bar{\Psi} = 1 - I_0 - \frac{\Delta r_p}{2w_b}$: The threshold between the low-alignment and high-alignment bundling regimes
- $\Psi_{\max} = 1 - I_0 + \frac{\Delta c - 2\Delta r_p}{2w_b}$: The maximum of the quadratic portion of r^{BP}, assuming it is attained.
- $\Psi_T = 1 + I_0 - \frac{\Delta c}{2w_b}$: The threshold beyond which r^{BP} ceases to be quadratic and becomes linear.
- $\Psi^{+} = 1 - I_0 + \frac{\Delta c}{2w_b}$: No bundling above this point

In addition to these definitions, we observe that r^{BP} is continuous in $\bar{\Psi}$ (whether it is quadratic or linear at that point). With these definitions, we immediately see several aspects:

- If $\frac{\Delta c}{2w_b} > I_0$, then $\Psi^{+} > 1$ so bundling occurs anytime when there is high alignment. At the same time, $\Psi_T < 1$, so r^{BP} is linear on the right part. In contrast, if $\frac{\Delta c}{2w_b} < I_0$, then r^{BP} is entirely quadratic and bundling does not occur under very high levels of alignment.

- If $\Delta c > \Delta r_p$, then $\bar{\Psi} < \Psi_{\max}$, else $\Psi_{\max} < \bar{\Psi}$.

- $\Psi_{\max} < \Psi^+$, so the maximum is attained unless the quadratic part is cut off by the linear part ($\Psi_T < \Psi_{\max}$).

- If $I_0 + \frac{\Delta c + 2\Delta r_p}{2w_b} < 1$, then alignment for bundling is bounded below ($\Psi^- > 0$), otherwise it is not.

- If $\frac{\Delta c}{2w_b} > \frac{\Delta r_p}{2w_b} + I_0$, then the linear part of r^{BP} is increasing, else it is decreasing. If it is increasing, it must be that $\bar{\Psi} < \Psi_{\max}$. Also, this condition is equivalent with $\Psi_T < \Psi_{\max}$.

Let's consider the cases for $r^{\text{FFS}} - r^{\text{BP}}$:

- If $\Psi > \bar{\Psi}$ and $T < T_{\max}$, then

$$r^{\text{FFS}} - r^{\text{BP}} = w_b \frac{(\Psi - (1 + \Delta c/(2w_b) - I_0))^2}{2}$$

- If $\Psi > \bar{\Psi}$ and $T = T_{\max}$, then

$$r^{\text{FFS}} - r^{\text{BP}} = 2w_b(1 - \Psi)\frac{\Delta c/(2w_b) - I_0}{2}$$

- If $\Psi < \bar{\Psi}$ and $T < T_{\max}$, then

$$r^{\text{FFS}} - r^{\text{BP}} = w_b \frac{(\Psi - (1 - I_0 - (\Delta c + 2\Delta r_p)/(2w_b)))^2}{2}$$

- If $\Psi < \bar{\Psi}$ and $T = T_{\max}$, then

$$r^{\text{FFS}} - r^{\text{BP}} = 2w_b(\Psi - (1 - (\Delta c + 2\Delta r_p)/(2w_b)))\frac{\Delta r_p/(2w_b) + I_0}{2}$$

We now distinguish cases where variants of Figure 2.6 would be qualitatively different:

(1a) $\Psi^+ < 1,\ \Psi^- > 0$, then $\bar{\Psi}$ can be anywhere relative to $\Psi_{\max}$.

(1b) $\Psi^+ < 1,\ \Psi^- < 0$, then $\bar{\Psi}$ can be anywhere relative to $\Psi_{\max}$.

(2a+2b) Then, $\Psi^+ > 1,\ \Psi_T < \Psi_{\max}$, and the linear part of r^{BP} is increasing. We further explore what happens to $\bar{\Psi}$ and Ψ^-: We have $\Psi_T > \bar{\Psi}$ when $2I_0 > \frac{\Delta c - \Delta r_p}{2w_b}$ and this holds when the linear part of r^{BP} is increasing because $\Delta c > \Delta r_p$. We can have Ψ^- either positive or negative, both are possible. The derivative of r^{BP} is continuous at Ψ_T.

(3a+3b) $\Psi^+ > 1,\ \Psi_{\max} < \Psi_T$, and the linear part of r^{BP} is decreasing. We further explore what happens to $\bar{\Psi}$ and Ψ^-. We have $\frac{\Delta c}{2w_b} < I_0 + \frac{\Delta r_p}{2w_B}$. Since $\Psi^+ > 1$, we have $\frac{\Delta c}{2w_b} > I_0$. The relationship between Δc and Δr_p is unclear. In case $\Delta c < \Delta r_p$ (so $\bar{\Psi} > \Psi_{\max}$), the inequality $\bar{\Psi} > \Psi_T$ can only hold if $2I_0 < \frac{\Delta c - \Delta r_p}{2w_b}$, which is impossible because $\Delta r_p > \Delta c$. Hence $\bar{\Psi} < \Psi_T$. It seems that Ψ^- can be either positive or negative.

B.1.2 Solutions of the Quality Model

We assume $0 < w_q^{\text{FFS}} \leq w_q^{\text{BP}}$ are hospital quality concerns under FFS, respectively BP, and $w_b > 0$.

Proposition B.1.1 (Optimal solution)**.** *The optimal solution cases in the Quality model are as follows:*

1. *If*

$$\begin{aligned} \frac{\Delta r_p}{w_b} + \frac{\Delta c + 2\Delta r_p}{w_q^{BP}} &> 2\frac{w_b}{w_q^{BP}}(1 - I_0 - \Psi), \\ \Delta c &> 2(w_b + w_q^{BP})(-1 + I_0 + \Psi), \\ \Delta c &\leq 2I_0 w_q^{BP} + 2w_b(1 + I_0 - \Psi) \end{aligned} \tag{B.1}$$

then

$$\begin{aligned} i_h &= 1 \\ i_p &= \frac{1}{2w_b(1-\Psi)}(\Delta r_p - \frac{T}{1-\Psi} + 2w_b I_0) \\ T &= T_{low} \equiv \frac{(1-\Psi)}{w_b}\frac{\Delta r_p(w_q^{BP}/w_b) + \Delta c + 2\Delta r_p + 2w_b(-1+I_0+\Psi)}{2w_b + w_q^{BP}} \end{aligned} \tag{B.2}$$

2. *If*

$$\begin{aligned} \frac{\Delta r_p}{w_b} + \frac{\Delta c + 2\Delta r_p}{w_q^{BP}} &> 2\frac{w_b}{w_q^{BP}}(1 - I_0 - \Psi), \\ \Delta c &> 2(w_b + w_q^{BP})(-1 + I_0 + \Psi), \\ \Delta c &\geq 2I_0 w_q^{BP} + 2w_b(1 + I_0 - \Psi) \end{aligned} \tag{B.3}$$

then

$$\begin{aligned} i_h &= 1 \\ i_p &= 0 \\ T &= T_{\max} \equiv \frac{(1-\Psi)}{w_b}(\frac{\Delta r_p}{w_b} + 2I_0) \end{aligned} \tag{B.4}$$

3. *If*

$$\begin{aligned} \frac{\Delta r_p}{w_b} + \frac{\Delta c + 2\Delta r_p}{w_q^{BP}} &< 2\frac{w_b}{w_q^{BP}}(1 - I_0 - \Psi) \textit{ or} \\ \Delta c &< 2(w_b + w_q^{BP})(-1 + I_0 + \Psi) \textit{ and} \\ 2w_b(1 - I_0 - \Psi) &\geq \Delta r_p \end{aligned} \tag{B.5}$$

then

$$\begin{aligned} & i_h \in [[\frac{2w_b(\Psi - I_0) - \Delta r_p}{\Psi}]^+, 1] \\ & i_p = \frac{1}{2w_b(1-\Psi)}(\Delta r_p + 2w_b(I_0 - (1-i_h)\Psi)) \\ & T = 0 \end{aligned} \tag{B.6}$$

Then $I^\sharp = I_0 + \frac{\Delta r_p}{2w_b}$.

4. *If*

$$\begin{aligned} \frac{\Delta r_p}{w_b} + \frac{\Delta c + 2\Delta r_p}{w_q^{BP}} < 2\frac{w_b}{w_q^{BP}}(1 - I_0 - \Psi) \textit{ or} \\ \Delta c < 2(w_b + w_q^{BP})(-1 + I_0 + \Psi) \textit{ and} \\ 2w_b(1 - I_0 - \Psi) < \Delta r_p < 2w_q^{BP}(-1 + I_0 + \Psi) - \Delta c. \end{aligned} \tag{B.7}$$

then

$$\begin{aligned} & i_h = \frac{1 - I_0}{\Psi} + \frac{\Delta c + \Delta r_p}{2w_q^{BP}\Psi} \\ & i_p = 1 \\ & T = 0 \end{aligned} \tag{B.8}$$

Then $I^\sharp = I_0 - \frac{\Delta c + \Delta r_p}{2w_q^{BP}}$.

5. *If*

$$\begin{aligned} \frac{\Delta r_p}{w_b} + \frac{\Delta c + 2\Delta r_p}{w_q^{BP}} < 2\frac{w_b}{w_q^{BP}}(1 - I_0 - \Psi) \textit{ or} \\ \Delta c < 2(w_b + w_q^{BP})(-1 + I_0 + \Psi) \textit{ and} \\ \Delta r_p > 2w_q^{BP}(-1 + I_0 + \Psi) - \Delta c. \end{aligned} \tag{B.9}$$

then

$$\begin{aligned} i_h &= 1 \\ i_p &= 1 \\ T &= 0 \end{aligned} \tag{B.10}$$

Then $I^{\sharp} = 1 - \Psi$.

B.1.3 Full Quality Model

In this section, we will provide the full range of solutions for the quality model, without restrictions on $w_b,\ w_q^{\text{FFS}},\ w_q^{\text{BP}}$, except that $w_q^{\text{BP}} \geq w_q^{\text{FFS}}$. We present these cases only for completeness and do not refer to them in the main text. In the main text, we provide only the part of results that are interesting and that provides insights.

Lemma B.1.1 (Full case of Lemma 2.4.1)**.** *There are 7 different FFS cases to characterize under the full quality model:*

(HM) If

$$\Psi < \bar{\Psi} \text{ and } \frac{w_b}{w_q^{FFS}} \Delta c + \Delta r_p \geq 0 \tag{B.11}$$

, then

$$i_h = 1, \quad i_p = \frac{2I_0 + \Delta r_p}{2w_b(1 - \Psi)} \tag{B.12}$$

(HH) If

$$\Psi > \bar{\Psi} \text{ and } \Psi > 1 - I_0 + \frac{\Delta c}{2w_q^{FFS}}, \tag{B.13}$$

then

$$i_h = 1, \quad i_p = 1 \tag{B.14}$$

(LL) If

$$\frac{\Delta r_p}{2w_b} \leq \Psi - I_0 \textit{ and } \frac{\Delta c}{2w_q^{FFS}} \leq I_0 - \Psi, \tag{B.15}$$

then

$$i_h = 0, \quad i_p = 0 \tag{B.16}$$

(LH) If

$$\frac{\Delta r_p}{2w_b} \geq 1 - I_0 \textit{ and } 1 - \Psi + \frac{\Delta c}{2w_q^{FFS}} \leq 0, \tag{B.17}$$

then

$$i_h = 0, \quad i_p = 1 \tag{B.18}$$

(LM) If

$$\Psi - I_0 \leq \frac{\Delta r_p}{2w_b} \leq 1 - I_0 \textit{ and } \frac{w_b}{w_q^{FFS}}\Delta c + \Delta r_p \leq 0, \tag{B.19}$$

then

$$i_h = 0, \quad i_p = \frac{\frac{\Delta r_p}{2w_b} + I_0 - \Psi}{1 - \Psi} \tag{B.20}$$

(MH) If

$$0 < 1 - I_0 + \frac{\Delta c}{2w_q^{FFS}} < \Psi \textit{ and } \frac{w_b}{w_q^{FFS}}\Delta c + \Delta r_p \geq 0, \tag{B.21}$$

then

$$i_h = \frac{1 - I_0 + \frac{\Delta c}{2w_q^{FFS}}}{\Psi}, \quad i_p = 1 \tag{B.22}$$

(ML) If

$$0 < \Psi - I_0 + \frac{\Delta c}{2w_q^{FFS}} < \Psi \textit{ and } \frac{w_b}{w_q^{FFS}}\Delta c + \Delta r_p \leq 0, \tag{B.23}$$

then

$$i_h = \frac{\frac{\Delta c}{2w_q^{FFS}} - I_0 + \Psi}{\Psi}, \quad i_p = 0 \tag{B.24}$$

Now, we explore the required r^{BP} and the payer's and physicians' surpluses. For that, we need to consider all combinations of FFS and BP, that is, the solutions from Proposition 2.4.1 and Lemma B.1.1. Note that the more likely case with $w_q^{\text{FFS}} > 0$ was covered in Proposition 2.4.1 and its proof. This proposition only additionally considers the cases with $w_q^{\text{FFS}} < 0$.

Proposition B.1.2 (When do they bundle and how much does it cost?)**.** *There are 35 possible cases, from which bundling could occur in 25.*

B.1.4 Variation-focused Model

In this section, we cast the provider's bundling dilemma as a decision problem, present a general model and solve a special case of it. Our approach considers the difference in the provider's income between the current (fee-for-service) payment system and the bundled payment opportunity after an appropriate care redesign. The care redesign problem is explicitly captured to recognize that the structure of the bundled payments program will affect the provider's response. In modeling bundled payments, we draw on the current BPCI initiative by CMS.

The presented model currently entails three key simplifying assumptions:

- No coordination among providers
- Only one condition
- Short-term perspective: No ”learning for the future” motivations or ”changing landscape” concerns, other than what can be represented as current fixed costs or benefits.

Therefore, we can view this model as appropriate for a specialized ”all-in” hospital, such as a surgery or rehabilitation center.

General Framework

We assume that the hospital is weighing the expected patient-related net margin (that is, revenue minus expenses) currently under fee-for-service I^{FFS} and the income it would receive under bundled payments I^{BP}. The hospital is risk averse, and it will therefore adopt bundled payments only if $\mathbf{E}[U(I^{\text{BP}}] > \mathbf{E}[U(I^{\text{FFS}})] > 0$, where U is a risk-averse utility function. The utility function that we assume in this work is the CARA utility function $U(x) := -\exp(-\alpha x)$ for some parameter α, which is in accordance with previous literature (e.g. (Fuloria and Zenios 2001)). Drawing on the framework from Ma (1994), we assume that trying to implement bundled payments, the hospital may exert quality efforts q and cost-reduction efforts ρ and these enter the income function under bundled payments I^{BP}. In contrast, we assume that the income function under FFS is already optimized, but we still retain it in the model to facilitate comparisons.

We assume that the profit margin function under FFS is

$$I^{\text{FFS}} = N(\sum_{i=1}^{m} \nu_i r_i - \sum_{i=1}^{m} \nu_i c_{0,i}), \tag{B.25}$$

where N is the number of patients, assumed constant, m is the number of care pathways, ν_i is the fraction of patients following pathway i ($\sum_i \nu_i = 1$), r_i is the FFS reimbursement for pathway i and $c_{0,i}$ being the average cost of care on pathway i. Next, we assume that the income function under bundled payments, as a function of ρ, q, has the form

$$I^{\text{BP}}(\rho, q) = N(r^{\text{BP}} - \sum_{i=1}^{m} \tilde{\nu}_i(\rho) \cdot \tilde{c}_i(\rho, q) - \tilde{\nu}_\Omega(\rho, q)\tilde{c}_\Omega)r^{\text{BP}}. \tag{B.26}$$

Here, $\tilde{\nu}_i, \tilde{c}_i$ are the new, optimized patient fractions and costs under bundled payments, $\tilde{\nu}_\Omega$ is the number of patients who will suffer from complications and c_Ω are the costs of complications. The reimbursement under bundled payments would be $r^{\text{BP}} = (1-d)\sum_{i=1}^{m} \nu_i r_{0,i}$, but let's keep it as it is for now. We assume that the complications "pathway" Ω would have been paid for under FFS, so it does not appear in Equation (B.25). We also assume that $\tilde{\nu}_\Omega$ is on top of the regular care, that is, we still have $\sum_{i=1}^{m} \tilde{\nu}_i = 1$. Regarding the costs, we assume that $\tilde{c}_i$ and $\tilde{c}_\Omega$ are random, normally distributed with $\tilde{c}_i(\rho, q) \sim \mathcal{N}(c_i(\rho, q), \sigma_i^2),\ \tilde{c}_\Omega \sim \mathcal{N}(c_\Omega, \zeta^2)$. However, we note that the assumption of $\tilde{c}_\Omega$ being Γ-distributed would still lead to a computationally (if not analytically) tractable model and would have a nice actuarial interpretation. We do not consider any uncertainty in original FFS costs because we assume the hospital has these "under control", as opposed to the new setting after the care redesign. In the following, we focus on the bundled payment component and we derive the utility of the bundled profit $I^{\text{BP}}(\rho, q)$. But note that (B.25) and so to compare the utility under FFS vs bundled payments, it is enough to compare the resulting bundled-payments utility to (B.25).

We now compute the expression $\mathbf{E}[U(I^{\text{BP}})]$ applying the transformation $x \mapsto -\frac{\log(-x)}{\alpha}$, and per patient (that is, dividing by N) to obtain the certainty equivalent yields the objective function for hospital to optimize with respect to the effort parameters associated with the care redesign (ρ and q):

$$F(\rho, q) = r^{\text{BP}} - \sum_{i=1}^{m} \tilde{\nu}_i(\rho) \cdot c_i(\rho, q) - \frac{\alpha}{2} \sum_{i=1}^{m} (\tilde{\nu}_i(\rho, q)\sigma_i)^2 - \tilde{\nu}_\Omega(\rho, q) c_\Omega - \frac{\alpha}{2} (\tilde{\nu}_\Omega(\rho, q)\zeta)^2 \tag{B.27}$$

After optimizing, the hospital knows if it breaks even and perhaps even profits from bundled payments, as well as what efforts are necessary to get the most from bundling. This functional form is flexible and allows hospitals to apply their own estimated parameters. If needed, this form can be further extended by considering separate effort parameters for different pathways. We assume that the hospital will try to solve (B.27) numerically. However, in the following, we also derive a closed-form solution for a simplified case, which allows us to state some general insights.

Simplified Model

All parameters w with a subscript in the following correspond to some weights which could be specified or estimated.We consider the following simplifying assumptions:

- We will assume, without losing much generality, that the effort setting under FFS is $q = 0, \rho = 0$, $0 \leq \rho \leq 1$, and $Q_0 \leq Q(q, \rho)$ where Q_0 corresponds to some lower bound on quality measures tracked by CMS and Q the function capturing this, and $0 \leq q \leq q_H$ performance based on the exerted quality and cost-reduction efforts. Assume also $q = 1$ corresponds to a ”maximal quality”. Similarly, $\rho = 1$ corresponds to the point of diminishing returns on cost reduction and streamlining, and we require $\rho \geq 0$ because of the widely believed assumption that FFS reimbursements encourage so much waste that new payment models cannot encourage hospitals to do even worse in this aspect under new payment models, that is, their cost reduction efforts cannot profitably be lower than under FFS.

- Rate of adverse events $\tilde{\nu}_\Omega = \tilde{\nu}_{\Omega,0}(w_\Omega \cdot (q_H - q) + w_{\Omega,\rho})$.

- We only consider two pathways, one “costly” and one “cheaper” because this should be already sufficient to capture the richness of the care pathway redesign. We assume that a hospital tries to stream some of the patients from the costlier pathway to the cheaper one. Let the costly pathway be $i = 1$ and the cheaper pathway $i = 2$.

- $\tilde{\nu}_i = \nu_i + w_{\delta,i}\rho$.

- Consider the following functional forms for c_i:

$$c_i = c_{0,i} + w_\gamma \rho - w_\xi q + \gamma_{0,i}, \tag{B.28}$$

here $\gamma_{0,i}$ is the fixed cost of bundling per patient assigned to this pathway. In the above equation, we could also in principle allow for an additional term wtih ρ^2. This function does not include quality, but quality is instead reflected in the probability of complications. In this model, quality essentially only counteracts the increased number of complications due to care redesign, up to a certain upper bound on quality. I note that considering $c_i = c_{0,i} + w_\gamma \rho - w_\xi q + \gamma_{0,i}$ instead, that is, incorporating a linear term for costs, also leads to a reasonably solvable model.

For convenience, we make a number of additional assumptions which somewhat declutter the final result, but they could be easily avoided; we also introduce some additional notation for this case $m = 2$:

- Assume $\sigma_i \equiv \sigma$.

- Set $2\zeta = c_\Omega$, this corresponds to the random cost of complications varying mostly between $[0, 4c_\Omega]$. We can then simplify the certainty equivalent for the cost of complications from $(w_\Omega \cdot (q_H - q) + w_{\Omega,\rho})c_\Omega + \frac{\alpha}{2}((w_\Omega \cdot (q_H - q) + w_{\Omega,\rho})\zeta)^2$ to just $\frac{-2}{\alpha} + \frac{\alpha}{8}((w_\Omega \cdot (q_H - q) + w_{\Omega,\rho})c_\Omega + \frac{4}{\alpha})^2$ and let just $K_\alpha := \frac{-2}{\alpha}$.

- Now, define $\tilde{\nu}_i = \nu_i \pm \rho\delta$, with plus for the cheaper pathway and minus for the costly pathway, that is, we are redirecting patients from the costlier pathway to the cheaper pathway.

- Let $\gamma := \frac{c_{0,1} - c_{0,2}}{2}$; note that this term is larger than zero under our assumption that the first pathway is the costly one.

- $\gamma_0 := \nu_1\gamma_{0,1} + \nu_2\gamma_{0,2}$ (average fixed cost per patient)

- At this point, we are not looking at the boundary conditions on ρ and q, but these will often be active. For instance, in this baseline model that we don't have quality in the cost, q will be probably typically at its upper bound.

With all these assumptions and some additional calculations, the objective function now reads

$$
\begin{aligned}
F(\rho, q) &= r^{\text{BP}} - \sum_{i=1}^{m} \tilde{\nu}_i(\rho) \cdot c_i(\rho) - \frac{\alpha}{2}\sum_{i=1}^{m}(\tilde{\nu}_i(\rho)\sigma_i)^2 - \tilde{\nu}_\Omega(\rho, q)c_\Omega - \frac{\alpha}{2}(\tilde{\nu}_\Omega(\rho, q)\zeta)^2 \\
&= r^{\text{BP}} - (\nu_1 - \rho\delta)(c_{0,1} + \gamma_{0,1}) - (\nu_2 + \rho\delta)(c_{0,2} + \gamma_{0,2}) + w_\gamma\rho \\
&- \frac{\alpha}{2}[((\nu_1 - \rho\delta)\sigma)^2 + ((\nu_2 + \rho\delta)\sigma)^2] + (R_0 - w_\Omega q + w_{\Omega,\rho}\rho))^2\zeta^2] - \tilde{\nu}_\Omega(\rho, q)c_\Omega \\
&= r^{\text{BP}} - \nu_1 c_{0,1} - \nu_2 c_{0,2} - \nu_1\gamma_{0,1} - \nu_2\gamma_{0,2} \\
&+ \rho\delta(c_{0,1} - c_{0,2}) - \rho\delta\gamma_{0,1} - \rho\delta\gamma_{0,2} \\
&+ w_\gamma\rho - w_\xi q \\
&- \frac{\alpha}{2}[(\nu_1^2 + \nu_2^2)\sigma^2 + 2\rho\delta(\nu_2 - \nu_1)\sigma^2 + 2\rho^2\delta^2\sigma^2] \\
&- \frac{-2}{\alpha} - \frac{\alpha}{8}((R_0 - w_\Omega q + w_{\Omega,\rho}\rho)c_\Omega + \frac{4}{\alpha})^2 \\
&= r^{\text{BP}} - \bar{c}_0 - \gamma_0 + \rho\frac{\delta\gamma}{2} + w_\gamma\rho - w_\xi q \\
&- \alpha\sigma^2(\rho\delta + \frac{\nu_2 - \nu_1}{4})^2 - \Omega(\rho, q) \\
&+ \frac{2}{\alpha} - \frac{\sigma^2\alpha}{4}
\end{aligned}
\tag{B.29}
$$

where some constants are omitted (including a constant term with σ^2).

I maximized (B.29), using Mathematica for some of the tedious computations. The resulting formula $F(\bar{\rho}, \bar{q})$ with optimized parameters $\bar{\rho}, \bar{q}$ follows:

$$\begin{aligned} F =& r^{\mathrm{BP}} - \gamma_0 - \bar{c}_0 + \frac{(\nu_1 - \nu_2)(\gamma\delta + w_\gamma)}{4\delta} - \frac{w_\xi(\alpha c_\Omega R_0 + 4)}{\alpha c_\Omega w_\Omega} + \frac{w_\xi w_{\Omega,\rho}(\nu_1 - \nu_2)}{4\delta w_\Omega} + \frac{2w_\xi^2}{\alpha c_\Omega^2 w_\Omega^2} \\ &+ \frac{(w_\gamma w_\Omega - w_\xi w_{\Omega,\rho} + \gamma\delta w_\Omega)^2}{4\alpha^2\delta^2\sigma^2 w_\Omega^2} + \frac{2}{\alpha} - \frac{\sigma^2\alpha}{4} \end{aligned} \tag{B.30}$$

The optimized ρ and q values are also available. The formula if we incorporate q in the cost function is slightly more complicated, but has a similar flavor. Again, we note this result may be somewhat different if the boundary conditions on q or ρ are met. From (B.30), we now directly see the profit from the bundled payment contract and we can compare this equation to the profit under FFS in (B.25). The equation is very sensitive to the value of α.

With the start of bundling, many analysts have suggested that there is a tradeoff between too little variation (because there is than nothing left to optimize) and too much variation (too much risk born by the provider). In our model, this tradeoff corresponds to postulating $\gamma = K_\gamma\sigma$ for some constant K_g, in other words, it ties the cost variation between the care pathways to the variation within them. This tradeoff is a testable assumption which may depend on disease. Even for DRGs that the K_g assumption holds, the "sweet-spot theorem" may not. Indeed, the sweet-spot theorem essentially says that F will be quadratic in $\sigma > 0$. The subterm of F containing σ is $\frac{1}{4}(K_\gamma(\nu_1 - \nu_2)\sigma - \sigma^2\alpha + \frac{2Kw_\gamma}{\alpha^2\delta\sigma} + \frac{w_\gamma^2}{\alpha^2\delta^2\sigma^2})$ Assuming now $w_\gamma = 0$ (no traditional cost cutting), this leads to the sweet spot $\sigma = \frac{K_\gamma(\nu_1 - \nu_2)}{2\alpha}$. Notably, this sweet spot exists only for $\nu_1 > \nu_2$, that is, the volume on the costly pathway is substantial. Furthermore, this vaulue very much depends on how risk averse the hospital is. The formula $\gamma = K_\gamma\sigma$ is an empirical hypothesis which we can test.

If we include quality in the cost function, the functional form containing σ becomes

$$\frac{1}{4}(K_\gamma(\nu_1-\nu_2)\sigma-\sigma^2\alpha+\frac{2K(w_\gamma-(w_{\Omega,\rho}/w_\Omega)w_\xi)}{\alpha^2\delta\sigma}+\frac{(w_\gamma-w_\xi(w_{\Omega,\rho}/w_\Omega))^2}{\alpha^2\delta^2\sigma^2}).$$

It is hard to say if we can now ignore the $\frac{1}{\sigma}$ and $\frac{1}{\sigma^2}$.

Regarding the volume, typically the fixed costs may depend on volume ($\gamma_0=\gamma_t/n$ for some total γ_t) and the hospital risk aversion will depend on volume: $\alpha=\alpha_0/n$ But we assume it is constant within the same hospital, which justifies the CARA assumption. Actually, $\alpha=\frac{\alpha_0}{n^\beta}$ seems more reasonable? We might have $\beta<1$ (observed in farmers; IRRA), but also $\beta>1$? This does not behave nicely as it indicates that the profit per patient goes to infinity as n increases. This probably has to do with the boundary conditions which ρ will hit as n increases, respectively α decreases.

B.1.5 Supplementary Results for the Coproduction Model

.

Lemma B.1.2 characterizes the equilibrium solution when bundling occurs and highlights the role of gainsharing in the base case.

Lemma B.1.2 (Optimal equilibrium under BP in the Base Model)**.** *When bundling occurs as outlined in Theorem 2.3.1, the solution becomes as follows:*

$$\begin{aligned}
&T=\min(\Delta r_p+2I_0w_b,\frac{1}{2}(-2w_b(1-I_0)+\Delta c+2\Delta r_p))>0\\
&i_h^\sharp=1 \qquad\qquad \text{(B.31)}\\
&i_p^\sharp=\max(0,\frac{1}{4w_b(1-\Psi)}(2(1+I_0)w_b-\Delta c))
\end{aligned}$$

We see from Lemma B.1.2 that gainsharing amount, T, is always positive. Positive gainsharing helps hospital to incentivize physicians to reduce the level of care intensity. The proof is parallel to the proof of Lemma B.1.4 (presented later in Appendix B below), and we therefore omit it.

Lemma B.1.3 (Equilibrium solutions under FFS in the General Model)**.** *The equilibrium solutions under FFS are as follows:*

I. If $\Psi \leq \bar{\Psi}$, then

$$i_h^* = 1,\ i_p^* = \frac{1}{1-\Psi}(I_0 + \frac{\Delta r_p}{2w_b}) \tag{B.32}$$

II. If $\bar{\Psi} \leq \Psi \leq \hat{\Psi}$, then

$$i_h^* = 1,\ i_p^* = 1 \tag{B.33}$$

III. If $\Psi \geq \hat{\Psi}$, then

$$i_h^* = \frac{1}{\Psi}(I_0 - \frac{\Delta c}{2w_q}),\ i_p^* = 1 \tag{B.34}$$

Lemma B.1.4 (Equilibrium solutions under BP in the General Model, case $T > 0$)**.** *Under BP, there are two feasible solutions with $T > 0$:*

I.

$$\begin{aligned} i_h^\sharp &= 1 \\ i_p^\sharp &= \frac{2(w_b + I_0(w_b + w_q)) - \Delta c}{2(2w_b + w_q)(1-\Psi)} \\ T &= \frac{w_b(\Delta c - 2(1-I_0)w_b)}{2w_b + w_q} + \Delta r_p \end{aligned} \tag{B.35}$$

This solution is equilibrium if

$$\begin{aligned} &2(w_b + I_0(w_b + w_q)) > \Delta c > 2(w_q(I_0 + \Psi - 1) + w_b(I_0 + 2\Psi - 1)) \textit{ and} \\ &\Delta r_p(2 + w_q/w_b) + \Delta c > 2(1 - I_0)w_b. \end{aligned} \tag{B.36}$$

II.

$$\begin{aligned} i_h^\sharp &= 1 \\ i_p^\sharp &= 0 \\ T &= 2I_0 w_b + \Delta r_p =: T_{\max}. \end{aligned} \tag{B.37}$$

This condition requires

$$\Delta c > 2(w_b(1 + I_0) + I_0 w_q). \tag{B.38}$$

Lemma B.1.5 (Equilibrium solutions under BP in the General Model, case $T = 0$)**.** *Let $\Psi_1 = \bar{\Psi}$ and $\Psi_2 = 1 - I_0 + \frac{\Delta c + \Delta r_p}{2w_q}$. Under BP, there are three feasible solutions with $T = 0$: As the first step, we show that not all all parties will benefit if $T = 0$.*

I. If $\Psi \leq \Psi_1$, then

$$i_h^\sharp = 1,\ i_p^\sharp = \frac{1}{1-\Psi}(I_0 + \frac{\Delta r_p}{2w_b}) \tag{B.39}$$

II. If $\Psi_1 \leq \Psi \leq \Psi_2$, then

$$i_h^\sharp = 1,\ i_p^\sharp = 1 \tag{B.40}$$

III. If $\Psi \geq \Psi_1$, then

$$i_h^\sharp = \frac{1}{\Psi}(I_0 - \frac{\Delta c + \Delta r_p}{2w_q}),\ i_p^\sharp = 1. \tag{B.41}$$

Corollary B.1.1 (Intensity under the General Model)**.** *The realized care intensity under BP, $I^\sharp$, is less than that under the FFS, and is given by*

$$I^\sharp = \begin{cases} 0 & \text{if } \Delta c > 2(1 + I_0 + w_q I_0) \\ I_0 - \frac{\Delta c - 2(1-I_0)w_b}{2(2w_b + w_q)} & \text{if } \Delta c < 2(1 + I_0 + w_q I_0). \end{cases} \tag{B.42}$$

for the cases that bundling is preferable.

B.1.6 Supplementary Results for the Physician-Driven Model

We list additional results from the physician-driven model, parallel to the results from the coproduction model. The proofs follow the same logic as in the proofs of the coproduction model results, and hence we omit them.

Lemma B.1.6 (Status-quo intensity)**.** *The equilibrium intensity under FFS, $I^* \equiv i_p^*$, is given by*

$$I^* = \begin{cases} I_0 + \frac{\Delta r_p - \Psi}{2w_b} & \textit{if } \Psi \leq \bar{\Psi} \\ 0 & \textit{otherwise} \end{cases}$$

where

$$\bar{\Psi} := \Delta r_p + 2w_b I_0. \tag{B.43}$$

Corollary B.1.2 (Intensity under BP vs. FFS)**.** *The equilibrium intensity under BP, $I^\sharp$, is less than that under the FFS, I^*. Furthermore, $I^\sharp$ is given by the following expression:*

$$I^\sharp \equiv i_p^\sharp = \max(I_0 + \frac{\Delta r_p - \Psi - T}{2w_b}, 0) \leq \max(I_0 + \frac{\Delta r_p - \Psi}{2w_b}, 0) = I^*. \tag{B.44}$$

Proposition B.1.3 (Quality under BP vs. FFS)**.** *Compared with FFS, the quality of care under BP may decrease or increase, depending on the physician integration level, Ψ. In particular:*

I. If integration is low, i.e., $\Psi < \Delta r_p$, overprovision of services characterizes FFS. In such a case, BP will improve quality if

$$2(1 - I_0)w_b + 2\Delta r_p > \Delta c + 3\Psi \tag{B.45}$$

and worsen it otherwise.

II. If integration is high, i.e., $\Psi > \Delta r_p$, then underprovision of services characterizes FFS. Then BP will further decrease intensity and hence quality.

B.1.7 Comparing the Coproduction and the Physician-driven model: Case $\Psi \to 0$

Let's examine the special case with $\Psi \to 0$ when we expect the Coproduction and Physician-driven models be somewhat comparable.

Corollary B.1.3 (Scenarios in the Coproduction model for $\Psi \to 0$)**.**

- *FFS:*

$$I^* = \begin{cases} I_0 + \Delta r_p/2, & \textit{if}\, 1 - I_0 > \frac{\Delta r_p}{2} \\ 1, & \textit{if}\, 1 - I_0 < \frac{\Delta r_p}{2} \end{cases} \tag{B.46}$$

- *BP*

$$\begin{aligned} i_h &= 1 \\ T &= 2 \cdot \min\{\frac{1}{2}(I_0 + \frac{\Delta c + 2\Delta r_p}{2} - 1), I_0 + \frac{\Delta r_p}{2}\} \\ i_p &= I_0 + \frac{\Delta r_p}{2} - \frac{T}{2} \end{aligned} \tag{B.47}$$

We will get $i_p = 0$ iff $\Delta c/2 > 1 + I_0$

We notice that in both the Coproduction and the Physician-driven model, as $\Psi \to 0$, if Δc is high enough, there will be a bound on T and $i_p^\sharp = 0$, otherwise $i_p^\sharp \in (0, 1)$.

Corollary B.1.4 (Comparing surplus)**.** *Surplus varies somewhat across the two scenarios.*

Corollary B.1.5 (Comparing quality and intensity)**.** *Under FFS, there was always overtreatment. In both models, Coproduction and Physician-driven, the quality can either increase or decrease.*

Corollary B.1.6 (When do they bundle?)**.**

- *Coproduction: Under the following conditions:*

$$
\begin{aligned}
&\Delta c > 2(1 - I_0 + w_q I_0) \textit{ and} \\
&\Delta c > \Delta r_p(2 + w_q) + 2(1 - I_0)(3 + 2w_q) \textit{ and } 1 - I_0 > \frac{\Delta r_p}{2} \textit{ or} \\
&\frac{(\Delta c + (2(1 - I_0)(1 + w_q))^2}{4(2 + w_q)} > \Delta c + \Delta r_p + (1 - 2I_0)w_q \textit{ and } 1 - I_0 < \frac{\Delta r_p}{2}
\end{aligned}
\tag{B.48}
$$

- *Physician-driven: The case-independent conditions are as follows:*

$$w_q(\Delta c + 3\Delta r_p - 2(2 - I_0)w_q) + (\Delta c + 2\Delta r_p - 1 - 4(1 - I_0)w_q) - 2(1 - I_0) > 0 \tag{B.49}$$

and

$$2I_0(1 + w_q) + \Delta r_p > 1 \tag{B.50}$$

. *Next come four case-dependent conditions on physician surplus, which we do not write out in detail anymore.*

B.1.8 Physician-driven Model in the Principal-agent Framework

In this section, we explore how we can profitably cast the physician-driven model from Section 2.4.4 into the principal-agent framework. Indeed, observe that in the physician-driven model, we model the utility of both the physician and the hospital, but it is only the physician who chooses the intensity while the hospital only enter the game through the gainsharing amount T. We can therefore see the hospital as the principal and the physician as an agent in the principal-agent framework. Here, we prove the concept by modeling a simple symmetric-information principal-agent setup and use it to derive the optimal gainsharing contract. At the end, we discuss how the proof of concept could be further extended to answer other questions of interest.

To derive the optimal gainsharing contract for the physician-driven model, we adapt the setup from Section 2.4.4 but with several additional assumptions to simplify the exposition:

- Assume that the hospital is mandated to bundle, so its utility function equals $F_h^{\text{FFS}} = F_h^{\text{BP}}$. On the other hand, the physicians have the option to not participate and leave with F_p^{FFS} as their utility. Equivalently, they can always choose to take $T = 0$ in F_p^{BP}. This assumption corresponds to the real-world situation where the CMS forces hospitals as the main BP participant, but the hospital needs to negotiate with other parties to make them involved.
- Assume $\Phi = 0$, so the hospital can only influence the physicians through gainsharing.
- Assume $I_0 + \Delta r_p/(2w_b) < 1$

Now, instead of linear gainsharing as in Section 2.4.4, the hospital can now offer an arbitrary payoff function $T(I)$. We now follow the standard principal-agent framework under the simplest setting of perfect information (Macho-Stadler and Pérez-Castrillo 2020) to solve for T and derive the optimal hospital utility. First, we assess the value of the physician's reservation utility, i.e. when the physician does not accept the contract, opting for $T = 0$. As in the original physician-driven model, we find that $I^* = I_0 + \Delta r_p/(2w_b)$ and $F_p^{\text{FFS}}(I^*) = r_{2,p} + I_0\Delta r_p + \frac{\Delta r_p^2}{4w_b}$. The participation constraint for this problem then reads as follows:

$$\begin{aligned}
&\max_{T,I} - (\Delta c + \Delta r_p)I - T(I) \\
&\qquad \text{s.t.} \ - w_b(I - I_0)^2 + \Delta r_p I + r_{2,p} + T(I) \geq r_{2,p} + I_0\Delta r_p + \frac{\Delta r_p^2}{4w_b},
\end{aligned} \tag{B.51}$$

where the right-hand side of the constraint is the reservation utility. We can substitute

$J := I - I_0$ and solve for an equivalent problem:

$$\begin{aligned} &\max_{T^*,J} -(\Delta c + \Delta r_p)J - T^*(J) \\ &\quad\text{s.t.} \ -w_b J^2 + \Delta r_p J + T^*(J) \geq \frac{\Delta r_p^2}{4w_b}, \end{aligned} \tag{B.52}$$

where $T^*(J) = T(I - I_0)$ is a substituted version. We observe that the inequality in (B.52) can always be made binding, otherwise the hospital can just decrease T^* while increasing its utility. This yields the optimal gainsharing contract

$$T^*(J) = w_b J^2 - \Delta r_p J + \frac{\Delta r_p^2}{4w_b} \tag{B.53}$$

or equivalently:

$$T(I) = w_b(I - I_0)^2 - \Delta r_p(I - I_0) + \frac{\Delta r_p^2}{4w_b}. \tag{B.54}$$

Solving then for the optimal $I^\sharp$ by the hospital yields $I^\sharp = I_0 - \frac{\Delta c}{2w_b}$ (or 0 if this term was lower than 0). Compared to the option without gainsharing ($T = 0$), the hospital gains an additional amount

$$F_h^{\text{BP}}(I^\sharp, T(I^\sharp)) - F_h^{\text{FFS}}(I^*) = \frac{(\Delta c + \Delta r_p)(4I_0 w_b + \Delta c + \Delta r_p)}{4w_b} > 0. \tag{B.55}$$

We have thus derived the optimal gainsharing contract for the hospital, extending the previously-assumed linear model.

We end with two additional potential applications:

- We could assume that the physician effort is unobserved and that the hospital cost. What would be the implications for the optimal contract?
- We could use the same framework to derive the optimal contract for the more complex coproduction model in Section 2.3.

B.2 Proofs

B.2.1 Initial Model

Proof. Proof of Lemmas 2.2.1, 2.2.2, and 2.2.3: Differentiating F_h^{FFS}, we see that this derivative is always positive, so the hospital will choose $i_h^* = 1$, regardless of the physicians' response. The physician utility function under $i_h = 1$ is then $F_p^{\text{FFS}}(i_p) = -w_b((1-\Psi)i_p - I_0)^2 + i_p\Delta r_p(1-\Psi) + r_{2,p}$. Optimizing this with respect to $i_p \in [0,1]$ yields $i_p = 1$ if $\Psi \geq \bar{\Psi}$ and $i_p = \frac{I_0+\frac{\Delta r_p}{2w_b}}{1-\Psi}$ otherwise. Given that in both of these cases $i_p \geq 0$, we have $I^* = 0 \cdot \Psi + i_p \cdot (1-\Psi)$, which completes the proof of Lemma 2.2.1.

Lemma 2.2.2 follows from the observation that $I(1, i_p) \leq I(1,1) = 1 - \Psi$ and the fact that $i_h^* = 1$. Lemma 3 follows from the optimal value of intensity as cited in Lemma 1 combined with the boundary value from Lemma 2.

□

We next present the proof of Proposition 2.3.2, which characterizes the optimal solution and is used in proving Theorem 2.3.1.

Proof. Proof of Proposition 2.3.2: We are seeking a sequential equilibrium of a two-stage game using backward induction. Therefore, we first compute the physicians' (second-stage) best response. Optimizing F_p^{BP} for given i_h and T as a function of i_p yields the optimal response $\iota_p(i_h, T) = \arg\max_{i_p \in [0,1]} F_p^{\text{BP}}(i_h, i_p, T)$, which is specified as

$$\iota_p := \iota_p(i_h, T) = \frac{1}{2w_b(1-\Psi)}(\Delta r_p - \frac{T}{1-\Psi} + 2w_b(I_0 - (1-i_h)\Psi)) \qquad \text{(B.56)}$$

when the expression in (B.56) lies within of $(0,1)$. In the case when the expression in (B.56) lies outside $(0,1)$, the physician response is equal to 0 (if (B.56) is below 0) or to 1 (if (B.56) is above 1).

In calculating hospital's response, we first consider the case where $\iota_p(i_h, T)$ is strictly between 0 and 1 and address the boundary cases later. Plugging ι_p into F_h^{BP} and differentiating with respect to i_h and T yields

$$\begin{aligned} \frac{\partial}{\partial i_h} F_h^{\mathsf{BP}}(i_h, \iota_p(i_h, T), T) &= \frac{T\Psi}{1-\Psi}, \\ \frac{\partial}{\partial T} F_h^{\mathsf{BP}}(i_h, \iota_p(i_h, T), T) &= \frac{-2T + (1-\Psi)(\Delta c + 2\Delta r_p - 2w_b(1 - I_0 - i_h\Psi))}{2w_b(1-\Psi)^2}. \end{aligned} \tag{B.57}$$

This implies that (ignoring the physicians' boundary conditions for now) for $T > 0$, the hospital will aim to increase i_h as much as possible, up to $i_h = 1$. For $T = 0$, the hospital is indifferent over $i_h \in [0, 1]$, so we can just work with the case $i_h = 1$ without loss of generality. We can now derive a condition separating $T = 0$ from $T > 0$ when $i_h = 1$. For this, we check the derivative of $F_h^{\mathsf{BP}}(i_h, \iota_p(i_h, T), T)$ with respect to T evaluated at $T = 0$ and $i_h = 1$ when $i_p = 1$. If the derivative is positive, we will have $T > 0$, otherwise $T = 0$. This yields the following condition for $T > 0$:

$$\frac{\Delta c + 2\Delta r_p}{2w_b} > 1 - I_0 - \Psi. \tag{B.58}$$

Next, we focus on deriving the optimal solution values under two cases, $T > 0$ and $T = 0$. In both cases, we will also address the boundary cases for i_p.

Case $T > 0$: When $T > 0$, $i_h = 1$. The optimal value is computed by setting the second equation in (B.57) to zero. Doing so, we get the following:

$$T = \frac{1}{2}(1-\Psi)(\Delta c + 2\Delta r_p - 2w_b(1 - I_0 - \Psi)). \tag{B.59}$$

Now, we need to check the boundary conditions for ι_p. To ensure that $\iota_p \geq 0$ and the hospital is not gainsharing with physicians more than it needs, we must have

$$T \leq (1-\Psi)(2I_0 w_b + \Delta r_p). \tag{B.60}$$

We conclude that T is the minimum of (B.59) and (B.60). In particular, if $\Delta c > 2w_b(1 + I_0 - \Psi)$, then (B.60) holds and otherwise (B.59) holds. Finally, we need to ensure $\iota_p < 1$, which requires

$$T \geq (1 - \Psi)(\Delta r_p - 2w_b(1 - I_0 - \Psi)). \tag{B.61}$$

Comparing the above with (B.59) implies the condition

$$\frac{\Delta c}{2w_b} > -1 + I_0 + \Psi. \tag{B.62}$$

If $i_p = 1$, then we can assume that the hospital sets $T = 0$ without any loss of generality. Hence, we have $T > 0$ if and only if (B.58) and (B.62) hold.

Case $T = 0$: Recall that $T = 0$ only occurs if either (B.58) or (B.62) does not hold. For $T = 0$, the hospital could, in principle, arbitrarily set i_h, which would always be countered by the physicians so that the overall intensity (and hospital utility) remains the same. However, this arbitrariness of i_h may not entirely be the case when the physicians' response is bounded. We know from (B.56) that i_h increases with i_p. Hence, let i_h^- be maximal such that $\iota_p(i_h^-, 0) = 0$ and i_h^+ be minimal such that $\iota_p(i_h^+, 0) = 1$. Then the hospital profit function $F_h^{\text{BP}}(i_h, \iota(i_h, 0), 0)$ is constant for $i_h \in [i_h^-, i_h^+]$, but can be further increased for $i_h > i_h^+$ (indeed, the derivative of $\frac{\partial}{\partial i_h} F_h^{\text{BP}}(i_h, 1, 0) = (\Delta c + \Delta r_p)\Psi > 0$). We therefore distinguish two cases: $i_h^+ \geq 1$ (in which case any $i_h \in [i_h^-, i_h^+]$ is optimal) or $i_h^+ < 1$ (in which case $i_h = 1$ is optimal). The condition $i_h^+ < 1$ is equivalent to $\iota(1, 0) > 1$ (so that the boundary condition becomes binding). In other words,

$$\frac{\Delta r_p}{2w_b} > 1 - I_0 - \Psi, \tag{B.63}$$

or, equivalently, $\Psi > \bar{\Psi}$. We note that i_h^- corresponds to the level of i_h such that the physicians would choose $\iota_p(i_h^-, 0) = 0$. This implies

$$i_h^- = \frac{1}{\Psi}(\Psi - I_0 - \frac{\Delta r_p}{2w_b}). \tag{B.64}$$

To summarize, when $T = 0$, the two cases are

1. $\Psi > \bar{\Psi}$, and then $i_h = 1$, the physicians respond with $i_p = 1$ and $I^\sharp = 1 - \Psi$
2. $\Psi \leq \bar{\Psi}$, and then $i_h \in [i_h^-, 1]$, the physicians respond with $\iota_p(i_h)$ and $I^\sharp = I_0 + \frac{\Delta r_p}{2w_b}$.

This finalizes the cases, and concludes the proof.

□

Proof. Proof of Theorem 2.3.1 and Proposition 2.3.2: In proving these results, we will show that the payer, the hospital, and the physicians all benefit (or at least do not lose) from bundled payments when condition (2.6) holds. For physicians and the hospital, we assume their benefit must be strictly positive, and for the payer, it must be at least non-negative. We consider two cases separately, $T = 0$ and $T > 0$.

Case $T = 0$**:** When $T = 0$ and $\Psi > \bar{\Psi}$, the resulting practiced intensity I under BP is the same as in the FFS case. Hence, the physicians payoff function would be the same as in the FFS and hence they would not be interested in bundling. On the other hand, when $\Psi < \bar{\Psi}$, the resulting intensity and payoffs (F_h^{BP}, and F_p^{BP}) will be the same regardless of what i_h the hospital chooses. This is because whatever $i_h < 1$ the hospital chooses, the physicians can choose i_p such that the intensity will be the same as when the hospital chooses $i_h = 1$ under FFS.

Case $T > 0$**:** From the analysis of case $T = 0$ above, we know that physicians will not be interested in bundling unless $T > 0$. In the proof of Proposition 2.3.2, we have seen that $T > 0$ if and only (B.58) and (B.62) hold. Hence, it remains to characterize when the physicians, the hospital, and the payer all benefit under $T > 0$. Starting with the physicians, we need to show that $F_p^{\text{BP}} - F_p^{\text{FFS}}$ is positive. We analyze this by considering all possible three cases with respect to the integration level:

- $\Psi > \bar{\Psi}$ and $i_p^\sharp > 0$: Then

$$F_p^{\text{BP}} - F_p^{\text{FFS}} = \frac{(\Delta c - 2w_b(-1 + I_0 + \Psi))^2}{16w_b} > 0, \tag{B.65}$$

  which follows from (B.62).

- $\Psi > \bar{\Psi}$ and $i_p^\sharp = 0$: Then

$$F_p^{\text{BP}} - F_p^{\text{FFS}} = w_b(1 - \Psi)^2 > 0. \tag{B.66}$$

- $\Psi < \bar{\Psi}$ and $i_p^\sharp > 0$: Then

$$F_p^{\text{BP}} - F_p^{\text{FFS}} = \frac{(\Delta c - 2\Delta r_p + 6w_b(1 - I_0 - \Psi))(\Delta c + 2\Delta r_p - 2w_b(1 - I_0 - \Psi))}{16w_b} > 0 \tag{B.67}$$

  because the first product term in the numerator is positive as a result of the condition in Lemma 2.2.3 and the second product term is also positive due to (B.58).

- $\Psi < \bar{\Psi}$ and $i_p^\sharp = 0$: Then

$$F_p^{\text{BP}} - F_p^{\text{FFS}} = \frac{(2I_0w_b + \Delta r_p)((2 - I_0 - 2\Psi)2w_b - \Delta r_p)}{4w_b} > 0 \tag{B.68}$$

  because the first product term in the numerator is clearly positive and the second product term in the numerator is also positive due to (2.4).

We also need to ensure that r^{BP} is high enough so that hospital is better off under BP. From this definition of r^{BP}, we need to determine the required r^{BP}, such that $F_h^{\text{BP}} > F_h^{\text{FFS}}$. Depending on the level of integration (i.e. $\Psi > \bar{\Psi}$ vs. $\Psi < \bar{\Psi}$), F_h^{FFS} takes a different expression and we explore the two cases separately. Finally, we will need to show that the payer benefits or at least does not lose. For that, we define $\Delta\pi := r^{\text{BP}} - r_h - r_{2,p} - I^*\Delta r_p$, the payer surplus, which is exactly the difference between the bundled payment and the previous total payment under FFS. We analyze the following cases:

- If $\Psi > \bar{\Psi}$ and $i_p^\sharp > 0$, then

$$\Delta\pi = -\frac{(\Delta c + 2w_b(1 - I_0 - \Psi))^2}{8w_b} < 0 \tag{B.69}$$

- If $\Psi > \bar{\Psi}$ and $i_p^\sharp = 0$, then

$$\Delta\pi = (2I_0 w_b - \Delta c)(1 - \Psi) \tag{B.70}$$

  Then, (B.70) is negative because $i_p^\sharp = 0$ requires $\frac{\Delta c}{2w_b} \geq I_0 + (1 - \Psi) > I_0$.

- If $\Psi \leq \bar{\Psi}$ and $i_p^\sharp > 0$, then

$$\Delta\pi = -\frac{(\Delta c + 2w_b(1 - I_0 - \Psi))^2}{8w_b} < 0 \tag{B.71}$$

- If $\Psi \leq \bar{\Psi}$ and $i_p^\sharp = 0$, then

$$\Delta\pi = -\frac{(2I_0 w_b + \Delta r_p)(\Delta c + \Delta r_p - 2w_b(1 - \Psi))}{2w_b} < 0. \tag{B.72}$$

  Then, (B.72) is negative because of Condition (B.58), which completes the proof.

□

Proof. Proof of Corollary 2.3.2: This proof follows directly from the proof above because Σ in Corollary 2.3.2 is equal to $\Delta\pi$ defined above.

Proof. Proof of Proposition 2.2.3: This result follows directly by computing the intensity and ΔQ from Proposition 2.3.2. Without loss of generality, assume $i_p^\sharp > 0$. Then, intensity under BP is given by:

$$I^\sharp = I(1, \iota_p(1,T)) = I(1, \iota_p^{\text{FFS}} - \frac{T}{2w_b(1-\Psi)^2}) = I_0 + \frac{\Delta r_p}{2w_b} - \frac{T}{2w_b(1-\Psi)}. \tag{B.73}$$

For computing ΔQ, we consider the following two cases separately:

1. $\Psi \geq \bar{\Psi}$:

$$\begin{aligned}\Delta Q &= |I^* - I_0| - |I^\sharp - I_0| = |1 - \Psi - I_0| - |\frac{\Delta r_p}{2w_b} - \frac{T}{2w_b(1-\Psi)}| \\ &= |1 - \Psi - I_0| - |\frac{\Delta r_p}{2w_b} - \frac{1}{2}(I_0 + \frac{\Delta c + 2\Delta r_p}{2w_b} - (1-\Psi))| \\ &= |1 - \Psi - I_0| - \frac{1}{2}|I_0 + \frac{\Delta c}{2w_b} - (1-\Psi)|.\end{aligned} \tag{B.74}$$

From (B.74), (2.11) directly follows.

2. $\Psi < \bar{\Psi}$:

$$\begin{aligned}\Delta Q &= |I^* - I_0| - |I^\sharp - I_0| = \frac{\Delta r_p}{2w_b} - |\frac{\Delta r_p}{2w_b} - \frac{T}{2w_b(1-\Psi)}| \\ &= \frac{\Delta r_p}{2w_b} - |\frac{\Delta r_p}{2w_b} - \frac{1}{2}(I_0 + \frac{\Delta c + 2\Delta r_p}{2w_b} - (1-\Psi))| \\ &= \frac{\Delta r_p}{2w_b} - \frac{1}{2}|I_0 + \frac{\Delta c}{2w_b} - (1-\Psi)|.\end{aligned} \tag{B.75}$$

from which (2.12) follows, which completes the proof.

□

Proof. Proof of Theorem 2.2.2:

We first describe our approach and then connect to the findings in the theorem. Recall that the hospital can adjust both i_h and T in solving the first stage of a two-stage problem, where it uses the physician response function in its utility function: $F_h(i_h, \iota_p(i_h,T), T)$. Recall also the characterization of the practiced intensity under BP, $I^\sharp$, from Equation (2.10) and consider a case where $I^\sharp < I_0$. Suppose in this case that the payer wants the hospital to increase the practiced intensity to $I_2 > I^\sharp$ so that a desired quality level I_2 is achieved.

If the hospital is to achieve this intensity, then it could adjust its gainsharing amount for the desired intensity, denoted by T_2, to

$$T_2 := T^\sharp - 2w_b(1-\Psi)(I_2 - I^\sharp), \tag{B.76}$$

where $T^\sharp$ is what the gainsharing amount would be when there are no constraints on quality under bundled payments. Plugging (B.76) back into the condition on r^{BP} in Proposition 2.3.1 results in the following minimum payment value by the payer to induce the hospital to improve quality from $I^\sharp$ to I_2:

$$\Delta r^{\text{BP}} \equiv r_2^{\text{BP}} - r_\sharp^{\text{BP}} = \frac{(I_2 - I^\sharp)}{1-\Psi}((1-\Psi)(\Delta c + 2\Delta r_p - 2w_b(1 - I_0 - (I_2 - I^\sharp) - \Psi)) - 2T^\sharp) = 2(I_2 - I^\sharp)^2 w_b. \tag{B.77}$$

Notice that r_2^{BP} is then the bundled payment that the payer needs pay to ensure the intensity $I_2 > I^\sharp$. Also note that the hospital, physicians, and the payer are all still incentivized to bundle.

Under FFS, the hospital is more constrained in affecting the quality. The only time that the hospital can impact intensity to improve quality is when $I^* = 1 - \Psi < I_0$ (otherwise it is always optimal to set $i_h = 1$ for the hospital). The extra cost to the hospital is then $\Delta r^{\text{FFS}} = \Delta c(I_2 - I^*)$ while the overall new reimbursement from the payer is $r_2^{\text{FFS}} = r_h + r_{2,p} + I_2 \Delta r_p$.[1]

We know from the previous results that the payer benefits at the unconstrained levels of intensity (i.e., I^* for FFS and $I^\sharp$ for BP), that is, $r^{\text{BP}} < r^{\text{FFS}}$. At the same time, the payer reimbursement at the optimal intensity I_0 is $r_{I_0}^{\text{BP}} - r_{I_0}^{\text{FFS}} = (\Delta c - \Delta r_p)(-1 + I_0 + \Psi)$. Since the payment is linear in intensity, we can derive that when $\Delta r_p < \Delta c$, the payer may reach high quality more easily under FFS than under bundled payments. In other words, when the physician reimbursement differential between the two pathways is lower than the hospital cost differential, FFS may offer higher quality for lower cost. This is unsurprising

[1] Observe that if we juxtapose the curves defined by $r_2^{\text{FFS}}(I_2)$ and $r_2^{\text{BP}}(I_2)$, we obtain Figures 2.4 and 2.5.

because under FFS, increasing intensity (hence in this case quality) increases physician reimbursement but not hospital reimbursement. Hence, if Δr_p is relatively low compared to Δc, quality improvements under FFS are relatively more attractive. In contrast, for $\Delta r_p > \Delta c$, the bundled payments will always offer better quality at a lower cost. Finally, if $I^* > I_0$, that is, FFS leads to overtreatment, BP may lessen overtreatment, but may also lead to undertreatment in which case the BP quality may be lower. Hence, the quality contrast between BP and FFS is a priori unclear (Case 1). This discussion finalizes the proof of Theorem 2.2.2. □

B.2.2 The Quality Model

Proof. Proof of Lemma 2.4.1 In this proof, we analyze the behavior of $F_h^{\text{FFS}},\ F_h^{\text{BP}}$, and explore their maxima. An optimal solution is a two-dimensional maximum of these two functions. Hence, we differentiate F_h^{FFS}, F_h^{BP}, and compute the best response in the interior. There is no joint best response in the interior for $\iota_h(\iota_p(i_h)) = i_h$ and $\iota_p(\iota_h(i_p)) = i_p$. Next, we check the boundary solutions, when$i_h = 0$ or $i_h = 1$ and $i_p = 0$ or $i_p = 1$ occurs. Specifically, we have the following cases:

- $i_h = 1,\ i_p < 1$: Then we have

$$\iota_p(1) = \frac{2I_0 w_b + \Delta r_p}{2w_b(1-\Psi)}, \tag{B.78}$$

and

$$I = I_0 + \frac{\Delta r_p}{2w_b}. \tag{B.79}$$

Hence, by the condition on $\iota_p(1)$ that $\iota_p(1) < 1$, this scenario ($i_h = 1,\ i_p < 1$) holds whenever

$$\Delta r_p \leq 2w_b(1 - \Psi - I_0), \tag{B.80}$$

as before, else $i_p \geq 1$. Formally,

$$\iota_h(\iota_p(1)) = 1 + \frac{w_b \Delta c + w_q^{\text{FFS}} \Delta r_p}{2 w_b w_q^{\text{FFS}} \Psi}, \tag{B.81}$$

which is always larger than 1, as needed.

- $i_h = 1,\ i_p = 1$: This joint response occurs when (B.80) does not hold. Then the hospital needs

  $$\frac{1 - I_0}{\Psi} + \frac{\Delta c}{2 w_q^{\text{FFS}} \Psi} \geq 1, \tag{B.82}$$

  in other words,

  $$\frac{\Delta c}{2 w_q^{\text{FFS}}} + 1 - I_0 \geq \Psi. \tag{B.83}$$

- $i_h = 0$. This case is impossible. Indeed, if $i_p = 0$, this gives the condition

  $$\frac{\Delta r_p}{2 w_b} < -(I_0 - \Psi) \tag{B.84}$$

  for the physicians and

  $$I_0 - \Psi > \frac{\Delta c}{2 w_q^{\text{FFS}}} \tag{B.85}$$

  for the hospital, and obviously they cannot both hold simultaneously. The condition on $i_p = 1$ gives

  $$\frac{\Delta c}{2 w_q^{\text{FFS}}} + (1 - I_0) \leq 0 \tag{B.86}$$

  for the hospital, which obviously does not hold. Finally, if (B.86) does not hold and we have $i_p < 1$, then the condition for $i_h = 0$ is $w_b \Delta c + w_q^{\text{FFS}} \Delta r_p \leq 0$, which never holds.

- Case $0 < i_h < 1$. Then we must have $i_p = 1$ or $i_p = 0$. Assume first that $i_p = 1$.

This assumption implies

$$1 + \frac{w_b \Delta c + w_q^{\text{FFS}} \Delta r_p}{2 w_b w_q^{\text{FFS}} (1 - \Psi)} \geq 1, \tag{B.87}$$

a condition that is automatically satisfied. Second, for the hospital we must have

$$0 < \frac{2 w_q^{\text{FFS}} (1 - I_0) + \Delta c}{2 w_q^{\text{FFS}} \Psi} < 1, \tag{B.88}$$

the last condition holds for $w_q^{\text{FFS}} > 0$ if

$$\Delta c < 2 w_q^{\text{FFS}} (-1 + I_0 + \Psi) \tag{B.89}$$

and for $w_q^{\text{FFS}} < 0$, the required condition is

$$-2 w_q^{\text{FFS}} (1 - I_0) > \Delta c > -2 w_q^{\text{FFS}} (1 - I_0 - \Psi). \tag{B.90}$$

Now consider $i_p = 0$. Then on the physician side, we need $\iota_p(\iota_h(0)) \leq 0$. This condition is impossible.

This captures all feasible cases in lemma, and concludes the proof. □

Proof. Proof of Proposition B.1.1:

The physicians' best response function remains the same as in the base model:

$$\iota_p(i_h, T) = \frac{1}{2 w_b (1 - \Psi)} (\Delta r_p - \frac{T}{1 - \Psi} + 2 w_b (I_0 - (1 - i_h) \Psi)). \tag{B.91}$$

For the hospital, the derivative with respect to i_h also remains the same:

$$\frac{\partial}{\partial i_h} F_h^{\text{BP}} (i_h, \iota_p(i_h, T), T) = T \frac{\Psi}{1 - \Psi}. \tag{B.92}$$

Hence, $i_h = 1$ whenever $T > 0$. We first consider $T > 0$. The optimal T is given by

$$T_{\text{low}} = (1-\Psi)\frac{w_q^{\text{BP}}\Delta r_p + w_b(\Delta c + 2\Delta r_p) - 2w_b^2(1 - I_0 - \Psi)}{2w_b + w_q^{\text{BP}}}. \tag{B.93}$$

Formula for T_{low} in (B.93) is larger than zero if

$$\frac{\Delta r_p}{w_b} + \frac{\Delta c + 2\Delta r_p}{w_q^{\text{BP}}} > 2\frac{w_b}{w_q^{\text{BP}}}(1 - I_0 - \Psi). \tag{B.94}$$

In order for condition in (B.94) to have an impact on the value of i_p, that is to change it to for $\iota_p < 1$, we need

$$\Delta c > 2(w_b + w_q^{\text{BP}})(-1 + I_0 + \Psi). \tag{B.95}$$

Finally, $\iota_p = 0$ will occur when the gainsharing amount is set to a maximum denoted by $T_{\max}$ such that

$$T_{\max} = (1-\Psi)(2I_0 w_b + \Delta r_p). \tag{B.96}$$

This condition prevails when

$$\Delta c \geq 2I_0 w_q^{\text{BP}} + 2w_b(1 + I_0 - \Psi). \tag{B.97}$$

Next, we consider the $T = 0$ case. Under the optimal intensity i_h, the derivative with respect to i_h is 0, so there will be an "indifference region", an interval of values for i_h such that the resulting intensity is the same for all i_h in this interval. Consider first the case when $\iota_p(i_h, 0) \in (0, 1)$ for $i_h \in [0, 1]$: Then it does not matter what i_h is chosen, in other words, the indifference region for i_h is the entire unit interval. It is easy to see that ι_p is increasing with i_h, so we can consider $\iota_p(1, 0)$ and $\iota_p(0, 0)$. We have that $\iota_p(1, 0) = 1$ when

$$\frac{\Delta r_p}{2w_b} \geq 1 - \Psi - I_0. \tag{B.98}$$

We have $\iota_p(0,0) = 0$ when

$$\frac{\Delta r_p}{2w_b} \leq \Psi - I_0 \tag{B.99}$$

Consider the smallest point i_h^+ such that $\iota_p(i_h^+, 0) = 1$ and largest point i_h^- such that $\iota_p(i_h^-, 0) = 0$. In both points (i_h^- and i_h^+), the derivative w.r.t. i_h is $(\Delta c + \Delta r_p + \frac{w_q^{\text{BP}} \Delta r_p}{w_b})\Psi > 0$. Hence, the hospital will never choose a quantity below i_h^-, i.e., below the indifference interval. However, the other point, i_h^+, is more interesting. The hospital will seek a point above i_h^+, in particular, it would ideally choose the point

$$i_h^* = \frac{2w_q^{\text{BP}}(1 - I_0) + \Delta c + \Delta r_p}{2w_q^{\text{BP}}\Psi}, \tag{B.100}$$

unless

$$\frac{\Delta c + \Delta r_p}{2w_q^{\text{BP}}} \geq -1 + I_0 + \Psi, \tag{B.101}$$

in which case $i_h^* = 1$.

The respective intensities for i_h indifferent, at i_h^*, and at 1 are then given by,

$$I^\sharp = I_0 + \frac{\Delta r_p}{2w_b} \qquad \text{(at } i_h \text{ indifferent)}$$

$$I^\sharp = I_0 - \frac{\Delta c + \Delta r_p}{2w_q^{\text{BP}}} \qquad \text{(at } i_h^*\text{; follows from (B.100))}$$

$$I^\sharp = \Psi \qquad \text{(at } i_h\text{=1)}$$

□

Proof. Proof of Proposition 2.4.1: We need to ensure that both physicians and hospital benefit, while the payer does not lose under BP, compared with FFS. That is, we need:

$$F_p^{\text{BP}}(i_h^\sharp, i_p^\sharp, T^\sharp) > F_p^{\text{FFS}}(i_h^*, i_p^*)\ , F_h^{\text{BP}}(i_h^\sharp, i_p^\sharp, T^\sharp) > F_h^{\text{FFS}}(i_h^*, i_p^*) \text{ and } r^{\text{BP}} < r_h + r_{2,p} + I^* \Delta r_p.$$

Similar to the analysis before, we consider the cases $T = 0$ and $T > 0$ separately.

Under $T = 0$, bundling does not occur, and we derive the conditions under $T > 0$. In the following, we explore the two cases.

Case $T = 0$: As we discussed in an earlier proof, there are three cases, depending on conditions (B.98) and (B.101):

1. The case when the hospital is indifferent for varying i_h. Because, hospital payoff is the same irrespective of i_h, we simply take $i_h = 1,\ i_p = \iota_p(1, 0)$
2. The hospital exerts the influencing effort i_h above the indifference region where $i_h < 1$ also holds. Then $i_p = 1$ and i_h given by (B.100).
3. The hospital exerts the maximal i_h where $i_h = 1, i_p = 1$.

We match the above cases to those in Lemma 2.4.1, and then we determine if the comparison of the BP and FFS payoffs can give rise to bundling. Note that, showing the intensity being the same under the FFS and the BP is enough to show that bundling will not occur.

The hospital is indifferent for all i_h under bundled payments only when $\Psi \leq \bar{\Psi}$. This fact leads to the same intensity $I^* = I^\sharp = I_0 + \frac{\Delta r_p}{2w_b}$ under BP and FFS, so there will be no bundling.

The case in (B.9) can on the one hand arise from $\Psi \leq \bar{\Psi}$ if the first and third conditions in (B.9) hold and $1 - I_0 - \Psi > 0$. In this case, bundling would lead to a reduction in intensity to the extent that physicians would lose and therefore bundling will not occur. The case (B.9) can also arise from $\bar{\Psi} \leq \Psi \leq \hat{\Psi}$ if the second and third conditions in (B.9) hold. In this case, intensity is the same, $1 - \Psi$, under both FFS and BP, so bundling will not occur. Case (B.9) can also arise from $\Psi \geq \hat{\Psi}$, again, if the second and third conditions hold. This case will not be appealing for the physicians to bundle either.

Finally, the case (B.7) can only arise from $\Psi \geq \hat{\Psi}$. When $\Psi \geq \hat{\Psi}$, it follows that the physicians actually lose, so bundling will not occur. We also check if (B.7) and (B.1) might occur as two simultaneous equilibria. Suppose $F_h^{\text{BP}}(1, \iota_p(1, T), T) > F_h^{\text{BP}}(1, \iota_p(1, 0), 0)$

and $F_h^{\text{BP}}(1, \iota_p(1, T), T) < F_h^{\text{BP}}(\iota_h(1), 1, 0)$ co-occur. In this case, both solutions would be local optima and we can compare the corresponding objective values. However, we will argue that this does not occur. First, they could only occur together under the scenario $\bar{\Psi} \leq \Psi \leq \hat{\Psi}$. But then, one finds out that conditions for (B.7) actually cannot occur under $\bar{\Psi} \leq \Psi \leq \hat{\Psi}$. Indeed, we then have that the condition $\Psi \leq \hat{\Psi}$ conflicts with the right-hand side of the third equation in (B.7).

Case $T > 0$: Again, we need to first compute $F_p^{\text{BP}}(i_h^\sharp, i_p^\sharp, T^\sharp) - F_p^{\text{FFS}}(i_h^*, i_p^*)$ to determine the physician benefit/loss. First, consider $\Psi \geq \hat{\Psi}$. In this case, bundling is not possible because the second condition in (B.1) cannot hold. Let us then consider the payoff difference under the two cases $\Psi \leq \bar{\Psi}$ and $\bar{\Psi} \leq \Psi \leq \hat{\Psi}$.

- For $\Psi \leq \bar{\Psi}$: If T reaches $T_{\max}$, the smallest amount of gainsharing such that $\iota_p(1, T_{\max}) = 0$, the physicians clearly benefit. On the other hand, if it stays at at a level T_{low} below $T_{\max}$, we write:

$$\begin{aligned} &F_p^{\text{BP}}(1, \iota_p(1, T_{\text{low}}), T_{\text{low}}) - F_p^{\text{FFS}}(1, \iota_p^{\text{FFS}}(1)) \\ &= -\frac{w_q \Delta r_p + w_b(\Delta c + 2\Delta r_p) + 2w_b^2(-1 + I_0 + \Psi)}{4w_b(2w_b + w_q)^2} \cdot \\ &\qquad (w_q \Delta r_p + 6w_b^2(-1 + I_0 + \Psi) + w_b(-\Delta c + 2\Delta r_p + 4w_q(-1 + I_0 + \Psi))). \end{aligned} \tag{B.102}$$

The first product term is clearly positive which follows from the first inequality in

(B.1). As for the second one, we show that it is less than 0 such that

$$\begin{aligned}
&w_q\Delta r_p + 6w_b^2(-1+I_0+\Psi) + w_b(-\Delta c + 2\Delta r_p + 4w_q(-1+I_0+\Psi)) \\
&\leq w_q\Delta r_p + 6w_b^2(-1+I_0+\Psi) + w_b(-2w_b(-1+I_0+\Psi) + 2\Delta r_p + 2w_q(-1+I_0+\Psi)) \\
&\leq w_q\Delta r_p + 6w_b^2(-1+I_0+\Psi) + w_b(\Delta r_p + 2w_q(-1+I_0+\Psi)) \\
&\leq 4w_b^2(-1+I_0+\Psi) < 0.
\end{aligned} \tag{B.103}$$

Hence, overall, the physicians will benefit.

- In case $\bar{\Psi} \leq \Psi \leq \hat{\Psi}$, physicians will also clearly benefit.

Now, we compute $r^{\text{BP}} \geq -(F_h^{\text{BP}}(i_h^\sharp, i_p^\sharp, T^\sharp) - F_h^{\text{FFS}}(i_h^*, i_p^*, T^*))$, where r^{BP} is excluded from the F_h^{BP} formula. Of course, we need to distinguish the cases for different I^* low-integration vs. moderate-integration levels and T low or high.

To see how the payer likes the scenario in this bullet point, we would compute

$$r^{\text{BP}} - r_h - r_{2,p} - \Delta r_p I^*. \tag{B.104}$$

In contrast to the base model, not all solutions that are appealing for the physicians are also appealing for the payer. The payer conditions are relatively complicated and we do not list all of them. However, for an illustration, we provide below the payer surplus (or deficit) when $\Psi \leq \bar{\Psi}$ and conditions from (B.1) both hold:

$$\begin{aligned}\Delta\pi = {} & \frac{1}{4w_b^2(2w_b + w_q^{\text{BP}})}(w_q^{\text{BP}} w_q^{\text{FFS}}(2I_0 + \Delta r_p)^2 \\ & - 2w_b(2I_0 + \Delta r_p)(2I_0(-1 + w_q^{\text{BP}})w_q^{\text{FFS}} - w_q^{\text{BP}}\Delta c - (w_q^{\text{BP}} + w_q^{\text{FFS}})\Delta r_p) \\ & + 4w_b^3(2I_0^2 w_q^{\text{FFS}} - I_0\Delta c + (\Delta c + 2\Delta r_p)(-1 + \Psi)) \\ & + 4w_b^4(-1 + I_0 + \Psi)^2 \\ & + w_b^2(4I_0^2(-4 + w_q^{\text{BP}})w_q^{\text{FFS}} - 4w_q^{\text{BP}}\Delta r_p + (\Delta c + 2\Delta r_p)^2 + I_0(8\Delta c - 4w_q^{\text{BP}}\Delta c + 8\Delta r_p - 8w_q^{\text{FFS}}\Delta r_p)\end{aligned} \tag{B.105}$$

□

B.2.3 The Salary Model

Proof. Proof of Lemma 2.4.2: The hospital's best response in the interior is

$$i_h^* = 1 - I_0 + \frac{\Delta c - \Delta r_p}{2w_q}. \tag{B.106}$$

The solution is equal to one if

$$\Delta c - \Delta r_p \geq 2w_q I_0 \tag{B.107}$$

and equal to zero if

$$\Delta r_p \geq 2w_q(1 - I_0) + \Delta c. \tag{B.108}$$

Plugging in the formula for I, this yields the desired result.

□

Proof. Proof of Propositions 2.4.2 and 2.4.3: Taking a similar approach as in Lemma 2.4.2, the optimal hospital solution under bundled payments is

$$i_h^\sharp = 1 - I_0 + \frac{\Delta c}{2w_q^{\text{FFS}}}. \tag{B.109}$$

This is always larger than 0, but the case $i_h^\sharp \geq 1$ must be again treated separately. In particular $i_h^\sharp = 1$ if

$$\Delta c \geq 2w_q^{\text{FFS}} I_0. \tag{B.110}$$

We first compute the difference $F_h^{\text{BP}}(i_h^\sharp) - r^{\text{BP}} - F_h^{\text{FFS}}(i_h^*)$, which yields several cases. This yields the lower bound on r^{BP} that the hospital will require for bundling to occur. We then check if the payer is not worse off, i.e. whether r^{BP} satisfies $r^{\text{BP}} < r_h + r_{2,p} + I^* \Delta r_p$ (note that r^{BP} is a strict lower bound, so for the payer not to be worse off, the inequality has to be strict). Let us consider the cases:

- If

  $$-2w_q(1 - I_0) \leq \Delta c - \Delta r_p \text{ and } \Delta c \leq 2w_q I_0, \tag{B.111}$$

  then

  $$r^{\text{BP}} > r_h + r_{2,p} + \Delta r_p I_0 - \frac{\Delta r_p(2\Delta c - \Delta r_p)}{4w_q}. \tag{B.112}$$

  The payer can save as much as

  $$\frac{\Delta r_p^2}{4w_q} > 0. \tag{B.113}$$

- If

  $$-2w_q(1 - I_0) \leq \Delta c - \Delta r_p \leq 2w_q I_0 \text{ and } \Delta c \geq 2w_q I_0, \tag{B.114}$$

  then

  $$r^{\text{BP}} > r_h + r_{2,p} + \frac{(2I_0 w_q - \Delta c + \Delta r_p)^2}{4w_q}. \tag{B.115}$$

The payer can save as much as

$$\frac{\Delta r_p^2 - (\Delta c - 2I_0 w_q)^2}{4w_q} \geq 0, \tag{B.116}$$

which is positive because

$$\Delta c - 2w_q I_0 \leq \Delta r_p. \tag{B.117}$$

- If

$$\Delta c - \Delta r_p \geq 2w_q I_0, \tag{B.118}$$

then

$$r^{\text{BP}} > r_h + r_{2,p}. \tag{B.119}$$

In this case, the payer cannot generate enough to recuperate the costs, and hence, bundling will not occur.

- If

$$-2w_q(1 - I_0) \geq \Delta c - \Delta r_p \text{ and } \Delta c \geq 2w_q I_0, \tag{B.120}$$

then

$$r^{\text{BP}} \geq r_h + r_{2,p} - w_q(1 - 2I_0) - \Delta c + \Delta r_p. \tag{B.121}$$

The payer saves

$$\Delta c + (1 - 2I_0)w_q > 0, \tag{B.122}$$

This is positive because

$$\Delta c \geq 2w_q I_0. \tag{B.123}$$

- If

$$-2w_q(1 - I_0) \leq \Delta c - \Delta r_p \leq 2w_q I_0 \text{ and } \Delta c \leq 2w_q I_0, \tag{B.124}$$

then

$$r^{\text{BP}} \geq r_h + r_{2,p} + \frac{(2I_0 w_q - \Delta c)^2}{4w_q}. \tag{B.125}$$

The payer saves

$$\frac{(\Delta c - 2I_0 w_q)^2}{4w_q} > 0. \tag{B.126}$$

- If

$$-2w_q(1 - I_0) \geq \Delta c - \Delta r_p \text{ and } \Delta c \leq 2w_q I_0, \tag{B.127}$$

then

$$r^{\text{BP}} \geq r_h + r_{2,p} - \Delta r_p + \frac{(2(1 - I_0)w_q - \Delta c)^2}{4w_q}. \tag{B.128}$$

The payer saves

$$\frac{(\Delta c + 2(1 - I_0)w_q)^2}{4w_q} > 0. \tag{B.129}$$

This completes the analysis of all of the cases, and concludes the proof.

□

Comparing the Salary and Quality models

Proof. Proof of Theorem 2.4.1: Our objective is to identify the conditions under which bundling would be preferred in at least one of the following two models: quality model and salary model.

We first analyze the cases when both models lead to bundling. Consider the case in the quality model where $\Psi \to 1$. We then always have $I_0 + \Psi - 1 > 0$. Note that (B.1) is impossible because the second and the third conditions contradict each other. If both (B.3) and $\Psi \leq \bar{\Psi}$ hold, then from (2.37)

$$\Delta c < \Delta r_p + 2w_b I_0 < 2(w_b + w_q^{\text{BP}})I_0, \tag{B.130}$$

which contradicts the second condition in (B.3). Hence, if bundling occurs under the salaried model, then we must have $\bar{\Psi} \leq \Psi \leq \hat{\Psi}$ and (B.3). Note that under $\Psi \to 1$, neither condition (B.1) nor $\Psi \leq \bar{\Psi}$ will ever hold in the quality model. Therefore, there are cases when bundling occurs under the quality model but not in the salary model.

Next, we demonstrate a case where the Salary model can lead to bundling but not the Quality model. Specifically, we are focusing on the case $\bar{\Psi} \leq \Psi \leq \hat{\Psi}$, and we see that $\Delta c \geq 2w_q I_0$, which leaves us with cases (B.120) and (B.114). Also, we see that bundling can occur in the salary model under the parameter values that would not lead to bundling in the quality model, namely when $2w_q I_0 > \Delta c$. □

B.2.4 The Full Quality Model

Proof. Proof of Lemma B.1.1 As before, we seek Nash equilibria,

- There is no interior solution. That is, there is no point, $(i_h, i_p) \in (0,1)^2$, such that $\iota_p(\iota_h(i_p)) = i_p$ and vice versa.

- There are two possibilities for a solution with $i_h = 1$. This requires $\iota_p(1) = i_p^*$ and $\iota_h(i_p^*) \geq 1$ for some i_p^*. Consider the three potential cases for $i_h = 1$:

(H1.1) $i_p \in (0,1)$, then $\iota_p(1) = \frac{2I_0 + \Delta r_p}{2w_b(1-\Psi)}$. Hence, this must be lower than 1, which requires $\frac{\Delta r_p}{2w_b} - (1 - I_0 - \Psi) < 0$, which is equivalent to $\Psi < \bar{\Psi}$. Furthermore, it must hold that $\iota_h(i_p) \geq 1$. This requires $\frac{w_b}{w_q^{\text{FFS}}}\Delta c + \Delta r_p \geq 0$.

(H1.2) $i_p = 1$. From the above, we deduce that $\iota_p(1) \geq 1$ which happens if $\Psi > \bar{\Psi}$. Furthermore, we need to have $\iota_h(1) \geq 1$. That happens if $\frac{\Delta c}{2w_q^{\text{FFS}}} > -(1 - I_0 - \Psi)$, that is $\Psi > 1 - I_0 + \frac{\Delta c}{2w_q^{\text{FFS}}}$.

(H1.3) There is no solution with $i_p = 0$ because $\iota_p(1) > 0$

- There are solutions with $i_h = 0$. Again, we review different physicians' choices separately:

(H0.1) We can have $i_p = 0$ if $\frac{\Delta r_p}{2w_b} \leq \Psi - I_0$. Then (to ensure $\iota_h(0) \leq 0$), we also need $\frac{\Delta c}{2w_q^{\text{FFS}}} \leq I_0 - \Psi$.

(H0.2) We can have $i_p = 1$ if $\frac{\Delta r_p}{2w_b} \geq 1 - I_0$. Then (to ensure $\iota_h(1) \leq 0$), we also need $1 - I_0 + \frac{\Delta c}{2w_q^{\text{FFS}}} \leq 0$.

(H0.3) We can have $i_p \in (0,1)$, if $\iota_p(0) \in (0,1)$. From the previous, this requires $1 - I_0 > \frac{\Delta r_p}{2w_b} > \Psi - I_0$. Then (to ensure $\iota_h(\iota_p(1)) \leq 0$), we also need $\frac{w_b}{w_q^{\text{FFS}}}\Delta c + \Delta r_p \leq 0$.

- There are solutions with $i_h \in (0,1)$. Here, it remains to consider two cases for physician response because the interior solution is impossible:

(Hi.1) $i_p = 1$. This will happen if $\iota_p(\iota_h(1)) \geq 1$ and $\iota_h(1) \in (0,1)$. For the first condition, we need $\frac{w_b}{w_q^{\text{FFS}}}\Delta c + \Delta r_p \geq 0$. For the second one, we need $0 < 1 - I_0 + \frac{\Delta c}{2w_q^{\text{FFS}}} < \Psi$.

(Hi.2) $i_p = 0$. This will happen if $\iota_p(\iota_h(0)) \leq 0$ and $\iota_h(0) \in (0,1)$. The first happens if $\frac{w_b}{w_q^{\text{FFS}}}\Delta c + \Delta r_p \leq 0$ while the second one requires $0 < \Psi - I_0 + \frac{\Delta c}{w_q^{\text{FFS}}} < \Psi$.

□

Here, we will provide bundled payments solutions under the full quality model. We postpone quantifying r^{BP} to the next proposition.

Proposition 2.4.1 (Complete Characterization)**.** *Not taking feasibility into account, there are five cases of bundling solutions:*

(G0) If

$$\begin{aligned} &1 < I_0 + \Psi + \frac{\Delta c + 2\Delta r_p}{2w_b} + \frac{w_q^{BP}}{2w_b^2}\Delta r_p \textit{ and} \\ &I_0\frac{w_q^{BP}}{w_b} + (1 + I_0 - \Psi) > \frac{\Delta c}{2w_b} \textit{ and} \\ &\Delta c > 2(w_b + w_q^{BP})(1 - I_0 - \Psi), \end{aligned} \tag{B.131}$$

then

$$\begin{aligned}
&i_h = 1 \\
&i_p = \frac{2I_0 w_q^{BP} + 2w_b(1 + I_0 - \Psi) - \Delta c}{2(2w_b + w_q^{BP})(1 - \Psi)} \\
&T = (1 - \Psi)\frac{2w_b^2(-1 + I_0 + \Psi + \frac{\Delta c + 2\Delta r_p}{2w_b}) + w_q^{BP}\Delta r_p}{2w_b + w_q^{BP}}
\end{aligned} \tag{B.132}$$

(GM) If

$$\begin{aligned}
&1 < I_0 + \Psi + \frac{\Delta c + 2\Delta r_p}{2w_b} + \frac{w_q^{BP}}{2w_b^2}\Delta r_p \textit{ and} \\
&I_0\frac{w_q^{BP}}{w_b} + (1 + I_0 - \Psi) \leq \frac{\Delta c}{2w_b} \textit{ and} \\
&\Delta c > 2(w_b + w_q^{BP})(1 - I_0 - \Psi),
\end{aligned} \tag{B.133}$$

then

$$\begin{aligned}
&i_h = 1 \\
&i_p = 0 \\
&T = 2w_b(1 - \Psi)(I_0 + \frac{\Delta r_p}{2w_b})
\end{aligned} \tag{B.134}$$

(N0) If

$$\begin{aligned}
&1 \geq I_0 + \Psi + \frac{\Delta c + 2\Delta r_p}{2w_b} + \frac{w_q^{BP}}{2w_b^2}\Delta r_p \textit{ or} \\
&\Delta c \leq 2(w_b + w_q^{BP})(1 - I_0 - \Psi) \textit{ and} \\
&\Psi \leq \bar{\Psi}
\end{aligned} \tag{B.135}$$

then

$$
\begin{aligned}
& i_h \in [\max(0, \frac{\Psi - I_0 - \frac{\Delta r_p}{2w_b}}{\Psi}), 1] \\
& i_p \in [\max(0, \frac{\frac{\Delta r_p}{2w_b} - (1 - i_h)\Psi + I_0}{(1 - \Psi}), 1] \\
& T = 0
\end{aligned} \tag{B.136}
$$

(NM) *If*

$$
\begin{aligned}
& 1 \geq I_0 + \Psi + \frac{\Delta c + 2\Delta r_p}{2w_b} + \frac{w_q^{BP}}{2w_b^2}\Delta r_p \textit{ or} \\
& \Delta c \leq 2(w_b + w_q^{BP})(1 - I_0 - \Psi) \textit{ and} \\
& \Psi > \bar{\Psi} \textbf{\textit{ and}} \\
& \frac{\Delta c + \Delta r_p}{2w_q^{BP}} < 1 - I_0 - \Psi
\end{aligned} \tag{B.137}
$$

then

$$
\begin{aligned}
& i_h = \frac{1}{\Psi}(1 - I_0 + \frac{\Delta c + \Delta r_p}{2w_q^{BP}}) \\
& i_p = 1 \\
& T = 0
\end{aligned} \tag{B.138}
$$

(N1) *If*

$$
\begin{aligned}
& 1 \geq I_0 + \Psi + \frac{\Delta c + 2\Delta r_p}{2w_b} + \frac{w_q^{BP}}{2w_b^2}\Delta r_p \textit{ or} \\
& \Delta c \leq 2(w_b + w_q^{BP})(1 - I_0 - \Psi) \textit{ and} \\
& \Psi > \bar{\Psi} \textbf{\textit{ and}} \\
& \frac{\Delta c + \Delta r_p}{2w_q^{BP}} \geq 1 - I_0 - \Psi
\end{aligned} \tag{B.139}
$$

then

$$\begin{aligned} i_h &= 1 \\ i_p &= 1 \\ T &= 0 \end{aligned} \tag{B.140}$$

Proof. Proof of Proposition 2.4.1 As in the previous BP models, we solve for a two-stage equilibrium. In the first stage, the hospital sets i_h and T while in the second stage, physicians set i_p. We solve it by backward induction, but we distinguish three cases depending on $T > 0$ and $i_p = 1, i_p = 0, i_p \in (0,1)$. Note that for $T > 0$, we have the derivative of $F_h > 0$, so the hospital will choose $i_h = 1$.

Let's compute the hospital response in face of this.

1. $i_p \in (0,1)$: Then the derivatives of F_h w.r.t. i_h and T are $\frac{T\Psi}{(1-\Psi)}$ as before and respectively, the optimal T is

   $$T_0 = (1-\Psi)\frac{2w_b^2(-1+I_0+\Psi+\frac{\Delta c+2\Delta r_p}{2w_b}) + w_q^{\text{BP}}\Delta r_p}{2w_b + w_q^{\text{BP}}}. \tag{B.141}$$

   Hence, T_0 will be positive unless

   $$1 \geq I_0 + \Psi + \frac{\Delta c + 2\Delta r_p}{2w_b} + \frac{w_q^{\text{BP}}}{2w_b^2}\Delta r_p, \tag{B.142}$$

   from which we deduce the first condition. Now, we only need to ensure that under these conditions, $i_p \in (0,1)$. Plugging T_0 and $i_h = 1$ back into ι_p gives:

   $$i_p^\sharp = 2w_b\frac{I_0\frac{w_q^{\text{BP}}}{w_b} - \frac{\Delta c}{2w_b} + (1+I_0-\Psi)}{2(2w_b+w_q^{\text{BP}})(1-\Psi)}. \tag{B.143}$$

Hence, $i_p^\sharp$ will be within $(0, 1)$ if

$$\begin{aligned} &I_0 \frac{w_q^{\text{BP}}}{w_b} + (1 + I_0 - \Psi) > \frac{\Delta c}{2w_b} \text{ and} \\ &\Delta c > 2(w_b + w_q^{\text{BP}})(1 - I_0 - \Psi). \end{aligned} \tag{B.144}$$

We next consider the other cases.

2. $i_p = 0$: Then the hospital will just stop with T such that i_p hits 0, and it will not go further. This means

$$T_{\max} = 2w_b(1 - \Psi)(I_0 + \frac{\Delta r_p}{2w_b}), \tag{B.145}$$

as before.

3. $T = 0$: This case may mean, that $\Delta c \leq 2(w_b + w_q^{\text{BP}})(1 - I_0 - \Psi)$ and i_p stays at 1 or it may mean that T_0 would be negative, i.e.,

$$1 < I_0 + \Psi + \frac{\Delta c + 2\Delta r_p}{2w_b} + \frac{w_q^{\text{BP}}}{2w_b^2} \Delta r_p. \tag{B.146}$$

In any case, as we discussed elsewhere, there will be an indifference zone for i_h for the ranges of i_p such that $\iota_p(i_h) \in (0, 1)$, but if this range does not cover the entire range of i_h, the hospital can also choose to go above or below the indifference zone. We have $\iota_p(i_h) = \frac{\frac{\Delta r_p}{2w_b} - (1 - i_h - I_0)\Psi}{1 - \Psi}$. This term is increasing as i_h increases. The derivative of F_h^{BP} w.r.t. i_h is decreasing with i_h and is positive at i_h^-, the largest point such that $i_p = 0$. Hence, the hospital will always prefer i_h in the indifference zone over lower i_h. The derivative is also positive at i_h^+, the largest point in the indifference zone (unless i_h^+ exceeds 1). The indifference zone spans up to $i_h = 1$ if $\iota_p(1, 0) \leq 1$, which is the case if $\frac{\Delta r_p}{w_b} - (1 - \Psi - I_0) \leq 0$, that is, $\Psi \leq \bar{\Psi}$. Otherwise, the hospital has the chance to improve its utility by increasing i_h beyond the indifference zone.

The optimal i_h is then

$$i_h = \frac{1}{\Psi}(1 - I_0 + \frac{\Delta c + \Delta r_p}{2w_q^{\mathrm{BP}}}). \tag{B.147}$$

This expression is larger than one (hence the optimum being one) if $\frac{\Delta c+\Delta r_p}{2w_q^{\mathrm{BP}}} \geq 1 - I_0 - \Psi$ and otherwise the optimum is the quantity from (B.147).

□

Proof. Proof of Proposition B.1.2 We consider all pairs of cases from the above two propositions and ask the following questions: 1) Is the combination feasible? That is, are the values of the parameters that give rise to the combination compatible? 2) If bundling is to occur, would physicians benefit?, 3) If bundling is to occur, would the payer benefit? Note that we are unable to fully evaluate the feasibility of all the cases. There is also a condition on r^{BP}, that is case dependent, but we do not list it below. However, the formula for the payer's profit includes it and assumes that the payer pays the hospital this minimal r^{BP}. We label the cases in format `(FFS case-BP case)`.

Consider the cases:

(HM-G0) : The physician surplus is

$$\begin{aligned}
&-\frac{1}{4w_b(2w_b + w_q^{\mathrm{BP}})^2)}(-4I_0^2(w_b^2(4 + (-8 + w_b)w_b) + 2w_b(2 + (-4 + w_b)w_b)w_q^{\mathrm{BP}} \\
&+ (-1 + w_b)^2(w_q^{\mathrm{BP}})^2) \\
&+ (w_q^{\mathrm{BP}}\Delta r_p + w_b(\Delta c + 2\Delta r_p) + 2w_b^2(-1 + \Psi)) \\
&\cdot (w_q^{\mathrm{BP}}\Delta r_p + w_b(-\Delta c + 2\Delta r_p + 4w_q^{\mathrm{BP}}(-1 + \Psi)) + 6w_b^2(-1 + \Psi)) \\
&+ 4I_0w_b((w_q^{\mathrm{BP}})^2\Delta r_p + w_bw_q^{\mathrm{BP}}(\Delta c + 4\Delta r_p) + w_b^2(\Delta c + 4\Delta r_p + 4w_q^{\mathrm{BP}}(-1 + \Psi)) \\
&+ 6w_b^3(-1 + \Psi)))
\end{aligned} \tag{B.148}$$

This expression needs to be positive. The payer surplus is

$$\begin{aligned}
&\frac{1}{4w_b^2(2w_b + w_q^{\text{BP}})}(w_q^{\text{BP}} w_q^{\text{FFS}}(2I_0 + \Delta r_p)^2 \\
&- 2w_b(2I_0 + \Delta r_p)(2I_0(-1 + w_q^{\text{BP}})w_q^{\text{FFS}} - w_q^{\text{BP}}\Delta c - (w_q^{\text{BP}} + w_q^{\text{FFS}})\Delta r_p) \\
&+ 4w_b^3(2I_0^2 w_q^{\text{FFS}} - I_0\Delta c + (\Delta c + 2\Delta r_p)(-1 + \Psi)) \\
&+ 4w_b^4(-1 + I_0 + \Psi)^2 \\
&+ w_b^2(4I_0^2(-4 + w_q^{\text{BP}})w_q^{\text{FFS}} - 4w_q^{\text{BP}}\Delta r_p + (\Delta c + 2\Delta r_p)^2 \\
&+ I_0(8\Delta c - 4w_q^{\text{BP}}\Delta c + 8\Delta r_p - 8w_q^{\text{FFS}}\Delta r_p) + 4w_q^{\text{BP}}\Delta r_p\Psi)),
\end{aligned} \tag{B.149}$$

which again has to be positive.

(HM-GM) : The physician surplus is

$$(I_0^2(4 - 8w_b) - 4I_0 w_b(\Delta r_p + 2w_b(-1 + \Psi)) - \Delta r_p(\Delta r_p + 4w_b(-1 + \Psi)))/(4w_b) \tag{B.150}$$

This needs to be positive. The payer surplus is

$$\begin{aligned}
&\frac{1}{4w_b^2}(4I_0^2(-w_b^2 w_q^{\text{BP}} + (-1 + w_b)^2 w_q^{\text{FFS}}) \\
&+ \Delta r_p(w_q^{\text{FFS}}\Delta r_p + 2w_b(\Delta c + \Delta r_p) + 4w_b^2(-1 + \Psi)) \\
&+ 4I_0(w_q^{\text{FFS}}\Delta r_p + w_b(\Delta c + \Delta r_p - w_q^{\text{FFS}}\Delta r_p) + 2w_b^3(-1 + \Psi)))
\end{aligned} \tag{B.151}$$

which again has to be positive.

(HM-N0) : The physician surplus here would be zero, so bundling will not be feasible.

(HM-NM) : The physician surplus would be negative, so bundling will not be feasible.

(HM-N1) : The physician surplus would be negative, so bundling will not be feasible.

(HH-G0) : This solution seems feasible. Furthermore, the physician surplus will always be positive. The payer surplus will be

$$\begin{aligned}&\frac{1}{4(2w_b+w_q^{\text{BP}}))}(\Delta c^2+4w_b^2(-1+I_0+\Psi)^2+\\&+4w_q^{\text{BP}}(-1+I_0+\Psi)(-\Delta c+w_q^{\text{FFS}}(-1+I_0+\Psi))\\&+4w_b(-1+I_0+\Psi)(-\Delta c+2w_q^{\text{FFS}}(-1+I_0+\Psi))),\end{aligned} \tag{B.152}$$

which must be positive for bundling to occur.

(HH-GM) : This solution might be jointly feasible. The physicians will benefit. The payer surplus will be

$$-I_0^2(w_q^{\text{BP}}-w_q^{\text{FFS}})-2I_0(w_b+w_q^{\text{FFS}})+(1-\Psi)(\Delta c+(1-\Psi)w_q^{\text{FFS}}). \tag{B.153}$$

This seems most likely to be negative, so bundling would most likely not happen.

(HH-N0) : This will not be possible because the condition for (N0) clashes with $\Psi \geq \bar{\Psi}$, which is needed for the HH case.

(HH-NM : Then the physician surplus is

$$\frac{1}{4w_b}(-\Delta c-\Delta r_p-2w_b(1-I_0-\Psi))(\Delta c+3\Delta r_p-2w_b(1-I_0-\Psi)). \tag{B.154}$$

The first product term will be negative, and the second one positive, overall therefore negative, so the physicians will not want to bundle.

(HH-N1) : Then the physician surplus will be zero, so the physicians will not want to bundle.

(LL-G0) : The physician surplus is

$$\begin{aligned}
&\frac{1}{4(2w_b + w_q^{\text{BP}})^2)}(4(w_q^{\text{BP}})^2\Delta r_p(1-2\Psi) + 4w_b^3(-3+I_0^2+I_0(6-14\Psi)+\Psi(6+\Psi))\\
&+ w_b(4I_0^2(w_q^{\text{BP}})^2 + \Delta c^2 + 4(w_q^{\text{BP}})^2\Psi^2 - 4I_0 w_q^{\text{BP}}(\Delta c + 2w_q^{\text{BP}}\Psi)\\
&+ 4w_q^{\text{BP}}(\Delta c + 4\Delta r_p - (\Delta c + 8\Delta r_p)\Psi))\\
&+ 4w_b^2(4\Delta r_p - 8\Delta r_p\Psi - \Delta c(-1+I_0+\Psi) + 2w_q^{\text{BP}}(-1+I_0^2+I_0(2-6\Psi)+\Psi(2+\Psi)))),
\end{aligned} \tag{B.155}$$

which has to be positive for bundling to occur. The payer surplus is

$$\begin{aligned}
&\frac{1}{4(2w_b + w_q^{\text{BP}})}(4I_0^2 w_q^{\text{BP}} w_q^{\text{FFS}}\\
&+ \Delta c^2 - 4w_q^{\text{BP}}\Delta r_p + 4w_q^{\text{BP}}(\Delta c + 2\Delta r_p)\Psi + 4w_q^{\text{BP}} w_q^{\text{FFS}}\Psi^2\\
&+ 4w_b^2(-1+I_0+\Psi)^2\\
&- 4I_0 w_q^{\text{BP}}(\Delta c + 2w_q^{\text{FFS}}\Psi)\\
&+ 4w_b(2I_0^2 w_q^{\text{FFS}} - \Delta c - 2\Delta r_p + \Psi(3\Delta c + 4\Delta r_p + 2w_q^{\text{FFS}}\Psi) - I_0(\Delta c + 4w_q^{\text{FFS}}\Psi)))
\end{aligned} \tag{B.156}$$

(LL-GM) : The physician surplus is

$$\Delta r_p + I_0 w_b(2-4\Psi) - 2\Delta r_p\Psi + w_b\Psi^2. \tag{B.157}$$

The payer surplus is

$$-I_0^2(w_q^{\text{BP}} - w_q^{\text{FFS}}) - \Delta r_p - 2I_0(w_b(1-\Psi) + w_q^{\text{FFS}}\Psi) + \Psi(\Delta c + 2\Delta r_p + w_q^{\text{FFS}}\Psi). \tag{B.158}$$

(LL-N0) : Then, physicians will always benefit. The payer might benefit if the following is positive:

$$
\begin{aligned}
&I_0^2 w_q^{\text{FFS}} - \frac{\Delta r_p(w_q^{\text{BP}}\Delta r_p + 2w_b(\Delta c + \Delta r_p))}{4w_b^2} \\
&+ (\Delta c + \Delta r_p)\Psi + w_q^{\text{FFS}}\Psi^2 - I_0(\Delta c + \Delta r_p + 2w_q^{\text{FFS}}\Psi).
\end{aligned}
\tag{B.159}
$$

Note that $w_q^{\text{FFS}} < 0$ in this FFS case.

(LL-NM) : Similarly as before in the NM case, the physician surplus will be negative, so there will be no bundling.

(LL-N1) : The physician surplus will be

$$(1 - 2\Psi)(\Delta r_p + (2I_0 - 1)w_b), \tag{B.160}$$

which needs to be positive for bundling. The payer surplus will be

$$w_q^{\text{FFS}}(I_0 - \Psi)^2 - (\Delta c + \Delta r_p)(1 - 2\Psi) - w_q^{\text{BP}}(1 - I_0 - \Psi)^2 \tag{B.161}$$

(LH-G0) : The physician surplus is

$$
\begin{aligned}
&\frac{1}{4(2w_b + w_q^{\text{BP}})^2)}(w_b(-2(-1 + I_0)(w_b + w_q^{\text{BP}}) + \Delta c)^2 \\
&- 4(w_b(2(-1 + I_0)w_b(3w_b + 2w_q^{\text{BP}}) + (w_b + w_q^{\text{BP}})\Delta c) + (2w_b + w_q^{\text{BP}})^2\Delta r_p)\Psi \\
&- 4w_b^2(3w_b + 2w_q^{\text{BP}})\Psi^2)
\end{aligned}
\tag{B.162}
$$

The payer surplus is

$$\begin{aligned}&\frac{1}{4(2w_b+w_q^{\text{BP}})}(\Delta c^2+4w_b^2(-1+I_0+\Psi)^2\\&+4w_q^{\text{BP}}((-1+I_0)^2w_q^{\text{FFS}}+\Delta c-I_0\Delta c+\Delta r_p\Psi)\\&+4w_b(2(-1+I_0)^2w_q^{\text{FFS}}+\Delta c-I_0\Delta c+\Delta c\Psi+2\Delta r_p\Psi))\end{aligned} \tag{B.163}$$

(LH-GM) : The physician surplus is

$$w_b(1-(I_0+\frac{2\Delta r_p}{2w_b})\Psi). \tag{B.164}$$

The payer surplus is

$$w_q^{\text{FFS}}-2I_0(w_b+w_q^{\text{FFS}})-I_0^2(w_q^{\text{BP}}-w_q^{\text{FFS}})+\Delta c+\Psi(2I_0w_b+\Delta r_p). \tag{B.165}$$

(LH-N0) : The physicians will always benefit. The payer surplus is as follows:

$$(1-I_0)^2w_q^{\text{FFS}}+(\Delta c+\Delta r_p)(1-I_0-\frac{\Delta r_p}{2w_b})-\frac{w_q^{\text{BP}}\Delta r_p^2}{4w_b^2}. \tag{B.166}$$

This will be negative, note that $1-I_0-\frac{\Delta r_p}{2w_b}<0$ because of the condition of (LH).

(LH-NM) : In this case, the physician surplus will be negative. Hence, no bundling.

(LH-N1) : The physician surplus is

$$-\Psi(\Delta r_p-w_b(2-2I_0-\Psi)). \tag{B.167}$$

The payer surplus is

$$w_q^{\text{FFS}}(1-I_0)^2+\Psi(\Delta c+\Delta r_p)-w_q^{\text{BP}}(1-I_0-\Psi)^2. \tag{B.168}$$

(LM-G0) The physician surplus is

$$\frac{1}{4w_b(2w_b+w_q^{\text{BP}})^2}(w_q^{\text{BP}}\Delta r_p + w_b(\Delta c + 2\Delta r_p) + 2w_b^2(-1+I_0+\Psi))$$
$$(w_q^{\text{BP}}\Delta r_p + 6w_b^2(-1+I_0+\Psi) + w_b(-\Delta c + 2\Delta r_p + 4w_q^{\text{BP}}(-1+I_0+\Psi))) \tag{B.169}$$

The payer surplus is

$$\begin{aligned}&\frac{1}{4w_b^2(2w_b+w_q^{\text{BP}})}(w_q^{\text{BP}}w_q^{\text{FFS}}\Delta r_p^2 + 2w_b\Delta r_p(w_q^{\text{FFS}}\Delta r_p + w_q^{\text{BP}}(\Delta c+\Delta r_p))\\&+4w_b^3(\Delta c+2\Delta r_p)(-1+I_0+\Psi)\\&+4w_b^4(-1+I_0+\Psi)^2\\&+w_b^2(\Delta c^2+4\Delta c\Delta r_p+4\Delta r_p(\Delta r_p+w_q^{\text{BP}}(-1+I_0+\Psi))))\end{aligned} \tag{B.170}$$

(LM-GM) : The physician surplus is

$$-\frac{1}{4w_b}(2I_0w_b+2\Delta r_p)(\Delta r_p-2w_b(2-I_0-2\Psi)). \tag{B.171}$$

The payer surplus is

$$\begin{aligned}&\frac{1}{w_b^2}(-4I_0^2w_b^2w_q^{\text{BP}}+4I_0w_b^2(\Delta c+\Delta r_p-2w_b(1-\Psi))\\&+\Delta r_p(w_q^{\text{FFS}}\Delta r_p+2w_b(\Delta c+\Delta r_p)-4w_b^2(1-\Psi)).\end{aligned} \tag{B.172}$$

(LM-N0) : Here, the physicians will not benefit, so this scenario is not possible.

(LM-NM) : Here, the physicians will lose money, so this scenario is not possible.

(LM-N1) : Again, the physicians will be losing money.

(MH-G0) : The physician surplus is

$$\begin{aligned}&\frac{1}{4(2w_b + w_q^{\text{BP}})^2(w_q^{\text{FFS}})^2}(2(w_q^{\text{BP}})^2 w_q^{\text{FFS}} \Delta r_p(\Delta c - 2w_q^{\text{FFS}}(-1 + I_0 + \Psi))\\ &- 4w_b^3(-\Delta c^2 + 3(w_q^{\text{FFS}})^2(-1 + I_0 + \Psi)^2)\\ &+ w_b((w_q^{\text{BP}})^2\Delta c^2 + (w_q^{\text{FFS}})^2\Delta c^2 - 4w_q^{\text{BP}} w_q^{\text{FFS}}(-2\Delta c\Delta r_p + w_q^{\text{FFS}}(\Delta c + 4\Delta r_p)(-1 + I_0 + \Psi)))\\ &+ 4w_b^2(-w_q^{\text{FFS}}(-2\Delta c\Delta r_p + w_q^{\text{FFS}}(\Delta c + 4\Delta r_p)(-1 + I_0 + \Psi))\\ &+ w_q^{\text{BP}}(\Delta c^2 - 2(w_q^{\text{FFS}})^2(-1 + I_0 + \Psi)^2)))\end{aligned} \tag{B.173}$$

The payer surplus is

$$\begin{aligned}&\frac{1}{4(2w_b + w_q^{\text{BP}})w_q^{\text{FFS}}}(w_q^{\text{FFS}}\Delta c^2 + 4w_b^2 w_q^{\text{FFS}}(-1 + I_0 + \Psi)^2\\ &- 2w_b(\Delta c + 2\Delta r_p)(\Delta c - 2w_q^{\text{FFS}}(-1 + I_0 + \Psi))\\ &- w_q^{\text{BP}}(\Delta c^2 + 2\Delta c\Delta r_p - 4w_q^{\text{FFS}}\Delta r_p(-1 + I_0 + \Psi)))\end{aligned} \tag{B.174}$$

(MH-GM) : The physician surplus is:

$$\Delta r_p(1 - \Psi) - I_0^2 w_b - I_0(\Delta r_p - 2w_b(1 - \Psi)) + \frac{\Delta c(w_b\Delta c + 2w_q^{\text{FFS}}\Delta r_p)}{4(w_q^{\text{FFS}})^2} \tag{B.175}$$

The payer surplus is

$$\begin{aligned}&\frac{-1}{4w_q^{\text{FFS}}}(4I_0^2 w_q^{\text{BP}} w_q^{\text{FFS}} + \Delta c^2 + 2\Delta c\Delta r_p\\ &- 4I_0 w_q^{\text{FFS}}(\Delta c + \Delta r_p - 2w_b(1 - \Psi)) + 4w_q^{\text{FFS}}\Delta r_p(1 - \Psi)).\end{aligned} \tag{B.176}$$

(MH-N0) : The physicians will always benefit. The payer surplus will be

$$-\frac{1}{4w_b^2 w_q^{\text{FFS}}}(w_q^{\text{BP}} w_q^{\text{FFS}}\Delta r_p^2 + 2w_b w_q^{\text{FFS}}\Delta r_p(\Delta c + \Delta r_p) + w_b^2\Delta c(\Delta c + 2\Delta r_p)). \tag{B.177}$$

(MH-NM) : The physician surplus will be

$$\frac{1}{4w_b(w_q^{\text{FFS}})^2}((w_b + w_q^{\text{FFS}})\Delta c + 3w_q^{\text{FFS}}\Delta r_p)(w_b\Delta c - w_q^{\text{FFS}}(\Delta c + \Delta r_p)). \qquad \text{(B.178)}$$

The payer surplus will be

$$\frac{-1}{4w_b^2 w_q^{\text{FFS}}}(-2w_b w_q^{\text{FFS}}(\Delta c + \Delta r_p)^2 + w_q^{\text{BP}} w_q^{\text{FFS}}(\Delta c + \Delta r_p)^2 w_b^2 \Delta c(\Delta c + 2\Delta r_p)). \qquad \text{(B.179)}$$

(MH-N1) : The physician surplus will be:

$$\frac{-1}{4(w_q^{\text{FFS}})^2}(-\Delta c - 2w_q^{\text{FFS}}(1 - I_0 - \Psi))(2w_q^{\text{FFS}}\Delta r_p + w_b(\Delta c - 2w_q^{\text{FFS}}(1 - I_0 - \Psi))). \qquad \text{(B.180)}$$

The payer surplus will be:

$$-\frac{1}{4w_q^{\text{FFS}}}(\Delta c(\Delta c + 2\Delta r_p) + 4w_q^{\text{FFS}}(\Delta c + \Delta r_p)(1 - I_0 - \Psi) + 4w_q^{\text{BP}} w_q^{\text{FFS}}(1 - I_0 - \Psi)^2) \qquad \text{(B.181)}$$

(ML-G0) : The physician surplus will be

$$\begin{aligned}
&\frac{1}{4(2w_b + w_q^{\text{BP}})^2(w_q^{\text{FFS}})^2}(2(w_q^{\text{BP}})^2 w_q^{\text{FFS}}\Delta r_p(\Delta c - 2w_q^{\text{FFS}}(-1 + I_0 + \Psi)) \\
&- 4w_b^3(-\Delta c^2 + 3(w_q^{\text{FFS}})^2(-1 + I_0 + \Psi)^2) \\
&+ w_b((w_q^{\text{BP}})^2\Delta c^2 + (w_q^{\text{FFS}})^2\Delta c^2 - 4w_q^{\text{BP}} w_q^{\text{FFS}}(-2\Delta c\Delta r_p + w_q^{\text{FFS}}(\Delta c + 4\Delta r_p)(-1 + I_0 + \Psi))) \\
&+ 4w_b^2(-w_q^{\text{FFS}}(-2\Delta c\Delta r_p + w_q^{\text{FFS}}(\Delta c + 4\Delta r_p)(-1 + I_0 + \Psi)) \\
&+ w_q^{\text{BP}}(\Delta c^2 - 2(w_q^{\text{FFS}})^2(-1 + I_0 + \Psi)^2))).
\end{aligned} \qquad \text{(B.182)}$$

The payer surplus will be

$$\begin{aligned}&\frac{1}{4(2w_b + w_q^{\text{BP}})w_q^{\text{FFS}}}(w_q^{\text{FFS}}\Delta c^2 + 4w_b^2 w_q^{\text{FFS}}(1 - I_0 - \Psi)^2 \\ &- 2w_b(\Delta c + 2\Delta r_p)(\Delta c + 2w_q^{\text{FFS}}(1 - I_0 - \Psi)) \\ &- w_q^{\text{BP}}(\Delta c^2 + 2\Delta c\Delta r_p + 4w_q^{\text{FFS}}\Delta r_p(1 - I_0 - \Psi)))\end{aligned} \tag{B.183}$$

(ML-GM) : The physician surplus will be

$$\Delta r_p(1 - \Psi) - I_0^2 w_b - I_0(\Delta r_p - 2w_b(1 - \Psi)) + \frac{\Delta c(w_b\Delta c + 2w_q^{\text{FFS}}\Delta r_p)}{4(w_q^{\text{FFS}})^2}. \tag{B.184}$$

The payer surplus will be

$$\begin{aligned}&\frac{-1}{4w_q^{\text{FFS}}}(4I_0^2 w_q^{\text{BP}} w_q^{\text{FFS}} + \Delta c^2 + 2\Delta c\Delta r_p - 4I_0 w_q^{\text{FFS}}(\Delta c + \Delta r_p - 2w_b(1 - \Psi)) \\ &+ 4w_q^{\text{FFS}}\Delta r_p(1 - \Psi)).\end{aligned} \tag{B.185}$$

(ML-N0) : The physicians will always benefit. The payer surplus will be

$$\frac{-1}{4w_b^2 w_q^{\text{FFS}}}(w_q^{\text{BP}} w_q^{\text{FFS}}\Delta r_p^2 + 2w_b w_q^{\text{FFS}}\Delta r_p(\Delta c + \Delta r_p) + w_b^2\Delta c(\Delta c + 2\Delta r_p)). \tag{B.186}$$

(ML-NM) : The physician surplus will be

$$\frac{1}{4w_b(w_q^{\text{FFS}})^2}((w_b + w_q^{\text{FFS}})\Delta c + 3w_q^{\text{FFS}}\Delta r_p)(w_b\Delta c - w_q^{\text{FFS}}(\Delta c + \Delta r_p)). \tag{B.187}$$

The payer surplus will be

$$\frac{-1}{4w_b^2 w_q^{\text{FFS}}}(-2w_b w_q^{\text{FFS}}(\Delta c+\Delta r_p)^2 + w_q^{\text{BP}} w_q^{\text{FFS}}(\Delta c+\Delta r_p)^2 + w_b^2 \Delta c(\Delta c+2\Delta r_p)). \tag{B.188}$$

(ML-N1) : The physician surplus will be

$$\frac{-1}{4(w_q^{\text{FFS}})^2}(-\Delta c - 2w_q^{\text{FFS}}(1-I_0-\Psi))(2w_q^{\text{FFS}}\Delta r_p + w_b(\Delta c - 2w_q^{\text{FFS}}(1-I_0-\Psi))). \tag{B.189}$$

The payer surplus will be

$$\frac{-1}{4w_q^{\text{FFS}}}(\Delta c(\Delta c+2\Delta r_p) + 4w_q^{\text{FFS}}(\Delta c+\Delta r_p)(1-I_0-\Psi) + 4w_q^{\text{BP}} w_q^{\text{FFS}}(1-I_0-\Psi)^2) \tag{B.190}$$

□

B.2.5 Observable Coproduction Model

Proof. Proof of Lemma B.1.3. We seek a Nash equilibrium for i_p and i_h. Equilibrium conditions require:

$$\iota_p^{\text{FFS}}(\iota_h^{\text{FFS}}(i_p)) = i_p \tag{B.191}$$

$$\iota_h^{\text{FFS}}(\iota_p^{\text{FFS}}(i_h)) = i_h \tag{B.192}$$

where ι_p^{FFS} ι_h^{FFS} are the best response functions such that these responses respectively maximize F_p and F_h. We write the corresponding first order conditions as

$$\iota_p^{\text{FFS}}(i_h) = \{i_h : \frac{\partial}{\partial i_p} F_p^{\text{FFS}}(i_h, i_p) = 0\} = \frac{2w_b(I_0 - (1-i_h)\Psi) + \Delta r_p}{2w_b(1-\Psi)}. \tag{B.193}$$

$$\iota_h^{\text{FFS}}(i_p) = \{i_p : \frac{\partial}{\partial i_h} F_h^{\text{FFS}}(i_h, i_p) = 0\} = \frac{2w_q(\Psi - I_0 + i_p(1-\Psi)) + \Delta c}{2w_q\Psi} \tag{B.194}$$

In the above equilibrium conditions, observe that there exists no solution with both the hospital and physician responses are in the interior (i.e., $0 < i_p, i_h < 1$). We therefore consider solutions with at least one of i_p, i_h at the boundary. We find three non-dominated solutions on the boundary: $i_p < 1, i_h = 1$; $i_p = i_h = 1$; and $i_p = 1, i_h < 1$. The respective solutions are given in (B.32), (B.33), and (B.34) in Lemma 2.3.3. Because the second derivatives are negative, that is

$$\frac{\partial^2}{\partial i_p^2} F_p^{\text{BP}}(i_h, i_p) = -2w_b(1-\Psi)^2 < 0 \tag{B.195}$$

$$\frac{\partial^2}{\partial i_h^2} F_h^{\text{BP}}(i_h, i_p) = -2w_q\Psi^2 < 0 \tag{B.196}$$

hold, both F_p^{FFS} and F_h^{BP} are strictly concave with respect to i_p and i_h. Hence, the solutions presented are also global optimums.

□

Proof. Proof of Lemmas B.1.4 and B.1.5. We seek an equilibrium of a two-stage game where the hospital announces T in Stage 1, and then the hospital and the physicians simultaneously choose their efforts $i_p,\ i_h$ in Stage 2. We proceed by backward induction in three steps as follows:

1. Stage 2: solving for i_p and i_h given arbitrary $T \geq 0$,

2. Stage 1: finding optimal T if $i_p^\sharp$ and $i_h^\sharp$ from Stage 2 are given,

3. Substitute optimal T back into Stage 2 solutions and verify that the solutions are feasible.

Stage 2: Assume that $T \geq 0$ from the first stage is given. Then, Stage 2 reduces to a two-player game with i_h and i_p being the decision variables of the hospital and the physicians, respectively. In such a game, both parties attempt to maximize their response functions given the other party's response. Then, the optimal response functions for the hospital and physicians respectively can be computed as

$$\begin{aligned}
\iota_h(i_p, T) &= \arg\max_{i_h} \; F_h^{\text{BP}}(i_h, i_p, T) = \frac{\Delta c + \Delta r_p + 2w_q(\Psi + i_p(1-\Psi) - I_0) - T}{2w_q\Psi} \\
\iota_p(i_h, T) &= \arg\max_{i_p} \; F_p^{\text{BP}}(i_h, i_p, T) = \frac{\Delta r_p + 2w_b(I_0 + i_h - \Psi) - T}{2(1-\Psi)}
\end{aligned} \tag{B.197}$$

when they are interior solutions, i.e., $0 < \iota_h < 1$ or $0 < \iota_p < 1$ respectively; and are equal to 0 or 1 otherwise. Equilibrium conditions are given as:

$$\iota_h(\iota_p(i_h), T) = i_h, \tag{B.198}$$

$$\iota_p(\iota_h(i_p), T) = i_p. \tag{B.199}$$

Based on equilibrium conditions, hospital and physician responses cannot be simultaneously in the interior range, i.e., $0 < i_p, i_h < 1$. This leaves us with eight alternative solutions where either i_p or i_h is at the boundary.

Stage 1: Let candidate solutions $i_p^\sharp$ and $i_h^\sharp$ from Stage 2 be given. We let $F_\tau(T) := F_h^{\text{BP}}(i_h^\sharp, i_p^\sharp, T)$ be the hospital's conditional payoff function, and observe that in this stage, we are seeking the quantity $\arg\max_T F_\tau$. We solve for this quantity by reviewing the eight possible combinations of $i_p^\sharp$ and $i_h^\sharp$ (depending on $i_p^\sharp = 0,\ i_p^\sharp = 1,\ i_p^\sharp \in (0,1)$ and similar for $i_h^\sharp$), which reveals the following:

- If both $i_p^\sharp$ and $i_h^\sharp$ are at the boundary or if $i_p^\sharp = 1$, it can be verified that $\frac{d}{dT} F_\tau < 0$ for any $T > 0$, so the hospital will choose $T = 0$. We notice that the game in this case is the same as in FFS solutions covered in Lemma 2.3.3, with solutions described in Lemma B.1.3. However, in the current case, Δc is replaced by $\Delta c + \Delta r_p$ in the definition of F_h^{FFS} from (2.2). By substituting this replacement into the conditions of Lemma B.1.3, we get conditions (B.39), (B.40), and (B.41). By substituting into the definitions of $\bar{\Psi}$ and $\hat{\Psi}$, we get Ψ_1 and Ψ_2 respectively, which concludes the proof of Lemma B.1.5. In the following, we will focus on the cases with $T > 0$ for Lemma B.1.4.

- Solutions with $i_h^\sharp = 0$ are dominated by solutions with $i_h^\sharp = 1$ (i.e., F_h^{BP} is larger for $i_h^\sharp = 1$), so they will not occur in the equilibrium.

- $i_p^\sharp = 0$ implies that T is larger than certain threshold, and the hospital will set exactly this threshold (because F_h^{BP} decreases beyond this threshold). Let $T_{\max}$ be the T threshold when $i_h^\sharp = 1$.

- Among the set of solutions with $i_p^\sharp = 0$, $i_h^\sharp \in (0, 1]$, and $T \geq T_{\max}$, the physicians are indifferent, but the hospital attains maximal F_h^{BP} for $T = T_{\max}$ and $i_h = 1$. Therefore, the solutions with $i_h^\sharp \in (0, 1),\ i_p^\sharp = 1$ are infeasible.

Then, given $i_h = 1$ and the response function for $i_p > 0$ given from (B.197), we find from the first order condition that the optimal T is as follows:

$$T = \frac{w_b(\Delta c - 2(1 - I_0)w_b)}{2w_b + w_q} + \Delta r_p. \tag{B.200}$$

This solution is valid for $i_p > 0$, while for $i_p = 0$, the solution becomes

$$T = T_{\max} = 2I_0 w_b + \Delta r_p. \tag{B.201}$$

Substituting $T > 0$ back: This analysis leaves us with only two equilibrium cases when $T > 0$: The first case as $0 < T < T_{\max}$, and the other one as $T = T_{\max}$. For the first case, as derived earlier, only the solutions with $i_h = 1$ are feasible. This then results in T

determined according to (B.200) and i_p determined from (B.197) with T substituted, which is given in the first part of Lemma in (B.35). The conditions for the first set of solutions in Lemma B.1.4, namely conditions in (B.36), correspond to the following requirements:

- $0 < i_p < 1$ and
- $T > 0$

respectively. For the second case with $T = T_{\max}$, we have shown that $i_p = 0$ and $i_h = 1$. Condition (B.38) then corresponds to $i_p \leq 0$ and equivalently $T \geq T_{\max}$.

□

Proof. Proof of Corollary B.1.1. First, we observe that we can directly compute Condition (B.42) by substituting $i_h^\sharp$ and $i_p^\sharp$ in the definition of I in the two cases described in Lemma B.1.4.

Second, we need to show that $I^\sharp < I^*$. We proceed by cases. For the case $I^\sharp = 0$, we observe that $I^* > 0$ when bundling is preferable, i.e., when $\Psi \leq \hat{\Psi}$. Therefore, $I^\sharp < I^*$ in this case.

If $I^\sharp > 0$, it follows from Lemmas B.1.3 and B.1.4 that $i_p^* - i_p^\sharp = \frac{T}{2(1-\Psi)} > 0$. At the same time, when bundling is preferable, we have $i_h^* = i_h^\sharp = 1$. Therefore, substituting these values into the definition of I, we derive the following:

$$I^\sharp = I(i_h^\sharp, i_p^\sharp) = (1-\Psi)i_p^\sharp = (1-\Psi)(i_p^* - \frac{T}{2(1-\Psi)}) < (1-\Psi)i_p^* = I(i_h^*, i_p^*) = I^*. \tag{B.202}$$

This completes the proof. □

We next present the proofs of the results presented as part of the General Case in Section 2.3.3. The Base Model results follow as special cases of the proofs for the General Case.

Proof. Proof of Lemma 2.3.3. The analog of this result in the Base Case is Lemma 2.2.3. The solutions can be derived from Lemma B.1.3 by substituting i_p^* and i_h^* in the definition of I. □

Proof. Proof of Theorem 2.3.2. The analog of this result in the Base Case is Theorem 2.3.1. Our strategy to prove the theorem will be as follows:

1. Analyze all equilibrium solutions from Lemmas B.1.4 and B.1.5
2. For each of the equilibrium solutions, identify if the payer, the hospital, and the physicians all benefit from bundling as stated in the theorem. If they benefit, under what additional conditions?
3. The specific conditions that need to be tested for each equilibrium solution are as follows:
   (a) $F_p^{\text{FFS}}(i_h^*, i_p^*) < F_p^{\text{BP}}(i_h^\sharp, i_p^\sharp, T)$ (the physicians benefit)
   (b) $F_h^{\text{FFS}}(i_h^*, i_p^*) < F_h^{\text{BP}}(i_h^\sharp, i_p^\sharp, T)$ (the hospital benefits)
   (c) $r^{\text{BP}} < r^{\text{FFS}}$ (the payer benefits).

We begin with the equilibrium solutions from Lemma B.1.5, the case $T = 0$. We compare the BP solutions from Lemma B.1.5 to the solutions under FFS from Lemma B.1.3. We find four distinct regions of the solution space:

1. $\Psi < \bar{\Psi}$: Then the FFS and BP solutions are equal ($i_p^* = i_p^\sharp$, $i_h^* = i_h^\sharp$), therefore $F_p^{\text{FFS}} = F_p^{\text{BP}}$, so physicians do not benefit.
2. $\bar{\Psi} \leq \Psi < \hat{\Psi}$: Then the FFS and BP solutions are equal, so physicians do not benefit.
3. $\hat{\Psi} \leq \Psi < \Psi_2$: Then the following holds:

$$F_p^{\text{BP}} - F_p^{\text{FFS}} = (\frac{1}{4w_q^2})(\Delta c - 2w_q(-1+I_0+\Psi))(2w_q\Delta r_p + w_b(\Delta c + 2w_q(-1+I_0+\Psi))). \tag{B.203}$$

   We find that the first product term is positive, the second one is negative by the condition $\Psi \geq \hat{\Psi}$, and the third one is positive because of the condition $\Psi \geq \hat{\Psi}$ (which implies $-1 + I_0 + \Psi > 0$). Therefore, $F_p^{\text{BP}} - F_p^{\text{FFS}} < 0$, so the physicians lose.

4. $\Psi_2 \leq \Psi$: Then

$$F_p^{\text{BP}} - F_p^{\text{FFS}} = -\frac{\Delta r_p(2\Delta c + \Delta r_p + 2w_q + \Delta r_p)}{4w_q} < 0, \tag{B.204}$$

so the physicians lose.

We conclude that under either of the cases with $T = 0$, the parties do not prefer bundling.

Next, we analyze the first equilibrium solution from Lemma B.1.4, the case $0 < T < T_{\max}$ from (B.35). We first exclude the region with high integration ($\Psi > \hat{\Psi}$) from the consideration. Indeed, in this case, physicians set $i_p^\sharp = 1$, so the hospital can set $T = 0$ without any negative consequences. But for $T = 0$, we have already showed the parties do not prefer bundling.

We next investigate the cases of low and moderate integration ($\Psi \leq \hat{\Psi}$). We start with the payer savings condition. Under low integration ($\Psi < \bar{\Psi}$), the payer achieves the following savings:

$$r^{\text{FFS}} - r^{\text{BP}} = \frac{(w_b(\Delta c - 2(1-I_0)w_b) + (2w_b + w_q))^2}{4w_b^2(2w_b + w_q)}. \tag{B.205}$$

This term is always positive, so the payer benefits. Under moderate integration, the payer achieves the following savings:

$$r^{\text{FFS}} - r^{\text{BP}} = \frac{(\Delta c + 2(1-I_0)(w_b + w_q))^2}{4(2w_b + w_q)} - (\Delta c + \Delta r_p + 2(1-I_0)w_q)\Psi + w_q\Psi^2. \tag{B.206}$$

The term in (B.206) is lower than (B.205). Then, when we solve for the constraint $r^{\text{FFS}} - r^{\text{BP}} > 0$ from (B.206), we find condition (2.26) from the theorem.

We next focus on the hospital benefit part. Under low integration, the hospital benefits under the following condition:

$$F_h^{\text{BP}}(i_h^\sharp, i_p^\sharp, T) - F_h^{\text{FFS}}(i_h^*, i_p^*) = \frac{(\Delta r_p(2w_b + w_q) + w_b(\Delta c - 2(1-I_0)w_b))^2}{4w_b^2(2w_b + w_q)}. \tag{B.207}$$

This term is always positive, equal to the payer benefit. Therefore, the hospital benefits. Under moderate integration, the hospital condition is the same as the payer condition from (B.206). Therefore, the hospital benefits exactly when the payer condition (2.26) satisfied.

Next, we focus on the physician benefit part. We analyze function $F_p^{\text{BP}}(i_h^\sharp, i_p^\sharp, T) - F_p^{\text{FFS}}(i_h^*, i_p^*)$ to find the following:

$$\frac{\partial}{\partial \Psi}[F_p^{\text{BP}}(i_h^\sharp, i_p^\sharp, T) - F_p^{\text{FFS}}(i_h^*, i_p^*)] = \begin{cases} 0 & \text{if } \Psi \leq \bar{\Psi} \\ \Delta r_p + 2w_b(-1 + I_0 + \Psi) & \text{if } \bar{\Psi} \leq \Psi \leq \hat{\Psi}. \end{cases} \tag{B.208}$$

The second term is positive by the definition of $\bar{\Psi}$. Therefore, the function is first constant and then increasing. Therefore, we only need to show that the function $F_p^{\text{BP}}(i_h^\sharp, i_p^\sharp, T) - F_p^{\text{FFS}}(i_h^*, i_p^*)$ is positive when substituting $\Psi = \bar{\Psi}$. After this substitution, we find the following:

$$\begin{aligned} &F_p^{\text{BP}}(i_h^\sharp, i_p^\sharp, T) - F_p^{\text{FFS}}(i_h^*, i_p^*)|_{\Psi=0} \\ &= \frac{1}{4(2w_b + w_q)^2} \cdot (\Delta c + 2w_b(1 - I_0)(3w_b + w_q) - (2w_b + w_q)\Delta r_p) \cdot (\Delta c + \Delta r_p(2 + w_q/w_q) - 2(1 - \end{aligned} \tag{B.209}$$

By the definition of $\bar{\Psi}$, we have $\Delta r_p \leq 2w_b(1 - I_0)$ at $\Psi = \bar{\Psi}$, so the first term is positive. Next, the second term is positive because of Condition (2.24). Therefore, the physicians always benefit when $T > 0$. To finalize the list of conditions for Theorem 2.3.2, we need to include the conditions that characterize the equilibrium condition from Lemma B.1.4:

1. Condition for $i_p^\sharp < 1$. From this condition, we derive Condition (2.25) of the theorem.

2. From condition $T > 0$, we derive Condition (2.24) of the theorem.

Finally, we still need to investigate the second case in Lemma B.1.4, the case $T = T_{\max}$. This case is actually subsumed by the proof of the case with $0 < T < T_{\max}$ at the limit $T \to T_{\max}$. Specifically, we have $i_p^\sharp < 1$ and $T > 0$ automatically and Condition (2.26) is still applicable as is. □

Proof. Proof of Proposition 2.3.3. The analog of this result in the Base Case is Proposition 2.3.1.

Let $\Delta Q := |I^\sharp - I_0| - |I^* - I_0|$. To prove the proposition, we need to characterize the cases when the quality under BP deteriorates ($\Delta Q > 0$) and when the quality improves ($\Delta Q < 0$). Overall, we find six cases (the combination of $i_p^\sharp > 0$ vs $i_p^\sharp = 0$ and $\Psi < \bar{\Psi}$ vs $\bar{\Psi} < \Psi < \hat{\Psi}$ vs $\Psi > \hat{\Psi}$). We also need to distinguish whether $I^* - I_0$ is positive (overtreatment under FFS) or negative (undertreatment). We first list several facts that we will use repeatedly:

- *Intensity* under BP will decrease. This fact is from Corollary B.1.1.
- There is overtreatment under FFS when $\Psi < \bar{\Psi}$. This follows from Lemma 2.3.3.
- We can determine the values of $I^\sharp$ from B.1.4 and the values of I^* from Lemma 2.3.3.

We now focus on the following feasible cases:

1. $\Psi < \bar{\Psi}$ and overtreatment under BP (i.e., $\Delta c < 2(1 - I_0)w_b$): Then BP will improve quality because we know that the intensity under BP is lower than under FFS, so there is less undertreatment. This condition is covered by (2.27) of the proposition.

2. $\Psi < \bar{\Psi}$ and undertreatment under BP (i.e., $\Delta c \geq 2(1 - I_0)w_b$). Then the following holds:
$$\Delta Q = \frac{1}{2}(\frac{\Delta c - 2(1 - I_0)w_b}{2w_b + w_q} - \frac{\Delta r_p}{w_b}) \tag{B.210}$$
This condition is equivalent to Condition (2.27) of the proposition.

3. $\bar{\Psi} < \Psi < \hat{\Psi}$ and there is undertreatment under FFS (i.e., $1 - \Psi < I_0$): Then there is undertreatment under FFS too and the quality differential must decrease. We observe this decrease because the intensity under BP is lower than under FFS.

4. If $\bar{\Psi} < \Psi < \hat{\Psi}$ and there is overtreatment under FFS (i.e., $1 - \Psi > I_0$): The quality differential is
$$\Delta Q = \frac{\Delta c - 2(1 - I_0)(3w_b + w_q)}{2(2w_b + w_q)} + \Psi. \tag{B.211}$$
This condition is equivalent to Condition (2.28).

5. $\Psi > \hat{\Psi}$: Then there is undertreatment already under FFS. Since we have observed that the intensity decreases under BP, there will be even more undertreatment under BP. Hence, BP will decrease quality.

□

Proof. Proof of Proposition 2.3.4. The analog of this result in the Base Case is Proposition 2.3.2. We derive the statements about the shape of the savings function straightforwardly by analyzing Σ from (2.29) as a univariate function of Ψ. Therefore, we are left to derive equations in (2.29). Here, we follow the definition of $\Sigma = r^{\text{FFS}} - r^{\text{BP}}$ and the definitions of r^{FFS} and r^{BP}. If we combine expand the latter definitions, we get the following:

$$\begin{aligned} r^{\text{FFS}} &= r_h + r_{2,p} + I^* \Delta r_p \\ r^{\text{BP}} &= -(F_h^{\text{BP}}(i_h^\sharp, i_p^\sharp, T) - r^{\text{BP}} - F_h^{\text{FFS}}(i_h^*, i_p^*)). \end{aligned} \tag{B.212}$$

We next use the equilibrium results from Lemmas B.1.3 and B.1.4 on I^*, i_h^*, i_p^*, $i_h^\sharp$, $i_p^\sharp$, and T, substitute these into the formulas for r^{FFS} and r^{BP}, and substitute r^{BP} and r^{FFS} into the definition of Σ. We then derive the two equations in (2.29) by using the equilibrium solutions for $\Psi \leq \bar{\Psi}$ and $\bar{\Psi} < \Psi < \hat{\Psi}$ respectively.

□

B.2.6 Physician-driven Model

Proof. Proof of Theorem 2.4.2. The proof of this theorem is similar to the proof of Theorem 2.3.2. We employ the following strategy:

1. Exhibit the equilibrium solutions
2. Use the equilibrium solutions to characterize when bundling is preferable, i.e., the hospital, the physicians, and the payer all benefit.

We first solve for the equilibrium solutions. Recall that in this game, the hospital first chooses the gainsharing amount T, and physicians then select their response i_p. We solve this game by backward induction. Given T, the physicians respond optimally as follows:

$$i_p^\sharp = I_0 + \frac{\Delta r_p - \Psi - T}{2w_b}. \tag{B.213}$$

If we substitute (B.213) back into the hospital objective function, we express optimal T as follows:

$$T = \frac{1}{2}(\Delta c + 2\Delta r_p - 2(1 - I_0)w_b - \Psi). \tag{B.214}$$

We briefly enumerate the equilibrium solutions (they are similar to the equilibrium solutions in Lemmas B.1.4 and B.1.5 in the main model):

- We get $T = 0$ if $\Delta c + 2\Delta r_p \leq 2(1 - I_0)w_b + \Psi$ (the condition for $i_p < 1$, which was required in the main model, is always satisfied)
- We get $T = T_{\max} = 2I_0 w_b + \Delta r_p - \Psi$ and $i_p = 0$ if $\Delta c + \Psi > 2(1 + I_0)w_b$
- Otherwise, T is given by (B.214). In particular, the following condition holds:

$$\Delta c + 2\Delta r_p > 2(1 - I_0)w_b + \Psi. \tag{B.215}$$

Having established the equilibrium solutions, we proceed to prove the theorem. In the proof, we review the equilibrium solutions one by one. In case $T = 0$, we see similar BP solutions as the FFS solutions but with $\Delta c + \Delta r_p$ substituted in the hospital profit function for Δc. There are three cases, and neither of them is beneficial for all parties. We omit a detailed proof because the structure of solutions is similar to the case of $T = 0$ in the proof of Theorem 2.3.2.

We next consider cases for $0 < T < T_{\max}$. Depending on i_p^*, we derive three cases:

- If $i_p^* = 1$, then the physician surplus is

$$\frac{(2(1-I_0)w_b + \Delta c + \Psi)^2}{16w_b}, \tag{B.216}$$

which is positive. The payer and hospital surplus is

$$\Sigma = \frac{(\Delta c + \Psi + 2(1-I_0)w_b)^2}{8w_b}, \tag{B.217}$$

which is also positive. Hence, under this condition, all parties prefer bundling.

- If $0 < i_p^* < 1$, then the physician surplus is

$$-\frac{(-6(1-I_0)w_b - \Delta c + 2\Delta r_p - 3\Psi)(-2(1-I_0)w_b + \Delta c + 2\Delta r_p - \Psi)}{16w_b}. \tag{B.218}$$

The second term is positive by condition (B.215), while the first one is negative because $\Psi + 2(1-I_0)w_b > \Delta r_p$ which follows from condition $i_p^* < 1$.

The payer and hospital surplus is

$$\Sigma = \frac{(\Delta c + 2\Delta r_p - \Psi - 2(1-I_0)w_b)^2}{8w_b}, \tag{B.219}$$

which is positive.

- For the case $i_p^* = 0$, we observe that we always have $i_p^\sharp \leq i_p^*$, and the inequality would be strict if $T > 0$. The hospital will therefore choose $T = 0$, which will not lead to physician benefit, as we have already shown.

To prove the theorem in case $T = T_{\max}$, we would follow a similar but simpler structure. We find that the conditions are special cases of the case with $0 < T < T_{\max}$ and in particular, the condition for $T = T_{\max}$ is already subsumed in Equation (2.45). We omit a detailed proof.

Reviewing these conditions, we infer that bundling is preferred exactly when $T > 0$, which is equivalent to the condition listed in Theorem 2.4.2. □

Proof. Proof of Proposition 2.4.5. The proof is analogous to the proof of Proposition 2.3.3 and follows from a direct comparison of $|I^\sharp - I_0|$ with $|I^* - I_0|$. In the current case, $i_p^\sharp = I^\sharp$ and $i_p^* = I^*$. We omit the details. □

APPENDIX C

FLEXIBLE BED MANAGEMENT: SUPPLEMENTAL CONTENT

C.1 Proof of Theorem 3.4.1

Our proof strategy extends the proof from Koole (1995). Without the loss of generality, we only need to show that the threshold policy is optimal when deciding to assign a patient of class 1 to unit 2. The argument for class 2 is symmetric. We will denote by $V(x_1, x_2, n_{11}, n_{12}, n_{22}, n_{21})$ the value function with x_i being the number of waiting patients of class i and n_{ij} being the number of patients of class i served in unit j. The goal is to minimize the value function. We denote by V^k the expected value over the next k transitions and by W^k the value after the transition but before the action.

We prove the theorem in two steps:

1. Show that the bed in unit 2 is always used when the primary patient class (class 2) is waiting and the same holds for bed in unit 1. (Lemma C.1.1)Victoria Beach

2. Show that the value function is monotone (non-increasing) and submodular in the number of patients of class 1 waiting. (Lemmas C.1.2 and C.1.4)

Assume that all three lemmas hold. Then the theorem follows from the following considerations:

- If unit 1 is empty, then waiting patients of class 1 are assigned there by Lemma C.1.1.

- If unit 1 is occupied, then by submodularity, we have $V(x_1+2, 0, 1, 0, 0, 0) - V(x_1+1, 0, 1, 0, 0, 0) \geq V(x_1+1, 0, 1, 1, 0, 0) - V(x_1, 0, 1, 1, 0, 0)$. By monotonicity then, either we always have $V(x_1 + 1, 0, 1, 0, 0, 0) \leq V(x\_1, 0, 1, 1, 0, 0)$ for any x_1 or there exists x_1 such that $V(x_1 + 1, 0, 1, 0, 0, 0) > V(x_1, 0, 1, 1, 0, 0)$ and this holds for any $y_1 > x_1$ as well, so it is optimal to assign the patient of class 1 to unit 2 if and only if there are $x_1 + 1$ or more patients of class 1 waiting.

- We observe that this is a threshold policy.

Therefore, in the rest of the proof, we focus on proving the Lemmas.

Lemma C.1.1. *In a setting with two patient classes and two units, a set of sufficient conditions for a primary pair to be assigned whenever possible is as follows:*

$$b_1 = b_2 \equiv b \tag{a}$$

$$\mu_1 = \mu_2 \equiv \mu \tag{b}$$

$$\frac{\pi_{ji}}{b_j} \geq 1 \textit{ for } i \neq j \tag{c}$$

Proof. Proof of Lemma C.1.1. Without loss of generality, we can assume that $\pi_{11} = \pi_{22} = 0$. We need to prove the following:

$$\begin{aligned}
&V^k(x_1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22}) < V^k(x_1+1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) \text{ and} \\
&V^k(x_1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22}) < V^k(x_1, x_2, n_{11}, n_{12}+1, n_{21}, n_{22})
\end{aligned} \tag{C.1}$$

for all k, x_i, and n_{ij} with $n_{21} + n_{11} = 0$ for Class 1 and similarly for Class 2. We will proceed by induction on k, simultaneously for patient class 1 and 2. We will only show the proof for patient class 1 but will refer to the induction hypothesis for $k' < k$ for either class. We can assume that at each transition, we assign at most one patient, either because a new patient has arrived or because a patient occupying a unit has left. If it was optimal to make multiple assignments, we could have made one of the assignments earlier while saving on boarding costs, which would contradict optimality.

Let β be the maximum transition rate (for uniformization), $\beta = \lambda_1 + \lambda_2 + 2\mu$. We start the induction with case $k = 0$. Then, V^0 is the average cost for one period:

$$\begin{gathered} V^0(x_1, x_2, 1, n_{12}, n_{21}, n_{22}) - V^0(x_1 + 1, x_2, 0, n_{12}, n_{21}, n_{22}) = (\pi_{11} - b)/\beta = -b/\beta < 0, \\ V^0(x_1, x_2, 1, n_{12}, n_{21}, n_{22}) - V^0(x_1, x_2, n_{11}, 1, n_{21}, n_{22}) = -\pi_{12}/\beta < 0 \end{gathered} \tag{C.2}$$

which was to be shown

Next, we consider $k \geq 0$. We write down the V^{k+1} under all three cases from (C.1). The value function if the open unit is left empty is as follows:

$$\begin{aligned} & V^{k+1}(x_1 + 1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) = \\ & (b(x_1 + x_2 + 1) + \pi_{12}n_{12} + \pi_{21}n_{21})/\beta + \\ & (1/\beta)(\lambda_1 W^k(x_1 + 2, x_2, n_{11}, n_{12}, n_{21}, n_{22}) + \\ & \lambda_2 W^k(x_1 + 1, x_2 + 1, n_{11}, n_{12}, n_{21}, n_{22}) + \\ & \mu \cdot n_{12} W^k(x_1 + 1, x_2, n_{11}, (n_{12} - 1)^+, n_{21}, n_{22}) + \\ & \mu \cdot n_{22} W^k(x_1 + 1, x_2, n_{11}, n_{12}, n_{21}, (n_{22} - 1)^+) + \\ & \beta'/\beta \, W^k(x_1 + 1, x_2, n_{11}, n_{12}, n_{21}, n_{22})), \end{aligned} \tag{C.3}$$

where β' is the residual dummy transition rate; $\beta' = 2\mu - \mu(n_{21} + n_{22})$. The value function if the patient is assigned to the primary unit is as follows:

$$
\begin{aligned}
&V^{k+1}(x_1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22}) = \\
&(b(x_1+x_2) + \pi_{12}n_{12} + \pi_{21}n_{21})/\beta + \\
&(1/\beta)(\lambda_1 W^k(x_1+1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22}) + \\
&\lambda_2 W^k(x_1, x_2+1, n_{11}+1, n_{12}, n_{21}, n_{22}) + \\
&\mu \cdot 1 \cdot W^k(x_1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) + \\
&\mu \cdot n_{12} W^k(x_1, x_2, n_{11}+1, (n_{12}-1)^+, n_{21}, n_{22}) + \\
&\mu \cdot n_{22} W^k(x_1, x_2, n_{11}+1, n_{12}, n_{21}, (n_{22}-1)^+) + \\
&\beta''/\beta \, W^k(x_1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22})),
\end{aligned} \tag{C.4}
$$

with $\beta'' = 2\mu - \mu(1 + n_{21} + n_{22})$. The value function if the patient is assigned to the secondary unit is as follows (this case requires $n_{12} + n_{22} = 0$):

$$
\begin{aligned}
&V^{k+1}(x_1, x_2, n_{11}, n_{12}+1, n_{21}, n_{22}) = \\
&(b(x_1+x_2) + \pi_{12}(n_{12}+1) + \pi_{21}n_{21})/\beta + \\
&(1/\beta)(\lambda_1 W^k(x_1, x_2, n_{11}, n_{12}+1, n_{21}, n_{22}) + \\
&\lambda_2 W^k(x_1, x_2+1, n_{11}, n_{12}+1, n_{21}, n_{22}) + \\
&\mu \cdot 1 \cdot W^k(x_1, x_2, n_{11}, n_{12}, n_{21}, n_{22})) + \\
&\beta''/\beta \, W^k(x_1, x_2, n_{11}, n_{12}+1, n_{21}, n_{22})).
\end{aligned} \tag{C.5}
$$

Note that $\beta' - \beta'' = \mu$.

To prove (C.1), we will first compare the second to third value function ((C.4) to (C.5)) and then first to second ((C.4) to (C.3)).

We compare (C.4) and (C.5) term-by-term:

- The difference in the penalty term is π_{12} in favor of (C.4).

- For λ_1: By induction hypothesis (IH), we derive:

$$\begin{aligned}
W^k(x_1+1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22}) &= \min(V^k(x_1, x_2, n_{11}+2, n_{12}, n_{21}, n_{22}), \\
&V^k(x_1, x_2, n_{11}+1, n_{12}+1, n_{21}, n_{22}), \\
&V^k(x_1+1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22})) \\
&\leq V^k(x_1, x_2, n_{11}+1, n_{12}+1n_{21}, n_{22}) \\
W^k(x_1+1, x_2, n_{11}, n_{12}+1, n_{21}, n_{22}) &= V^k(x_1, x_2, n_{11}+1, n_{12}+1, n_{21}, n_{22}).
\end{aligned} \tag{C.6}$$

  Hence, the term from (C.4) is smaller or equal.

- For λ_2, recall that for a comparison with $n_{12}+n_{22}=0$ is necessary. Then, the induction hypothesis yields $W^k(x_1, x_2+1, n_{11}, n_{12}+1, n_{21}, n_{22}) = V^k(x_1, x_2, n_{11}, n_{12}+1, n_{21}, n_{22}+1)$. By the induction hypothesis, this quantity is larger or equal to $V^k(x_1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22}+1)$, which is by the induction hypothesis equal to $W^k(x_1, x_2+1, n_{11}+1, n_{12}, n_{21}, n_{22})$.

- For the departure scenarios, recall that $\sum_{ij} n_{ij} = 0$. Hence, both departure terms are equal to $\mu \cdot W^k(x_1, x_2, 0, 0, 0, 0)$ and are equal.

- The residual terms are $(\beta''/\beta) \cdot W^k(x_1, x_2, 1, 0, 0, 0)$ and $(\beta''/\beta) \cdot W^k(x_1, x_2, 0, 1, 0, 0)$. Write W^k as a minimum over different actions leading to V^k:

$$\begin{aligned}
W^k(x_1, x_2, 1, 0, 0, 0) = \min(\quad & V^k(x_1-1, x_2, 1, 1, 0, 0), \\
& V^k(x_1, x_2-1, 1, 0, 0, 1), \\
& V^k(x_1, x_2, 1, 0, 0, 0)) \\
W^k(x_1, x_2, 0, 1, 0, 0) = \min(\quad & V^k(x_1-1, x_2, 1, 1, 0, 0), \\
& V^k(x_1, x_2-1, 0, 1, 1, 0), \\
& V^k(x_1, x_2, 0, 1, 0, 0)).
\end{aligned}$$

Here, the first terms ($V^k(x_1 - 1, x_2, 1, 1, 0, 0)$) are equal. For the second terms, $V^k(x_1, x_2 - 1, 1, 0, 0, 1) < V^k(x_1, x_2 - 1, 0, 1, 1, 0)$ by a two-fold application of the induction hypothesis. And for the third term, $V^k(x_1, x_2, 1, 0, 0, 0) < V^k(x_1, x_2, 0, 1, 0, 0)$ by induction hypothesis. Hence, the first minimum is applied over a set of pointwise lower values. We conclude that $W^k(x_1, x_2, 1, 0, 0, 0) \leq W^k(x_1, x_2, 0, 1, 0, 0)$.

Combining these term-by-term findings, we derive:

$$V^{k+1}(x_1, x_2, n_{11} + 1, n_{12}, n_{21}, n_{22}) - V^{k+1}(x_1, x_2, n_{11}, n_{12} + 1, n_{21}, n_{22}) \leq -\pi_{12}/\beta < 0 \tag{C.7}$$

Next, we compare the second value function against the first value function ((C.4) against (C.3)). We again compare term-by-term:

- For the penalty term, the difference is b/β is in favor of (C.4).
- For λ_1 term,

$$\begin{aligned} W^k(x_1 + 2, x_2, 0, n_{12}, 0, n_{22}) &= \min(V^k(x_1 + 1, x_2, 1, n_{12}, 0, n_{22}), V^k(x_1 + 2, x_2 - 1, 0, 0, 0, 1)) \\ &\geq \min(V^k(x_1 + 1, x_2, 1, n_{12}, 0, n_{22}), V^k(x_1 + 1, x_2 - 1, 1, 0, 0, 1)) \\ &\geq W^k(x_1 + 1, x_2, 1, n_{12}, 0, n_{22}). \end{aligned} \tag{C.8}$$

  where we used the induction hypothesis in limiting the cases for the first equality, used the induction hypothesis for the second inequality, and in the third inequality observed that both states under the minimum can be reached from state $(x_1 + 1, x_2, 1, n_{12}, 0, n_{22})$.

- For λ_2 term, there are two cases:
  1. If $n_{12} + n_{22} = 0$, then $W^k(x_1, x_2 + 1, n_{11} + 1, n_{12}, n_{21}, n_{22}) = V^k(x_1, x_2, n_{11} + 1, n_{12}, n_{21}, n_{22} + 1)$ by induction hypothesis while $\lambda_2 W^k(x_1 + 1, x_2 + 1, n_{11}, n_{12}, n_{21}, n_{22}) =$

$\min(V^k(x_1, x_2+1, n_{11}+1, n_{12}, n_{21}, n_{22}), V^k(x_1+1, x_2, n_{11}, n_{12}, n_{21}, n_{22}+1))$ and either of these terms is lower than $V^k(x_1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22}+1)$ by induction hypothesis

2. If $n_{12} + n_{22} = 1$, then the states that can be reached from $(x_1, x_2+1, n_{11}+1, n_{12}, n_{21}, n_{22})$ in (C.4) are the same that can be reached from $(x_1+1, x_2+1, n_{11}, n_{12}, n_{21}, n_{22})$ in (C.3), with the exception of $(x_1+1, x_2, n_{11}, n_{12}, n_{21}+1, n_{22})$. Then, we can compare $V^k(x_1, x_2+1, n_{11}+1, n_{12}, n_{21}, n_{22})$ against $V^k(x_1+1, x_2, n_{11}, n_{12}+1, n_{21}, n_{22})$. We can assume that we follow the same actions in $(x_1+1, x_2, n_{11}, n_{12}+1, n_{21}, n_{22})$ as we do in $(x_1, x_2+1, n_{11}+1, n_{12}, n_{21}, n_{22})$ (the transition rates are the same), until the patient in unit 1 departs, and then we assign the additional patient from class 1 to unit 1, and we reach the same state in both cases. Over that period (say, k transitions), we incur additional cost $k \cdot b$ in case (C.4) while cost $k \cdot \pi_{21}$ in case (C.3). By condition (c), we conclude that the expected value for (C.4) is superior or equal.

- For departure term combined with the residual term, the terms that differ are as follows:

$$(1/\beta)(\mu W^k(x_1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) + \beta'' W^k(x_1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22}))$$

for (C.4) against the following:

$$(1/\beta)(\beta' W^k(x_1+1, x_2, n_{11}, n_{12}, n_{21}, n_{22}))$$

for (C.3). Recall that $\beta' = \mu + \beta''$, which implies:

$$\begin{aligned}
\beta' W^k(x_1+1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) = & \quad \mu W^k(x_1+1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) \\
& + \beta'' W^k(x_1+1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) \\
\geq & \quad \mu W^k(x_1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) \\
& + \beta'' W^k(x_1+1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) \\
= & \quad \mu W^k(x_1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) \\
& + \beta'' V^k(x_1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22}) \\
\geq & \quad \mu W^k(x_1+1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) \\
& + \beta'' W^k(x_1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22})
\end{aligned}$$

Here, the second line is obvious (the state with more waiting patients has lower value, everything else being equal), the third line follows from the induction hypothesis, and the fourth line follows because $V^k(x_1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22})$ is one of the terms entering the minimum over which $W^k(x_1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22})$ is computed.

Combining all term-by-term steps, we conclude that $V^{k+1}(x_1, x_2, n_{11}+1, n_{12}, n_{21}, n_{22}) - V^{k+1}(x_1+1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) \leq -\pi_{12} < 0$, which we wanted to show.

□

Lemma C.1.2. *The value function is monotone (nondecreasing) in x_i. That is:*

$$\begin{aligned}
V(x_1+1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) \geq V(x_1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) \\
V(x_1, x_2+1, n_{11}, n_{12}, n_{21}, n_{22}) \geq V(x_1, x_2, n_{11}, n_{12}, n_{21}, n_{22})
\end{aligned} \tag{C.9}$$

for any $x_1, x_2, n_{11}, n_{12}, n_{22}, n_{21} \geq 0$.

Proof. Proof of Lemma C.1.2 We again prove the lemma by induction on k, showing the proof for class 1. For $k = 0$, the following holds:

$$\begin{aligned}&V^0(x_1+1, x_2, n_{11}, n_{12}, n_{22}, n_{21})\\ &= b{\cdot}(1{+}x_1{+}x_2){+}\pi_{12}n_{12}{+}\pi_{21}n_{21} > b{\cdot}(x_1{+}x_2){+}\pi_{12}n_{12}{+}\pi_{21}n_{21} = V^0(x_1, x_2, n_{11}, n_{12}, n_{22}, n_{21}),\end{aligned} \tag{C.10}$$

so the statement is true for $k = 0$.

For $k > 0$, we write

$$\begin{aligned}&V^{k+1}(x_1, x_2, n_{11}, n_{12}, n_{21}, n_{22}) =\\ &(b(x_1+x_2) + \pi_{12}n_{12} + \pi_{21}n_{21})/\beta +\\ &(1/\beta)(\lambda_1 W^k(x_1+1, x_2, n_{11}, n_{12}, n_{21}, n_{22})+\\ &\lambda_2 W^k(x_1, x_2+1, n_{11}, n_{12}, n_{21}, n_{22})+\\ &\mu \cdot n_{11} W^k(x_1, x_2, (n_{11}-1)^+, n_{12}, n_{21}, n_{22})+\\ &\mu \cdot n_{12} W^k(x_1, x_2, n_{11}, n_{12}, (n_{21}-1)^+, n_{22})+\\ &\mu \cdot n_{21} W^k(x_1, x_2, n_{11}, (n_{12}-1)^+, n_{21}, n_{22})+\\ &\mu \cdot n_{22} W^k(x_1, x_2, n_{11}, n_{12}, n_{21}, (n_{22}-1)^+)+\\ &\beta'''/\beta\, W^k(x_1, x_2, n_{11}, n_{12}, n_{21}, n_{22})),\end{aligned} \tag{C.11}$$

and similarly with $x_1' := x_1 + 1$ (we call this case C' and the case with x_1 as C) We distinguish two cases:

1. If $x_1 > 0$, then any action taken in case C' can be also taken in case C. Let * denote the variables before the action and ** after the optimal action taken in case C'. Then, the following holds:

$$\begin{aligned}&W^k(x_1^*+1, x_2^*, n_{11}^*, n_{12}^*, n_{21}^*, n_{22}^*) = V^k(x_1^{**}+1, x_2^{**}, n_{11}^{**}, n_{12}^{**}, n_{21}^{**}, n_{22}^{**}) >\\ &\qquad V^k(x_1^{**}+1, x_2^{**}, n_{11}^{**}, n_{12}^{**}, n_{21}^{**}, n_{22}^{**}) \geq W^k(x_1^*, x_2^*, n_{11}^*, n_{12}^*, n_{21}^*, n_{22}^*),\end{aligned} \tag{C.12}$$

where the second inequality holds by induction hypothesis and the third inequality because the action was taken to be optimal for case C' but not necessarily for case C.

2. If $x_1 = 0$, then the previous argument holds except when the optimal action for case C' is to assign a patient of class 1 but there has not been an arrival from class 1. Then in case C, it is impossible to take the assignment action. Then in case C, consider the case of not doing anything. Then the state after the action is the same for C' and C except that in C', there is one additional patient assigned in one of the beds. Then, it is possible in state C to replicate the actions from state C' until the departure of the patient assigned in case C', which in C will correspond to a transition without a state change if the two cases were coupled on the same probability space. In this case, the penalty accumulated over this trajectory will be the same between C' and C if the patient was originally assigned to unit 1 and will be higher for C' if the patient was assigned to unit 2 (accruing an additional misallocation penalty). In either case, we conclude that the value function in state $(1, x_2, n_{11}, n_{12}, n_{21}, n_{22})$ was higher (worse) than in state $(0, x_2, n_{11}, n_{12}, n_{21}, n_{22})$.

In both cases, we conclude that (C.9) holds.

□

Next, we show a lemma that will simplify the proof of the submodularity condition.

Lemma C.1.3. *Let the following hold:*

$$\begin{aligned} f(x+1,0) + f(x+1,1) &\leq f(x+2,0) + f(x,1) \\ f(x+1,0) + f(x,1) &\leq f(x,0) + f(x+1,1). \end{aligned} \tag{C.13}$$

Then the following holds:

$$f(x+1,0) - f(x,1) \leq f(x+2,0) - f(x+1,1) \tag{C.14}$$

Proof. Proof of Lemma C.1.3 Clearly, the first condition in (C.13) is just rearranged (C.14). Hence, the statement is trivial. □

Lemma C.1.4. *The value function is submodular in x_i. That is:*

$$V(x_1+2,0,1,0,0,0)-V(x_1+1,0,1,0,0,0) \geq V(x_1+1,0,1,1,0,0)-V(x_1,0,1,1,0,0) \tag{C.15}$$

for any $x_1 \geq 0$ and mutatis mutandis for x_2 and for $n_{21} = 1$ instead of $n_{11} = 1$.

Proof. Proof of Lemma C.1.4 Observe that by Lemma C.1.1, we can focus on the cases with $x_2 = 0$ and $n_{11} + n_{21} = 1$. We proof the statement by induction on k, showing that if the statement holds for V^{k+1}, it also holds for W^k and V^k, in parallel for class 1 and class 2. By Lemma C.1.3, we will not show the submodularity directly, but instead show the pair of conditions from (C.13), namely:

$$\begin{aligned}&\tilde{V}(x_1+2,0)+\tilde{V}(x_1,1) \equiv V(x_1+2,0,1,0,0,0)+V(x_1,0,1,1,0,0)\\&\geq V(x_1+1,0,1,1,0,0)+V(x_1+1,0,1,0,0,0) \equiv \tilde{V}(x_1+1,1)+\tilde{V}(x_1+1,0)\end{aligned} \tag{C.16}$$

$$\begin{aligned}&\tilde{V}(x_1,0)+\tilde{V}(x_1+1,1) = V(x_1,0,1,0,0,0)+V(x_1+1,0,1,1,0,0)\\&\geq V(x_1,0,1,1,0,0)+V(x_1+1,0,1,0,0,0) = \tilde{V}(x_1,1)+\tilde{V}(x_1+1,0),\end{aligned} \tag{C.17}$$

where we write $\tilde{V}(y,i) := V(y,0,1,i,0,0)$ and similarly with $\tilde{W}$. Given the conditions on equal service times, it is clear that the case with $n_{21} = 1$ is equivalent to $n_{11} = 1$, with value functions equal up to a constant multiple of π_{21}.

First, the following holds for $k = 0$:

$$\begin{aligned}&\tilde{V}^0(x_1+2,0)+\tilde{V}^0(x_1,1) = b(2x_1+2)+\pi_{12} = \tilde{V}^0(x_1+1,1)+\tilde{V}^0(x_1+1,0)\\&\tilde{V}^0(x_1,0)+\tilde{V}^0(x_1+1,1) = b(2x_1+1)+\pi_{12} = \tilde{V}^0(x_1,1)+\tilde{V}^0(x_1+1,0).\end{aligned} \tag{C.18}$$

This shows (C.16) and (C.17) for $k = 0$.

Next, let $k > 0$, the statement (C.15) holds for V^k, and we show that it also holds for W^k and V^{k+1}. We then proceed to show the following statements (in this order):

1. Show that (C.16) holds for W^k
2. Show that (C.17) holds for W^k
3. Show that Lemma C.1.5 holds for k
4. Show that Lemma C.1.6 holds for k
5. Show that (C.16) holds for V^{k+1}
6. Show that (C.17) holds for V^{k+1}

Then the lemma will be proven.

We first show that the condition from (C.16) holds for W^k. There are cases depending on the optimal action under $\tilde{x}_1 = x + 2$:

- If the optimal action is to assign to the secondary unit under $\tilde{x}_1 = x_1 + 2$, then the following holds:

  $$\tilde{W}^k(x_1{+}2,0){+}\tilde{W}^k(x_1,1) = \tilde{V}^k(x_1{+}1,1){+}\tilde{V}^k(x_1,1) = \tilde{W}^k(x_1{+}1,1){+}\tilde{W}^k(x_1,1), \tag{C.19}$$

  which is the first equation in (C.16). Here, we have used that $\tilde{W}(y,1) = \tilde{V}(y,1)$ for any y because both beds are occupied.

- If the optimal action is to *not* assign to the secondary unit under $\tilde{x}_1 = x_1 + 2$, then the following holds:

  $$\begin{aligned}\tilde{W}^k(x_1+1,0) + \tilde{W}^k(x_1+1,1) &\leq \tilde{V}^k(x_1+1,0) + \tilde{V}^k(x_1+1,1)\\ &\leq \tilde{V}^k(x_1+2,0) + \tilde{V}^k(x_1,1) = \tilde{W}^k(x_1,1) + \tilde{W}^k(x_1+2,0),\end{aligned} \tag{C.20}$$

where we used the definition of W^k as a minimum over V^k in the first inequality and the inductive hypothesis for the second inequality.

Next we show that condition (C.17) holds for W^k. We again distinguish cases depending on the optimal action:

- If the optimal action under x_1 is to wait, then the following holds:

$$\begin{aligned}\tilde{W}^k(x_1+1,0)+\tilde{W}^k(x_1,1) &\leq \tilde{V}^k(x_1+1,0)+\tilde{V}^k(x_1,1) \leq \tilde{V}^k(x_1,0)+\tilde{V}^k(x_1+1,1)\\ &= \tilde{W}^k(x_1,0)+\tilde{W}^k(x_1+1,1).\end{aligned} \tag{C.21}$$

  Here, the first inequality followed from the definition of W^k and the second inequality from the inductive hypothesis.

- If the optimal action under x_1 is to assign to the secondary department, then the following holds:

$$\begin{aligned}\tilde{W}^k(x_1+1,0)+\tilde{W}^k(x_1,1) &\leq \tilde{V}^k(x_1,1)+\tilde{V}^k(x_1,1) \overbrace{\leq}^{(*)} \tilde{V}^k(x_1-1,1)+\tilde{V}^k(x_1+1,1)\\ &= \tilde{W}^k(x_1,0)+\tilde{W}^k(x_1+1,1),\end{aligned} \tag{C.22}$$

  where $(*)$ follows from a successive application of inductive hypothesis for (C.16) and then (C.17).

We would next prove Lemmas C.1.5 and C.1.6. The proofs are fairly long, so we provide them at the end of the proof of this Lemma C.1.4.

Next, we show that condition (C.16) holds for $\tilde{V}^{k+1}$ if it holds for $\tilde{V}^k$ and $\tilde{W}^k$. We write

$$\tilde{V}^{k+1}(x+1,0)+\tilde{V}^{k+1}(x+1,1)=$$

$$(c(x_1+1,0,1,0,0,0)+c(x_1+1,0,1,1,0,0))/\beta+\frac{1}{\beta}(\tag{C.23}$$

$$\lambda_1(W^k(x_1+2,0,1,0,0,0)+W^k(x_1+2,0,1,1,0,0))+ \tag{C.24}$$

$$\lambda_2(W^k(x_1+1,1,1,0,0,0)+W^k(x_1+1,1,1,1,0,0)+ \tag{C.25}$$

$$\mu(W^k(x_1+1,0,0,0,0,0)+W^k(x_1+1,0,0,1,0,0))+ \tag{C.26}$$

$$\mu(W^k(x_1+1,0,1,0,0.0)+W^k(x_1+1,0,1,0,0.0))+ \tag{C.27}$$

$$(\beta-\lambda_1-\lambda_2-2\mu)(W^k(x_1+1,0,1,0,0.0)+W^k(x_1+1,0,1,1,0.0))) \tag{C.28}$$

We demonstrate that inequality (C.16) holds for each of the lines (C.23)–(C.28):

- For (C.23):

$$(c(x_1{+}1,0,1,0,0,0){+}c(x_1{+}1,0,1,1,0,0))=b(x_1{+}1{+}x_1{+}1){+}\pi_{12}=b(x_1{+}2){+}b(x_1){+}\pi_{12}$$
$$=(c(x_1+2,0,1,0,0,0)+c(x_1,0,1,1,0,0)) \tag{C.29}$$

- For (C.24), the inequality

$$W^k(x_1+2,0,1,0,0,0)+W^k(x_1+2,0,1,1,0,0)\leq$$
$$W^k(x_1+3,0,1,0,0,0)+W^k(x_1+1,0,1,1,0,0),$$

follows directly from the induction hypothesis for (C.16) and W^k.

- For (C.25), the following holds:

$$W^k(x_1+1,1,1,0,0,0)+W^k(x_1+1,1,1,1,0,0) \tag{C.30}$$

$$=V^k(x_1+1,0,1,0,0,1)+V^k(x_1+1,1,1,1,0,0) \tag{C.31}$$

$$=V^k(x_1+1,0,1,0,0,1)-c\pi_{12}+V^k(x_1+1,1,1,0,0,1) \tag{C.32}$$

$$W^k(x_1+2,1,1,0,0,0)+W^k(x_1,1,1,1,0,0) \tag{C.33}$$

$$=V^k(x_1+2,0,1,0,0,1)+V^k(x_1,1,1,1,0,0) \tag{C.34}$$

$$=V^k(x_1+2,0,1,0,0,1)1-c\pi_{12}+V^k(x_1,1,1,0,0,1). \tag{C.35}$$

  That (C.32) is less or equal than (C.35) follows from Lemma C.1.5.

- For (C.26) with $x_1 > 0$, a patient of class 1 will be assigned to unit 1 and the statement holds by induction hypothesis for V^k.

- For (C.26) with $x_1 = 0$, we have

$$\begin{gathered}W^k(1,0,0,0,0,0)+W^k(1,0,0,1,0,0)=V^k(0,0,1,0,0,0)+V^k(0,0,1,1,0,0)\\ \overbrace{\leq}^{*} V^k(1,0,1,0,0,0)+V^k(0,0,0,1,0,0)=W^k(2,0,0,0,0,0)+W^k(0,0,0,1,0,0)\end{gathered} \tag{C.36}$$

  The inequality $(*)$ is proved similarly by induction as later inequality (C.55), so we skip a detailed proof.

- For (C.27) with $x_1 \geq 0$, we have

$$W^k(x_1{+}1,0,1,0,0.0)+W^k(x_1{+}1,0,1,0,0.0)\leq W^k(x_1{+}2,0,1,0,0.0)+W^k(x_1,0,1,0,0.0) \tag{C.37}$$

  and the inequality holds by invoking both (C.16) and (C.17) for W^k and summing the two inequalities.

- For (C.28), the inequality (C.16) holds because we have already shown that if it holds for V^k, it holds for W^k.

Finally, we show that condition (C.17) holds for $\tilde{V}^{k+1}$ if it holds for $\tilde{V}^k$ and $\tilde{W}^k$.

$$\tilde{V}^{k+1}(x+1,0) + \tilde{V}^{k+1}(x,1) =$$

$$(c(x_1+1,0,1,0,0,0) + c(x_1,0,1,1,0,0))/\beta + \frac{1}{\beta}(\tag{C.38}$$

$$\lambda_1(W^k(x_1+2,0,1,0,0,0) + W^k(x_1+1,0,1,1,0,0))+ \tag{C.39}$$

$$\lambda_2(W^k(x_1+1,1,1,0,0,0) + W^k(x_1,1,1,1,0,0))+ \tag{C.40}$$

$$\mu(W^k(x_1+1,0,0,0,0,0) + W^k(x_1,0,0,1,0,0))+ \tag{C.41}$$

$$\mu(W^k(x_1+1,0,1,0,0.0) + W^k(x_1,0,1,0,0.0))+ \tag{C.42}$$

$$(\beta - \lambda_1 - \lambda_2 - 2\mu)(W^k(x_1+1,0,1,0,0.0) + W^k(x_1,0,1,1,0.0))) \tag{C.43}$$

We demonstrate that inequality (C.17) holds for each of the lines (C.39)–(C.43):

- We derive the inequality for (C.39) directly from the inductive hypothesis for (C.17) for W^k.

- For (C.40), we have

$$W^k(x_1+1,1,1,0,0,0) + W^k(x_1,1,1,1,0,0) \tag{C.44}$$

$$= V^{k-1}(x_1+1,0,1,0,0,1) + V^{k-1}(x_1,1,1,1,0,0) \tag{C.45}$$

$$W^k(x_1,1,1,0,0,0) + W^k(x_1+1,1,1,1,0,0) \tag{C.46}$$

$$= V^{k-1}(x_1,0,1,0,0,1) + V^{k-1}(x_1+1,1,1,1,0,0) \tag{C.47}$$

That (C.45) is less or equal to (C.47) follows from Lemma C.1.6.

- For (C.41) for $x_1 > 0$, a patient from class 1 will be placed in unit 1 in all cases and then we use the inductive hypothesis for (C.17) for V^k.
- For (C.41) for $x_1 = 0$, we have

$$W^k(1,0,0,0,0,0) + W^k(0,0,0,1,0,0) = V^k(0,0,1,0,0,0) + V^k(0,0,0,1,0,0)$$
$$\overbrace{\leq}^{(*)} V^k(0,0,0,0,0,0)+V^k(0,0,1,1,0,0) = W^k(0,0,0,0,0,0)+W^k(1,0,0,1,0,0) \tag{C.48}$$

and inequality $(*)$ can be proved by induction on k (the non-trivial cases of the V^k expansion are the departures; these end up resulting in the same set of scenarios, so both sides are then equal).

- For(C.42), we have

$$W^k(x_1{+}1,0,1,0,0.0)+W^k(x_1,0,1,0,0.0) \leq W^k(x_1{+}2,0,1,0,0.0)+W^k(x_1,0,1,0,0.0), \tag{C.49}$$

which follows by combining (C.16) and (C.17) for W^k and applying the inductive hypothesis for these cases.

- For (C.43), the inequality (C.17) holds because we have already shown that if it holds for V^k, it holds for W^k.

We conclude with proving Lemmas C.1.5 and C.1.6.

Lemma C.1.5.

$$V^k(x_1{+}1,x_2{+}1,1,0,0,1){-}V^k(x_1,x_2{+}1,1,0,0,1) \leq V^k(x_1{+}2,x_2,1,0,0,1){-}V^k(x_1{+}1,x_2,1,0,0,1) \tag{C.50}$$

for any $x_1 \geq 0, x_2 \geq 0$

Proof. Proof. The equation clearly holds for $k = 0$. For $k > 0$, we expand V^k as usual. For λ_1, λ_2, both units are occupied, apply induction. For μ, the terms are always equal for $x_1 \geq 1$.The only difficult cases are the following:

1. $x_2 = 0$ and the patient corresponding to n_{22} departs: Then the following two sets of equalities and inequalities hold:

$$\begin{aligned}
&W^k(x_1+1,1,1,0,0,0) - W^k(x_1,1,1,0,0,0)\\
&= V^k(x_1+1,0,1,0,0,1) - V^k(x_1,0,1,0,0,1)\\
&= V^k(x_1+1,0,1,1,0,0) - V^k(x_1,0,1,1,0,0)\\
&= W^k(x_1+1,0,1,1,0,0) - W^k(x_1,0,1,1,0,0)\\
&W^k(x_1+2,0,1,0,0,0) - W^k(x_1+1,0,1,0,0,0)\\
&\overbrace{\geq}^{(*)} W^k(x_1+1,0,1,1,0,0) - W^k(x_1,0,1,1,0,0)
\end{aligned}$$

For $(*)$, we have used the induction hypothesis for W^k and (C.16). This shows the inequality for this part of the V^k expansion.

2. $x_1 = 0$ and n_{11} departs. We need to show the following:

$$W^k(1,x_2+1,0,0,0,1)+W^k(1,x_2,0,0,0,1) \leq W^k(2,x_2,0,0,0,1)+W^k(0,x_2+1,0,0,0,1) \tag{C.51}$$

The following holds:

$$W^k(1,x_2+1,0,0,0,1) + W^k(1,x_2,0,0,0,1) = V^k(0,x_2+1,1,0,0,1) + V^k(0,x_2,1,0,0,1) \tag{C.52}$$

$$W^k(2,x_2,0,0,0,1) + W^k(0,x_2+1,0,0,0,1) = V^k(1,x_2,1,0,0,1) + W^k(0,x_2+1,0,0,0,1) \tag{C.53}$$

There are two cases:

(a) Either $W^k(0, x_2+1, 0, 0, 0, 1) = V^k(0, x_2, 0, 0, 1, 1)$. Then

$$\begin{aligned}
&V^k(1, x_2, 1, 0, 0, 1) + W^k(0, x_2+1, 0, 0, 0, 1) \\
&= V^k(1, x_2, 1, 0, 0, 1) + V^k(0, x_2, 0, 0, 1, 1) \qquad \text{(C.54)} \\
&= V^{k-1}(1, x_2, 1, 0, 0, 1) + V^k(0, x_2, 1, 0, 0, 1) + \frac{\pi_{21}}{\mu}
\end{aligned}$$

Then comparing with (C.52), the term $V^k(0, x_2, 1, 0, 0, 1)$ is canceled and it remains to be seen that $V^k(0, x_2+1, 1, 0, 0, 1) \leq V^k(1, x_2, 1, 0, 0, 1) + \frac{\pi_{21}}{\mu}$. This follows by a coupling argument where we follow a suboptimal strategy for $V^k(0, x_2+1, 1, 0, 0, 1)$ to mimic what the other side is doing, but when there would be assigned of class 1 to unit 1, we will instead assign class 2, incurring an additional loss of at most $\frac{\pi_{21}}{\mu}$ on average.

(b) If $W^k(0, x_2+1, 0, 0, 0, 1) = V^k(0, x_2+1, 0, 0, 0, 1)$, then we need to compare the following terms:

$$V^k(0, x_2+1, 1, 0, 0, 1) + V^k(0, x_2, 1, 0, 0, 1) \leq V^k(1, x_2, 1, 0, 0, 1) + V^k(0, x_2+1, 0, 0, 0, 1), \qquad \text{(C.55)}$$

and show that the first is less or equal. We can prove this by induction on k, drawing on previous induction hypotheses. We expand V as usual:

$$\begin{aligned}
&V^{k+1}(0, x_2+1, 1, 0, 0, 1) + V^{k+1}(0, x_2, 1, 0, 0, 1) = \\
&(c(0, x_2+1, 1, 0, 0, 1) + c(0, x_2, 1, 0, 0, 1))/\beta + \frac{1}{\beta}(\qquad \text{(C.56)} \\
&\lambda_1(W^k(1, x_2+1, 1, 0, 0, 1) + W^k(1, x_2, 1, 0, 0, 1)) + \qquad \text{(C.57)} \\
&\lambda_2(W^k(0, x_2+2, 1, 0, 0, 1) + W^k(0, x_2+1, 1, 0, 0, 1)) + \qquad \text{(C.58)} \\
&\mu(W^k(0, x_2+1, 0, 0, 0, 1) + W^k(0, x_2, 0, 0, 0, 1)) + \qquad \text{(C.59)} \\
&\mu(W^k(0, x_2+1, 1, 0, 0, 0) + W^k(0, x_2, 1, 0, 0, 0))). \qquad \text{(C.60)}
\end{aligned}$$

Inequalities for lines (C.56) and (C.58) are obvious: the first case follows by simple algebra, while the second case from the induction hypothesis for (C.55). For (C.57), we have:

$$\begin{aligned}
&W^k(1, x_2+1, 1, 0, 0, 1) + W^k(1, x_2, 1, 0, 0, 1) = \\
&\qquad V^k(1, x_2+1, 1, 0, 0, 1) + V^k(1, x_2, 1, 0, 0, 1) \leq \\
&\qquad V^k(2, x_2, 1, 0, 0, 1) + V^k(0, x_2+1, 1, 0, 0, 1) = \\
&\qquad\qquad W^k(2, x_2, 1, 0, 0, 1) + W^k(1, x_2+1, 0, 0, 0, 1),
\end{aligned} \tag{C.61}$$

where the inequality holds because of the inductive hypothesis of (C.50) with $x_1 = 0$.

For (C.60) combined with (C.59) , we have, if $x_2 = 0$:

$$\begin{aligned}
&W^k(0,1,0,0,0,1)+W^k(0,0,0,0,0,1)+W^k(0,1,1,0,0,0)+W^k(0,0,1,0,0,0) \\
&= W^k(0,1,0,0,0,1)+V^k(0,0,0,0,0,1)+V^k(0,0,1,0,0,1)+V^k(0,0,1,0,0,0) \\
&\qquad = A + V^k(0,0,1,0,0,0) = A + W^k(0,0,1,0,0,0) \leq \\
&\qquad\qquad = A + W^k(1,0,1,0,0,0) \\
&= W^k(1,0,1,0,0,0)+V^k(0,0,0,0,0,1)+V^k(0,0,1,0,0,1)+W^k(0,1,0,0,0,1) \\
&W^k(1,0,1,0,0,0)+W^k(0,1,0,0,0,0)+W^k(1,0,0,0,0,1)+W^k(0,1,0,0,0,1),
\end{aligned} \tag{C.62}$$

where $A = V^k(0,0,0,0,0,1) + V^k(0,0,1,0,0,1) + W^k(0,1,0,0,0,1)$ and the inequality follows from monotonicity.

For (C.60) combined with (C.59) , we have, if $x_2 > 0$:

$$\begin{aligned}
&W^k(0, x_2+1, 0, 0, 0, 1) + W^k(0, x_2, 0, 0, 0, 1) \\
&\qquad + W^k(0, x_2+1, 1, 0, 0, 0) + W^k(0, x_2, 1, 0, 0, 0) = \\
&W^k(0, x_2{+}1, 0, 0, 0, 1){+}W^k(0, x_2, 0, 0, 0, 1){+}V^k(0, x_2, 1, 0, 0, 1){+}V^k(0, x_2{-}1, 1, 0, 0, 1) \\
&= B + W^k(0, x_2, 0, 0, 0, 1) + V^k(0, x_2-1, 1, 0, 0, 1) \\
&\le B + V^k(0, x_2, 0, 0, 0, 1) + V^k(0, x_2-1, 1, 0, 0, 1) \\
&\le B + V^k(1, x_2-1, 1, 0, 0, 1) + V^k(0, x_2, 0, 0, 0, 1) \\
&= V^k(0, x_2, 1, 0, 0, 1) + W^k(0, x_2+1, 0, 0, 0, 1) + \\
&\quad V^k(1, x_2-1, 1, 0, 0, 1) + V^k(0, x_2, 0, 0, 0, 1) \\
&= W^k(1, x_2, 0, 0, 0, 1) + W^k(0, x_2+1, 0, 0, 0, 1) + \\
&\qquad W^k(1, x_2, 1, 0, 0, 0) + W^k(0, x_2+1, 0, 0, 0, 0), \quad \text{(C.63)}
\end{aligned}$$

where $B = V^k(0, x_2, 1, 0, 0, 1) + W^k(0, x_2+1, 0, 0, 0, 1)$ and the second inequality follows from monotonicity.

□

Lemma C.1.6.

$$V^k(x_1{+}1, x_2, 1, 0, 0, 1){-}V^k(x_1, x_2, 1, 0, 0, 1) \le V^k(x_1{+}1, x_2{+}1, 1, 0, 0, 1){-}V^k(x_1, x_2{+}1, 1, 0, 0, 1) \quad \text{(C.64)}$$

Proof. Proof. We prove the statement by induction on k. We expand V^{k+1} as usual. The terms corresponding to arrivals and to the cost are straightforward.

The difficult cases are the boundary cases for departures:

1. $x_1 = 0$ and n_{11} departs: We have the following

$$
\begin{aligned}
& W^k(0, x_2+1, 0, 0, 0, 1) - W^k(0, x_2, 0, 0, 0, 1) \\
& \overbrace{\leq}^{(*)} W^k(0, x_2+1, 0, 0, 1, 1) - W^k(0, x_2, 0, 0, 1, 1) \\
& = V^{k-1}(0, x_2+1, 0, 0, 1, 1) - V^{k-1}(0, x_2, 0, 0, 1, 1) \\
& = V^{k-1}(0, x_2+1, 1, 0, 0, 1) - V^{k-1}(0, x_2, 1, 0, 0, 1) \\
& W^k(1, x_2+1, 0, 0, 0, 1) - W^k(1, x_2, 0, 0, 0, 1) \\
& = V^{k-1}(0, x_2+1, 1, 0, 0, 1) - V^{k-1}(0, x_2, 1, 0, 0, 1)
\end{aligned}
$$

In $(*)$, we have used the induction hypothesis for x_2 and W^k.

2. $x_2 = 0$ and n_{22} departs: We have

$$
\begin{aligned}
& W^k(x_1+1, 0, 1, 0, 0, 0) - W^k(x_1, 0, 1, 0, 0, 0) \\
& \overbrace{\leq}^{(*)} W^k(x_1+1, 0, 1, 1, 0, 0) - W^k(x_1, 0, 1, 1, 0, 0) \\
& = V^{k-1}(x_1+1, 0, 1, 1, 0, 0) - V^{k-1}(x_1, 0, 1, 1, 0, 0) \\
& = V^{k-1}(x_1+1, 0, 1, 0, 0, 1) - V^{k-1}(x_1, 0, 1, 0, 0, 1) \\
& W^k(x_1+1, 1, 1, 0, 0, 0) - W^k(x_1, 1, 1, 0, 0, 0) \\
& = V^{k-1}(x_1+1, 0, 1, 0, 0, 1) - V^{k-1}(x_1, 0, 1, 0, 0, 1)
\end{aligned}
$$

In $(*)$, we have used the induction hypothesis.

□

□

C.2 Proof of Theorem 3.4.2

To prove Theorem 3.4.2, we need to show that there is a set of parameters such that:

1. The value of secondary assignment is positive when considering only the expected value of the flexible class, $V_2(\lambda_{21})$ (otherwise, the statement of the theorem is trivial), and

2. The value of secondary assignment is negative when considering the expected value of both classes, $V(\lambda_{21})$.

Indeed, we can formulate the theorem as comprising the following two statements:

$$\begin{aligned} 0 &> -\frac{\partial}{\partial \lambda_{21}} V_2(\lambda_{21}), \\ 0 &> \frac{\partial}{\partial \lambda_{21}}(-V_2(\lambda_{21}) + V_1(\lambda_{21})), \end{aligned} \tag{C.65}$$

under a certain nonempty set of parameters. We describe the proof for the case with $\mu_{ij} = \mu_{i'j'}$ for $i, i' \in I,\ j, j' \in J$, but this condition can be relaxed. We start the proof by separately considering the case with reservation ($\lambda_{21} = 0$) and without reservation ($\lambda_{21} > 0$).

In the case with reservation, we observe the following system dynamics:

- Class 2 experiences M/M/n_2.

- Class 1 experiences M/M/1

- Waiting and boarding times follow from standard queueing theory.

- No misallocation penalties are incurred.

Next, in the case without reservation, Class 1 has non-preemptive priority for Unit 1 but Class 2 can assign a (λ_{21}/λ_2) fraction of its arrivals to Unit 1 to be served when no patients of Class 1 are waiting. Under a fixed λ_{21}, Class 2 experiences waiting times that can be derived using queueing models with priority (Adan and Resing (2002), Section 9). We use superscript R to denote the reservation case and N the non-reservation case. We then have the expected value functions as follows:

- For the reservation case:

$$V = b \cdot (\lambda_1 \, \mathbf{E} \, W_1^R(M/M/1) + \lambda_2 \, \mathbf{E} \, W_2^R(M/M/n_2)) \tag{C.66}$$

The waiting times W_i^R follow from the basic queueing theory:

  - Class 1 experiences M/M/1. Hence:

$$\mathbf{E} \, W_1^R = \frac{\rho_1}{1-\rho_1} \cdot (1/\mu_1). \tag{C.67}$$

  - Class 2 experiences $M/M/n_2$. Hence:

$$\mathbf{E} \, W_2^R(\lambda_2) = \Pi_W(\lambda_2) \cdot \frac{1}{1-\lambda_2/(n_2\mu_2)} \cdot \frac{1}{n_2\mu_2} \tag{C.68}$$

    where $\Pi_W(\lambda_2)$ is the so-called delay probability (see Adan and Resing (2002) formula (5.1) for the exact expression).

- For the no-reservation case:

$$V = b \cdot (\lambda_1 \, \mathbf{E} \, W_1^N + \lambda_{21} \, \mathbf{E} \, W_{21}(\lambda_{21}) + \lambda_{22} \, \mathbf{E} \, W_{22}(\lambda_{22}) + \lambda_{21}\pi_{21}, \tag{C.69}$$

where we can write $\mathbf{E} \, W_2^N = \lambda_{21} \, \mathbf{E} \, W_{21}(\lambda_{21}) + \lambda_{22} \, \mathbf{E} \, W_{22}(\lambda_{22})$. We also have the following expression for the waiting times:

  - From eq (9.1) in Adan and Resing (2002):

$$\mathbf{E} \, W_1^N = \frac{\rho_1/\mu_1^2 + \lambda_{21}/\mu_2^2 \cdot}{1-\rho_1} = \mathbf{E} \, W_1^R + \frac{\lambda_{21}/\mu_2^2}{1-\rho_1} \tag{C.70}$$

$$
\begin{aligned}
b\,\mathbf{E}\,W_2^N(\lambda_{22}) &= b(\lambda_{22}\,\mathbf{E}\,W_{22}(\lambda_{22}) + \lambda_{21}\,\mathbf{E}\,W_{21}(\lambda_{21}))\\
&= b\frac{\lambda_{22}}{\lambda_2}\Pi_W(\lambda_{22})\cdot\frac{1}{1-\lambda_{22}/(n_2\mu_2)}\cdot\frac{1}{n_2\mu_2} + \frac{(1-\lambda_{22})}{\lambda_2}\cdot\\
&\quad b(\frac{\rho_1/\mu_1^2+(1-\lambda_{22})/\mu_2^2\cdot}{(1-\rho_1)(1-\rho_1-(1-\lambda_{22})/\mu_2)} + \pi_{21}/b)\\
&= b\cdot(\frac{1}{n_2\mu_2}\cdot\frac{\lambda_{22}}{\lambda_2}\frac{\lambda_{22}}{1-\lambda_{22}/(\mu_2 n_2)}\\
&\quad(\frac{(n_2^{n_2}/n_2!)\lambda_{22}^{n_2}}{(1-\lambda_{22}/(\mu_2 n_2))\sum_{\nu=0}^{n_2-1}\frac{n_2^{\nu}\lambda_{22}^{\nu}}{\nu!}+(n_2^{n_2}/n_2!)\lambda_{22}^{n_2}})\\
&\quad+(1-\frac{\lambda_{22}}{\lambda_2})\frac{\rho_1/\mu_1^2+(1-\lambda_{22})/\mu_2^2\cdot}{(1-\rho_1)(1-\rho_1-(1-\lambda_{22})/\mu_2)} + \pi_{21}/b).
\end{aligned}
\tag{C.71}
$$

We now demonstrate a case where the value of secondary assignment is positive for C_2 but negative overall. We first derive a condition for the value for C_2 to be positive. We consider a limiting scenario $\lambda_{22} \sim n_2\mu_2$, and subsequently $\lambda_2 \sim \lambda_{22}$ with λ_{21} being small. Continuing the calculation from (C.71), we approximate as follows:

$$
b\,\mathbf{E}\,W_2^N(\lambda_{22}) \approx b(\frac{\lambda_{22}}{\lambda_2}\cdot\frac{1}{1-\lambda_{22}/(\mu_2 n_2)}+(1-\frac{\lambda_{22}}{\lambda_2})\frac{\rho_1/\mu_1^2+(1-\lambda_{22})/\mu_2^2\cdot}{(1-\rho_1)(1-\rho_1-(1-\lambda_{22})/\mu_2)})+(1-\frac{\lambda_{22}}{\lambda_2})\pi_{21}.
\tag{C.72}
$$

In this formula, if we have $\rho_2 \equiv \lambda_2/(\mu_2 n_2) \gg \rho_1$, the first term predominates. Hence, the value of C_2 is maximized with $\lambda_{22} < \lambda_2$, which is equivalent to $\lambda_{21} > 0$. At the same time, the gain for C_2 can be made arbitrarily small by adjusting π_{21}. Finally, we know from (C.70) that the loss for C_1 is positive and does not depend on π_{21}. Hence, we have demonstrated that we can adjust parameters such that the value of secondary assignment for C_2 is positive, the loss from this secondary assignment for C_1 outweighs the gain for C_2, so that the overall objective value is negative if secondary assignment is allowed. This proves that the reservation of the bed for C_1 can improve the overall objective value.

We conclude the theorem by observing that $-\frac{\partial}{\partial\lambda_{21}}V_2(\lambda_{21}) = \frac{\pi_{21}}{\lambda_2}+b\frac{\partial}{\partial\lambda_{21}}[(1-\lambda_{21})\,\mathbf{E}\,W_{22}(\lambda_{21})+$

$\lambda_{21}\,\mathbf{E}\,W_{21}(\lambda_{21})]$ by (C.69) and that $-\frac{\partial}{\partial\lambda_{21}}V_2(\lambda_{21}) = -b\frac{1/\mu_2^2}{1-\rho_1} = -\Xi$, which follows from (C.69) and (C.70).

C.3 Proof of Proposition 3.5.1

Each instance of the GREAT-RL policy is characterized by its scheduling function σ and admission function α. It is easy to see the following settings of σ and α reproduce the Reserve-k-beds policy and Threshold policy respectively:

- Reserve-k-beds policy: $\alpha(i,j) = 1\{(1-\chi_{i',j})o_{i',j} < \kappa_j - k\},\ \sigma(i,j) = -\pi_{ij}$
- Threshold policy: $\alpha(i,j) = 1\{Q_i > \tau_{ij}\},\ \sigma(i,j) = -\pi_{ij}$

□

C.4 Simulation Setup: Scenario Generation

This section details how, given parameter ranges from Table 3.5, we generate a set of scenarios for the numerical study.

First, we generate parameter sets by Latin Hypercube sampling from the four-dimensional hypercube characterized by the parameter intervals in Table 3.5. The algorithmic steps to generate a scenario from a parameter set are as follows:

1. Generate I patient classes and its single primary unit, so that $I = J$. Without loss of generality, assign labels so that patient class i has primary unit $j = i$.
2. Set the number of beds per unit.
3. Randomly generate secondary pairs from all potential non-primary pairs, independently with a specified probability
4. Save the configuration, derived from steps 1–3, for all scenarios within a sample of scenarios for one evaluation. For the configuration, generate common random numbers as follows:

- Random numbers for service time variation $\Delta\mu_i$ from $N(0,1)$ for each class
- Random numbers for utilization variation Δu_i from $U(0,1)$ for each class

5. Boarding-to-misallocation ratio determines the boarding penalty given the misallocation penalty equal to 1.0
6. service rate for each class: $\mu_i = 1 + \zeta \cdot \Delta\mu_i$
7. Primary utilization for each class, ρ_i, drawn from the Beta distribution using Δu, with the beta parameter provided by variation in utilization β and alpha parameter set such that the mean utilization equals ρ.
8. Arrival rate per class, $\lambda_i = \rho_i \cdot \mu_i \cdot \kappa_i$

In total, we draw 24 parameter sets (scenarios) and generate 10 instances for each scenario.

C.5 Reservation Policies

In this section, we revisit the concept of reservation policies from Section 10 and introduce more advanced reservation policies.

Definition A *reservation policy* allows for secondary assignments but "reserves" a certain number of beds for use by patients from the primary class. More formally, if a primary pair becomes available for assignment, it is assigned immediately. Otherwise, secondary pair (i, j) will be assigned if o_{ij} and O_j are not "too high," i.e., some beds may be reserved for the primary class when occupancy is high. Section 10 described the Reserve-k-beds policy, here we propose further extensions.

Static Reservation Policy ("RS"). This approach expands the Reserve-k-beds policy to consider separate k for each unit, an approach adapted from the n-class static single-resource capacity control algorithm in revenue management. Consider all eligible patient classes and group the classes into patient class sets that share the same misallocation penalty. For example, if there are two primary classes, they are grouped into a single class set with the combined arrival process. We define the revenue from each patient assignment inversely proportional to the patient class' misallocation penalty. If classes vary in boarding penalties, these differences would also be incorporated into the revenue definition. To compute the reservation levels for each class set, we define a capacity control model, solved through dynamic programming, as in (Talluri and Van Ryzin 2006). To ensure that each patient class has at least one eligible unit, we further introduced an adjustment to account for the increasing boarding cost over time. This results in Algorithm 4. Here, i_p denotes the patient class for patient p. We specify the revenue function ρ as follows

$$R(i, j; \phi) = r_{ij}/\mu_{ij} + \phi(b_i), \tag{C.73}$$

where ϕ is a function to translate the boarding penalty to revenue. We will later describe several possible forms of this function.

Algorithm 4: Static Reservation Policy

```
Input: $R_{ij} := R(i, j, \phi)$ revenue function with $\phi$ given / precomputed
$B_{ij} \leftarrow$ ComputeBookingLimits() // See Algorithm 5
while Patient $p$ arrives or Unit $j$ frees a bed do
    if arrival then
        $j' \leftarrow \arg\max_j R_{i_p j}$ for $j$ with $B_{i_p j} > O_j$;
        if $j'$ exists then
            Assign $i_p$ to $j'$
    else if departure then
        $i' \leftarrow \arg\max_i (B_{ij}, R_{ij})$ for $i$ with waiting patients;
        if $B_{i'j} > O_j$ then
            Assign patient of class $i'$ in $j$
```

Our implementation of function `ComputeBookingLimits` is based on the static booking limits algorithm from Revenue Management (see Talluri and Van Ryzin (2006), Section 2.2.2) and is described in Algorithm 5. By $\{i\}$, we refer to the patient class set that contains class i. To determine ϕ from (C.73) while also ensuring that all patient classes can be assigned, we proceed iteratively:

Algorithm 5: ComputeBookingLimits

```
Output: $B_{ij}$, the booking limits
for $j \in J$ do
    Order $\{i\}$ in decreasing order according to $R_{ij}$;
    // Computes horizon
    $T_j \leftarrow$
      $ExpectedTimeUntilABedBecomesEmptyIfOccupiedByPrimaryClasses(j)$;

    for $i \in \{i\}$ do
        // Dynamic programming
        $D_i \sim \text{Pois}(\lambda_i, T_j)$ // Demand distribution
        $V_i(x) = \mathbf{E}_D[r_{ij} \min(D_i, (x - y_{i-1})^+) + V_{i-1}(x - \min(D_i, (x - y_{i-1})^+))]$;
        $y_i = \max\{x : r_{i+1,j} < V_i(x) - V_{i-1}(x)\}$;
    for $i \in \{i\}$ do
        $B_{ij} = \kappa_j - y_i$
```

1. Initialize the patient “revenue” for each department based on the misallocation policy and arbitrary constants for the boarding penalty $\phi(b)$. Then, compute the booking limits as described in Algorithm 5.

2. Generate new booking limits iteratively:

   (a) Compute new ϕ from equation (C.73) based on the estimated waiting time of each class from the M/M/c model

   (b) Recompute the booking limits using using the new revenue function

   (c) Stop when a stopping criterion is met. This stopping criterion could be a limit on the number of iterations or a simulated annealing-style stopping when the booking limits “do not change much” anymore.

3. Once the iteration stopped, the final booking limits are the ones to use for the reservation policy.

The entire calculation in steps 1 to 3 takes place before the simulation is started, so the resulting policy is static. We propose two directions for further improvement:

1. Improve the aforementioned iterative algorithm with a better search: For instance, one could implement a tabu-search-style iteration over the space of the booking limits.

2. Instead of iterating, determine the booking limits using machine learning or optimization based on actual simulation results.

Dynamic Reservation Policy ("RD"). Next, we develop a policy adapted from what is called in revenue management as n-class *dynamic* single-resource capacity control. We define the algorithm formally in Algorithm 6. Heuristically, the policy can be defined as follows:

1. Whenever a patient arrives or departs, if the arriving patient has a primary unit available or the unit opened has a primary patient class waiting, match the patient to its primary unit.

2. Otherwise, compute the marginal revenue and booking limit for all patient-department pairs for (if patient arrival) all open eligible units for the patient or (if a unit has a patient departed) all pairs of patient classes with waiting patients that are eligible for the unit. If there are multiple patient-unit pairs where the booking limit allows assignment, choose the pair with the highest marginal revenue. We compute the booking limit and the marginal revenue using an algorithm based on the dynamic reservation dynamic program from (Talluri and Van Ryzin 2006), with the following major modifications:

    - The horizon of the dynamic program is determined using the expected service time of the patient class considered.

- The expected revenue from the assignment is based on the service-time-adjusted misallocation penalty and expected boarding penalty incurred by the patient class if the patient is not assigned.
- The expected opportunity cost is based on the dynamic program, accounting for the arrivals and revenues of the other patient classes that are eligible for the unit.

To state the algorithm formally, we define the revenue function as follows:

$$R(i, j, S) = r_{ij}/\mu_{ij} + \hat{b}_i(S), \tag{C.74}$$

where S is the state of the system and $\hat{b}_i(S)$ is the estimated boarding cost incurred by a patient of class i until the next bed is available. The adjusted revenue function $R'(i, j, S)$ is defined in Algorithm 7 and captures the opportunity cost of the patient assignment. The revenue to go is computed through dynamic programming, based on the following recursion:

$$V(t, x) = \sum_{i' \in I'} \lambda'_{i'}(r(i', j, S) - (V_{t+1}(x) - V_{t+1}(x-1)))^+ \tag{C.75}$$

where I' is the set of other patient classes that can use department j and λ'_i is the arrival rate after uniformization, t represents time steps, and x the number of patients in unit j. We then compute the dynamic booking limits similarly as the dynamic booking limits algorithm in Talluri and Van Ryzin (2006), Section 2.5.

C.6 Threshold Policies

In this section, we revisit the concept of threshold policies from Section 3.4.1 and introduce more advanced policies.

Definition. A *threshold policy* schedules waiting patients to a secondary unit if the boarding queue is long and keeps them waiting otherwise. Formally, if primary pair (i, j) be-

Algorithm 6: Dynamic Reservation Policy

```
Input: $R(i, j, S)$ revenue function; $R'(i, j, S)$ adjusted revenue function
while Patient $p$ arrives or unit $j$ frees a bed do
    if arrival then
        if Primary unit $j'$ for $p$ has empty beds then
            Assign $p$ in $j'$
        else
            $j' \leftarrow \arg\max_j R'(i_p, j, S)$ for $j$ with empty beds;
            if $j'$ exists and $R'(i_p, j', S) > 0$ then
                Assign $i_p$ to $j'$
    else if departure then
        if Primary class $i$ for $j$ has waiting patients then
            Assign $i$ in $j$
        else
            $i' \leftarrow \arg\max_i R'(i, j, S)$ for $i$ with waiting patients;
            if $R'(i', j, S) > 0$ then
                Assign patient of class $i'$ in $j$
```

Algorithm 7: Adjusted revenue function $\rho'(i, j, S)$

```
Input: $S$ system state, $i$ patient class, $j$ department
Output: $R'(i, j, S)$ adjusted revenue evaluation
// Computes horizon
$T \leftarrow$ ExpectedTimeUntilDepartureWithPatientScheduled($j$);
$V(t, \xi) \leftarrow$ ComputeRevenueToGo($i$, $j$, $S$, $T$), $t = 1, \ldots, T,\ \xi = 0, \ldots, (\kappa_j - O_j)$
    // See (C.75)
$x \leftarrow \kappa_j - O_j$;
$\Delta V \leftarrow V(1, x) - V(1, x - 1)$;
// Adjusted revenue
$\rho'(i, j, S) \leftarrow \rho(i, j, S) - \Delta V$
```

comes available for assignment, it is assigned immediately. Otherwise, the pair is assigned if Q_i is higher than a certain threshold, $Q_i > \tau(i, j)$. The threshold may be static or depend on the system state, parameterized by the threshold function $\tau(i, j, S)$. We formalize this class of policies in Algorithm 8. In the main manuscript, we described a first-order threshold policy, which we henceforth call **Threshold policy T1**. We next discuss further options for the threshold function τ.

Algorithm 8: Threshold Policy

```
Input: Threshold function $\tau$
if patient $p$ arrives then
    if A primary unit $j$ for $p$ has free beds then
        Assign $p$ to $j$
    else if $J' := \{j : \tau(i_p, j') < Q_{i_p}\} \neq \emptyset$ then
        Assign $i_p$ into $j' \in J'$ with $\pi_{i_p,j'}$ minimal
else if bed becomes available in unit $j$ then
    if Primary patient $p$ for $j$ is waiting then
        Assign $p$ to $j$
    else if $I' := \{\iota : \tau(\iota, j) < Q_\iota\} \neq \emptyset$ then
        Assign $i' \in I'$ into $j$ for $i'$ with $\pi_{i',j}$ minimal
```

Threshold policy T2: We consider a patient class for which a patient arrived or a unit cleared. First, we compute time until either of the primary units for this patient class opens, T_o. In policy T2, we base this calculation on the current state of the hospital rather than a static calculation as in T1. After T_o, we assume that the patient class is assigned with a certain probability based on other primary classes for the units. This probability can be estimated by Little's Law, which stipulates that the long-term fraction of each class i in the unit should be equal to λ_i/μ_i, assuming none of the classes have multiple eligible units. Aggregating over all primary units of this class, we then compute the expected time until being assigned to any primary unit, during which patients will be boarding. This expected boarding time is multiplied by the boarding penalty to derive the expected boarding cost per patient if not assigned. Hence, the threshold function for policy T2 is defined as follows:

$$\begin{aligned} \tau_{i,j}^{(2)}(t) &= \lfloor \frac{\pi_{ij} + \Omega_{i,j}^{(1)}}{B_i^{(2)}(t)} \rfloor \\ B_i^{(2)}(t) &= b_i / (\sum_{j'} \chi_{i,j'} (\sum_{p \in U_{j'}(t)} \mu_p) \cdot \frac{\lambda_i/\mu_i}{\sum_{\iota \in I} \chi_{\iota,j'} \lambda_\iota / \mu_\iota}), \end{aligned} \tag{C.76}$$

where $\Omega_{ij}^{(1)}$ refers to the estimate from policy T1 and $U_j(t)$ are patients present in unit j at time t.

Threshold policy T3: This policy assumes that the considered class competes for beds with all other classes that are eligible for a given unit. This contrasts with T2, which only assumes competition with other primary classes. Therefore, the definition of the threshold function is as follows:

$$\begin{aligned} \tau_{i,j}^{(3)}(t) &= \lfloor \frac{\pi_{ij} + \Omega_{i,j}^{(1)}}{B_i^{(3)}(t)} \rfloor \\ B_i^{(3)}(t) &= b_i / (\sum_{j'} \chi_{i,j'} (\sum_{p \in U_{j'}(t)} \mu_p) \cdot \frac{\lambda_i/\mu_i}{\sum_{\iota \in I} \eta_{\iota,j'} \lambda_\iota / \mu_\iota}) \end{aligned} \tag{C.77}$$

Threshold policy T4: In this policy, we base the boarding cost calculation on steady state waiting times. First, we determine the steady state assignment by linear programming, for instance by minimizing excess capacity as in Kilinc et al. (2019). Let steady state assignment fractions be $\mathbf{y} = y_{ij}$. Given the assignments, we approximate waiting times using the M/M/c model for each class, assuming the number of available beds for each class corresponding to the optimal assignment, with these beds dedicated for the respective classes. The waiting time is then estimated as $W_i^{(M/M/c)} := W^{M/M/\sum_j y_{ij}}(\lambda_i, \mu_i)$, where the right-hand side can be computed as in the standard M/M/c model (formula (5.3) in

(Adan and Resing 2002)).Hence, we get the following threshold function:

$$\begin{aligned} \tau_{i,j}^{(4)} &= \lfloor \frac{\pi_{ij} + O_{i,j}^{(1)}}{B_i^{(4)}(t)} \rfloor \\ B_i^{(4)} &= b_i \cdot W_i^{(M/M/c)} \end{aligned} \tag{C.78}$$

C.7 Assumptions for Structural Results

We discuss four assumptions required by Theorems 3.4.1 and 3.4.2 from Section 3.4 of the main text.

Assumption 1. *The boarding penalty per unit of time is the same for all patient classes: $b_i = b$ for $i \in I$ for constant boarding-to-misallocation ratio b.*

Observe that the boarding penalty represents three factors: 1) that a boarding patient consumes scarce ED resources, 2) that they are not necessarily getting the ideal treatment (which could be afforded in the patient's primary unit), and 3) they are potentially preventing other patients from receiving timely service in the ED. The assumption indicates that the first and third factors predominate as these factors are likely to be similar for all patients while the second factor could vary.

Assumption 2. *Each patient class has one primary unit.*

This assumption reflects that in most hospitals, a unit corresponds to a different clinical specialty (e.g., neurology, cardiology, orthopedics), and there is a single ideal unit for each patient class.

Assumption 3. *The service time rate, μ_{ij} depends on the patient class served but not the unit. That is, $\mu_i \equiv \mu_{ij} = \mu_{ij'}$ for $j, j' \in J,\ i \in I$.*

This assumption suggests that the patient treatment and its duration are only dependent on the patient condition, and different units will treat the same patient class with the same efficiency.

Assumption 4. *The service time rate, μ_{ij} depends on the unit but not the patient class served. That is, $\mu_{ij} = \mu_{i'j}$ for $i, i' \in I,\ j \in J$.*

This condition indicates a specialized hospital or hospital subdivision where patient classes are all clinically similar.

C.8 Counterexamples: Specific Parameterization

Here, we provide the counterexample parameterization for Propositions 3.4.1 and 3.4.2.

Counterexample against Reserve-k-beds Policy. The parameters are as follows:

- Three units (U_1, U_2, U_3), each with one bed
- $1/\lambda_1 = 1/\lambda_2 = 1.1,\ 1/\lambda_3 = 100$
- $b_1 = 20,\ b_2 = 1,\ b_3 = 20$
- $\mu_1 = \mu_2 = 1,\ \mu_3 = 0.01$

Counterexample against Threshold Policy. The parameters are as follows:

- Three units (U_1, U_2, U_3), each with one bed
- $1/\lambda_1 = 1.1,\ 1/\lambda_2 = 4,\ 1/\lambda_3 = 20$
- $b_1 = 0.5,\ b_2 = 50.0,\ b_3 = 0.1$
- $\mu_1 = 0.1,\ \mu_2 = 0.5,\ \mu_3 = 0.1$
- Patient class 1 is secondary for U_2, with $\pi_{12} = 1$. Patient class 2 is primary for both U_2 and U_3.

C.9 Counterexamples: Generalized Framework

Here, we describe a generalization of counterexamples from Section 3.4.3 against a broader family of reservation and threshold policies, which we term *local policies*. To define the local policy, consider the flexible bed management queueing system as a bipartite graph. The nodes in this graph are 1) units and 2) patient classes, with an edge between unit U and a patient class C if (C, U) is an eligible pair. A policy is *local* if a decision at a node (e.g., a threshold for a patient class queue or a reservation level in a unit) is only dependent on state variables at the node itself and on adjacent nodes. We will say that a queueing system has *long-distance dependencies* if the graph has paths of length two or longer. We argue that local reservation and threshold policies may perform poorly in systems with long-distance dependencies. This can be seen by generalizing the counterexamples from the Appendix C.8, where there are three connected units and patient classes, the third patient class has very long interarrival times, so the third unit effectively serves as additional service capacity for patient classes from other units when the third patient class is not being served. Then, a policy that is local does not make decision at unit 1 with awareness of the availability of beds in unit 3, which will make the policy suboptimal.

C.10 Simulation Program Design

We implemented the simulation module in Python and now describe it in detail.

We begin with the input and output. The input of each simulation instance is a scenario parameterized as in Section 3.6.1 and a parameter indicating the policy. The output includes the objective value and various secondary measures such as the percentage of secondary ("off-service") assignments and mean boarding time. The output also includes the detailed sequence of arrivals and departures. We run the simulation for a period of user's choice and compute metrics over this entire period. We use common random numbers for arrival and service times for all policies within each simulation scenario. The input is provided as two lists: one of policies and one of scenarios. The list of scenarios must reference scenarios defined as files in the YAML configuration language. A single YAML configuration file contains all the parameters other than the policy required to run the simulation:

- Definition of the units: Their names and capacity

- Definition of the patient classes: The class' boarding penalty rate, misallocation penalty for each acceptable unit and arrival and service distributions.

- Stopping criteria: Either the maximum number of (virtual) days to run the simulation or the maximum number of arrivals to generate

The high-level classes to specify policies are `Policy` and its subclass, `MarkovPolicy`, which handle all interactions with the other classes. Subclassing the `MarkovPolicy` class, the user only needs to indicate the behavior of the policy on the arrival or on the departure. The user can access the state of the system and through the `Hospital` class, which we discuss in detail in the next paragraph. We provide the `StaticThresholdPolicy` class for easy experimentation with threshold approximations, where it is enough to specify the computation of thresholds. Subclassing this class, the user only needs to specify how to compute the thresholds.

The high-level design of the rest of the simulation code is as follows. The main script reads the specified policies and settings and initializes the `Sim` class, which manages the runtime. The `Sim` class constructs the policy and controls the simulation. The three main components are the following objects:

1. An object of class `Settings`, which is responsible for reading and initializing the settings from the YAML configuration file as well as saving the arrival and service times across different policies applied for these settings (i.e., the common random numbers).

2. An object of class `Hospital`, which encapsulates the state of the system. The

hospital further manages objects of classes `Patient` and `Department`, which capture states of particular patients and units.

3. An object of class `Timer`, which triggers events (arrivals and departures) and mediates the interaction between the state of the system, the policies, and the simulation runtime. The timer also stops the simulation once the stopping criterion is met.

The code is available upon request.

C.11 Additional Results for Section 1.5

We first discuss the results of Section 3.6.3 in more detail. We derive the findings and in particular Figure 3.2 through a series of numerical experiments. Our first experiment considers nine different scenarios, varying across three levels of the utilization and three levels of the boarding-to-misallocation ratio ("3-by-3"), on a grid with $\rho \in \{0.5, 0.85, 0.95\}$ (low, moderate, and high utilization) and boarding-to-misallocation ratio in $\{0.1, 0.5, 1.0\}$ (low, moderate, high). The other parameters were held constant, with the standard deviation of service times equal to zero and variation of utilization equal to 1.0. Table C.1 lists the results. Since the GREAT-RL policy always dominates, the results are presented as percent differences between the average per-patient value of the GREAT-RL policy and the other policies. Gcμ performs well under high-utilization scenarios, which is consistent with the theoretical results from Van Mieghem (1995). The Strict policy performs poorly in every scenario except for the lowest utilization and boarding-to-misallocation ratio.

In our second numerical experiment, we fix utilization equal to 0.85, the standard deviation of service times to 0.0 and the variation in utilization to 1.0, while varying the boarding-to-misallocation ratio. Figure C.1 visualizes the relative value of different policies as a function of the boarding-to-misallocation ratio. Again, the GREAT-RL policy dominates overall. GCμ approaches the GREAT-RL performance as the boarding component predominates, and the Strict performance approaches the performance of GREAT-RL

Table C.1: Percentage difference in average per-patient value between a benchmark policy and the GREAT-RL policy on the 3-by-3 set of scenarios.

Scenario	Policy			
Utilization / b-m ratio	GCμ	Greedy	LEWC-p	Strict
Low, low	91%	65%	62%	0%
Low, moderate	19%	3%	3%	215%
Low, high	18%	3%	4%	490%
Moderate, low	47%	45%	69%	185%
Moderate, moderate	8%	14%	28%	419%
Moderate, high	1%	14%	23%	554%
High, low	1%	18%	40%	144%
High, moderate	1%	30%	48%	216%
High, high	0%	23%	45%	204%

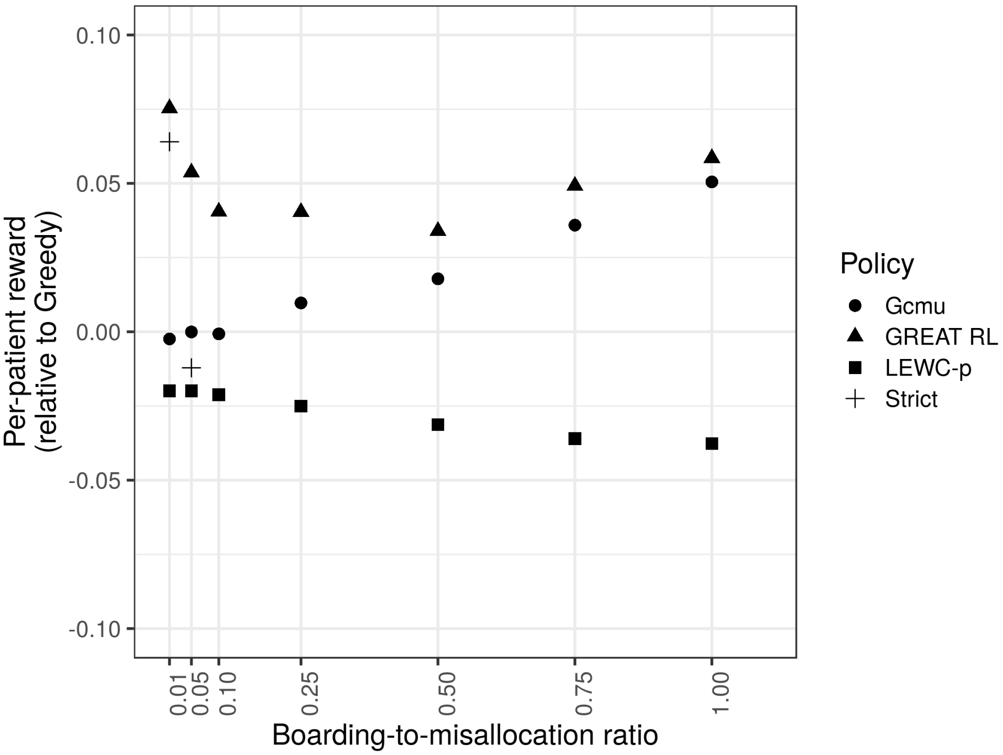


Figure C.1: Average per-patient value by boarding-to-misallocation ratio, relative to Greedy policy.

Note: For higher boarding-to-misallocation ratios, the Strict policy performance is off-the-chart negative, thus the corresponding point is not displayed.

as the boarding component vanishes.

In our third numerical experiment, we study the optimal policies in terms of the two components of the objective function, the secondary assignment penalties and boarding time. Figures C.2 and C.3 demonstrate the breakdowns of average secondary assignment fraction and the average boarding time for six of the 3x3 scenarios from Section 3.6.3. In the subfigure header, the first number corresponds to the utilization and the second one to the boarding-to-misallocation ratio. All results are averaged over 100 instances. The GREAT-RL algorithm achieves boarding times similar to Gcμ (which does not consider misallocation penalties) while attaining lower secondary assignment rates than Gcμ and all other policies except for the Strict policy.

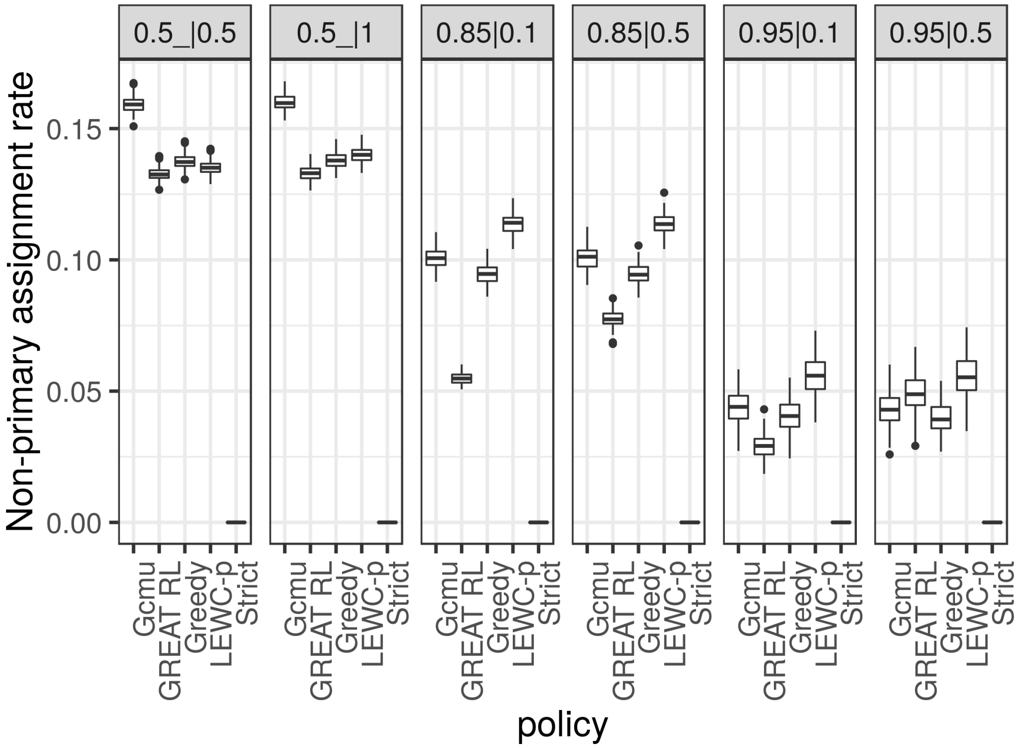


Figure C.2: Average secondary assignment rate on 3x3 scenarios

C.12 Results for additional Reservation and Threshold policies

Figure C.4 demonstrates the performance of the GREAT-RL policy and reservation and threshold policies from Appendices C.5 and C.6 on the hypercube set of scenarios from Figure 3.2. The GREAT-RL policy dominates when measured by either the mean or median performance across scenarios. The dynamic reservation policy also performs well but is

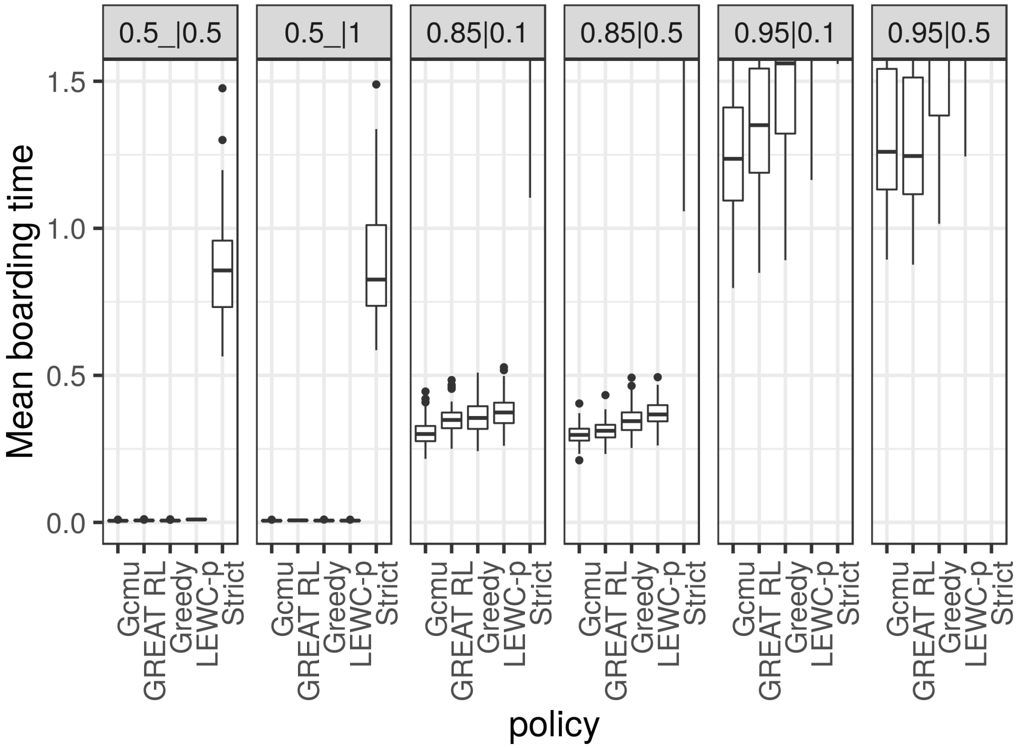


Figure C.3: Average boarding time on 3x3 scenarios

more complex than the typical policy learned by the GREAT-RL algorithm.

C.13 Sensitivity Analysis for Unequal Boarding Time

To assess the sensitivity of the GREAT-RL policy to the variation in boarding times, we assess robustness on several scenarios with unequal boarding times, with the following parameters: The baseline utilization beta equals 1.0, there is no service time variation, utilization equals 0.85, and boarding-to-misallocation ratio equals 0.1 before applying boarding penalty variation. There are four settings of the boarding penalty standard deviation, with five scenarios for each, then evaluated over ten instances. The averages are reported in Figure C.5. We observe no substantial differences in the performance of the GREAT-RL policy.

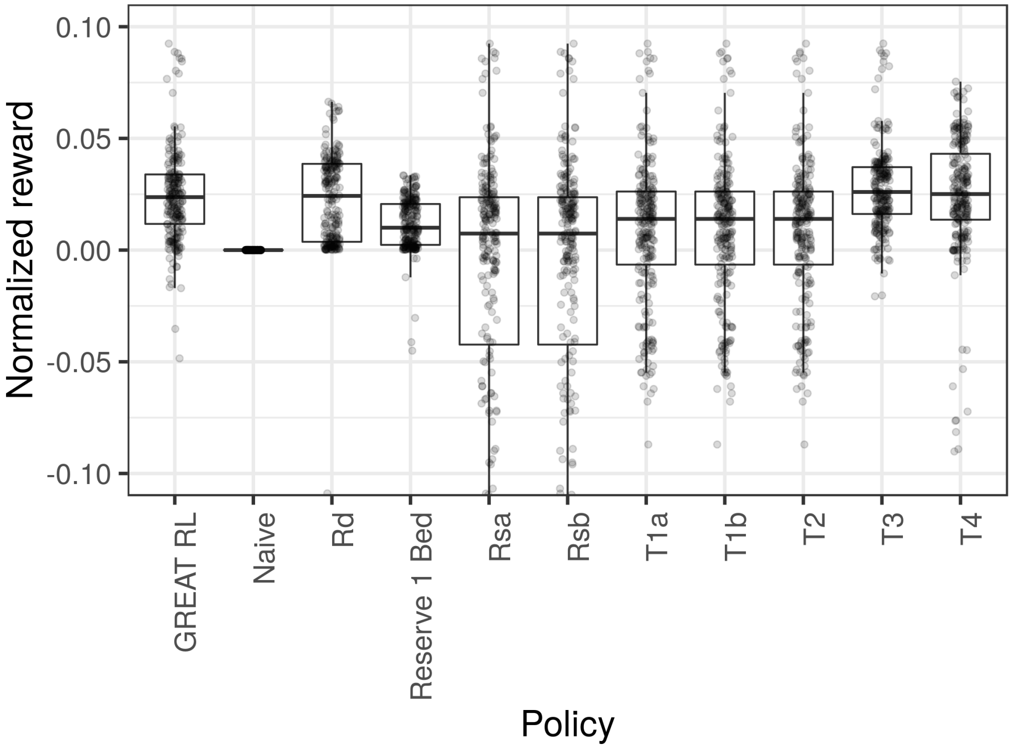


Figure C.4: Distribution of relative normalized rewards by policy for reservation and threshold policies

C.14 Patient Transfers

While this has not been the case in the hospital we worked with, one hospital in the literature reported assigning patients to secondary units and later transferring them when a primary unit becomes available (Thompson et al. 2009). Thompson et al. (2009) formulates a model and objectives different from the flexible bed management setup, but our setup could be extended to handle such "patient transfers" scenarios. Specifically, we can estimate the probability P_i^τ that a patient of class i will be transferred to a primary unit j_1 before finishing the service in secondary unit j_2 and, given the transfer, the average amount of time T_i^τ that the patient would spend in the primary rather than secondary unit. We can then use these estimates to adjust the misallocation penalties in the flexible bed

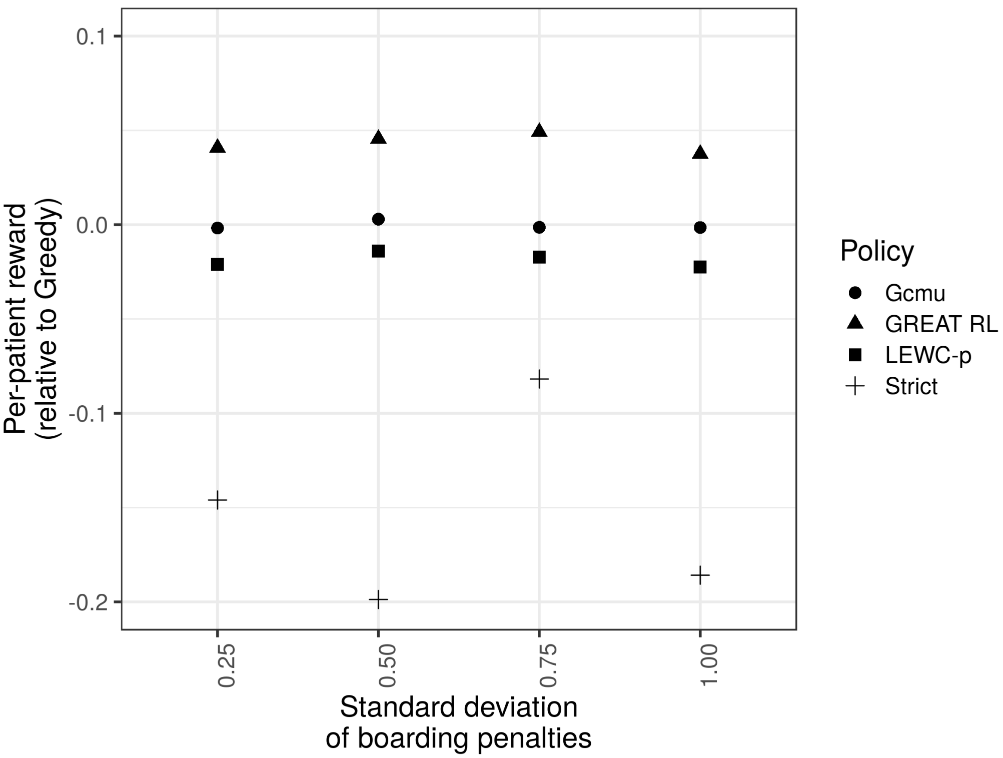


Figure C.5: Average relative normalized reward by variation in boarding penalty for scenario

management model, hence using adjusted penalties $\pi'_{i,j}$ defined as

$$\pi'_{i,j} := \pi_{i,j} - P_i^\tau \cdot T_i^\tau \cdot (\pi_{i,j_2} - \pi_{i,j_1}). \tag{C.79}$$

C.15 Choice of k for the Reserve k Beds Policy

We offer general guidelines and simulation insights on how to choose k for the Reserve k Beds policy. First, we observe that $k = 0$ translates to the Greedy policy while $k \to \infty$ to the Strict policy. Next, we note that from the M/M/c queueing theory one would expect that in scenarios with many beds, Reserve k Beds policies with larger values of k will indeed quickly behave like the Strict policy because there are rarely many patients waiting. Figure C.6 demonstrates these theoretical insights in a simulation study. Note that the units in these scenarios had between 10 and 30 beds. First, the Reserve 1 Bed and Reserve 2 Beds

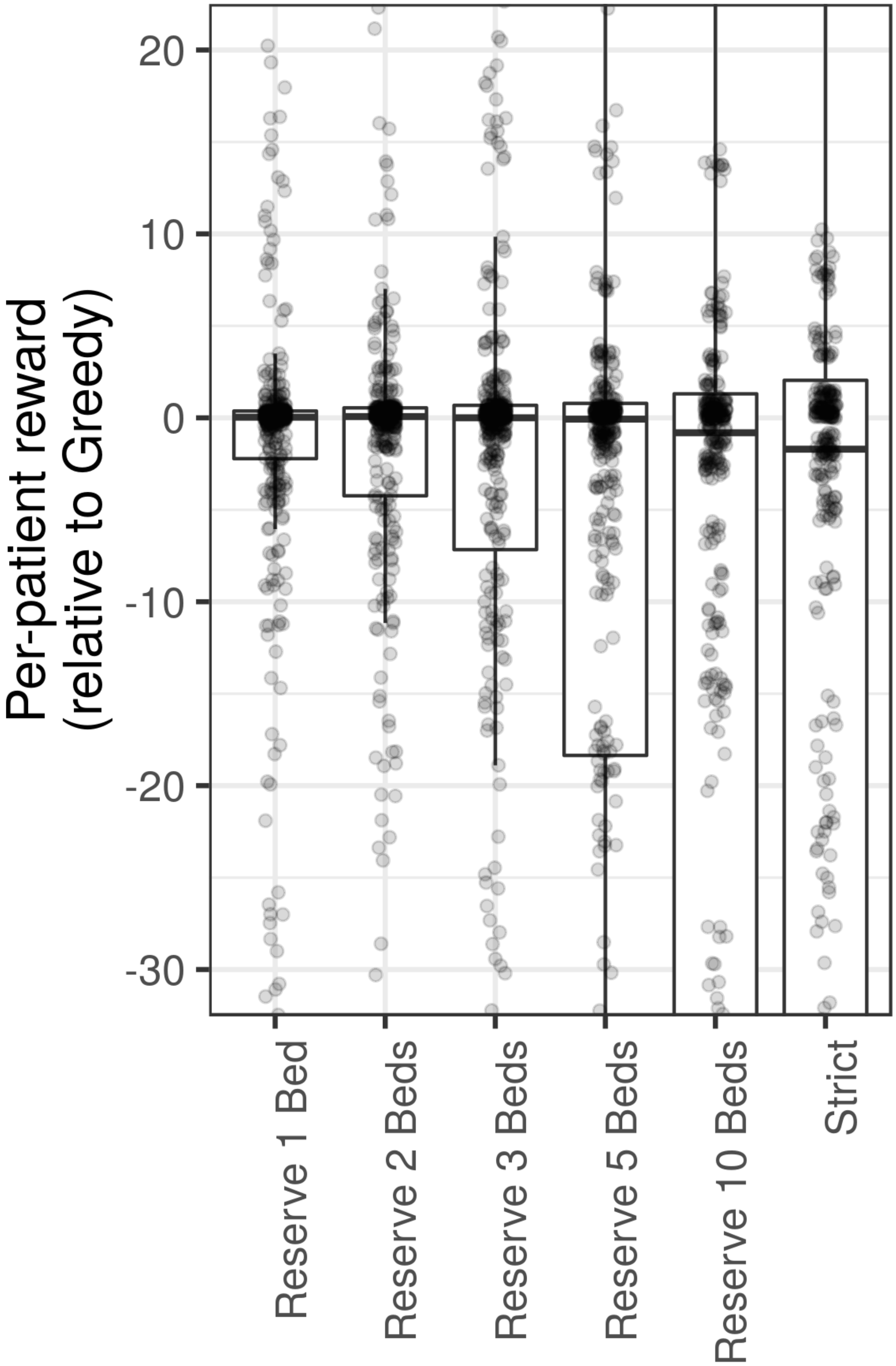


Figure C.6: Distribution of relative normalized rewards for different values of k in Reserve k Beds policies.

offer the highest median normalized rewards, higher than the Greedy policy, which serves as a baseline. Second, the Reserve k Beds policies do not perform well in scenarios with very high utilization, which can be seen by the very heavy bottom tail of scenarios with low performance. Third, the Reserve k Beds policies with high k do indeed perform similarly as the Strict Policy.

Thus, heuristically, we recommend choosing $k = 1$ except for scenarios with very high utilization where Reserve k Beds policies should not be used at all. When it is possible, we further recommend running an actual simulation to establish the most appropriate value of k more accurately.

BIBLIOGRAPHY

Adan, Ivo, Jacques Resing. 2002. *Queueing theory*. Eindhoven University of Technology.

Adida, Elodie, Hamed Mamani, Shima Nassiri. 2016. Bundled payment vs. fee-for-service: Impact of payment scheme on performance. *Management Science* Forthcoming.

Adler-Milstein, Julia, David Bates, Ashish Jha. 2013. Operational health information exchanges show substantial growth, but long-term funding remains a concern. *Health Affairs* **32**(8) 1486–92.

Adler-Milstein, Julia, Ashish Jha. 2014. Health information exchange among US hospitals: Who's in, who's out, and why? *Healthcare* **2**(1) 26–32.

Agency for Healthcare Research and Quality. 2014. HCUP State emergency department databases (SEDD). `http://www.hcup-us.ahrq.gov`.

Agha, Leila. 2014. The effects of health information technology on the costs and quality of medical care. *Journal of Health Economics* **34** 19–30.

Altman, Stuart. 2012. The lessons of Medicare's Prospective Payment System show that the bundled payment program faces challenges. *Health Affairs* **31**(9) 1923–1930.

American Association of Medical Colleges. 2014. Why teaching hospitals are important to all Americans. Online; accessed 2/28/2015. URL `https://www.aamc.org/advocacy/campaigns_and_coalitions/gmefunding/factsheets/253374/teaching-hospitals.html`.

American Hospital Association. 2009. Teaching hospitals: Their impact on patients and the future health care workforce. Tech. rep., American Hospital Association.

American Hospital Association. 2013. Moving towards bundled payment. Issue brief, American Hospital Association. Obtained from `http://www.aha.org/content/13/13jan-bundlingissbrief.pdf`.

American Hospital Association. 2014a. The AHA annual survey database.

American Hospital Association. 2014b. AHA Trendwatch Chartbook. Online; accessed 3/17 2015. URL `http://www.aha.org/research/reports/tw/chartbook/index.shtml`.

Anderson, David, Guodong Gordon Gao, Bruce Golden. 2014. Life is all about timing: An examination of differences in treatment quality for trauma patients based on hospital arrival time. *Production and Operations Management* **23**(12) 2178–2190.

Andritsos, Dimitrios, Christopher Tang. 2018. Incentive programs for reducing readmissions when patient care is co-produced. *Production and Operations Management* **27**(6) 999–1020.

Angst, Corey, Sarv Devaraj, Carrie Queenan, Brad Greenwood. 2011. Performance effects related to the sequence of integration of healthcare technologies. *Production and Operations Management* **20**(3) 319–333.

Appari, Ajit, Eric Johnson, Denise Anthony. 2013. Meaningful use of electronic health record systems and process quality of care: Evidence from a panel data analysis of US acute-care hospitals. *Health Services Research* **48**(2pt1) 354–375.

Armony, Mor, Carri W Chan, Bo Zhu. 2018. Critical care capacity management: Understanding the role of a step down unit. *Production and Operations Management* **27**(5) 859–883.

Arrow, Kenneth J, Theodore Harris, Jacob Marschak. 1951. Optimal inventory policy. *Econometrica: Journal of the Econometric Society* 250–272.

Ash, Joan, Marc Berg, Enrico Coiera. 2004. Some unintended consequences of information technology in health care: The nature of patient care information system-related errors. *Journal of the American Medical Informatics Association* **11**(2) 104–112.

Ata, Bars, Bradley Killaly, Tava Olsen, Rodney Parker. 2013. On hospice operations under medicare reimbursement policies. *Management Science* **59**(5) 1027–1044. doi:10.1287/mnsc.1120.1606.

Ayabakan, Sezgin, Zhiqiang Zheng, Indranil Bardhan, Kirk Kirksey. 2013. Can health information sharing reduce duplicate testing? A longitudinal analysis of patient switching behavior across multiple hospitals. *Working Paper* .

Ayal, Moshe, Abraham Seidman. 2009. An empirical investigation of the value of integrating enterprise information systems: the case of medical imaging informatics. *Journal of Management Information Systems* **26**(2) 43–68.

Ayanian, J. Z., J. S. Weissman. 2002. Teaching hospitals and quality of care: a review of the literature. *Milbank Quarterly* **80**(3) 569–593.

Bahmani, Bahman, Michael Kapralov. 2010. Improved bounds for online stochastic matching. *Algorithms–ESA 2010* 170–181.

Bailey, James, Rebecca Pope, Elizabeth Elliott, Jim Wan, Teresa Waters, Mark Frisse. 2013. Health information exchange reduces repeated diagnostic imaging for back pain. *Annals of Emergency Medicine* **62**(1) 16–24.

Bartel, Ann, Carri Chan, Song-Hee Hailey Kim. 2014. Should hospitals keep their patients longer? The role of inpatient and outpatient care in reducing readmissions. Working Paper 20499, National Bureau of Economic Research.

Bates, David W, Suchi Saria, Lucila Ohno-Machado, Anand Shah, Gabriel Escobar. 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs* **33**(7) 1123–1131.

Batt, Robert, Christian Terwiesch. 2012. Doctors under load: An empirical study of state-dependent service times in emergency care. Working paper, The Wharton School.

Batt, Robert, Christian Terwiesch. 2015. Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science* **61**(1) 39–59.

Baumann, Michael, Tania Strout. 2007. Triage of geriatric patients in the emergency department: validity and survival with the Emergency Severity Index. *Annals of Emergency Medicine* **49**(2) 234–240.

Bavafa, Hessam, Lorin Hitt, Christian Terwiesch. 2013. Patient portals in primary care: Impacts on patient health and physician productivity. *Working Paper* .

Bazzoli, Gloria, Stephen Shortell, Nicole Dubbs, Cheeling Chan, Peter Kralovec. 1999. A taxonomy of health networks and systems: bringing order out of chaos. *Health Services Research* **33**(6) 1683.

Ben-Assuli, Ofir, Moshe Leshno, Itamar Shabtai. 2012. Using electronic medical record systems for admission decisions in emergency departments: examining the crowdedness effect. *Journal of Medical Systems* **36**(6) 3795–3803.

Bendoly, Elliot, Karen Donohue, Kenneth Schultz. 2006. Behavior in operations management: Assessing recent findings and revisiting old assumptions. *Journal of Operations Management* **24**(6) 737–752.

Berenson, Robert, Paul Ginsburg, Jessica May. 2007. Hospital-physicians relations: cooperation, competition, or separation? *Health Affairs* **26**(1) w31–w43.

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

Bertsimas, Dimitris, Sanne De Boer. 2005. Simulation-based booking limits for airline revenue management. *Operations Research* **53**(1) 90–106.

Berwick, Donald, Andrew Hackbarth. 2012. Eliminating waste in US health care. *JAMA* **307**(14) 1513–1516.

Bhargava, Hemant, Abhay Mishra. 2014. Electronic medical records and physician productivity: Evidence from panel data analysis. *Management Science* **60**(10) 2543–2562.

Blei, David, Andrew Ng, Michael Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* **3** 993–1022.

Blischke, Wallace R, DNP Murthy. 1992. Product warranty managementi: A taxonomy for warranty policies. *European Journal of Operational Research* **62**(2) 127–148.

Bloom, Nicholas, Raffaella Sadun, John Van Reenen. 2013. Does management matter in healthcare. Working paper, London School of Economics.

Blumenthal, David. 2010. Launching HITECH. *New England Journal of Medicine* **362**(5) 382–385.

Blumenthal, David, Marilyn Tavenner. 2010. The Meaningful Use regulation for electronic health records. *New England Journal of Medicine* **363**(6) 501–504.

Boadway, Robin, Maurice Marchand, Motohiro Sato. 2004. An optimal contract approach to hospital financing. *Journal of Health Economics* **23**(1) 85–110.

Bodenheimer, Thomas. 2008. Coordinating care-a perilous journey through the health care system. *New England Journal of Medicine* **358**(10) 1064.

Bolton, Patrick, Mathias Dewatripont. 2005. *Contract theory*. MIT Press.

Borzekowski, Ron. 2009. Measuring the cost impact of hospital information systems: 1987–1994. *Journal of Health Economics* **28**(5) 938–949.

Bozarth, Cecil, Donald Warsing, Barbara Flynn, James Flynn. 2009. The impact of supply chain complexity on manufacturing plant performance. *Journal of Operations Management* **27**(1) 78–93.

Bozic, Kevin, Lorrayne Ward, Thomas Vail, Mervyn Maze. 2014. Bundled payments in total joint arthroplasty: targeting opportunities for quality improvement and cost reduction. *Clinical Orthopaedics and Related Research* **472**(1) 188–193.

Brill, Joel, Rajeev Jain, Peter Margolis, Lawrence Kosinski, Worthe Holt, Scott Ketover, Lawrence Kim, Laura Clote, John Allen. 2014. A bundled payment framework for colonoscopy performed for colorectal cancer screening or surveillance. *Gastroenterology* **146**(3) 849–853.

Brumelle, Shelby L, Jeffrey I McGill. 1993. Airline seat allocation with multiple nested fare classes. *Operations research* **41**(1) 127–137.

Budetti, Peter P, Stephen M Shortell, Teresa M Waters, Jeffrey A Alexander, Lawton R Burns, Robin R Gillies, Howard Zuckerman. 2002. Physician and health system integration. *Health Affairs* **21**(1) 203–210.

Buntin, Melinda Beeuwkes, Sachin Jain, David Blumenthal. 2010. Health information technology: laying the infrastructure for national health reform. *Health Affairs* **29**(6) 1214–1219.

Burns, Lawton, Jeffrey Alexander, Stephen Shortell, Howard Zuckerman, Peter Budetti, Robin Gillies, Teresa Waters. 2001. Physician commitment to organized delivery systems. *Medical Care* **39**(7) I–9.

Burns, Lawton Robert, Ralph Muller. 2008. Hospital-physician collaboration: Landscape of economic integration and impact on clinical integration. *Milbank Quarterly* **86**(3) 375–434.

Burt, Catharine, Linda McCaig, Roberto Valverde. 2006. Analysis of ambulance transports and diversions among us emergency departments. *Annals of emergency medicine* **47**(4) 317–326.

CAEP. 2009. Position statement on emergency department overcrowding. Position statement, Canadian Association of Emergency Physicians.

Campbell, Donald. 1988. Task complexity: A review and analysis. *Academy of Management Review* **13**(1) 40–52.

Campion, Thomas, Alison Edwards, Stephen Johnson, Rainu Kaushal. 2013. Health information exchange system usage patterns in three communities: Practice sites, users, patients, and data. *International Journal of Medical Informatics* **82**(9) 810–820.

Carr, Brendan, Adam Kaye, Douglas Wiebe, Vicente Gracias, William Schwab, Patrick Reilly. 2007. Emergency department length of stay: a major risk factor for pneumonia in intubated blunt trauma patients. *Journal of Trauma and Acute Care Surgery* **63**(1) 9–12.

Casalino, Lawrence, Elizabeth November, Robert Berenson, Hoangmai Pham. 2008. Hospital-physician relations: two tracks and the decline of the voluntary medical staff model. *Health Affairs* **27**(5) 1305–1314.

Chalfin, Donald, Stephen Trzeciak, Antonios Likourezos, Brigitte Baumann, Phillip Dellinger. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* **35**(6) 1477–1483.

Chan, Carri, Linda Green. 2013. Improving access to healthcare: Models of adaptive behavior. *Handbook of Healthcare Operations Management*. Springer, 1–18.

Chandra, Amitabh, David Cutler, Zirui Song, et al. 2012. Who ordered that? the economics of treatment choices in medical care. Mark Pauly, Thomas McGuire, Pedro Barros, eds., *Handbook of health economics*, vol. 2. Elsevier, 397–432.

Chandra, Amitabh, Jonathan Gruber, Robin McKnight. 2011. The importance of the individual mandate – evidence from Massachusetts. *New England Journal of Medicine* **364**(4) 293–295.

Chang, Tom, Mireille Jacobson. 2012. What do nonprofit hospitals maximize? Evidence from Californias seismic retrofit mandate.

Charles, Dustin, Meghan Gabriel, Talisha Searcy. 2015. Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008-2014. Onc data brief, The Office of the National Coordinator for Health Information Technology.

Charlson, Mary, Peter Pompei, Kathy Ales, Ronald MacKenzie. 1987. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases* **40**(5) 373–383.

Chen, Jinfa, David D Yao, Shaohui Zheng. 1998. Quality control for products supplied with warranty. *Operations Research* **46**(1) 107–115.

CMS. 2014. Bundled Payments for Care Improvement initiative (BPCI): Background on Model 2 for prospective participants. Tech. rep., CMS. Available from `https://innovation.cms.gov/Files/x/BPCI_Model2Background.pdf`.

CMS. 2016. Meaningful Use data: Public use files. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/PUF.html.

Connelly, Donald, Young-Taek Park, Jing Du, Nawanan Theera-Ampornpunt, Bradley Gordon, Barry Bershow, Raymond Gensinger, Michael Shrift, Daniel Routhe, Stuart Speedie. 2012.

The impact of electronic health records on care of heart failure patients in the emergency room. *Journal of the American Medical Informatics Association* **19**(3) 334–340.

Crainich, David, Hervé Leleu, Ana Mauleon. 2008. The optimality of hospital financing system: the role of physician–manager interactions. *International Journal of Health Care Finance and Economics* **8**(4) 245–256.

Curran, Denise, Jo Browning, Andrew Bryett, Catherine Love, Karen McConochie, Jackie Nankervis, Kristina O'Dwyer. 2005. A toolkit for developing a clinical pathway. Tech. rep., Queensland Health.

Custer, William, James Moser, Robert Musacchio, Richard Willke. 1990. The production of health care services and changing hospital reimbursement: the role of hospital–medical staff relationships. *Journal of Health Economics* **9**(2) 167–192.

Cutler, Joshua, Susan Seekins. 2014. Episode of care payments: A new payment model for specialty care services. Slide pack, Maine Heart Center. Obtained from `https://www.mainequalitycounts.org/image_upload/Cutler&Seekins%20MHC-BPCI%20QC'14.pdf`.

Dai, JG, Pengyi Shi. 2019. Inpatient overflow: An approximate dynamic programming approach. *Manufacturing & Service Operations Management* **21**(4) 894–911.

Dai, Jim, Pengyi Shi. 2017. A two-time-scale approach to time-varying queues in hospital inpatient flow management. *Operations Research* **65**(2) 514–536.

Daly, Rich. 2016. Physician recruitment competition spreads to urban areas: Analysis. *Healthcare Financial Management Association News* .

Davis, Karen, Kristof Stremikis, David Squires, Cathy Schoen. 2014. Mirror, mirror on the wall: How the performance of the U.S. health care system compares internationally. Tech. rep., Commonwealth Fund.

De Bleser, Leentje, Roeland Depreitere, Katrijn de Waele, Kris Vanhaecht, Joan Vlayen, Walter Sermeus. 2006. Defining pathways. *Journal of Nursing Management* **14**(7) 553–563.

De Brantes, F., D. W. Emery, J. M. Overhage, J. Glaser, J. Marchibroda. 2006. The potential of HIEs as infomediaries. *Journal of Healthcare Information Management* **21**(1) 69–75.

Deglise-Hawkinson, Jivan, Jonathan E Helm, Todd Huschka, David L Kaufman, Mark P Van Oyen. 2018. A capacity allocation planning model for integrated care and access management. *Production and Operations Management* **27**(12) 2270–2290.

DeLia, Derek, Joel Cantor. 2009. Emergency department utilization and capacity. The synthesis project, Robert Wood Johnson Foundation.

DeSalvo, Karen, Ahmed Haque. 2015. HHS and ONC invest $28 million in health information exchange grants. Online; accessed 3/31/2016. URL `http://www.healthit.gov/buzz-blog/from-the-onc-desk/health-information-exchange-grants/`.

DesRoches, Catherine, Eric Campbell, Christine Vogeli, Jie Zheng, Sowmya Rao, Alexandra Shields, Karen Donelan, Sara Rosenbaum, Steffanie Bristol, Ashish Jha. 2010. Electronic health records' limited successes suggest more targeted uses. *Health Affairs* **29**(4) 639–646.

Devaraj, Sarv, Rajiv Kohli. 2003. Performance impacts of information technology: Is actual usage the missing link? *Management Science* **49**(3) 273–289.

Ding, Ru, Melissa McCarthy, Jeffrey Desmond, Jennifer Lee, Dominik Aronsky, Scott Zeger. 2010. Characterizing waiting room time, treatment time, and boarding time in the emergency department using quantile regression. *Academic Emergency Medicine* **17**(8) 813–823.

Dobrzykowski, David, Monideepa Tarafdar. 2015. Understanding information exchange in healthcare operations: evidence from hospitals and patients. *Journal of Operations Management* URL http://dx.doi.org/10.1016/j.jom.2014.12.003.

Dobson, Allen, Joan DaVanzo, Steven Heath, Matthew Shimer, Gregory Berger, Anne Pick, Kevin Reuter, Audrey El-Gamil, Nikolay Manolov. 2012. Medicare payment bundling: Insights from claims data and policy implications. Report for aha and aamc, Dobson DaVanzo & Associates, LLC.

Dobson, Gregory, Hsiao-Hui Lee, Edieal Pinker. 2010. A model of ICU bumping. *Operations research* **58**(6) 1564–1576.

Dong, Lynn, Kate Fitch, Bruce Pyenson, Kathryn Rains-McNally. 2011. Evaluating bundled payment contracting. Tech. rep., Milliman.

Dor, Avi, Harry Watson. 1995. The hospital-physician interaction in US hospitals: evolving payment schemes and their incentives. *European Economic Review* **39**(3) 795–802.

Doyle, Joseph, John Graves, Jonathan Gruber. 2015. Uncovering waste in us healthcare. Nber working paper series, National Bureau of Economic Research.

Dranove, David, Chris Forman, Avi Goldfarb, Shane Greenstein. 2014. The trillion dollar conundrum: Complementarities and health information technology. *American Economic Journal: Economic Policy* **6**(4) 239–70.

Dranove, David, Zhe Jin. 2010. Quality disclosure and certification: theory and practice. *Journal of Economic Literature* 935–963.

Dummit, Laura, Grecia Marrufo, Jaclyn Marshall, Aylin Bradley, Laura Smith, Cornelia Hall, Youn Lee, Jon Kelly, Megan Hyland, Rebecca Cherry, Adaeze Akamigbo, Court Melin, Ellen Tan. 2015. Cms bundled payments for care improvement (bpci) initiative models 2-4: Year 1 evaluation & monitoring annual report. Tech. rep., The Lewin Group. Available from https://innovation.cms.gov/Files/reports/BPCI-EvalRpt1.pdf.

Dykstra, Richard, Joan Ash, Emily Campbell, Dean Sittig, Ken Guappone, James Carpenter, Joshua Richardson, Adam Wright, Carmit McMullen. 2009. Persistent paper: the myth of going paperless. *AMIA Annual Symposium Proceedings*, vol. 2009. American Medical Informatics Association, 158.

Edlow, Jonathan, Peter Panagos, Steven Godwin, Tamara Thomas, Wyatt Decker. 2009. Clinical policy: critical issues in the evaluation and management of adult patients presenting to the emergency department with acute headache. *Journal of Emergency Nursing* **35**(3) e43–e71.

Ellis, Randall, Thomas McGuire. 1986. Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of Health Economics* **5**(2) 129–151.

Emerman, Charles. 2012. National reporting of emergency department length of stay: Challenges, opportunities, and risks. *The Journal of the American Medical Association* **307**(5) 511–512.

Fee, Christopher, Helen Burstin, Judith Maselli, Renee Hsia. 2012. Association of emergency department length of stay with safety-net status. *The Journal of the American Medical Association* **307**(5) 476–482.

Finnell, John, Marc Overhage, Paul Dexter, Susan Perkins, Kathleen Lane, Clement McDonald. 2003. Community clinical data exchange for emergency medicine patients. *AMIA Annual Symposium Proceedings*, vol. 2003. American Medical Informatics Association, 235.

Fisher, Elliott, David Wennberg, Therese Stukel, Daniel Gottlieb, F. L. Lucas, Étoile Pinder. 2003. The implications of regional variations in Medicare spending. Part 1: The content, quality, and accessibility of care. *Annals of Internal Medicine* **138**(4) 273–287.

for Medicare & Medicaid Services, Centers. 2015. Medicare and Medicaid programs; electronic health record incentive program–stage 3 and modifications to meaningful use in 2015 through 2017. *Federal register* **80**(200) 62761.

Franczak, Michael, Madeline Klein, Flavius Raslau, Jo Bergholte, Leighton Mark, John Ulmer. 2014. In emergency departments, radiologists access to EHRs may influence interpretations and medical management. *Health Affairs* **33**(5) 800–806.

Friedberg, Mark, Peggy Chen, Chapin White, Olivia Jung, Laura Raaen, Samuel Hirshman, Emily Hoch, Clare Stevens, Paul Ginsburg, Lawrence Casalino, Michael Tutty, Carol Vargo, Lisa Lipinski. 2015. Effects of health care payment models on physician practice in the United States. Report, RAND and American Medical Association.

Frisse, Mark, Kevin Johnson, Hui Nian, Coda Davison, Cynthia Gadd, Kim Unertl, Pat Turri, Qingxia Chen. 2012. The financial impact of health information exchange on emergency department care. *Journal of the American Medical Informatics Association* **19**(3) 328–333.

Froimson, Mark, Adam Rana, Richard White, Amanda Marshall, Steve Schutzer, William Healy, Peggy Naas, Gail Daubert, Richard Iorio, Brian Parsley. 2013. Bundled payments for care improvement initiative: the next evolution of payment formulations: AAHKS bundled payment task force. *The Journal of Arthroplasty* **28**(8) 157–165.

Fuloria, Prashant, Stefanos Zenios. 2001. Outcomes-adjusted reimbursement in a health-care delivery system. *Management Science* **47**(6) 735–751.

Gaynor, Martin, Robert Town. 2012. The impact of hospital consolidation–update. The synthesis project, Robert Wood Johnson Foundation.

Gilboy, Nicki, Paula Tanabe, Debbie Travers, Alexander Rosenau. 2012. Emergency severity index (esi): A triage tool for emergency department care, version 4. implementation handbook 2012 edition. Implementation Handbook 12-0014, AHRQ.

Goh, Jie Mein, Guodong Gao, Ritu Agarwal. 2011. Evolving work routines: Adaptive routinization of information technology in healthcare. *Information Systems Research* **22**(3) 565–585.

Goldsmith, Jeff, Nathan Kaufman, Lawton Burns. 2016. The tangled hospital-physician relationship. `http://healthaffairs.org/blog/2016/05/09/the-tangled-hospital-physician-relationship/`.

Goldzweig, Caroline Lubick, Ali Towfigh, Margaret Maglione, Paul Shekelle. 2009. Costs and benefits of health information technology: New trends from the literature. *Health Affairs* **28**(2) w282–w293.

Goodhue, Dale. 1995. Understanding user evaluations of information systems. *Management Science* **41**(12) 1827–1844.

Goulding, Lucy, Joy Adamson, Ian Watt, John Wright. 2015. Lost in hospital: a qualitative interview study that explores the perceptions of NHS inpatients who spent time on clinically inappropriate hospital wards. *Health Expectations* **18**(5) 982–994.

Grady, D., R. F. Redberg. 2010. Less is more: how less health care can result in better health. *Archives of Internal Medicine* **170**(9) 749–750.

Grandusky, Rosemary, Kathy Kronenberg. 2006. Hospital-physician gainsharing. *Trustee Magazine* **59**(3).

Griffin, Jacqueline, Pinar Keskinocak, Cheryl Stokes, Nikki O'Hara, Atul Vats. 2012. Development of patient-bed assignment algorithms to support bed management processes for improvements in the rate of overflow assignments and average request to assign metrics. *Critical Care Medicine* **40**(12) U48–U48.

Griswold, Sharon, Carla Nordstrom, Sunday Clark, Theodore Gaeta, Michelle Price, Carlos Camargo. 2005. Asthma exacerbations in North American adults: Who are the frequent fliers in the emergency department? *Chest Journal* **127**(5) 1579–1586.

Grover, Atul, Peter Slavin, Peters Willson. 2014. The economics of academic medical centers. *New England Journal of Medicine* **370**(25) 2360–2362.

Gupta, Diwakar. 2013. Queueing models for healthcare operations. *Handbook of Healthcare Operations Management*. Springer, 19–44.

Gupta, Diwakar, Mili Mehrotra. 2014. Bundled payments for healthcare services: Proposer selection and information sharing. *Available at SSRN 2386124* .

Hafızoğlu, A Baykal, Esma S Gel, Pınar Keskinocak. 2016. Price and lead time quotation for contract and spot customers. *Operations Research* **64**(2) 406–415.

Halamka, John. 2013. Health information exchange for emergency department care is on the right trajectory. *Annals of Emergency Medicine* **62**(1) 25–27.

Han, Zheng, Mazhar Arikan, Suman Mallik. 2017. Can bundled payment cure the ills of fee-for-service? An equilibrium analysis. *Extended abstract* .

Handel, Daniel, Joshua Hilton, Michael Ward, Elaine Rabin, Frank Zwemer, Jesse Pines. 2010. Emergency department throughput, crowding, and financial outcomes for hospitals. *Academic Emergency Medicine* **17**(8) 840–847.

Haque, Rezwan. 2014. Technological innovation and productivity in service delivery: Evidence from the adoption of electronic medical records. Job market paper, Harvard Business School.

Harris, Jeffrey. 1977. The internal organization of hospitals: some economic implications. *The Bell Journal of Economics* 467–482.

HCUP. 2014. Description of data elements. Available at `http://www.hcup-us.ahrq.gov/db/vars/hosp_teach/kidnote.jsp`. Last accessed 3/30/2015.

Hebel, Esteban, Blackford Middleton, Maria Shubina, Alexander Turchin. 2012. Bridging the chasm: effect of health information exchange on volume of laboratory testing. *Archives of Internal Medicine* **172**(6) 517–519.

Helm, Jonathan, Shervin AhmadBeygi, Mark Van Oyen. 2011a. Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management* **20**(3) 359–374.

Helm, Jonathan E, Shervin AhmadBeygi, Mark P Van Oyen. 2011b. Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management* **20**(3) 359–374.

Herring, Andrew, Andrew Wilper, David Himmelstein, Steffie Woolhandler, Janice Espinola, David Brown, Carlos Camargo. 2009. Increasing length of stay among adult visits to U.S. emergency departments, 20012005. *Academic Emergency Medicine* **16**(7) 609–616.

Himmelstein, David, Adam Wright, Steffie Woolhandler. 2010. Hospital computing and the costs and quality of care: a national study. *The American Journal of Medicine* **123**(1) 40–46.

HIMSS. 2010. HIMSS Analytics database. http://www.himssanalytics.org/home/index.aspx. The Dorenfest Institute for H.I.T. Research and Education, HIMSS Foundation, Chicago, Illinois.

Hincapie, Ana Lucia, Terri Warholak, Anita Murcko, Marion Slack, Daniel Malone. 2011. Physicians' opinions of a health information exchange. *Journal of the American Medical Informatics Association* **18**(1) 60–65.

Hodgetts, Timothy J, Gary Kenward, Ioannis Vlackonikolis, Susan Payne, Nicolas Castle, Robert Crouch, Neil Ineson, Loua Shaikh. 2002. Incidence, location and reasons for avoidable in-hospital cardiac arrest in a district general hospital. *Resuscitation* **54**(2) 115–123.

Hoot, Nathan, Dominik Aronsky. 2008. Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine* **52**(2) 126–136.e1.

Horwitz, Leora, Jeremy Green, Elizabeth Bradley. 2010. Us emergency department performance on wait time and length of visit. *Annals of emergency medicine* **55**(2) 133–141.

Huang, Sean Sheng-Hsiu, Ian McCarthy. 2015. Hospital alignment with physicians as a bargaining response to commercial insurance markets. *Working Paper* Available at SSRN 2612879.

Hussey, Peter, Andrew Mulcahy, Christopher Schnyer, Eric Schneider. 2012. Bundled payment: Effects on health care spending and quality. Evidence report/technology assessment no. 208, Agency for Healthcare Research and Quality, Rockville, MD.

Hwang, Ula, John Concato. 2004. Care in the emergency department: how crowded is overcrowded? *Academic Emergency Medicine* **11**(10) 1097–1101.

Hydari, Muhammad Zia, Rahul Telang, William Marella. 2014. Saving patient ryancan advanced electronic medical records make patient care safer? *Available at SSRN 2503702* .

Institute of Medicine. 2006a. *Emergency care for children: growing pains*. National Academies Press.

Institute of Medicine. 2006b. *Emergency Medical Services At the Crossroads*. The National Academies Press.

Institute of Medicine. 2006c. *Hospital-based emergency care: At the breaking point*. The National Academies Press.

Jacob, Julie. 2015. On the road to interoperability, public and private organizations work to connect health care data. *Journal of American Medical Association* **314**(12) 1213–1215.

Jelovac, Izabela, Inés Macho-Stadler. 2002. Comparing organizational structures in health services. *Journal of Economic Behavior & Organization* **49**(4) 501–522.

Jha, Ashish, Catherine DesRoches, Eric Campbell, Karen Donelan, Sowmya Rao, Timothy Ferris, Alexandra Shields, Sara Rosenbaum, David Blumenthal. 2009. Use of electronic health records in U.S. hospitals. *New England Journal of Medicine* **360**(16) 1628–1638.

Jiang, Houyuan, Zhan Pang, Sergei Savin. 2012. Performance-based contracts for outpatient medical services. *Manufacturing & Service Operations Management* **14**(4) 654–669.

Johnson, Kevin B, Cynthia S Gadd, Dominik Aronsky, Kevin Yang, Lianhong Tang, Vicki Estrin, Janet K King, Mark Frisse. 2008. The MidSouth eHealth Alliance: use and impact in the first year. *AMIA Annual Symposium Proceedings*, vol. 2008. American Medical Informatics Association, 333.

Jones, Spencer, Mark Friedberg, Eric Schneider. 2011. Health information exchange, health information technology use, and hospital readmission rates. *AMIA Annual Symposium Proceedings*, vol. 2011. American Medical Informatics Association, 644.

Jones, Spencer, Robert Rudin, Tanja Perry, Paul Shekelle. 2014. Health information technology: An updated systematic review with a focus on Meaningful Use. *Annals of Internal Medicine* **160**(1) 48–54.

Karaca, Zeynal, Herbert Wong. 2013. Racial disparity in duration of patient visits to the emergency department: Teaching versus non-teaching hospitals. *Western Journal of Emergency Medicine* **14**(5) 529.

Karaca, Zeynal, Herbert Wong, Ryan Mutter. 2012. Duration of patients' visits to the hospital emergency department. *BMC Emergency Medicine* **12**(1) 15.

Karaesmen, Itir, Garrett Van Ryzin. 2004. Overbooking with substitutable inventory classes. *Operations Research* **52**(1) 83–104.

Kc, Diwas. 2014. Does multitasking improve performance? Evidence from the emergency department. *Manufacturing and Service Operations Management* **16**(2) 168–183.

Kennebeck, Stephanie Spellman, Nathan Timm, Michael Farrell, Andrew Spooner. 2012a. Impact of electronic health record implementation on patient flow metrics in a pediatric emergency department. *Journal of the American Medical Informatics Association* **19**(3) 443–447.

Kennebeck, Stephanie Spellman, Nathan Timm, Michael Farrell, Andrew Spooner. 2012b. Impact of electronic health record implementation on patient flow metrics in a pediatric emergency department. *Journal of the American Medical Informatics Association* **19**(3) 443–447.

Keskinocak, Pinar, Sridhar Tayur. 2004. Due date management policies. *Handbook of quantitative supply chain analysis*. Springer, 485–554.

Kilinc, Derya, Soroush Saghafian, Stephen Traub. 2019. Dynamic assignment of patients to primary and secondary inpatient units: Is patience a virtue. *Under review* .

Koenig, Lane, Allen Dobson, Silver Ho, Jonathan Siegel, David Blumenthal, Joel Weissman. 2003. Estimating the mission-related costs of teaching hospitals. *Health Affairs* **22**(6) 112–122.

Kohli, Rajiv, Sarv Devaraj. 2003. Measuring information technology payoff: A meta-analysis of structural variables in firm-level empirical research. *Information Systems Research* **14**(2) 127–145.

Kolstad, Jonathan. 2013. Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *American Economic Review* **103**(7) 2875–2910.

Komajda, M, F Follath, K obot Swedberg, J Cleland, JC Aguilar, A Cohen-Solal, R Dietz, A Gavazzi, WH Van Gilst, R Hobbs, et al. 2003. The Euroheart Failure Survey programme: a survey on the quality of care among patients with heart failure in Europe: Part 2: treatment. *European heart journal* **24**(5) 464–474.

Koole, Ger. 1995. A simple proof of the optimality of a threshold policy in a two-server queueing system. *Systems & Control Letters* **26**(5) 301–303.

Kuntz, Ludwig, Roman Mennicken, Stefan Scholtes. 2014. Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science* URL `http://dx.doi.org/10.1287/mnsc.2014.1917`. Published Online on May 19, 2014.

Kuperman, Gilad, Julie McGowan. 2013. Potential unintended consequences of health information exchange. *Journal of General Internal Medicine* **28**(12) 1663–1666.

LaCalle, Eduardo, Elaine Rabin. 2010. Frequent users of emergency departments: the myths, the data, and the policy implications. *Annals of Emergency Medicine* **56**(1) 42–48.

Lahiri, Atanu, Abraham Seidmann. 2012. Information hangovers in healthcare service systems. *Manufacturing and Service Operations Management* **14**(4) 634–653.

Lammers, E., J. Adler-Milstein, K. Kocher. 2014. Does health information exchange reduce redundant imaging? Evidence from emergency departments. *Medical Care* **52**(3) 227–234.

Lan, Yingjie, Huina Gao, Michael O Ball, Itir Karaesmen. 2008. Revenue management with limited demand information. *Management Science* **54**(9) 1594–1609.

Larkin, Gregory Luke, Cynthia Claassen, Andrea Pelletier, Carlos Camargo. 2006. National study of ambulance transports to United States emergency departments: importance of mental health problems. *Prehospital and Disaster Medicine* **21**(02) 82–90.

Le, Sidney, Renee Hsia. 2014. Timeliness of care in US emergency departments: An analysis of newly released metrics from the Centers for Medicare & Medicaid Services. *JAMA Internal Medicine* **174**(11) 1847–1849.

Lee, Donald, Stefanos Zenios. 2012. An evidence-based incentive system for Medicare's End-Stage Renal Disease Program. *Management Science* **58**(6) 1092–1105.

Lee, Jinhyung, Yong-Fang Kuo, James Goodwin. 2013a. The effect of electronic medical record adoption on outcomes in US hospitals. *BMC Health Services Research* **13**(1) 1.

Lee, Jinhyung, Jeffrey McCullough, Robert Town. 2013b. The impact of health information technology on hospital productivity. *The RAND Journal of Economics* **44**(3) 545–568.

Lee, Thomas H, Albert Bothe, Glenn D Steele. 2012. How geisinger structures its physicians compensation to support improvements in quality, efficiency, and volume. *Health Affairs* **31**(9) 2068–2073.

Leisch, Friedrich. 2004. Flexmix: A general framework for finite mixture models and latent class regression in r. *Journal of Statistical Software* **11**(1).

Lin, Mingfeng, Henry Lucas, Galit Shmueli. 2013. Research commentary-Too big to fail: large samples and the p-value problem. *Information Systems Research* **24**(4) 906–917.

Litwin, Adam, Ariel Avgar, Peter Pronovost. 2012. Measurement error in performance studies of health information technology: Lessons from the management literature. *Applied Clinical Informatics* **3**(2) 210–220.

Long, Elisa F, Kusum S Mathews. 2018. The boarding patient: Effects of ICU and hospital occupancy surges on patient flow. *Production and operations management* **27**(12) 2122–2143.

Ma, Ching-to Albert. 1994. Health care payment systems: cost and quality incentives. *Journal of Economics & Management Strategy* **3**(1) 93–112.

Ma, Will, David Simchi-Levi. 2017. Online resource allocation under arbitrary arrivals: Optimal algorithms and tight competitive ratios. *Working Paper* Available at SSRN: https://ssrn.com/abstract=2989332.

Macho-Stadler, Inés, David Pérez-Castrillo. 2020. *Complex Social and Behavioral Systems: Game Theory and Agent-Based Models*, chap. Principal-agent models. Springer, 287–303.

Mahajan, A. P. 2016. Health Information Exchange–obvious choice or pipe dream? *Journal of American Medical Association Internal Medicine* **176**(4) 429–430.

Manshadi, Vahideh H, Shayan Oveis Gharan, Amin Saberi. 2012. Online stochastic matching: Online actions based on offline statistics. *Mathematics of Operations Research* **37**(4) 559–573.

Marazzi, Alfio, Fred Paccaud, Christiane Ruffieux, Claire Beguin. 1998. Fitting the distributions of length of stay by parametric models. *Medical Care* **36**(6) 915–927.

Marx, John, Ron Walls, Robert Hockberger, eds. 2010. *Rosen's Emergency Medicine-Concepts and Clinical Practice*. 7th ed. Elsevier.

McCarthy, Melissa L, Scott L Zeger, Ru Ding, Scott R Levin, Jeffrey S Desmond, Jennifer Lee, Dominik Aronsky. 2009a. Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients. *Annals of emergency medicine* **54**(4) 492–503.

McCarthy, Melissa L, Scott L Zeger, Ru Ding, Scott R Levin, Jeffrey S Desmond, Jennifer Lee, Dominik Aronsky. 2009b. Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients. *Annals of emergency medicine* **54**(4) 492–503.

McClellan, Mark. 2011. Reforming payments to healthcare providers: The key to slowing healthcare cost growth while improving quality? *The Journal of Economic Perspectives* 69–92.

McClellan, Mark, John O'Shea, Christine Dang-Vu, Sara Bencic, Sarah Bleiberg, Judith Tobin. 2014. Specialty payment model opportunities assessment and design. Tech. rep., The Brookings Institution.

McClelland, Mark, Danielle Lazar, Vickie Sears, Marcia Wilson, Bruce Siegel, Jesse Pines. 2011. The past, present, and future of urgent matters: Lessons learned from a decade of emergency department flow improvement. *Academic Emergency Medicine* **18**(12) 1392–1399.

McCullough, Jeffrey, Stephen Parente, Robert Town. 2013. Health information technology and patient outcomes: The role of organizational and informational complementarities. Working Paper 18684, National Bureau of Economic Research.

McDonough, John, Brian Rosman, Mehreen Butt, Lindsey Tucker, Lisa Kaplan Howe. 2008. Massachusetts health reform implementation: Major progress and future challenges. *Health Affairs* **27**(4) w285–w297.

Mechanic, Robert, Stuart Altman. 2009. Payment reform options: episode payment is a good place to start. *Health Affairs* **28**(2) w262–w271.

Medicare. 2013. Measures displayed on Hospital Compare. Online; accessed 3/27 2015. URL `http://www.medicare.gov/hospitalcompare/Data/Measures-Displayed.html`.

Mehrotra, Ateev, Peter Hussey. 2015. Including physicians in bundled hospital care payments: Time to revisit an old idea? *JAMA* **313**(19) 1907–1908.

Members of the Emergency Medicine Practice Committee. 2012. Publishing wait times for emergency department care. Tech. rep., American College of Emergency Physicians.

Menon, Nirup, Rajiv Kohli. 2013. Blunting damocles sword: A longitudinal model of healthcare it impact on malpractice insurance premium and quality of patient care. *Information Systems Research* **24**(4) 918–932.

Menon, Nirup, Byungtae Lee. 2000. Cost control and production performance enhancement by IT investment and regulation changes: evidence from the healthcare industry. *Decision Support Systems* **30**(2) 153 – 169.

Menon, Nirup, Byungtae Lee, Leslie Eldenburg. 2000. Productivity of information systems in the healthcare industry. *Information Systems Research* **11**(1) 83–92.

Miller, Amalia, Catherine Tucker. 2014. Health information exchange, system size and information silos. *Journal of Health Economics* **33** 28–42.

Miller, Harold. 2015. Bundling badly: The problems with Medicares proposal for Comprehensive Care for Joint Replacement. Tech. rep., Center for Healthcare Quality and Payment Reform.

Moskop, John, David Sklar, Joel Geiderman, Raquel Schears, Kelly Bookman. 2009a. Emergency department crowding, part 1concept, causes, and moral consequences. *Annals of emergency medicine* **53**(5) 605–611.

Moskop, John C, David P Sklar, Joel M Geiderman, Raquel M Schears, Kelly J Bookman. 2009b. Emergency department crowding, part 1: Concept, causes, and moral consequences. *Annals of emergency medicine* **53**(5) 605–611.

Mukherji, Suresh, Thomas Fockler. 2014. Bundled payment. *Journal of the American College of Radiology* **11**(6) 566–571.

Murray, Scott, David Bates, Long Ngo, Jacob Ufberg, Nathan Shapiro. 2006. Charlson index is associated with one-year mortality in emergency department patients with suspected infection. *Academic Emergency Medicine* **13**(5) 530–536.

Murthy, DNP, WR Blischke. 1992a. Product warranty managementii: an integrated framework for study. *European Journal of Operational Research* **62**(3) 261–281.

Murthy, DNP, WR Blischke. 1992b. Product warranty managementiii: a review of mathematical models. *European Journal of Operational Research* **63**(1) 1–34.

Murthy, DNP, I Djamaludin. 2002. New product warranty: A literature review. *International Journal of Production Economics* **79**(3) 231–260.

National Center for Health Statistics. 2013. Health, United States, 2012: With special feature on emergency care. Report, Department of Health and Human Services, Hyattsville, MD.

National Center for Health Statistics. 2016. National hospital ambulatory medical care survey. Tech. rep., CDC.

National Quality Forum. 2009. National voluntary consensus standards for emergency care. A consensus report, NQF, Washington, DC.

Nelson, Barry. 2013. *Foundations and methods of stochastic simulation: a first course*. Springer Science & Business Media.

Obermeyer, Ziad, Ezekiel J Emanuel. 2016. Predicting the futurebig data, machine learning, and clinical medicine. *The New England Journal of Medicine* **375**(13) 1216.

Ouyang, Huiyin, Nilay Tank Argon, Serhan Ziya. 2020. Allocation of intensive care unit beds in periods of high demand. *Operations Research* **68**(2) 591–608.

Overhage, Marc, Paul Dexter, Susan Perkins, William Cordell, John McGoff, Roland McGrath, Clement McDonald. 2002. A randomized, controlled trial of clinical information shared from another institution. *Annals of Emergency Medicine* **39**(1) 14–23.

Overhage, Marc, William Tierney, Clem McDonald. 1995. Design and implementation of the indianapolis network for patient care and research. *Bulletin of the Medical Library Association* **83**(1) 48.

OMalley, Ann S, Amelia M Bond, Robert A Berenson. 2011. Rising hospital employment of physicians: better quality, higher costs. *Issue Brief Cent Stud Health Syst Change* **136**(136) 1–4.

Pallin, Daniel, Matthew Allen, Janice Espinola, Carlos Camargo, Stephen Bohan. 2013. Population aging and emergency departments: Visits will not increase, lengths-of-stay and hospitalizations will. *Health Affairs* **32**(7) 1306–1312.

Papanikolaou, P. N., G. D. Christidi, J. P. Ioannidis. 2006. Patient outcomes with teaching versus nonteaching healthcare: a systematic review. *PLoS Medicine* **3**(9) e341.

Papier, Felix, Ulrich W Thonemann. 2010. Capacity rationing in stochastic rental systems with advance demand information. *Operations research* **58**(2) 274–288.

Patel, Mitesh. 2016. Medical waste: why american health care is so expensive. URL `http://knowledge.wharton.upenn.edu/article/medical-waste-american-health-care-expensive/`. Accessed: 2017-02-07.

Pauly, Mark, Michael Redisch. 1973. The not-for-profit hospital as a physicians' cooperative. *The American Economic Review* **63**(1) 87–99.

Pines, Jesse, Sanjay Iyer, Maureen Disbot, Judd Hollander, Frances Shofer, Elizabeth Datner. 2008a. The effect of emergency department crowding on patient satisfaction for admitted patients. *Academic Emergency Medicine* **15**(9) 825–831.

Pines, Jesse, Sanjay Iyer, Maureen Disbot, Judd Hollander, Frances Shofer, Elizabeth Datner. 2008b. The effect of emergency department crowding on patient satisfaction for admitted patients. *Academic Emergency Medicine* **15**(9) 825–831.

Pines, Jesse, Anjeli Prabhu, Joshua Hilton, Judd Hollander, Elizabeth Datner. 2010. The effect of emergency department crowding on length of stay and medication treatment times in discharged patients with acute asthma. *Academic Emergency Medicine* **17**(8) 834–839.

Pitts, Stephen, Jesse Pines, Michael Handrigan, Arthur Kellermann. 2012. National trends in emergency department occupancy, 2001 to 2008: effect of inpatient admissions versus emergency department practice intensity. *Annals of Emergency Medicine* **60**(6) 679–686.

Powell, Adam, Sergei Savin, Nicos Savva. 2012. Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient. *Manufacturing and Service Operations Management* **14**(4) 512–528.

President's Council of Advisors on Science and Technology. 2010. Realizing the full potential of health information technology to improve healthcare for Americans: The path forward. Report to President, Executive Office of the President.

Proudlove, Nathan, Ruth Boaden, Julie Jorgensen. 2007. Developing bed managers: the why and the how. *Journal of nursing management* **15**(1) 34–42.

Rabin, Elaine, Keith Kocher, Mark McClelland, Jesse Pines, Ula Hwang, Niels Rathlev, Brent Asplin, Seth Trueger, Ellen Weber. 2012. Solutions to emergency department boardingand crowding are underused and may need to be legislated. *Health Affairs* **31**(8) 1757–1766.

Rahurkar, Saurabh, Joshua Vest, Nir Menachemi. 2015. Despite the spread of health information exchange, there is little evidence of its impact on cost, use, and quality of care. *Health Affairs* **34**(3) 477–483.

Reddi, BAJ, Roger Carpenter. 2000. The influence of urgency on decision time. *Nature Neuroscience* **3**(8) 827–830.

Reid, Proctor P, W Dale Compton, Jerome H Grossman, Gary Fanjiang, et al. 2005. *Building a better delivery system: a new engineering/health care partnership*. National Academies Press.

Ridgely, Susan, David De Vries, Kevin Bozic, Peter Hussey. 2014. Bundled payment fails to gain a foothold in california: The experience of the iha bundled payment demonstration. *Health Affairs* **33**(8) 1345–1352.

Risser, Daniel, Matthew Rice, Mary Salisbury, Robert Simon, Gregory Jay, Scott Berns. 1999. The potential for improved teamwork to reduce medical errors in the emergency department. *Annals of Emergency Medicine* **34**(3) 373–383.

Robinson, J. C., L. P. Casalino, R. R. Gillies, D. R. Rittenhouse, S. S. Shortell, S. Fernandes-Taylor. 2009. Financial incentives, quality improvement programs, and the adoption of clinical information technology. *Medical Care* **47**(4) 411–417.

Romeijn, H Edwin, Stefanos A Zenios. 2008. Introduction to the special issue on operations research in health care. *Operations Research* **56**(6) 1333–1334.

Rosen, Allison, Ana Aizcorbe, Alexander Ryu, Nicole Nestoriak, David Cutler, Michael Chernew. 2013. Policy makers will need a way to update bundled payments that reflects highly skewed spending growth of various care episodes. *Health Affairs* **32**(5) 944–951.

Rosenthal, Meredith, Adams Dudley. 2007. Pay-for-performance: will the latest payment trend improve care? *Journal of the American Medical Association* **297**(7) 740–744.

Ross, Stephen, Tiffany Radcliff, William LeBlanc, Miriam Dickinson, Anne Libby, Donald Nease. 2013. Effects of health information exchange adoption on ambulatory testing rates. *Journal of the American Medical Informatics Association* **20**(6) 1137–1142.

Ruben, Amarasingham, Plantinga Laura, Diener-West Marie, Gaskin Darrell, Neil Powe. 2009. Clinical information technologies and inpatient outcomes: A multiple hospital study. *Archives of Internal Medicine* **169**(2) 108–114.

Rucker, Donald, Roger Edwards, Helen Burstin, Anne O'Neil, Troyen Brennan. 1997. Patient-specific predictors of ambulance use. *Annals of Emergency Medicine* **29**(4) 484–491.

Rudin, R., A. Motala, C. Goldzweig, P. Shekelle. 2014. Usage and effect of health information exchange. *Annals of Internal Medicine* **161**(11) 803–811.

Rudin, Robert, Lynn Volk, Steven Simon, David Bates. 2011. What affects clinicians usage of health information exchange? *Applied Clinical Informatics* **2**(3) 250.

Sands, Daniel, David Rind, Cynthia Vieira, Charles Safran. 1997. Going paperless: Can it be done? *Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, 887.

Savaşaneril, Seçil, Paul M Griffin, Pınar Keskinocak. 2010. Dynamic lead-time quotation for an m/m/1 base-stock inventory queue. *Operations research* **58**(2) 383–395.

Schoen, C. 2016. *The Affordable Care Act and the US economy: a five-year perspective*. Commonwealth Fund.

Scott, Tim, Thomas Rundall, Thomas Vogt, John Hsu. 2005. Kaiser Permanente's experience of implementing an electronic medical record: a qualitative study. *BMJ* **331**(7528) 1313–1316.

Senathirajah, Mahil, Pam Owens, Ryan Mutter, Mika Nagamine. 2011. Special study on the meaning of the first-listed diagnosis on emergency department and ambulatory surgery records. HCUP methods series. report # 2011-03., Agency for Healthcare Research and Quality. Online. U.S. Agency for Healthcare Research and Quality. Available: http://www.hcup-us.ahrq.gov/reports/methods/methods.jsp.

Shapiro, Jason, Joseph Kannry, Andre Kushniruk, Gilad Kuperman. 2007. Emergency physicians perceptions of health information exchange. *Journal of the American Medical Informatics Association* **14**(6) 700–705.

Shih, Terry, Lena Chen, Brahmajee Nallamothu. 2015. Will bundled payments change health care? Examining the evidence thus far in cardiovascular care. *Circulation* **131**(24) 2151–2158.

Shortell, Stephen, Jeffery Alexander, Peter Budetti, Lawton Burns, Robin Gillies, Teresa Waters, Howard Zuckerman. 2001. Physician-system alignment: Introductory overview. *Medical Care* **39**(7) I–1.

Shumsky, Robert A, Fuqiang Zhang. 2009. Dynamic capacity management with substitution. *Operations research* **57**(3) 671–684.

Shwartz, Michael, Lisa Iezzoni, Mark Moskowitz, Arlene Ash, Eric Sawitz. 1996. The importance of comorbidities in explaining differences in patient costs. *Medical care* **34**(8) 767–782.

Shy, Bradley, Eugene Kim, Nicholas Genes, Tina Lowry, George Loo, Ula Hwang, Lynne Richardson, Jason Shapiro. 2016. Increased identification of emergency department 72-hour returns using multi-hospital health information exchange. *Academic Emergency Medicine* .

Sirovich, Brenda, Patricia Gallagher, David Wennberg, Elliott Fisher. 2008. Discretionary decision making by primary care physicians and the cost of U.S. health care. *Health Affairs* **27**(3) 813–823.

Skinner, Halcyon, Janice Blanchard, Anne Elixhauser. 2014. Overview of emergency department visits in the united states, 2011. Hcup statistical brief # 174, Agency for Healthcare Research and Quality, Rockville, MD. URL `http://www.hcup-us.ahrq.gov/reports/statbriefs/sb174-Emergency-Department-Visits-Overview.pdf`.

Sloan, Frank. 2000. *Handbook of Health Economics*, chap. Not-for-profit ownership and hospital behavior. Elsevier, 1141–1174.

Smith, Stephen A, Narendra Agrawal. 2000. Management of multi-item retail inventory systems with demand substitution. *Operations Research* **48**(1) 50–64.

Song, Hummy, Anita Tucker, Karen Murrell. 2014. The diseconomies of queue pooling: an empirical investigation of emergency department length of stay. *Working Paper* .

Song, Hummy, Anita L. Tucker, Ryan Graue, Sarah Moravick, Julius J. Yang. 2019. Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science* doi:10.1287/mnsc.2019.3395.

Sood, Neeraj, Peter J. Huckfeldt, Jos J. Escarce, David C. Grabowski, Joseph P. Newhouse. 2011. Medicares bundled payment pilot for acute and postacute care: Analysis and recommendations on where to begin. *Health Affairs* **30**(9) 1708–1717.

Speedie, Stuart, Young-Taek Park, Jing Du, Nawanan Theera-Ampornpunt, Barry Bershow, Raymon Gensinger, Daniel Routhe, Donald Connelly. 2014. The impact of electronic health records on people with diabetes in three different emergency departments. *Journal of the American Medical Informatics Association* **21**(e1) e71–e77.

Spencer, Rosemary, Enrico Coiera, Pamela Logan. 2004. Variation in communication loads on clinical staff in the emergency department. *Annals of Emergency Medicine* **44**(3) 268–273.

Spivey, Michael Z, Warren B Powell. 2004. The dynamic assignment problem. *Transportation Science* **38**(4) 399–419.

Stiell, Andrew, Alan Forster, Ian Stiell, Carl van Walraven. 2003. Prevalence of information gaps in the emergency department and the effect on patient outcomes. *Canadian Medical Association Journal* **169**(10) 1023–1028.

Stock, J. H., M. Yogo. 2002. Testing for weak instruments in linear iv regression.

Sun, Benjamin C, Renee Y Hsia, Robert E Weiss, David Zingmond, Li-Jung Liang, Weijuan Han, Heather McCreath, Steven M Asch. 2013. Effect of emergency department crowding on outcomes of admitted patients. *Annals of emergency medicine* **61**(6) 605–611.

Sutton, Richard S, Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Talluri, Kalyan T, Garrett J Van Ryzin. 2006. *The theory and practice of revenue management*, vol. 68. Springer Science & Business Media.

Tan, Tom Fangyun, Serguei Netessine. 2014. When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science* **60**(6) 1574–1593.

Tanabe, Paula, Rick Gimbel, Paul Yarnold, James Adams. 2004. The Emergency Severity Index (version 3) 5-level triage system scores predict ed resource consumption. *Journal of Emergency Nursing* **30**(1) 22–29.

Thomas Schneider, A. J., P. Luuk Besselink, Maartje E. Zonderland, Richard J. Boucherie, Wilbert B. van den Hout, Job Kievit, Paul Bilars, A. Jaap Fogteloo, Ton J. Rabelink. 2018. Allocating emergency beds improves the emergency admission flow. *Interfaces* **48**(4) 384–394.

Thompson, Steven, Manuel Nunez, Robert Garfinkel, Matthew Dean. 2009. OR Practice: efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operations Research* **57**(2) 261–273.

Thorn, Shirley, Michael Carter, James Bailey. 2013. Emergency physicians' perspectives on their use of health information exchange. *Annals of Emergency Medicine* **63**(3) 329–337.

Trzeciak, Stephen, Emanuel Rivers. 2003. Emergency department overcrowding in the United States: an emerging threat to patient safety and public health. *Emergency Medicine Journal* **20**(5) 402–405.

Tsai, Thomas, Karen Joynt, Robert Wild, John Orav, Ashish Jha. 2015. Medicare's bundled payment initiative: Most hospitals are focused on a few high-volume conditions. *Health Affairs* **34**(3) 371–380.

Tzeel, Albert, Victor Lawnicki, Kim Pemble. 2011. The business case for payer support of a community-based health information exchange: a Humana pilot evaluating its effectiveness in cost control for plan members seeking emergency department care. *American Health and Drug Benefits* **4**(4) 207–216.

Van Mieghem, Jan A. 1995. Dynamic scheduling with convex delay costs: The generalized c-mu rule. *The Annals of Applied Probability* 809–833.

Venkatesh, R, Vijay Mahajan. 2009. *Handbook of Pricing Research in Marketing*, chap. The Design and Pricing of Bundles: A Review of Normative Guidelines and Practical Approaches. Edward Elgar Publishing Company, 232–257.

Véricourt, Francis de, Otis B Jennings. 2011. Nurse staffing in medical units: A queueing perspective. *Operations Research* **59**(6) 1320–1331.

Vest, Joshua. 2009. Health information exchange and healthcare utilization. *Journal of Medical Systems* **33**(3) 223–231.

Vest, Joshua. 2010. More than just a question of technology: Factors related to hospitals adoption and implementation of health information exchange. *International Journal of Medical Informatics* **79**(12) 797–806.

Vest, Joshua, Zachary Grinspan, Lisa Kern, Thomas Campion, Rainu Kaushal. 2013. Using a health information exchange system for imaging information: patterns and predictors. *AMIA Annual Symposium Proceedings*, vol. 2013. American Medical Informatics Association, 1402.

Vest, Joshua, Rainu Kaushal, Michael Silver, Keith Hentel, Lisa Kern. 2014. Health information exchange and the frequency of repeat medical imaging. *American Journal of Managed Care* **20**(SP17) eSP16–eSP24.

Vest, Joshua, Thomas Miller. 2011. The association between health information exchange and measures of patient satisfaction. *Applied Clinical Informatics* **2**(4) 447.

Vest, Joshua, Hongwei Zhao, Jon Jasperson, Larry Gamm, Robert Ohsfeldt. 2011. Factors motivating and affecting health information exchange usage. *Journal of the American Medical Informatics Association* **18**(2) 143–149.

Vezyridis, Paraskevas, Stephen Timmons, Heather Wharrad. 2011. Going paperless at the emergency department: a socio-technical study of an information system for patient tracking. *international Journal of Medical Informatics* **80**(7) 455–465.

Walker, Jan, Eric Pan, Douglas Johnston, Julia Adler-Milstein, David Bates, Blackford Middleton. 2005. The value of health care information exchange and interoperability. *Health Affairs* .

Wasserstein, Ronald, Nicole Lazar. 2016. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician* **70**(2) 129–133.

Weber, Ellen, Suzanne Mason, Angela Carter, Ruth Hew. 2011. Emptying the corridors of shame: organizational lessons from England's 4-hour emergency throughput target. *Annals of Emergency Medicine* **57**(2) 79–88.

Weiss, Gail. 2003. Exclusive survey: Practice expenses. *Medical Economics* **80**(31).

Welch, Shari, Brent Asplin, Suzanne Stone-Griffith, Steven Davidson, James Augustine, Jeremiah Schuur. 2011. Emergency department operational metrics, measures and definitions: results of the second performance measures and benchmarking summit. *Annals of Emergency Medicine* **58**(1) 33–40.

Wennberg, John, Megan Cooper. 1996. *The Dartmouth atlas of health care*. American Hospital Publishing Chicago, Illinois.

Wholey, Douglas, Lawton Burns. 1991. Convenience and independence: Do physicians strike a balance in admitting decisions? *Journal of Health and Social Behavior* 254–272.

Wilensky, Gail, Nicholas Wolter, Michelle Fischer. 2007. Gain sharing: A good concept getting a bad name? *Health Affairs* **26**(1) w58–w67.

Wooldridge, Jeffrey. 2010. *Econometric analysis of cross section and panel data*. MIT Press.

Wright, Adam, Christine Soran, Chelsea Jenter, Lynn Volk, David Bates, Steven Simon. 2010. Physician attitudes toward health information exchange: results of a statewide survey. *Journal of the American Medical Informatics Association* **17**(1) 66–70.

Xu, Kuang, Carri W Chan. 2016. Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management* **18**(3) 314–331.

Yaraghi, N. 2015. A sustainable business model for health information exchange platforms: the solution to interoperability in healthcare IT. *The Brookings Institution* .

Yaraghi, Niam, Anna Ye Du, Raj Sharman, Ram Gopal, Ram Ramesh. 2015. Health information exchange as a multisided platform: Adoption, usage, and practice involvement in service co-production. *Information Systems Research* **26**(1) 1–18.

Zhan, Dongyuan, Amy R Ward. 2013. Threshold routing to trade off waiting and call resolution in call centers. *Manufacturing & Service Operations Management* **16**(2) 220–237.

Zhang, Dennis, Eric Park, Itai Gurvich, Jan Van Mieghem, Robert Young, Mark Williams. 2016. Hospital readmissions reduction program: An economic and operational analysis. *Forthcoming in Management Science* .