NORTHWESTERN UNIVERSITY

An Operations Management Approach to Evaluating Health Policy Interventions

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Operations Management

By

Eric Park

EVANSTON, ILLINOIS

December 2014

ABSTRACT

An Operations Management Approach to Evaluating Health Policy Interventions

Eric Park

This dissertation analyzes health policy interventions from an operations management perspective. To improve policy effectiveness, policymakers must have a good understanding of the health delivery process and the policy implications. We study how health policies connect to care delivery operations and why policies may not achieve the intended goal.

In Chapter 1, we study a county-wide ambulance routing policy intervention intended to reduce fraction of ambulances diverted by restricting the duration and frequency of emergency department (ED) ambulance diversion episodes. The econometric approach based on the framework of strategic communication between service providers (EDs) and potential customers (paramedics) identifies that EDs responded to the restrictions by improving their patient flow processes and paramedics responded to this change by reducing their level of compliance to the diversion thereby reducing the level of diversion. This effect was counteracted by an increase in the level of diversion because fewer EDs were simultaneously on diversion post-intervention. These mechanisms together explain the

discrepancy between 65% reduction in ED time on diversion and 10% reduction in fraction of ambulances diverted. The result encourages policymakers that a relatively light intervention can induce EDs to improve their processes. However, such intervention might not reduce diversions due to the complex interaction between EDs and paramedics.

In Chapter 2, we study the effectiveness of the Affordable Care Acts Hospital Readmissions Reduction Program in motivating hospitals to reduce excess readmissions by financially penalizing excess readmissions. We link readmissions performance to hospital financials through a hospital admissions model and identify the opposing effects of reducing readmissions on hospital net income: profit loss due to reduced patient volume and savings by avoiding penalty. The model predicts a window of readmission rate unique to each hospital where it is financially incentivized to reduce readmissions. However, not all hospitals are incentivized, because, either the readmission rate is outside of the window or a window does not exist. Reducing the Floor Adjustment Factor will widen the window and incentivize the former but not the later. Increasing the number of penalty applied conditions will create a window for the later and widen it for the former.

Acknowledgements

Within your life and your career, there arent many opportunities to express your gratitude for the people around you in a formal manner. I am pleased for such opportunity.

I thank my advisors Professors Sarang Deo and Itai Gurvich for providing the training towards an academic researcher. There is still a long way to go but I have been well prepared by their guidance. Especially, Prof. Deo for introducing the field of health care and empirical research which I would have not imagined to study before, and Prof. Gurvich for recruiting me to Kellogg. Though I have not pursued his recommended project, I am thankful for him offering the opportunity as I have learned how hard it is to ask good research questions.

I am thankful to Professors Achal Bassamboo and Antonio Moreno-Garcia for being on my dissertation committee and providing valuable feedback on my research. Professor Moreno-Garcia has always guided me towards the right direction when I encountered technical issues in my research.

I want to thank Professor Jan A. Van Mieghem for providing the opportunity and guidance on our joint research project with our colleagues formerly at the Northwestern Memorial Hospital. Working as a team on an academic research project from two different fields is even more challenging but can be much more rewarding. I also thank Professor Baris Ata for admitting a less than qualified candidate in myself to our PhD program and guiding through the tough first year and providing all possible help to pass the qualifiers.

I am appreciative of my fellow colleagues in the Kellogg operations PhD program. I have learned a lot from those before me: Tingliang, Seyed, Jingqi (special thanks for the econ discussions), my classmates: Seung bum and Ruomeng, and those behind me: Xiaoxian, William, Ahmet, Evan, Lu, Yonnis, Can, Dennis, Serasu, Pantelis, Keija, Zhiji. I also thank other Kellogg PhDs, Mark and TJ for the many hours we spent together on tackling our first year assignments and sharing our emotions going through the process of a PhD student.

Lastly, I am most thankful for my family. My wife Min Jung, who I have married during my doctoral studies, has always been the biggest supporter of me. She has sacrificed so much for being together with me and I cannot ask anything more. I am so graceful for having our first child, Lillian, during my degree. I love you both so much.

It has been an unbelievable journey for me and I am proud to put the title Dr. in front of my name and will make the most out of the opportunity that I am fortunate to have. Thank you.

Sep 26, 2014

On the flight back to my family

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Does Limiting Time on Ambulance Diversion Reduce Diversions? Signaling Equilibrium and Network Effect (joint with Sarang Deo and Itai Gurvich)

1.1. Introduction

Ambulance diversion is a phenomenon wherein overcrowded emergency departments (EDs) request the Emergency Medical Services (EMS) agency to be put on an *on-diversion* status for a specified period of time during which ambulance crews are advised to take patients to neighboring EDs. Ambulance diversion has been associated with longer ambulance transport times (Schull et al. 2003) and worse health outcomes (Shenoi et al. 2009) including higher mortality (Shen and Hsia 2011). Pham et al. (2006) provide a detailed review of various consequences of ambulance diversion. However, a growing body of evidence suggests that diversion has limited ability to reduce ED crowding as the latter is predominantly caused by inefficient patient-flow processes such as boarding of patients on ED beds while waiting to be transferred to an inpatient bed (Allon et al. 2013, McConnell et al. 2005, Hoot and Aronsky 2008). Accumulated evidence also points to a *network effect* of ambulance diversion: the likelihood of an ambulance being diverted at one ED is affected by diversion statuses of its neighboring EDs (McCarthy et al. 2007), diversion hours at neighboring EDs are correlated (Sun et al. 2006) and EDs in sparse networks (with fewer neighboring EDs) spend fewer hours on diversion (Allon et al. 2013).

Consequently, EMS agencies in several communities across the US have adopted a network-wide approach to reduce ambulance diversion by restricting how often and for how long EDs can be in an on-diversion status. Such policy interventions have reported substantial reduction (60%–80%) in time spent by EDs on diversion (Lagoe et al. 2003, Vilke et al. 2004, Patel et al. 2006, Asamoah et al. 2008), an extreme example being that of 100% reduction in Massachusetts following a complete state-wide ban on ambulance diversion (Burke et al. 2013). Yet, empirical evidence regarding the impact of such interventions on other operational measures is mixed. For instance, some interventions report a reduction in the time required to offload ambulance patients (Burke et al. 2013) while others report an increase (Asamoah et al. 2008). Also, one of the implicit objectives of these interventions is to stimulate improvement in internal patient flow process at the EDs. Indeed, several interventions are accompanied by optional best-practice recommendations to participating EDs with regards to patient-flow management (Burke et al. 2013, Vilke et al. 2004). Objective data on whether EDs actually implement these recommendations is, however, difficult to obtain (Burke et al. 2013).

In this paper, we adopt an empirical approach using detailed operational data on ambulance transports in a network of EDs toward three objectives: (i) to uncover underlying mechanisms of ambulance diversion, (ii) to understand the role of these mechanisms in determining the impact of community-wide policy interventions on operational measures (e.g. time on diversion, probability of diverting an ambulance, ambulance waiting time), and (iii) to infer whether the policy intervention succeeded in stimulating EDs to improve their patient flow processes. Next, we describe the conceptual model underlying our approach, which is based on the premise that diversion is a signal of crowding at the ED.

1.1.1. Conceptual model of diversion status as a signal of ED crowding

Our conceptual model builds on recent analytical work in operations management (Allon and Bassamboo 2009) that studies a service provider's decision to communicate the level of congestion in the system and the arriving customers' response to this communication through their joining behavior. Of particular relevance is Allon et al. (2011), which characterizes a non-cooperative equilibrium wherein the service provider uses a binary signal to communicate the state of the system. The two signals can be interpreted as "busy" (e.g. on-diversion) when the queue exceeds a certain threshold and "not busy" (e.g. off-diversion) otherwise. In equilibrium, customers follow a mixed strategy of whether to join the queue or not given the signal. Under some conditions, it is shown that this signal is informative in that the joining probability is lower when the signal is busy compared to when it is "not busy".

In line with the above model, we conceptualize that the ED chooses a threshold on the level of crowding (e.g. number of patients in the ED including in service and waiting) and uses an on-diversion signal (red light) when this threshold is exceeded and an off-diversion signal (green light) otherwise (Allon et al. 2013, Deo and Gurvich 2011, Do and Shunko 2013, Ramirez-Nafarrate et al. 2013). The ambulance crew does not know the exact level of crowding in the ED at the time of determining the destination of the patient. However, over time ("in equilibrium"), they may form an estimate of the crowding level associated with the on- and off-diversion signals. They evaluate the impact of the estimated level of

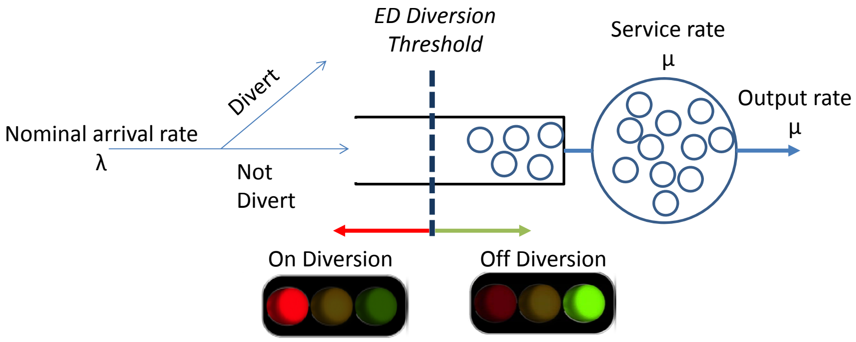


Figure 1.1. A conceptual queuing model of diversion

crowding on the ambulance waiting time at the ED and on the patient's quality of care and compare it with the impact of longer transportation time to reach an off-diversion neighboring ED. Based on this comparison they decide whether to comply with the ED's diversion signal and divert the ambulance or not. Thus, the diversion signal is only a recommendation and the paramedics are not required to comply with the signal.

We term the diversion signal to be *informative* if the estimated level of crowding associated with it is such that the paramedics comply with the signal and the likelihood that they divert an ambulance from an on-diversion ED is higher relative to an off-diversion ED. We also define *strength* of the diversion signal as the difference between the likelihood of diverting an ambulance from an on-diversion ED compared to an off-diversion ED.

Actions taken by the ED following the policy intervention to reduce time on diversion and the paramedics' response to these actions can affect the *strength* and the *informativeness* of the diversion signal. Suppose that the ED improves its processes (reflected by a higher service rate in Figure 1.1) without changing the diversion threshold. Then, the crowding level that the paramedics associate with the diversion signal will be lowered in the long run (in equilibrium). Consequently, they will be more likely to overrule the diversion status of the nearest ED and less likely to undertake longer transportation times to the neighboring ED. In other words, the diversion signal will become weaker. However, if the EDs respond by merely increasing the crowding threshold that triggers the on-diversion status, the paramedics will learn this in equilibrium and respond by decreasing the likelihood of overruling the diversion statutes; the signal will become stronger.

The outcome of the strategic interaction between the ED and the paramedics outlined above depends, also, on the diversion status of neighboring EDs. If several of these are on diversion simultaneously, a coordination guideline employed by many EMS agencies referred to as “All on Diversion, Nobody on Diversion” or simply ADND (Fatovich et al. 2005, Mihal and Moilanen 2005, Schneider et al. 2001, Vilke et al. 2004) provides more leeway to paramedics in overruling the diversion status and thus decreases the likelihood of diversion. However, the frequency with which this guideline is invoked may change post-intervention depending on the magnitude of the reduction in time on diversion at each ED.

1.1.2. Study preview

We apply the above conceptual framework to formulate hypotheses regarding the mechanism underlying ambulance diversion and test them using evidence from a policy intervention in Los Angeles (LA) County, California, implemented in April 2006. Pre-intervention, EDs did not face any restriction on the frequency and duration of diversion episodes. Post-intervention, the regulation mandated that the duration of an ED diversion episode should not exceed 1 hour and that the interval between two consecutive diversion episodes should be at least 15 minutes.

For our empirical analysis, we assemble a unique dataset comprising more than 46000 ambulance dispatches in a network of seven geographically proximate EDs over the years 2003-2009. The dataset consists of three different parts: (i) ambulance routing information such as the intended and the actual destination of the ambulance and the reason for diversion, if the two are different; (ii) information entered by the paramedics such as patient characteristics (age, gender and vital signs), and various time stamps such as the time of dispatch, the time of arrival at the scene and the time of arrival at the ED; and (iii) the start and end time of each diversion episode at each of the seven EDs over our study period.

We estimate a Probit model to study the paramedics' response to the diversion signal of the nearest ED and how this response changes after the policy intervention. We control for the status of neighboring EDs to account for the effect of the ADND policy described above. We also control for observable patient factors and fixed effects corresponding to the ED, time of day, day of week, month and year. Similarly, we estimate a two-part model (Probit and OLS) to study the effect of the diversion status of an ED and the policy intervention on the probability of delay (non-zero waiting time) and the expected ambulance waiting time conditional on being delayed at the ED.

We find that, pre- and post- intervention, the probability of an ambulance being diverted from an on-diversion ED was higher than that from an off-diversion ED when its neighboring EDs are off-diversion. This validates the role of diversion as an informative signal of ED crowding for paramedics. Second, an ambulance facing an on-diversion ED with off-diversion neighbors (so that diverting is possible) had a lower probability of being diverted after the intervention. Third, the likelihood of diverting an ambulance during an off-diversion period remained unchanged post-intervention. Fourth, the increase in the likelihood of diverting during on-diversion periods (compared to off-diversion) was lower post-intervention. This implies that the diversion signal was weaker in the equilibrium established between the EDs and the paramedics post-intervention. Finally, irrespective of pre- or post-intervention, ambulances were less likely to be diverted from the nearest ED if the nearest and neighboring EDs are all on diversion providing empirical evidence for the network effect (Allon et al. 2011, Deo and Gurvich 2011, Sun et al. 2006).

These findings uncover the complex multifaceted impact of the policy intervention on diversion probability and on time EDs spend on diversion. On the one hand, reduction in diversion hours at each ED had negative impact (decrease) on the overall diversion probability, as expected. This effect was reinforced by the "weakening" of the diversion signal, i.e., the probability of diversion from an on-diversion ED with off-diversion neighbors reduced post-intervention due to the paramedics' anticipation of reduced crowding at the EDs. On the other hand, the reduction in the likelihood of multiple neighboring EDs being on diversion simultaneously had a positive impact (increase) on the diversion probability due to the network effect described above. The net impact of these mechanisms was that time on diversion by 68% but the diversion probability reduced only by 8% post-intervention. We further find that, regardless of the intervention, ambulances waited less at an on-diversion ED compared to an off-diversion ED because of the lower arrival rate of ambulances during on-diversion periods. However, ambulance waiting time at an on-diversion ED increased because of the increased arrival rate (i.e., reduced diversion probability) post-intervention. Again, a combination of these opposing effects resulted in only 2% reduction of ambulance waiting time post-intervention.

We test the robustness of these findings to various changes in our models' specifications as well as for potential endogeneity resulting from other sources of information (apart from the diversion signal) that might be available to paramedics when they make their transport decisions.

1.1.3. Contribution to the related literature and policy implications

Previous studies of region-wide coordination programs to reduce ambulance diversion (Vilke et al. 2004, Asamoah et al. 2008, Patel et al. 2006, Lagoe et al. 2003, Burke et al. 2013) have used monthly data to report the impact of the intervention on diversion hours but did not directly test for the mechanisms underlying the observed changes. We contribute to this literature in two ways. First, we use the theoretical framework of signaling in queues (Allon and Bassamboo 2009) to construct hypotheses regarding: (i) the strength and the informativeness of the diversion signal for crowding in the ED based on paramedics' compliance to it, and (ii) the impact of the policy intervention on the signal's strength. This construction allows us to study rigorously whether EDs undertake process improvements in response to the policy intervention and how paramedics respond, in turn, to the EDs' actions. Second, we use individual ambulance transports (as against aggregated monthly data), which enables us to control for other factors that might affect the diversion probability.

Our empirical study also contributes to the nascent OM literature on signaling in queues by: (i) identifying an "informative" equilibrium in a novel application context and studying how this equilibrium is altered by an outside intervention. Our empirical analysis raises interesting theoretical questions beyond the immediate context of ambulance diversion: When is it beneficial for a service provider to increase its capacity in the presence of strategic communication with the customer regarding the extent of congestion? How does a change in service capacity alter the provider's communication strategy?

Finally, our work complements recent analytical work on ambulance diversion. Deo and Gurvich (2011) develop a stylized game theoretic queuing model of two EDs, where each ED chooses its own diversion threshold to minimize the average waiting time of its patients. They establish the existence of a *defensive equilibrium*, wherein both EDs are always on diversion but effectively do not divert any ambulances at all because of the ADND policy. Hagtvedt et al. (2009) reach a similar conclusion using an agent-based simulation model. Ramirez-Nafarrate et al. (2013) characterize the optimal threshold structure of the diversion policy and Do and Shunko (2013) propose a (partially) centralized routing policy that is pareto-improving for both EDs. All these models implicitly assume that paramedics perfectly comply with the diversion signal rather than interpreting it and responding to it based on the interpretation. We contribute to this literature by explicitly considering the decision making process of the paramedics in response to the ED diversion signals.

Our findings also have important policy implications. First, an intervention primarily aimed at reducing diversion hours can be successful at triggering improvements in patient flow processes at the EDs without introducing explicit incentives towards such process changes. Second, focusing on diversion hours as a performance metric might overestimate the efficacy of the intervention because it is likely to have a smaller impact on, arguably, more important driver of adverse patient outcomes namely the fraction of diverted ambulances. This differential impact is likely to be more substantive in settings where the magnitude of the network effect (multiple EDs simultaneously on diversion) pre-intervention is high. Third, increased efficiency of patient flow processes post-intervention might not always result in lower ambulance waiting time at the EDs because it can be outweighed by an increase in arrival rate of ambulances. This provides one potential explanation for the contradictory results in previous studies regarding changes in ambulance waiting time following similar community-wide interventions (Asamoah et al. 2008, Burke et al. 2013).

1.2. Study Setting

We study the effect of a diversion policy intervention, implemented by the Los Angeles County EMS (LACEMS) agency in April 2006. We focus on ambulance transports in a network of seven geographically proximate EDs located within a 5 mile radius in the North-West suburbs of Los Angeles. The average distance between EDs in the network is 4.3 miles and the travel time under regular traffic is 12.4 minutes. Our study network is relatively "closed" in that more than 80% of the ambulances originally intended for the EDs in the network were received within the network. Moreover, if one considers the set of the three closest EDs for each of the seven EDs, 19 out of 21 EDs in this set are in the study network and constitute 91% of all traffic interactions. Next, we describe various policies that govern ambulance traffic in this setting.

1.2.1. ED diversion policy

EDs may declare diversion status due to ED crowding, unavailability of critical equipment (e.g. CT scanner, operating room) or specialists (e.g. neurosurgeon, trauma care team), and internal disaster (e.g. flooding, fire, power outage). EDs communicate the start and end time of each diversion episode to the EMS agency in real time through the Reddinet electronic system (http://www.reddinet.com). Each ED maintains its own criteria for declaring an on-diversion status.

Until March 2006, the EMS agency did not restrict the usage of diversion status by the EDs in any way. In April 2006, LACEMS revised its diversion policy with the objective of reducing the EDs' use of the on-diversion status. Under the revised policy, an ED was allowed to remain on diversion for up to an hour, after which the status was automatically updated to off-diversion (or "open"). The EDs were required to be off-diversion for at least 15 minutes in between two on-diversion periods. The policy also included other provisions aimed at reducing the usage of diversion status by the EDs. It required that the ED nurse in-charge undertake preventive measures for relieving ED crowding such as expediting patient discharges and laboratory and radiological tasks. It mandated that the diversion status be updated by the hospital CEO or an appropriate administrative representative instead of the ED staff and empowered the EMS agency staff to perform unannounced site visits to verify compliance with the policy guidelines.

1.2.2. Ambulance transport policy

The LACEMS agency designates a limited number of hospitals, called "base hospitals", to coordinate the transport of ambulances from the scene (patient's initial location) to their final destination. A mobile intensive care nurse at the base hospital and the ambulance crew (jointly referred to as "paramedics"), assess the medical condition of the patient. Under normal circumstances, the paramedics direct all patients to the most accessible receiving ED (MAR) that is staffed and equipped to provide appropriate care for the patient's medical condition. The MAR is usually the one that has shortest travel time from the scene, which itself may depend on the distance, traffic and weather conditions.

The actual destination of an ambulance is determined by the patient's preference, the type of ambulance[1] and the discretion of the paramedics in the following hierarchical manner. Patients can request to be transported to a facility other than the MAR because of their preferred physician or health plan and the paramedics must comply with the patient's request. Even in the absence of an explicit patient preference, a basic life support (BLS) unit must be transported to the MAR irrespective of its diversion status. However, paramedics on an advanced life support (ALS) unit whose MAR is on diversion can transport the patient to an alternative receiving ED (REC) if it is deemed to be better for patient care. Thus, an ambulance is labeled as diverted if MAR $\neq$ REC. Because BLS ambulances cannot be diverted due to ED crowding, we focus only on the transport of ALS ambulances in our analysis.

1.3. Hypotheses

In this section, we draw upon our conceptual framework (§1.1.1) based on the strategic communication between EDs and paramedics to formulate hypotheses regarding the

[1] Ambulances are classified either as Advanced Life Support or Basic Life Support (BLS) units. BLS units are designed for inter-facility transportation and pre-hospital response to ill or injured patients. ALS units, which respond to 911 calls, extend BLS units to further support circulation and provide an open airway and adequate breathing. They can administer certain medications, have cardiac monitors, advanced cardiac life support equipment and blood glucose testing equipment.

impact of diversion status and policy intervention on the ambulance diversion probability and on the time ambulances wait to offload their patients. Recall that EDs decide on a threshold level of crowding and use the diversion status to signal that to the paramedics. Paramedics, in turn, use the MAR's signal to infer the underlying level of crowding, consequent ambulance waiting time and quality of care that would be provided to the patient. They then compare this with transport times to neighboring EDs and their diversion statuses to determine whether they should comply with the MAR's signal or overrule it, i.e., whether to divert the ambulance or not. Building on the conceptual model in the introduction, we first ask whether the signal is informative (decision-relevant) to paramedics?

Consider a situation where the paramedic is indifferent between transporting the patient to an *off-diversion* MAR or one of its neighboring off-diversion ED. Now, when the MAR is *on-diversion*, everything else being equal, this paramedic should be willing to transport the ambulance to the neighboring ED if she believes that the ambulance waiting time and consequent quality of care would be sufficiently worse at the on-diversion MAR. In that case, the likelihood of diverting the ambulance from the MAR would be greater during the on-diversion periods and we would deem the diversion signal to be informative. Hence, we arrive at the following hypothesis:

Hypothesis 1. *Both pre- and post-intervention, an ambulance is more likely to be diverted from an on-diversion MAR compared to an off-diversion MAR.*

The policy intervention imposes a constraint on the frequency and duration of diversion episodes. As discussed earlier, the EDs can respond to this constraint by following

LACEMS agency's recommendations and improving their patient flow processes. This will enable the EDs to avoid periods of severe crowding thereby reducing the fundamental need for diversion without changing the diversion threshold. EDs can choose process improvements that apply to all crowding levels (e.g. overall increase in staffing level, installation of IT solutions to improve visibility of available beds) or those that apply only to periods of high crowding (e.g. early discharge policies in the ICU when the number of boarding patients exceeds a certain threshold, increase in staffing level only during busy periods of the day). In both cases, because of the nonlinear dynamics of queuing systems, the impact of increase in service rate will be greater during on-diversion periods as they are more congested compared to off-diversion periods. In equilibrium, paramedics will learn to associate lower crowding level (compared to pre-intervention) with the on-diversion signal. Consequently, a paramedic who was pre-intervention indifferent between taking the ambulance to an on-diversion MAR or diverting it to a neighboring off-diversion ED, will prefer post-intervention the on-diversion MAR because of the lower crowding levels. This leads to the following hypothesis:

Hypothesis 2(A). *An ambulance is less likely to be diverted from an on-diversion MAR post-intervention compared to pre-intervention.*

The EDs can also reduce time spent on diversion by simply increasing the threshold that triggers the diversion signal. In contrast to the first strategy, this will increase the average crowding level during both on-diversion and off-diversion periods. However, due to the nonlinear dynamics typical of queuing systems, the impact of increased crowding will be much higher during on-diversion periods compared to off-diversion periods. Again, the paramedics will learn that on-diversion periods are associated with higher crowding post-intervention. Hence, the paramedic who was indifferent pre-intervention will prefer an off-diversion neighboring ED over the on-diversion MAR post-intervention. This leads to an alternate characterization of the post-intervention equilibrium as follows:

Hypothesis 2(B). *An ambulance is more likely to be diverted from an on-diversion MAR post-intervention.*

Next, we turn our attention to the likelihood of an ambulance being diverted when the MAR is off-diversion. Irrespective of whether the EDs undertake process improvement or not, the paramedics continue to compare off-diversion MAR with off-diversion neighboring EDs in this case. As a result, the underlying trade-off in their decisions does not change post-intervention. In other words, a marginal paramedic who was indifferent between transporting the ambulance to an off-diversion MAR vs. an off-diversion neighboring ED will continue to be indifferent post-intervention too, yielding the following hypothesis:

Hypothesis 3. *The likelihood of an ambulance being diverted when the MAR is off-diversion will remain unchanged post-intervention.*

Next, we focus on the equilibrium “strength” of the diversion signal which is captured by the difference in the likelihood of diverting an ambulance during on-diversion periods compared to off-diversion periods. If EDs undertake process-improvement efforts, then as previously argued, the impact on crowding and, consequently, on patient outcomes will be greater during on-diversion periods due to the nonlinear queuing dynamics. In this case, the paramedics will learn that on- and off-diversion periods are less differentiated post-intervention, compared to pre-intervention. In other words, the diversion signal will be weaker, which leads us to the following hypothesis.

Hypothesis 4(A). *The increase in likelihood of diverting an ambulance from an on-diversion MAR (compared to off-diversion) will be lower post-intervention compared to pre-intervention.*

Instead, if the EDs increase the diversion threshold, the expected crowding level during both off-diversion and on-diversion periods will increase. However, because of the nonlinearity of queuing dynamics, the latter will increase more than the former. In other words, the two periods will become more distinguishable from each other post-intervention compared to pre-intervention thus making the signal stronger in equilibrium. This yields the following alternate hypothesis:

Hypothesis 4(B). *The increase in likelihood of diverting an ambulance from an on-diversion MAR (relative to off diversion) will be higher post-intervention compared to pre-intervention.*

Queuing theory predicts that changes in arrival rates will lead to commensurate changes in ambulance waiting times (measured as the time between the ambulance's arrival at ED until the patient is offloaded and the ambulance is again available for dispatch). First, Hypothesis 1 states that arrival rate of ambulances will be lower during on-diversion periods compared to off-diversion periods. Moreover, the service rate during on-diversion period is likely to either remain unchanged or actually increase (based on speed-up behavior observed in Kc and Terwiesch (2009), Batt and Terwiesch (2012).). A combination of these two reinforcing effects leads us to the following hypothesis:

Hypothesis 5. *Both pre- and post-intervention, ambulance waiting times will be shorter during an on-diversion period compared to an off-diversion period.*

At first glance, the above hypothesis might seem counterintuitive. First, excessive crowding is likely to trigger diversion but the level of crowding is likely to come down as staff reacts and responds to the situation. Second, one might expect that lower waiting time will attract more ambulances to an on-diversion ED thus pushing up the waiting time until it is equal to that of an off-diversion ED. However, recall that the paramedics' decisions are not based solely on waiting time of the ambulance but on patient outcome, which is likely to depend on the sum of waiting time and transport time (§1.1.1).

If, in response to the policy intervention, the EDs only increased the ED-diversion threshold and this resulted in fewer ambulances arriving during on-diversion periods (Hypothesis 2(A)(B)), ambulance waiting times at an on-diversion REC should be lower post-intervention. If, instead, the EDs improve their patient flow processes and if paramedics responded to these changes by increasing the arrival rate (Hypothesis 2(A)(A)), the net change in ambulance waiting time would depend on which effect is dominant: if the effect of increased service rate is grater than that of increased arrival rate, it will lead to decreased waiting time, else it will lead to increased waiting time. This leads to the following pair of competing hypotheses:

Hypothesis 6(A). *Ambulance waiting times at an on-diversion ED will be longer post-intervention.*

Hypothesis 6(B). *Ambulance waiting times at an on-diversion ED will be shorter post-intervention.*

1.4. Methods

In this section, we describe various components of our dataset, construction of outcomes and predictors, and finally the empirical specifications using these measures to test the hypotheses developed in §1.3.

1.4.1. Data

We obtain three datasets from different sources in collaboration with the LACEMS agency. They span a period of 39 months pre-intervention (January 2003–March 2006) and 57 months post-intervention (April 2006–December 2009) and include information on more than 46000 ALS ambulance transport during this period to 7 geographically proximate EDs in northwest part of the LA county. We include more information on them below.

(i) The *ED diversion dataset* is obtained directly from the Reddinet system (§1.2.1). It includes the start and the end time of each ED diversion episode, measured at the granularity of minutes, announced by the EDs from January 2003 to December 2009.

(ii) The *ambulance routing dataset* is obtained from the base hospital that coordinates routing of ambulances to the seven EDs. For each ALS ambulance that was intended for one of the seven EDs in our network from January 2003 through December 2009, the dataset includes the following fields: ambulance sequence number, time at which the ambulance crew contacted the base hospital to know the diversion status of EDs and decide on the destination (called the incidence time), identity of the most accessible receiving ED (MAR) and the actual receiving ED (REC). It also reports a rationale for diversion if REC $\neq$ MAR, e.g. patient request, trauma patient, ED saturation.

(iii) The *ambulance sequence dataset*, spanning the periods from January 2003 through December 2006 and from January 2009 through December 2009 (data from January 2007–December 2008 are missing) is obtained from the Los Angeles City Fire Department, which operates ALS ambulances. It contains the age and gender of the patient along with vital statistics such as blood pressure and pulse rate. The dataset includes, in addition, various time stamps of key events during ambulance transport such as the arrival time at the scene, the arrival time to the REC, and the time at which the ambulance became available after offloading the patient.

We merge the ambulance routing and ambulance sequence datasets by the sequence number, which uniquely identifies each ambulance run. Then, for each ambulance sequence we add the diversion status of the MAR at its incidence time because it plays a critical role in the paramedics' decision of whether to divert the ambulance or not. Our final dataset includes 46842 ambulance transports; 22530 pre-intervention from January 2003 through March 2006 and 24312 post-intervention from April 2006 and through December 2009.

Table 1.1 summarizes the impact of the policy intervention on the main outcome variables. The policy intervention reduced the time spent by EDs on diversion by 68.2%; from 24.2% pre-intervention (2199 hours per year) to 7.7% post-intervention (750 hours per year). The magnitude of this reduction reflects one dimension of the effectiveness of the policy intervention and is comparable to earlier reports of similar interventions (Vilke et al.

Table 1.1. Descriptive statistics: ED Time-on-diversion & Fraction of Ambulances Diverted

	Pre-intervention	Post-intervention	Change
Fraction of ED time-on-diversion	0.242	0.077	-68.2%
Fraction of ambulances whose MAR is *on diversion*	0.190	0.068	-64.2%
Fraction of Ambulances diverted (overall)	0.269	0.247	-8.2%
Fraction diverted (MAR is *on-diversion*)	0.618	0.706	14.24%
Fraction diverted (MAR is *off-diversion*)	0.108	0.136	25.92%
Average ambulance waiting time (overall) (min)	15.44	15.14	-1.9%
Average waiting time (REC is *on-diversion*) (min)	13.44	14.39	6.32%
Average waiting time (REC is *off-diversion*) (min)	15.63	15.16	3.00%

2004, Asamoah et al. 2008, Patel et al. 2006). A natural consequence of this was a 64.2% reduction in the fraction of ambulances whose MAR was on-diversion at the incidence time from 19% pre-intervention to 6.8% post-intervention. However, the reduction in the fraction of diverted ambulances and their waiting time was not commensurate with the reduction in time on diversion: 8.2% and 1.9%, respectively. Further, the diversion probability during both on- and off-diversion periods actually increased by 14.24% and 25.92%, respectively and the average ambulance waiting time during on-diversion periods increased by 6.32% post-intervention. One of the objectives of our empirical analysis is to provide a detailed account of these seemingly discrepant observations based on the hypotheses developed above.

1.4.2. Outcome Variables

We use two outcome variables to understand the underlying mechanisms through which the policy intervention reduced the time on diversion, i.e., whether EDs improved their patient flow processes or increased the diversion threshold. First, we use a binary variable Div_k, which is 1 if ambulance k is diverted from the most accessible receiving ED (MAR), i.e., if REC $\neq$ MAR, and 0 otherwise. Second we use a continuous variable $Wait_k$ to represent the ambulance waiting time at its receiving ED (REC).

1.4.3. Predictor Variables

ED diversion statuses (MARd, RECd, dis$_i$d): Paramedics' decision of whether to divert the ambulance or not depends primarily on the diversion status of the MAR, which we denote using a binary variable $MARd_k$, which is set to 1 if the MAR is on diversion at the incidence time of ambulance k and 0 otherwise. To account for the fact that the probability of diverting an ambulance also depends on the diversion statuses of MAR's neighboring EDs, we introduce the diversion statuses of the three nearest EDs to the MAR at the incidence time of ambulance k, denoted by $dis1d_k, dis2d_k, dis3d_k$, where 1 is the closest ED and 3 is the 3^{rd} closest ED to the MAR.

In contrast, the time an ambulance waits to offload its patient depends on the diversion status of the receiving ED (REC) at its arrival time and does not depend on the status of the neighboring EDs. For the model that estimates the ambulance waiting time, we use a binary variable $RECd_k = 1$ to denote that the receiving ED of ambulance k was on diversion status at the *arrival time* of ambulance k to the ED. We set $RECd_k = 0$ otherwise.

Policy intervention (AFTER): We introduce a binary variable $AFTER_k$ that is set to 1 if ambulance k was incident after the policy intervention, i.e. on or after April 1^{st},

2006. We set this variable to 0 for ambulances that were incident pre-intervention, i.e., on or before March 31$^{\text{st}}$, 2006.

1.4.4. Control Variables

Patient request (Req): A patient's request to be taken to a specific ED takes precedence over all other considerations. Hence, we introduce a binary variable Req_k that is set to 1 if the patient had requested to be transported to an ED other than the MAR and is set to 0 otherwise.

Time fixed effects (T): Service rates in EDs and potential arrival rates to EDs are known to be non-stationary, i.e, to vary over time regardless of the diversion mechanism. We introduce a series of fixed effects to account for these variations. First, we use hour of day to account for the intraday non-stationarity in the arrival pattern and consequently in the crowding levels, e.g. late afternoons tend to be more crowded then early mornings. Second, we include day of week to account for any predictable variability in ED and inpatient departments' occupancy over the course of the week. Third, we include a fixed effect for each month to account for seasonal trends such as the annual flu epidemic. Last, we include a fixed effect for the year to control for any long term trend in diversion patterns that is common across all EDs. In some alternate specifications, we include a linear trend variable to capture some secular changes in diversion practices over time; see §1.6. Collectively, we denote these by T_k for ambulance k.

Patient characteristics (P): Patient's clinical characteristics can have a significant impact on paramedics' diversion decisions. Extremely urgent patients, for example, might not survive the additional transport time required to divert the ambulance away from an on-diversion MAR to an off-diversion neighboring ED and, hence, such patients are less likely to be diverted. We use vital statistics of the patients to capture this dimension and use a binary variable $SEVERE_k$ that is set to 1 if the patient in ambulance k satisfies any of the following: the systolic blood pressure is below 60 or above 180, the diastolic blood pressure is below 40 or above 110, and the pulse rate is below 50 or above 100. Further, older patients have a greater level of urgency, present higher load to the ED, and are admitted to hospitals more frequently (Caplan et al. 2004, Aminzadeh and Dalziel 2002). Hence, we employ a binary variable $ELDERLY_k = 1$ to denote if the age of the patient in ambulance k is greater than or equal to 75 years of age. We collectively denote these patient variables by P_k for ambulance k.

ED fixed effect (E): We include a fixed effect for each ED to account for various sources of unobserved heterogeneity, e.g. internal thresholds that trigger a diversion signal or the presence of a bed manager who coordinates the placement of patients from the ED to the inpatient wards. In the model for diversion probability, the fixed effect corresponds to the MAR and in the model for waiting time, it corresponds to the REC. These fixed effects are denoted by E_k for ambulance k.

1.4.5. Model specification for probability of diversion

Using the variables defined above, we formulate the following cross-sectional Probit model for the probability of diversion:

$$(1.1)\quad Div_k = \Phi\big(\beta_0 + \beta_1 MARd_k + \beta_2 AFTER_k + \beta_3 MARd_k \cdot AFTER_k$$

$$+\beta_4 MARd_k \cdot \mathit{dis1d}_k + \beta_5 MARd_k \cdot \mathit{dis2d}_k + \beta_6 MARd_k \cdot \mathit{dis3d}_k + \beta_7 Req_k + \eta_k E_k + \pi_k P_k + \tau_k T_k\big) + \epsilon_k,$$

where $\Phi(\cdot)$ represents the standard normal cumulative distribution function.

Note that coefficient β_1 and $\beta_1+\beta_3$ captures the informativeness of the diversion signal, i.e., the whether the likelihood of diverting an ambulance intended to an on-diversion MAR is greater than an off-diversion one, pre- and post-intervention, respectively. Thus, Hypothesis 1 corresponds to $\beta_1, \beta_1 + \beta_3 > 0$. The sum $\beta_2 + \beta_3$ captures the effect of the policy intervention on the likelihood of diversion from an on-diversion MAR. Improvement in the patient flow processes by the EDs would result in $\beta_2 + \beta_3 < 0$ (Hypothesis 2(A)) whereas a change in the diversion threshold by EDs would yield $\beta_2 + \beta_3 > 0$ (Hypothesis 2(B)). The coefficient β_2 captures the change, pre- vs. post-intervention, in the diversion probability from an off-diversion MAR. Hypothesis 3 corresponds to $\beta_2 = 0$. Finally, coefficient β_3 captures the change in the strength of the diversion signal from pre- to post-intervention. Hypothesis 4(A) implies $\beta_3 < 0$ whereas 4(B) implies $\beta_3 > 0$.

1.4.6. Model specification for ambulance waiting time

Next, we build an empirical specification for the time ambulances wait to offload their patients. As ALS ambulances receive the highest priority in the ED, a non-negligible fraction of these (roughly 5%) have zero delay. To handle this “zero-inflated” data, we use a two-part model (Duan et al. 1984). The first part is a Probit model that predicts whether an ambulance will experience delay (non-zero waiting time at the ED) and the second part is an OLS model that estimates the expected waiting time conditional on the ambulance being delayed. More formally, we sequentially estimate the following pair of equations:

$$(1.2)\quad \mathbb{P}(Wait_k > 0) = \Phi\big(\alpha_0 + \alpha_1 RECd_k + \alpha_2 AFTER_k + \alpha_3 RECd_k \cdot AFTER_k + \alpha_4 Req_k + \alpha_5 Div_k + \eta_k E_k + \pi_k P_k + \tau_k T_k\big) + \zeta_k,$$

$$(1.3)\quad Wait_k | Wait_k > 0 = \gamma_0 + \gamma_1 RECd_k + \gamma_2 AFTER_k + \gamma_3 RECd_k \cdot AFTER_k + \gamma_4 Req_k + \gamma_5 Div_k + \eta_k E_k + \pi_k P_k + \tau_k T_k + \xi_k,$$

where error terms ζ_k and ξ_k are uncorrelated. We explicitly test the validity of this assumption in the error structure in our empirical analysis.

The coefficients α_1 and γ_1 capture the impact of the REC's diversion status on the ambulance waiting time pre-intervention whereas $\alpha_1 + \alpha_3$ and $\gamma_1 + \gamma_3$ capture that impact post-intervention. According to Hypothesis 5, we expect $\alpha_1, \gamma_1 < 0$ and $\alpha_1 + \alpha_3, \gamma_1 + \gamma_3 < 0$. Similarly, coefficients $\alpha_2 + \alpha_3$ and $\gamma_2 + \gamma_3$ denote the effect of policy intervention on the waiting times during on-diversion periods. Hypothesis 6(A) and Hypothesis 6(B) correspond to $\gamma_2 + \gamma_3, \alpha_2 + \alpha_3 > 0$ and $\gamma_2 + \gamma_3, \alpha_2 + \alpha_3 < 0$, respectively.

1.5. Results

1.5.1. Probability of diversion

Table 1.2 displays the results for the Probit model that estimates the probability of diverting an incident ambulance (1.1). Since patient related variables are not available in our dataset for calendar years 2007 and 2008, we run two distinct specifications of this model. In specification (1), we omit the patient controls (severity and age) which allows us to use the entire dataset from January 2003 through December 2009. In specification (2)

Table 1.2. Probit model for estimating the likelihood of diverting an ambulance

	(1)	(2)
Intercept	-1.17*** (0.07)	-1.00*** (0.08)
$MARd$ (β_1)	1.87*** (0.04)	1.89*** (0.04)
$AFTER$ (β_2)	0.03 (0.05)	0.03 (0.05)
$MARd \cdot AFTER$ (β_3)	-0.14*** (0.05)	-0.13* (0.07)
$MARd + MARd \cdot AFTER$ ($\beta_1 + \beta_3$)	1.73*** (0.04)	1.77*** (0.06)
$AFTER + MARd \cdot AFTER$ ($\beta_2 + \beta_3$)	-0.11* (0.06)	-0.09 (0.08)
$MARd \cdot dis1d$ (β_4)	-0.34*** (0.04)	-0.34*** (0.04)
$MARd \cdot dis2d$ (β_5)	-0.37*** (0.04)	-0.37*** (0.04)
$MARd \cdot dis3d$ (β_6)	-0.12*** (0.04)	-0.12*** (0.04)
$Request$ (β_7)	2.77*** (0.04)	2.89*** (0.05)
Year	Yes	Yes
Month	Yes	Yes
Time of day	Yes	Yes
Day of week	Yes	Yes
Patient controls	No	Yes
MAR FE	Yes	Yes
Time Period	03 - 09	03-06, 09
N	46842	34021

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

we include patient controls and hence drop the observations for the years 2007 and 2008. The number of observations for the two specifications are 46842 and 34021 respectively.

Note that the coefficient estimates are robust to these changes in the specification. Both pre- and post-intervention, the probability of diversion is higher when the MAR is on-diversion (and neighboring EDs are off-diversion) relative to when it is off diversion ($\beta_1 > 0$ and $\beta_1 + \beta_3 > 0$). This supports Hypothesis 1 that the diversion signal is informative to the paramedics. Also, $\beta_2+\beta_3 < 0$, which lends support to Hypothesis 2(A) that EDs improved patient flow processes in response to the policy intervention and the paramedics responded by reducing the likelihood of diversion from an on-diversion MAR whose neighboring EDs are off-diversion. The fact that β_2 is statistically insignificant supports Hypothesis 3 that the probability of diversion during an off-diversion status did not change following the policy intervention. Finally, $\beta_3 < 0$ supports Hypothesis 4(A), i.e., that the difference in likelihood of diversion between on-diversion and off-diversion periods reduced after the policy intervention. In other words, the strength of the diversion signal reduced post-intervention.

We also note that the coefficient β_4, β_5 and β_6 are negative and statistically significant. In other words, an ambulance is less likely to be diverted from an on-diversion MAR if its neighboring EDs are also on-diversion. Thus, these coefficients provide empirical evidence for the network effect that has been discussed in both the emergency medicine (Sun et al. 2006, McCarthy et al. 2007) and the operations management literature (Allon et al. 2013).

Table 1.3. Two-part model for estimating the ambulance waiting time to offload patients

	Probit	OLS
Intercept	1.76^{***}	13.34^{***}
	(0.13)	(0.57)
$RECd$ (α_1, γ_1)	-0.32^{***}	-1.66^{***}
	(0.06)	(0.28)
$AFTER$ (α_2, γ_2)	0.26^{***}	0.52
	0.08	0.34
$RECd \cdot AFTER$ (α_3, γ_3)	-0.18	1.93^{**}
	(0.14)	(0.79)
$RECd + RECd \cdot AFTER$ $(\alpha_1{+}\alpha_3, \gamma_1{+}\gamma_3)$	-0.50^{***}	0.27
	(0.13)	(0.74)
$AFTER + RECd \cdot AFTER$ $(\alpha_2{+}\alpha_3, \gamma_2{+}\gamma_3)$	0.08	2.46^{***}
	(0.15)	(0.84)
$Request$ (α_4, γ_4)	0.03	0.31
	(0.10)	(0.39)
$Diverted$ (α_5, γ_5)	0.17^{**}	0.29
	(0.06)	(0.24)
Year	Yes	Yes
Month	Yes	Yes
Time of day	Yes	Yes
Day of week	Yes	Yes
Patient controls	Yes	Yes
REC FE	Yes	Yes
Time Period	03 - 06, 09	03-06, 09
N	24540	23441

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

1.5.2. Waiting time of ambulances

Table 1.3 includes the results for the two-part model that predicts the time an ambulance waits to offload its patient. Based on the coefficients of $RECd$ (γ_1, α_1), we conclude that ambulances are less likely to wait at their receiving ED pre-intervention $(\alpha_1 < 0)$ and their expected waiting time conditional on having to wait is lower $(\gamma_1 < 0)$ during an on-diversion period compared to an off-diversion period. The coefficient of $RECd + RECd \cdot AFTER$ in the Probit model $(\alpha_1 + \alpha_3 < 0)$ suggests that ambulances are less likely to wait during on-diversion periods post-intervention. However, the corresponding coefficient in the OLS model $(\gamma_1 + \gamma_3)$ is not significant. Together, these provide support for Hypothesis 5. Further, $\gamma_2 + \gamma_3 > 0$ implies that, conditional on delay, ambulances wait longer during an on-diversion period post-intervention. Finally, $\alpha_2 + \alpha_3$ is positive but statistically insignificant suggesting no change in the probability of delay during on-diversion periods post-intervention. Thus, we find partial support for Hypothesis 6(A) but not for Hypothesis 6(B). In other words, the increased arrival rate of ambulances during on-diversion periods post-intervention seems to have more than compensated for the increased service rate (improved patient flow processes) resulting in longer ambulance waiting time at the receiving ED.

1.5.3. Overall effect of the policy intervention

In this section, we use our empirical results to deconstruct the overall effect of the policy intervention on the probability of diversion and on the ambulance waiting time as reflected in the descriptive statistics (Table 1.1).

First, a reduction in time on diversion had a direct impact in reducing the probability of diversion. Second, the mechanism which facilitated this reduction, the improvement in the EDs' patient flow processes, had an indirect reinforcing effect in reducing the diversion probability. In particular, because the paramedics correctly interpreted that the diversion signal was weaker post-intervention, they further reduced the probability of diversion conditional on the on-diversion status of the MAR. However, the network dynamics had an opposing effect on probability of diversion. Reduction in time on diversion at each ED also reduced the likelihood of multiple EDs simultaneously being on diversion. As a result, the All on Diversion, None on Diversion (ADND) policy described earlier was applied less frequently, which actually led to an increase in the probability of diversion. We can infer that the second effect dominated the first one resulting in a net increase in the probability of diverting an ambulance from an on-diversion MAR post-intervention as reported in Table 1.1. Note that this effect is actually opposite of the intended effect of the policy intervention, which was to reduce the probability of diversion through a reduction in ED time on diversion. As a result, the net efficacy of the policy intervention in reducing diverted ambulances was lower than its efficacy in reducing ED time on diversion (8.2% vs. 68.2% as seen in Table 1.1).

The other impact of the policy intervention was on the time ambulances wait to offload patients. While one intuitively expects a reduction in this delay time post-intervention, our empirical results reveal a more complex picture. Because of the improvement in patient flow processes by EDs, diversion periods attracted more ambulance arrivals post-intervention compared to pre-intervention. This resulted in an 6.3% increase in ambulance waiting time during on-diversion episodes at the REC. As a result, despite a reduction in time on diversion across the network, the net reduction in average waiting time was only 2%. We believe that this mechanism provides a partial explanation for contradictory evidence on changes in ambulance waiting times found in prior studies based on monthly averages (Asamoah et al. 2008, Burke et al. 2013).

1.6. Robustness Checks

1.6.1. Model choice

In our econometric specification, the identification of the intervention's effect occurs purely through time using a binary variable $AFTER$, which indicates periods that are before and after April 1^{st}, 2006. To ascertain that we do not attribute the effects of exogenous temporal trends to the policy intervention, we run several alternative specifications of the base model as reported in column (1) of Table 1.2. The results of the alternative specifications are reported in Table 1.4.

First, we estimate the model after removing one month of data before and after the effective date of the policy intervention, i.e., March and April 2006 to minimize the effect of any changes that EDs might have made in anticipation of the intervention and to account for the gradual ramp up in their responses following the intervention (Burke et al. 2013). Second, we replace the year fixed effect with a linear trend variable, $Trend$ that counts the number of months elapsed since January 2003, which is the first month of observation in the dataset. These two specifications are displayed in columns (1) and (2) of Table 1.4. We find that the coefficient estimates for both specifications are very similar to those in Table 1.2. In column (2), we further find that the trend is positive and significant indicating that the probability of an ambulance being diverted is increasing over time

Table 1.4. Robustness checks: Probit models for estimating the likelihood of diverting an ambulance

	(1)	(2)	(3)
Intercept	-1.17***	-1.23***	-1.18***
	(0.08)	(0.06)	(0.08)
$MARd$ (β_1)	1.87***	1.87***	1.87***
	(0.04)	(0.04)	(0.04)
$AFTER$ (β_2)	0.02	0.02	0.03
	(0.06)	(0.03)	(0.05)
$MARd \cdot AFTER$ (β_3)	-0.16***	-0.15***	-0.12**
	(0.05)	(0.05)	(0.05)
$MARd + MARd \cdot AFTER$ ($\beta_1 + \beta_3$)	1.71***	1.72***	1.75***
	(0.04)	(0.04)	(0.04)
$AFTER + MARd \cdot AFTER$ ($\beta_2 + \beta_3$)	-0.14**	-0.12**	-0.09
	(0.07)	(0.05)	(0.06)
$MARd \cdot dis1d$ (β_4)	-0.34***	-0.34***	-0.36***
	(0.04)	(0.04)	(0.04)
$MARd \cdot dis2d$ (β_5)	-0.38***	-0.37***	-0.37***
	(0.04)	(0.04)	(0.04)
$MARd \cdot dis3d$ (β_6)	-0.12***	-0.12***	-0.12***
	(0.04)	(0.04)	(0.04)
$Request$ (β_7)	2.76***	2.77***	2.84***
	(0.04)	(0.04)	(0.04)
$Trend$		0.0025***	
		(0.0006)	
Year	Yes	No	Yes
Month	Yes	Yes	Yes
Time of day	Yes	Yes	Yes
Day of week	Yes	Yes	Yes
Patient controls	No	No	No
MAR FE	Yes	Yes	Yes
Time period	03 - Feb 06, May 06 - 09	03 - 09	Apr 03 - Mar 09
N	45566	46842	41055

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

irrespective of the intervention. The national increase in the number of ED visits and ED occupancy may be a driver for this trend (Pitts et al. 2012). In column (3), we revisit the base model but use data from 36 months before and after the policy intervention, i.e., April 2003 through March 2009 instead of including the calendar years January 2003 through December 2009. Again, the coefficient estimates are almost unchanged except for the coefficient of $AFTER + MARd \cdot AFTER$ ($\beta_2 + \beta_3$), which reduces slightly in magnitude. However, there is also an increase in the standard error because of a lower sample size thereby resulting in a loss of statistical significance (p-value = 0.17).

Next, we test the robustness of the network effect. Our base model includes separate interaction terms between the diversion status of the MAR and that of each neighboring ED. However, it does not effectively capture the ADND policy, which applies when several neighboring EDs are simultaneously on diversion. To accommodate the simultaneity condition, we add a binary variable $MARd \cdot dis1d \cdot dis2d$ to the base model that is set to 1 if the MAR and two nearest EDs to the MAR were on diversion simultaneously at the ambulance's incident time. We find this term to be significant and negative. In this expanded model the (separate) coefficients of each of the first two nearest EDs, $MARd \cdot dis1d$ and $MARd \cdot dis2d$ remain negative and statistically significant. All other coefficients are robust to this change in the specification. We also test whether the magnitude of the *network effect* changes after the policy intervention by interacting the binary variable $AFTER$ with the diversion statuses of the neighboring EDs $MARd \cdot dis1d$, $MARd \cdot dis2d$, $MARd \cdot dis3d$ in the base model. All three interaction terms are insignificant at the 10% level which indicates that the policy intervention did not affect the magnitude of the network effect. We do not include the details of these results for brevity.

Table 1.5. Robustness checks: Probit models for estimating the likelihood of ambulance waiting time being greater than a threshold

Threshold	20 minutes	15 minutes	10 minutes
Intercept	-1.32***	-0.37***	0.46***
	(0.08)	(0.07)	(0.08)
$RECd$	-0.08**	-0.26***	-0.50***
	(0.04)	(0.04)	(0.04)
$AFTER$	0.05	0.029	0.15***
	(0.05)	(0.04)	(0.05)
$RECd \cdot AFTER$	0.04	0.15	0.16
	(0.11)	(0.10)	(0.10)
$RECd + RECd \cdot AFTER$	-0.04	-0.11	-0.34***
	(0.10)	(0.09)	(0.12)
$AFTER + RECd \cdot AFTER$	0.09	0.18†	0.31***
	(0.13)	(0.11)	(0.09)
$Request$	0.07	0.049	0.06
	(0.06)	(0.05)	(0.06)
$Diverted$	0.02	0.06*	0.07**
	(0.03)	(0.03)	(0.03)
Year	Yes	Yes	Yes
Month	Yes	Yes	Yes
Time of day	Yes	Yes	Yes
Day of week	Yes	Yes	Yes
Patient controls	Yes	Yes	Yes
REC FE	Yes	Yes	Yes
Time Period	03 - 06, 09	03-06, 09	03 - 06, 09
N	24540	24540	24540

Robust standard errors in parentheses

† $p < 0.12$, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

To side-step the problem of zero-inflated waiting time data, which necessitates a two-part model, we consider Probit models for the likelihood that the ambulance waiting time is greater than a certain target. We estimate three variants of this model with targets of $10, 15$ and 20 minutes based on what is considered as acceptable in our setting (Eckstein and Chan 2004). Table 1.5 shows that our qualitative insights are virtually unchanged across all three specifications as indicated by the negative significant coefficients of $RECd$ and positive significant coefficients of $AFTER + RECd \cdot AFTER$. The coefficients are, however, smaller in absolute magnitude and less significant for the target of 20 minutes compared to 10 and 15 minutes. This suggests that the distribution of ambulance waiting times shifted post-intervention so that ambulances are less likely to experience shorter waits during on-diversion periods.

1.6.2. Exogeneity of the diversion signal

In practice, it is possible that paramedics are privy to crowding information through informal communication with the ED staff or based on recent ambulance transports, which is not observable to us as researchers. This additional information is likely to be correlated with the diversion signal as well as the paramedics' decision to divert or not thereby making the coefficients of the MAR diversion status variable $MARd$ and its interactions with the intervention indicator $AFTER$ and diversion status of the neighboring EDs ($dis1d$, $dis2d$, $dis3d$) biased.

We test for potential endogeneity by following the two-stage approach proposed in Smith and Blundell (1986). In the first stage, we fit linear probability models (See Wooldridge (2012) for the appropriateness of this approach for Probit models) with each of the five suspected discrete regressors as the dependent variable and all other exogenous regressors in the main equation as the independent variables along with a set of instrumental variables. As a basis of instrumental variables, we construct lagged diversion status variables ($L.MARd$, $L.dis1d$, $L.dis2d$, $L.dis3d$) using the fraction of time spent by the MAR and three nearest EDs to the MAR on-diversion during the same hour and day as the focal ambulance's incidence time (say Monday, 10:00-11:00) one week ago and on average over the previous four weeks. We then construct eight instrumental variables: $L.MARd$, $L.dis1d$, $L.dis2d$, $L.dis3d$, $L.MARd \cdot AFTER$, $L.MARd \cdot L.dis1d$, $L.MARd \cdot L.dis2d$, $L.MARd \cdot L.dis3d$. These instruments are likely to be correlated with the suspected endogenous variables but not with the diversion decision pertaining to the focal ambulance. We obtain the residuals from these first stage models and include them in the second stage Probit model for the diversion probability in addition to the original regressors. We do not find the residuals to be jointly statistically significant at the 10% level (p-value of 0.2436 for one week instruments and 0.1286 for four week average instruments) and hence cannot reject the null hypothesis that the suspected variables are exogenous (Baum 2007, Wooldridge 2012).

Using simulated likelihood to estimate a multivariate Probit model comprising system of six equations—one for diversion probability and five additional equations each for one of the suspected discrete regressors ($MARd$, $MARd \cdot AFTER$, $MARd \cdot dis1d$, $MARd \cdot dis2d$, $MARd \cdot dis3d$)–is computationally intractable (Cappellari and Jenkins 2003). However, we obtain estimates for five separate bivariate Probit models, where the main diversion probability equation is paired with each of the suspected discrete endogenous variable equations. Statistically insignificant correlation between the error terms in these models (p-value > 10%) provides further evidence that diversion status and its interactions are not endogenous.

1.7. Conclusion and Future Research

In this paper, we studied the mechanisms underlying ambulance diversion in a network of EDs and their role in determining the outcomes of a community-wide policy intervention. In particular, we developed a conceptual model of signaling in queues for the strategic interaction between EDs, who signal their level of crowding through the diversion signal, and paramedics, who decide whether to comply with that signal or not. We employ this framework to understand the impact on key operational measures–time on diversion, probability of diverting an ambulance and ambulance waiting time–of a community-wide diversion-policy intervention in LA County aimed at reducing the extent of ambulance diversion.

Our results suggest that EDs improved their patient flow processes to comply with the restrictions on the usage of ambulance diversion imposed by the new policy. On one hand, paramedics interpreted this change in equilibrium and responded by reducing the likelihood of diverting an ambulance from an on-diversion ED, which increased the effectiveness of the intervention. On the other hand, the reduced likelihood of multiple EDs being simultaneously on diversion led to paramedics' increasing the probability of diverting an ambulance from an on-diversion ED, which reduced the effectiveness of the intervention. The net effect of these mechanisms was that, while time-on-diversion reduced significantly after the policy intervention, the reduction in the fraction of ambulances diverted was much smaller. Similarly, increased arrival rate during on-diversion periods post-intervention more than outweighed the benefit of improved patient flow processes and effectively increased the offload time of ambulances during those periods.

Our empirical approach and results provide indirect evidence regarding the effectiveness of a relatively simple policy intervention–one that restricts time on diversion at EDs–in inducing process improvement. At the same time, it urges the relevant decision makers to be cautious in choosing appropriate performance measures to evaluate the effectiveness of the policy intervention. Focusing only on the reduction in time on diversion overestimates its effectiveness and overlooks the limited impact it had on other, clinically more relevant, operational measures such as the fraction of ambulances diverted and the ambulances' waiting times.

Our empirical setting and the challenges it presents provide a strong motivation for extending existing work on strategic delay announcement to a network of servers. In such a model the servers strategically decide on the delay signal based on their own congestion level and the anticipated congestion level of the other server. Customers decide which server to join based on the (vector of) signals and their decisions, in turn, impact the congestion level at the two servers.

In the specific context of ambulance diversion, our work underscores the importance of modeling paramedics (and not just EDs) as decision makers that need not comply with the diversion signal. For instance, analytical work on ambulance diversion suggests coordinated policies to reduce the inefficiency of decentralized diversion (e.g. Deo and Gurvich 2011, Do and Shunko 2013) and benchmarks them against an idealized network where a central controller can optimize the routing of ambulances. Our results highlight the need to re-evaluate the effectiveness of such regulations after incorporating the response of the paramedics and the mechanisms through which they update their interpretation of the diversion signals.

CHAPTER 2

Effectiveness of Hospital Readmissions Reduction Program: Are Hospitals Incentivized To Reduce Readmissions? (joint with Itai Gurvich, Jan A. Van Mieghem, Mark V. Williams, Robert S. Young, and Dennis Zhang)

Abstract

The Affordable Care Acts Hospital Readmissions Reduction Program (HRRP) penalizes hospitals with high excess readmission rates among Medicare beneficiaries. The program intends to motivate hospitals to reduce readmissions by providing financial incentives. Hospitals that elect to reduce readmissions may lessen the penalty incurred but at the same time will decrease revenue from patient admissions as they lose patient volume. To this end, how hospitals should respond to the program from a purely financial perspective is not clear. We analyze the hospitals financial incentive to reduce readmissions and forecast the effectiveness of the HRRP. We model hospitals as forward-looking net income maximizers and link their operational parameters to financials. Our financial model suggests that hospitals may or may not be incentivized to reduce readmissions by the HRRP. A hospital is financially incentivized to reduce readmissions only if its current readmission rate falls into a window of readmission rate that is unique to each hospital and depends on its own patient mix and characteristics. In a simulation of our framework

with a sample of 183 hospitals in California, 61.2 percent of the hospitals are incentivized to reduce while 5.5 percent are not incentivized despite incurring penalty. Patients under the worst quality of care may be more vulnerable as the program fails to incentivize hospitals that perform far worse than the target the program sets, the worst performers. The program also fails to incentivize those with small fraction of revenue contributed by Medicare beneficiaries and small fraction of Medicare revenue contributed by the three penalty applied conditions.

2.1. Introduction

Since October 1st, 2012 , hospitals with excessive 30 day readmissions for acute myocardial infarction (AMI), pneumonia (PN), and heart failure (HF) have been penalized under the Hospital Readmissions Reduction Program (HRRP) as mandated by the Affordable Care Act (ACA). Medicare has indicated that they view hospital readmissions as potential adverse events and a financial burden on the health care system, with nearly 20% of Medicare beneficiaries in 2003-2004 readmitted within 30 days of discharge at an estimated cost of $17 billion. The HRRP incentivizes hospitals to embrace efforts to reduce readmissions through improving patient care quality and transitions (Jencks et al. 2009, CMS et al. 2012).

In general, the HRRP penalizes excessive readmissions by adjusting all Medicare payments for target conditions to a hospital based on a ratio of excess readmissions for patients initially admitted for care of the target conditions. The data for the excessive readmissions is recorded in a 3 year window and then applied to the next fiscal year (FY) beginning 18 months after the data time period. The Excess Readmission Ratio is a measure of a hospitals performance (Risk-adjusted Predicted Readmissions) compared to the national average for the hospitals set of patients with these conditions (Risk-adjusted Expected Readmissions) (CMS 2013). Since hospitals are compared to their peers under the HRRP, this competitive aspect of the penalty structure induces hospitals to consider others readmissions reduction decision in their own readmission reduction decision. In FY 2013, more than 2,200 hospitals were penalized approximately $280 million for excess readmissions (Rau 2012). In FY 2014, a similar number of hospitals were penalized for excess readmissions, with a total penalty of $227 million (Rau 2013).

Despite these penalties, hospitals with excess readmissions, on the other hand, also face the financial disincentive of reducing readmissions from a hospital patient volume standpoint. With the high proportions of patients being readmitted, these readmissions constitute a significant amount of Medicare revenue to these hospitals (Berenson et al. 2012). Thus hospitals are caught in the dilemma of avoiding the readmissions penalty at the cost of reducing hospital volume and associated revenue from these readmissions. In addition, there are concerns that this financial dilemma disproportionately affects safety net hospitals and those that serve low-income populations who lack the financial resources to adjust to the HRRP requirements (Joynt and Jha 2011, 2012).

The objective of this analysis is to examine, through financial modeling, the financial incentives of hospitals to engage in readmissions reduction efforts, considering both the HRRP penalty and the potential of loss of revenue from lower hospital patient volume. Specifically, we model hospitals as net income maximizers and link operational parameters to hospital financials. We incorporate the HRRP penalty structure into the hospitals revenue in a game theoretic framework that allows us to take into account the effect of other hospitals readmissions reduction decisions, as manifest in the risk-adjusted Expected Readmissions metric, on a hospitals financials and eventually its decision to reduce readmissions. We examine the emerging equilibrium outcome between hospital decisions and analyze the implications in the effectiveness of the program.

We apply our framework to a sample of 183 California hospitals and simulate the hospitals readmission reduction decision to forecast the effectiveness of HRRP in inducing hospitals to reduce readmission. We also perform sensitivity analysis on the fiscal maximum penalty cap limit and number of penalty applied conditions.

2.2. Methods

2.2.1. HRRP Penalty Structure

According to the HRRP guideline, a hospitals excess readmissions penalty in a given fiscal year is assessed by the readmission performance of the three applicable conditions during a 3 year period lagged by 18 months from the applied fiscal year. For example, penalty in FY 2013 is assessed by performance measured between July 2008 and June 2011. The penalty calculation is based on a risk-adjusted performance metric for each condition known as the Excess Readmission Ratio which is equal to Risk-adjusted Predicted Readmissions divided by Risk-adjusted Expected Readmissions. In order to adjust for hospital specific risk factors, the two readmissions measures are estimated from a hierarchical logistic regression model using national level discharge data which covers all discharges from hospitals subject to the HRRP. Risk-adjusted Predicted Readmissions can be viewed as the hospitals actual readmissions performance measure after adjusting for the patient risk factors, while Risk-adjusted Expected Readmissions represents the benchmark performance measure to which the hospitals Risk-adjusted Predicted Readmissions rate is compared to. Risk-adjusted Expected Readmissions is an estimation of readmission performance of a national average hospital subject to the HRRP with the same patient risk factors as the evaluated hospital (CMSc).

If a conditions Excess Readmission Ratio exceeds 1, which means the predicted performance exceeds (is worse than) the benchmark performance, the hospital incurs a financial penalty proportional to the level of excessiveness (Excess Readmission Ratio − 1) multiplied by the Centers for Medicare & Medicaid Services (CMS) payment for Medicare patients diagnosed with the condition. The total penalty levied to the hospital is the sum of the penalty for the three index conditions. However, the total readmissions penalty has a maximum (equal to 1 Floor Adjustment Factor), which is equal to a predetermined percentage of all Medicare payments the hospital receives for treatment of all conditions. Under the current structure, this penalty cap increases from 1 percent in FY 2013 to 2 percent in FY 2014 and 3 percent in FY 2015 (CMS 2013).

2.2.2. Financial Model

We developed a financial model which represents patient admission related net income of a HRRP-enrolled hospital as a function of its readmission performance. We define net income as inpatient profit minus the HRRP penalty. Inpatient profit captures the hospitals profitability driven only by its operations which excludes external costs such as the HRRP penalty and is defined as inpatient revenue minus the operating expenses associated with admitted patients.

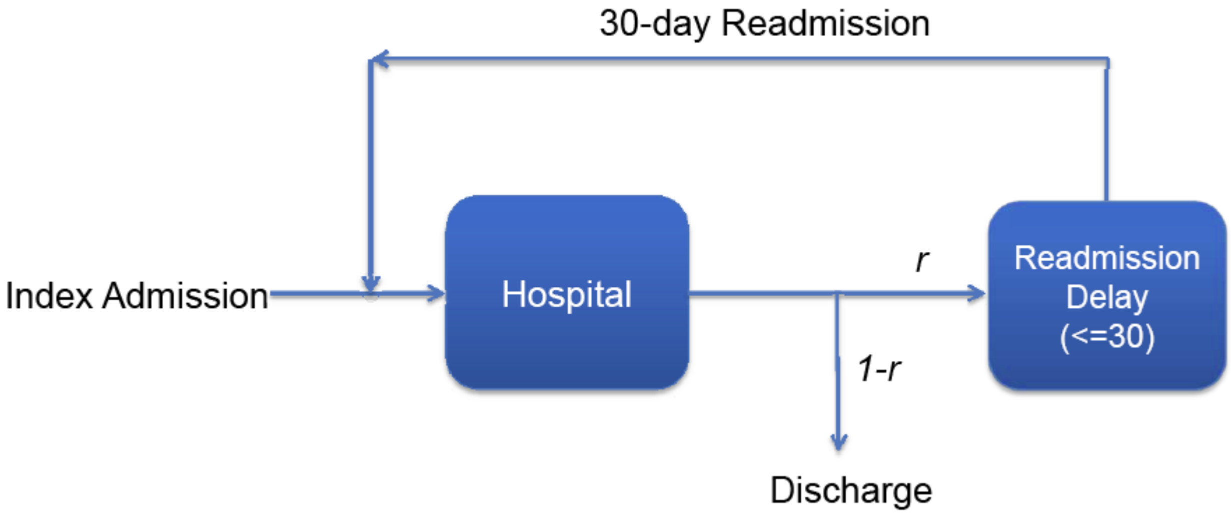


Figure 2.1. Patient flow process

First, using operations management modeling techniques, we model a hospitals inpatient admissions and discharges as a flow process and capture readmissions back to the hospital as a feedback loop with constant return rate r (Figure 2.1). In our model, r is represented by the Risk-adjusted Predicted Readmissions of the HRRP measures. From the model, we can express the hospitals patient volume as the following:

(2.1)

$$\begin{aligned} \textit{Patient Volume} &= \textit{Index Admission Rate} \cdot \frac{1}{1-r} \\ &= \textit{Index Admission Rate} \cdot \frac{1}{1 - \textit{Risk-adjusted Predicted Readmissions}}. \end{aligned}$$

Index Admission Rate is the rate of index admission patients arriving to the hospital. The $\frac{1}{1-r}$ term captures the nonlinear relationship between readmission rate and overall patient volume.

Next, we link the hospitals admission performance to its financials. We assume hospitals to receive a constant average payment for each inpatient admission from all payers (CMS, private insurers, self-paying patients or others). Within our model we aggregate the three HRRP applicable conditions into a single group and all other non-applicable conditions into another single group. This reduces the complexity of the penalty structure in the analysis and allows us to provide a clearer picture of the tradeoffs in reducing readmissions and insight of the individual hospitals financial incentive. We also assume that if the hospital decides to reduce readmissions, the rate of reduction in readmissions is constant through all condition types. In addition, we assume that admitted patients, regardless of their admission status, i.e., index admission or readmission, have the same profitability at the hospital level by assuming a constant operating margin, which is the ratio of the hospital level inpatient profit over inpatient revenue. Later in the simulation section, we perform a sensitivity analysis on patient profitability. These assumptions allow us to link a hospitals inpatient profit with its patient volume as following (CMS 2013):

(2.2)

$$
\begin{aligned}
\textit{Inpatient Profit} &= \textit{Inpatient Revenue} - \textit{Inpatient Operating Expenses} \\
&= \textit{Patient Volume} \cdot \textit{(Revenue per Patient} - \textit{Operating Expenses per Patient)} \\
&= \textit{Patient Volume} \cdot \textit{Profitability per Patient} \\
&= (\textit{Index Admission Rate} \cdot \frac{1}{1 - \textit{Risk-adjusted Predicted Readmissions}}) \\
&\cdot (\textit{Average Inpatient Revenue per Admission} \cdot \textit{Operating Margin}).
\end{aligned}
$$

We then explicitly model the HRRP penalty structure, reflecting the Floor Adjustment Factor and Risk-adjusted Expected Readmissions, and incorporate the penalty into the net income calculation as following:

(2.3)

$$\begin{aligned}
\textit{HRRP Penalty} &= \min[\textit{Medicare Inpatient Revenue} \cdot (1 - \textit{Floor Adjustment Factor}),\\
&\quad \textit{Medicare Inpatient Revenue from Applicable Conditions}\\
&\quad \cdot \max(\frac{\textit{Risk-adjusted Predicted Readmissions}}{\textit{Risk-adjusted Expected Readmissions}} - 1, 0)]\\
&= \textit{Inpatient Revenue} \cdot \textit{Fraction of Inp Rev Generated from Medicare Beneficiaries}\\
&\quad \cdot \min[(1 - \textit{Floor Adjustment Factor}),\\
&\quad \textit{Fraction of Medicare Inp Rev Generated from Applicable Conditions}\\
&\quad \cdot \max(\frac{\textit{Risk-adjusted Predicted Readmissions}}{\textit{Risk-adjusted Expected Readmissions}} - 1, 0)].
\end{aligned}$$

Finally, net income is inpatient profit minus the HRRP penalty. Since we can express inpatient revenue as

(2.4)

$$\begin{aligned}
\textit{Inpatient Revenue} &= \textit{Patient Volume} \cdot \textit{Revenue per Patient}\\
&= (\textit{Index Admission Rate} \cdot \frac{1}{1 - \textit{Risk-adjusted Predicted Readmissions}})\\
&\quad \cdot \textit{Average Inpatient Revenue per Admission},
\end{aligned}$$

we can substitute it into the HRRP penalty equation and express net income as

(2.5)

$$\begin{aligned}
&\textit{Net Income} = \textit{Inpatient Profit} - \textit{HRRP Penalty}\\
&\qquad = \textit{Index Admission Rate} \cdot \frac{1}{1 - \textit{Risk-adjusted Predicted Readmissions}}\\
&\qquad \cdot \textit{Average Inpatient Revenue per Admission}\\
&\qquad \cdot (\textit{Operating Margin} - \textit{Fraction of Inp Rev Generated from Medicare Beneficiaries}\\
&\qquad \cdot \min[(1 - \textit{Floor Adjustment Factor}),\\
&\qquad \textit{Fraction of Medicare Inp Rev Generated from Applicable Conditions}\\
&\qquad \cdot \max(\frac{\textit{Risk-adjusted Predicted Readmissions}}{\textit{Risk-adjusted Expected Readmissions}} - 1, 0)]).
\end{aligned}$$

The financial model captures the two opposing effects of reducing readmission rates (Risk-adjusted Predicted Readmissions) on hospital net income: inpatient profit loss due to less patients and penalty savings due to improved readmission performance measure. Using the framework we first predict individual hospital's incentive compatibility with the HRRP penalty.

It is important to note that we take a purely financial perspective in analyzing the effectiveness of the HRRP. We do not consider additional benefits of reducing readmission rate that are difficult to monetize, such as higher quality of service measures and better reputation. For example, a hospital may have a strategic advantage over competitors if its readmission rates are lower than the competitors and can take advantage in marketing practices. As the information on readmissions performance measures is publicly available, patients or their referring doctors may use readmission data in reference to choosing hospitals.

2.2.3. Strategic Hospital Decisions: Game Theoretic Perspective

Hospitals subject to the HRRP will weigh the two opposing effects of reducing readmissions on their financials and decide whether to and if so how much to reduce readmissions depending on the more dominant effect between the two. In the decision of reducing readmissions, hospitals are strategic in a game-theoretic perspective as they anticipate the effect of other hospitals readmissions reduction decision on their own net income.

The HRRP penalty structure adopts a relative performance measure by comparing a hospitals readmission performance to its peers in the program. Due to the national level discharge data pool and three year rolling structure, readmissions reduction actions taken by other hospitals will affect a hospitals benchmark performance measure, Risk-adjusted Expected Readmissions, in future years. Both Risk-adjusted Expected and Predicted Readmissions are publicly reported for all hospitals subject to the HRRP. A hospital can anticipate how other hospitals will react to the reported performance measures and forecast the impact of others reaction to its own future Risk-adjusted Expected Readmissions. Thus the action of reducing readmissions affects the hospitals own net income not only through the aforementioned two opposing effects but also through its own Risk-adjusted Expected Readmissions as it depends on the other hospitals reaction to the hospitals action of reducing readmissions. This requires a game-theoretic analysis that takes the decision of other hospitals into account in a hospitals optimal readmissions reduction decision.

In the financial model, we capture the action-reaction setting between hospitals through the mechanism of updating the Risk-adjusted Expected Readmissions. If a hospital reduces readmissions, the impact is reflected not only in its own but also in all other hospitals future Risk-adjusted Expected Readmissions. We follow the three year rolling performance measuring period and properly adjust the approximated Risk-adjusted Expected Readmissions in future years according to hospitals decisions.

Each hospital subject to the HRRP is maximizing its own net income conditional on others readmission reduction decision. Since all hospitals are deciding its own readmissions reduction simultaneously, they anticipate their peers decision when making their own. Given the anticipated peers decision each hospital will make an optimal decision where they achieve the maximum net income. As each and every hospital subject to the HRRP may correctly anticipate their peers decision (action) and decide in their own best interest (reaction), hospitals will reach equilibrium in their decisions.

Under the decisions made in equilibrium, in order to guarantee no hospital deviates from that decision, each and every hospital is better or at least as good as in their objective than in any other possible decision. Therefore, in the simulation analysis, we consider all combinations of possible decisions by the participating hospitals and compute the corresponding hospitals net incomes and identify the equilibrium outcome. In equilibrium, the net income a hospital achieves in its optimal decision is no less than that achieved in other decisions.

However, the realized equilibrium in reduction decisions is not guaranteed to be unique. For instance, it may be optimal for a hospital to reduce readmissions in one equilibrium while it may be better off to not reduce in another equilibrium where its peers decision are also different. Hence, one cannot predict whether a hospital will always be incentivized to reduce in all realized equilibria. Instead, we analyze the upper bound of the effectiveness of the HRRP, i.e., analyzing whether a hospital will be incentivized to reduce readmissions in any of the realized equilibria.

2.2.4. Simulation

In the simulation study we expand the framework to multiple hospitals subject to the HRRP and predict the hospitals readmission reduction action and the effectiveness of the HRRP, we model the exact penalty structure with the 18 month delay between the performance measurement and penalty application periods. We report a hospital to be incentivized if it has a non-zero probability of reducing readmissions in any of the realized equilibria. We assume hospitals may reduce readmissions in each and every year of the time horizon. We aggregate the readmission performances of the three applicable conditions into a single metric for each hospital by taking the average of Risk-adjusted Predicted (Expected) Readmissions for the three conditions weighted by their respective number of discharges as the predicted (expected) performance metric in the simulation. This modification implies that we assume hospitals to make a centralized decision in reducing readmissions, such that the rate of readmissions reduction is constant throughout the hospital regardless of the condition group.

We utilize three sets of data from distinct sources:

(i) The CMS Hospital Compare hospital level readmission database, from which we acquired the Risk-adjusted Predicted Readmissions rate, Risk-adjusted Expected Readmissions rate and number of discharges for Medicare patients with index admissions due to Acute Myocardial Infarction (AMI), Heart Failure (HF) and Pneumonia (PN) diagnoses (CMSa).

(ii) The inpatient CMS Medicare Provider Utilization and Payment Data was used to determine the fraction of Medicare revenue from the three applicable conditions (CMS 2013).

(iii) The Office of Statewide Health Planning & Development (OSHPD) database for the state of California. The annual inpatient revenue, inpatient profit, and Medicare inpatient revenue are reported in this database for 434 California hospitals (OSHPD).

We restrict our analysis to the 183 hospitals in California for which a complete set of data from all three sources is available

Using the framework of individual hospital's incentive compatibility with the HRRP, we perform simulations on the game-theoretic setting between the sample hospitals. We consider three distinct hospital objectives that depend on patient profitability to maximize from 2014 to 2020 with a discount factor of 0.99: net income with operating margin (inpatient profit HRRP penalty), net income with patient margin (inpatient profit only from patient treatment activity - HRRP penalty), and net revenue (inpatient revenue HRRP penalty).

From a researcher's perspective, profitability of an admitted patient which is captured by operating margin in Equation (2.5) is difficult to accurately measure. Also hospitals may have different criteria in viewing patient profitability. As a sensitivity analysis, we examine three profitability measures that hospitals might use from least to most profitable: operating margin, patient margin, and net revenue. Within the available hospital income statement from the OSHPD database, we use *Patient Margin* $(= \frac{\textit{Net Patient Revenue} - \textit{Patient Expenses}}{\textit{Inpatient Revenue}})$ as an alternative measure of patient profitability to *Operating Margin* $(= \frac{\textit{Inpatient Revenue} - \textit{Inpatient Operating Expenses}}{\textit{Inpatient Revenue}})$. Net patient revenue, which is the sum of gross patient revenue (sum of daily hospital, ambulatory, ancillary service patient revenue) and capitation premium revenue less deductions from revenue, and patient expenses, which is the sum of hospital, ambulatory, ancillary services operating expenses, only take operations from patient care services into consideration. On the other hand, inpatient revenue and inpatient operating expenses take both patient care services related and non-related health care operations into account. Examples for the latter include non-patient food sales, refunds and rebates, supplies sold to non-patients, and Medical Records abstract sales.

In the individual hospital's incentive compatibility analysis, we find that a hospitals optimal decision in each fiscal year is reduced to a binary choice: either completely avoiding the penalty (0) or incurring the penalty while maintaining a readmission rate above the higher range of the incentive-compatible region (1). This allows us to model the decision of a hospital maximizing net income over a long horizon as a vector of binary decisions for each fiscal year. For instance, a hospital with a 3 year time horizon has a total of 8 $(= 2^3)$ possible decisions: (0, 0, 0), (1, 0, 0), (0,1,0), (0,0,1), (1,0,1), (1,1,0), (0,1,1), (1,1,1). For a 7 year time horizon, there are 128 possible decisions.

2.3. Results

Table 2.1 summarizes descriptive statistics for the sample data. Predicted and Expected Readmission Rates are the aggregated readmission performance metric for each hospital calculated as the average of Risk-adjusted Predicted and Expected Readmissions for the three applicable conditions weighted by their respective number of discharges.

Table 2.1. Descriptive statistics of 183 sample California hospital data, 2013

Variables	Mean	Std. Dev.	Min.	Max.
CMS Hospital Compare database, Readmission data				
Predicted Readmission Rate	21.4	2.3	16.2	30.9
Expected Readmission Rate	21.6	1.5	17.1	26.5
AMI no. of discharges	158.2	115.3	25	599
AMI R-A Predicted Readmissions	20.2	3.1	12.3	35.8
AMI R-A Expected Readmissions	20.4	2.2	15.2	29.4
HF no. of discharges	364.8	209.1	31	1714
HF R-A Predicted Readmissions	24.5	2.5	18	32.2
HF R-A Expected Readmissions	24.7	1.3	20.7	28.5
PN no. of discharges	306.9	163.3	43	1146
PN R-A Predicted Readmissions	18.6	2.2	14	27.7
PN R-A Expected Readmissions	18.8	1.5	14.7	23.5
CMS Medicare provider utilization and payment data				
Fraction of Medicare revenue from 3 applicable conditions	13.3%	4.9%	3.8%	35.3%
OSHPD database, State of California				
Bed utilization	59.6%	13.7%	7.5%	97.6%
Fraction of revenue from Medicare	35.9%	10.3%	7.9%	63.2%
Operating margin	3.3%	8.0%	-29.1%	19.3%
Patient margin	40.0%	10.1%	-52.3%	69.5%

2.3.1. Individual Hospital Incentive Compatibility To HRRP

Using our financial model, we construct a graphical representation of the relationship between hospital net income and readmission rate Figure 2.2. The model provides a framework to predict whether or not a hospital is financially incentivized to reduce readmissions in order to maximize its net income. We modeled a hypothetical hospital using parameters derived from the average statistics from the California OSHPD and the CMS databases found in Table 2.1: the fraction of inpatient revenue received from Medicare beneficiaries of 35.9%, the fraction of Medicare revenue from three applicable conditions of 13.3%, an average operating margin of 3.3%, and a Risk-adjusted Expected Readmissions value of 0.216. We assume a Floor Adjustment Factor of 0.97, that is a 3 percent penalty cap.

Figure 2.2 depicts the hypothetical hospitals normalized net income and inpatient profit as a function of its readmission rate in a given fiscal year. The horizontal axis represents Risk-adjusted Predicted Readmissions. We normalize inpatient revenue to 1 when there are no readmissions, i.e., the readmission rate is zero. Hence, as the operating margin is 3.3%, the inpatient profit ($= \textit{Operating Margin} \cdot \textit{Inpatient Revenue}$) on the vertical axis is normalized to 0.033 when the readmission rate is zero which corresponds to the intercept of the vertical axis. Normalized inpatient profit is represented by the dash-dot line and increases as the readmission rate increases due to its non-linear proportionality with patient volume.

The Expected Readmission Rate value of 0.2159 is marked by the vertical dashed line in Figure 2.2. If the hospitals readmission rate exceeds the Expected Readmissions Rate, the hospital incurs a penalty proportional to the difference between the two rates. This is

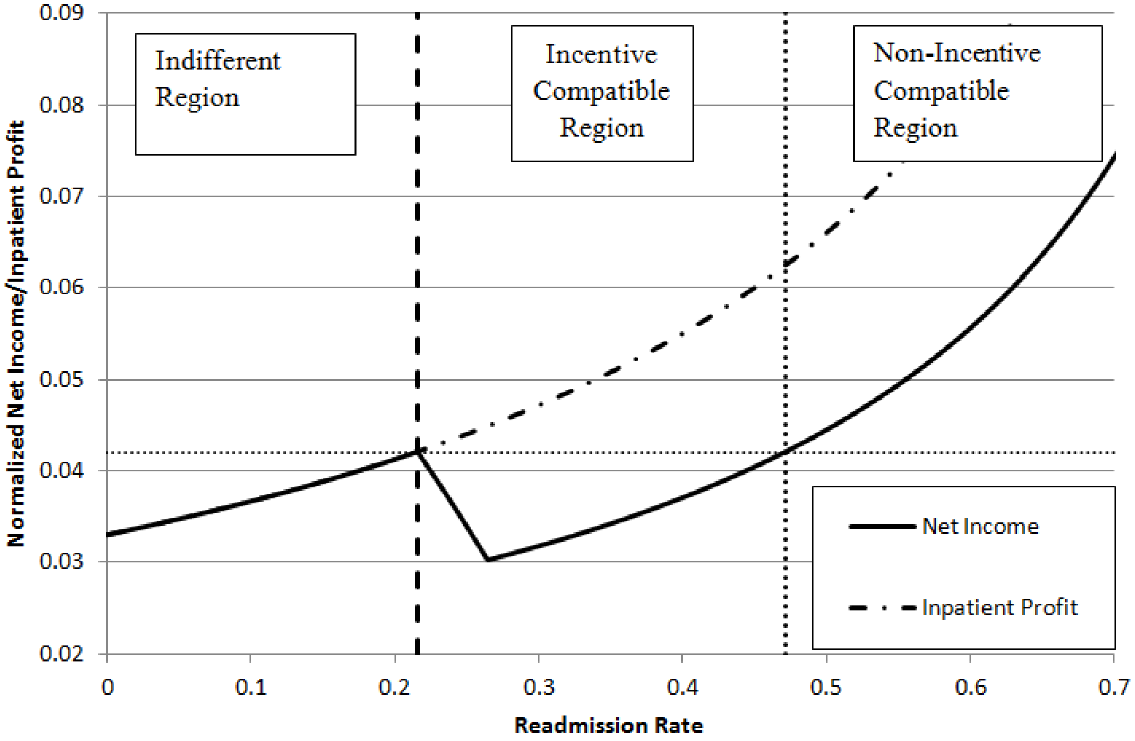


Figure 2.2. HRRP incentive compatibility: Average hospital

evident by the increasing gap between the inpatient profit (dash-dot line) and net income (solid line) as the readmissions rate increases away from the Expected Readmissions Rate. The Floor Adjustment Factor of the HRRP penalty structure forces a percentage-wise cap on the applicable penalty regardless of the hospitals readmission rate. The 3 percent penalty cap installed by the Floor Adjustment Factor of 0.97 is reached at the readmission rate of 0.2646, which is the lowest point of net income. At readmission rates above 0.2646, the HRRP penalty is capped at 3 percent of the inpatient profit (dash-dot line). Net income in this region is at the 97 percent level of the inpatient profit, hence, increases as the readmission rate increases. We can partition a hospitals readmission rate into three distinctive regions by its optimal strategy under the respective readmission rate.

If the hospitals performance readmission rate (Risk-adjusted Predicted Readmissions) is within the first region, the left of the three, it is already performing below or at the Expected Readmissions Rate, and therefore does not incur any HRRP penalty. Reducing readmissions will result only in profit loss without any penalty savings. The optimal strategy for the hospital is to remain at the current readmission rate. Hence, the hospital is not incentivized to reduce readmissions. We note this region as the *Indifferent Region*.

The revenue achieved at any readmission rate within the second region, the middle of the three, is not more than the revenue achieved at the Risk-adjusted Expected Readmissions, 0.2159. The upper bound of the region, 0.4718, is the point where the achieved net income is equivalent to the net income achieved at the Expected Readmissions Rate. The optimal strategy for a hospital performing (Risk-adjusted Predicted Readmissions) within this region is to reduce readmissions to the Expected Readmissions Rate, and generate the maximum achievable net income among the possible options of readmission rates equal to or below its current readmission rate. For example, a hospital currently performing at 0.4 readmission rate generates net income of 0.0371. Among readmission rates below 0.4, the Expected Readmissions Rate, 0.2159, generates the most net income at 0.0421. For such hospital, the penalty savings from reducing readmissions to the Expected Readmissions Rate outweigh the profit loss and the hospital is incentivized to reduce. We note this region as the *Incentive Compatible Region*. This is the window where the HRRP successfully incentivizes hospitals to reduce readmissions.

Finally in the last region, the right of the three, if the hospitals performance readmission rate (Risk-adjusted Predicted Readmissions) exceeds 0.4718, reducing readmissions leads to more profit loss than penalty savings. Hence, the hospital is not incentivized to reduce readmissions and the optimal strategy is to remain at the current readmissions rate. We note this region as the *Non-Incentive Compatible Region*.

2.3.2. All Conditions Applied Readmissions: Incentivizing Hospitals Without Incentive Compatible Region Under Current HRRP Structure

In the previous section, we have observed that there exists a window of readmission rates where a hospital will be incentivized to reduce readmissions. However, the existence of this window is not guaranteed for all hospitals. In the following example, we analyze the hospital with the maximum operating margin in our data set which has the following parameters: a lower fraction of inpatient revenue received from Medicare beneficiaries of 29.1% compared to the prior example; a fraction of Medicare revenue from the applicable conditions of 9.1%, as opposed to 13.3%; a higher operating margin of 19.3%; and a slightly lower Expected Readmissions Rate of 0.2061. The Floor Adjustment Factor (penalty cap) remains constant at 0.97.

Figure 2.3 depicts the hospitals normalized net income and inpatient profit as a function of its readmission rate. Normalized inpatient profit and net income under the current HRRP penalty structure are represented by the dash-dot line and solid line respectively. In contrast to Figure 2.2, the net income at any readmission rate above the Expected Readmissions Rate is larger than the net income achieved at the Expected Readmissions Rate; therefore, the hospital is not incentivized to reduce readmissions over any region.

For the HRRP to successfully incentivize hospitals to reduce readmissions, the net income achieved at readmission rates above the Expected Readmissions Rate should be less

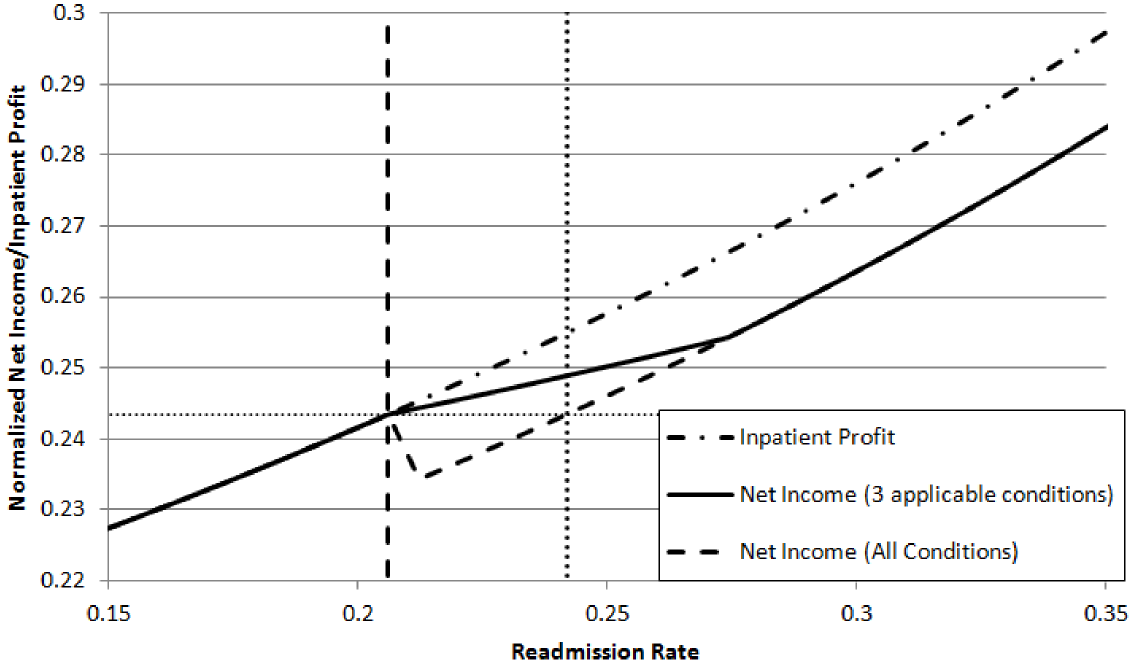


Figure 2.3. HRRP incentive compatibility: Hospital with large operating margin

than the net income achieved at the Expected Readmissions Rate. To accomplish the desired income loss required for incentivization, the marginal penalty amount should exceed the marginal inpatient profit once the hospitals readmission rate exceeds the Expected Readmissions Rate. In the current HRRP penalty structure, as we cannot artificially increase the fraction of Medicare revenue contributed by the current applied conditions by either increasing the amount revenue paid or incidence of the conditions, the only other option to increase the fraction of applicable Medicare revenue is increasing the number of applicable conditions. The counterfactual condition where the HRRP penalty is applicable to all conditions is represented by the dash line in Figure 2.3. Since we do not have the data for Expected Readmissions Rate for all conditions applied readmissions, we assume it to be constant with that of the three applicable conditions, 0.2061. Under the

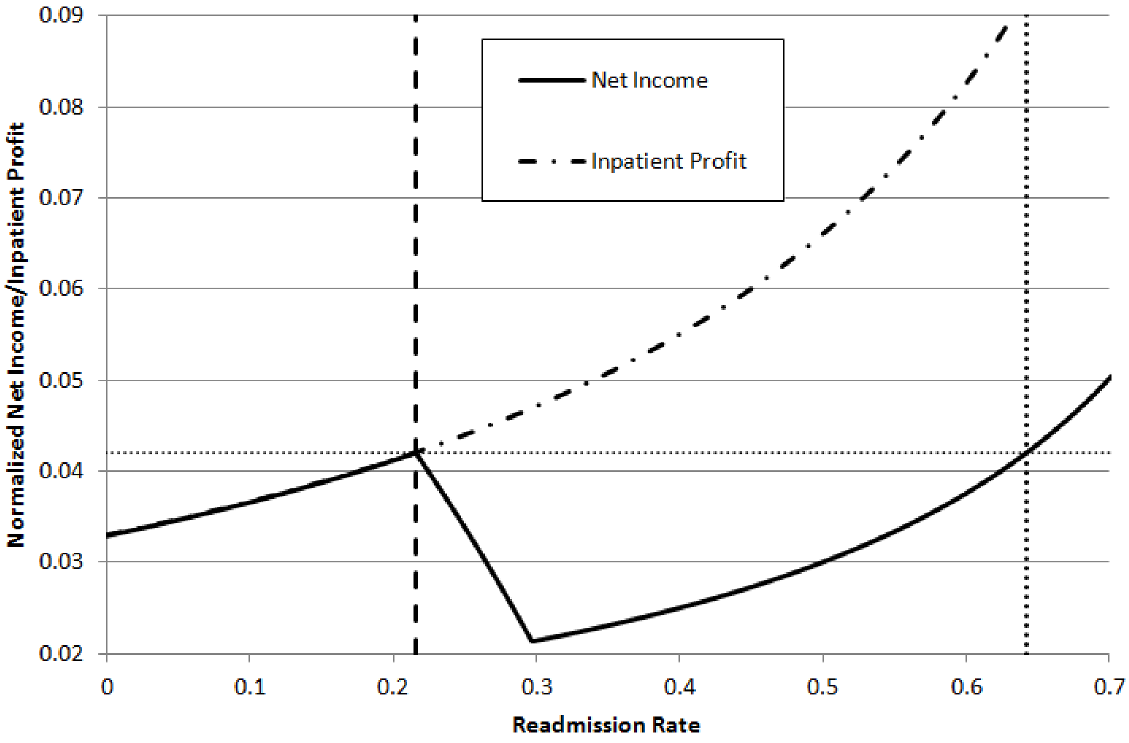


Figure 2.4. Wider incentive compatible region: Average hospital with Floor Adjustment Factor 0.95

all conditions applied penalty, the marginal penalty exceeds the marginal profit, thus the net income decreases as the readmission rate increases from the Expected Readmission Rate. In this example, the inclusion of all conditions applied to the HRRP penalty creates an incentive compatible region of [0.2061, 0.2419].

2.3.3. Reducing Floor Adjustment Factor (Increasing Penalty Cap) Widens The Incentive Compatible Region

The HRRP penalty has a gradually decreasing Floor Adjustment Factor and the impact of the Floor Adjustment Factor on the hospitals incentives is of interest from a policy makers perspective. For a hospital that has an incentive compatible region, decreasing the

Floor Adjustment Factor will widen the region, which effectively increases the likelihood of the hospital having its Risk-adjusted Predicted Readmissions within the region and being incentivized to reduce. In a counterfactual analysis, we take the same hospital from Figure 2.2 but with a 0.95 Floor Adjustment Factor instead of 0.97. In contrast to the incentive compatible region in Figure 2.2, the region widens to (0.2159, 0.6424] (Figure 2.4). From the program's perspective, the wider incentive compatible region is likely to capture more hospitals and incentivize them to reduce readmissions.

2.3.4. Simulation On The Effectiveness Of HRRP

If hospitals are maximizing operating margin net income under the current HRRP structure (Floor Adjustment Factor of 0.98 in FY 2014 and 0.97 in FY 2015 and beyond), 61.2 percent are financially incentivized to reduce readmissions while 5.5 percent are not despite incurring penalty (Table 2.2). If we look at the hospitals specific parameters, the hospitals that elect to not reduce are not incentivized because they lack an incentive compatible region. This is also evident from the sensitivity analysis (Table 2.2) where we widen the incentive compatible region by reducing the minimum Floor Adjustment Factor (maximum penalty cap) beyond the current 0.97 (3 percent) level. For instance, when the minimum Floor Adjustment Factor is set at 0.95, it decreases by 0.01 each year from 0.98 in FY 2014 to 0.95 in FY 2017 and remains at 0.95 afterwards. Even in the extreme case of a minimum Floor Adjustment Factor of 0.92, these hospitals remain unincentivized.

As a solution to incentivize hospitals without an incentive compatible region under the current structure, we analyze a penalty applied to readmissions associated with all

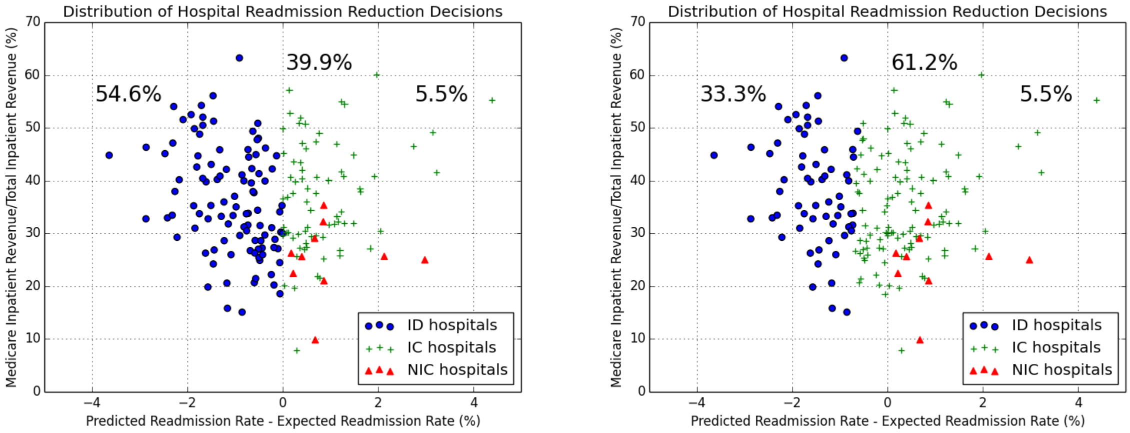
Table 2.2. HRRP effectiveness simulation: Percentage of 183 sample California hospitals by incentive compatibility

Net income (operating margin) maximizer						
Minimum Floor Adjustment Factor	3 applicable conditions			All conditions applied		
	Indiff	Inc Comp	Non-Inc Comp	Indiff	Inc Comp	Non-Inc Comp
0.97	33.3%	61.2%	5.5%	24.6%	75.4%	0.0%
0.96	33.3%	61.2%	5.5%	24.6%	75.4%	0.0%
0.95	33.3%	61.2%	5.5%	24.6%	75.4%	0.0%
0.94	33.3%	61.2%	5.5%	24.6%	75.4%	0.0%
0.93	33.3%	61.2%	5.5%	24.6%	75.4%	0.0%
0.92	33.3%	61.2%	5.5%	24.6%	75.4%	0.0%
Net income (patient margin) maximizer						
Minimum Floor Adjustment Factor	3 applicable conditions			All conditions applied		
	Indiff	Inc Comp	Non-Inc Comp	Indiff	Inc Comp	Non-Inc Comp
0.97	53.6%	3.8%	42.6%	35.5%	60.7%	3.8%
0.96	53.6%	3.8%	42.6%	35.5%	60.7%	3.8%
0.95	53.6%	3.8%	42.6%	26.2%	72.7%	1.1%
0.94	53.6%	3.8%	42.6%	26.2%	72.7%	1.1%
0.93	53.6%	3.8%	42.6%	25.7%	73.8%	0.5%
0.92	53.6%	3.8%	42.6%	25.7%	73.8%	0.5%
Net revenue maximizer						
Minimum Floor Adjustment Factor	3 applicable conditions			All conditions applied		
	Indiff	Inc Comp	Non-Inc Comp	Indiff	Inc Comp	Non-Inc Comp
0.97	54.6%	0.0%	45.4%	51.4%	24.0%	24.6%
0.96	54.6%	0.0%	45.4%	49.7%	30.1%	20.2%
0.95	54.6%	0.0%	45.4%	47.0%	36.6%	16.4%
0.94	54.6%	0.0%	45.4%	47.0%	38.8%	14.2%
0.93	54.6%	0.0%	45.4%	47.0%	39.3%	13.7%
0.92	54.6%	0.0%	45.4%	45.9%	41.0%	13.1%

conditions. Under the all-condition-applied readmissions penalty, the percentage of incentivized hospitals increase to 75.4 and no hospital remains unincentivized while incurring the penalty. The percentage of indifferent hospitals decreased from 33.3 to 24.6. As the all-condition-applied penalty incentivizes the previously unincentivized hospitals to reduce, the reduction in readmissions at those hospitals reduces the anticipated Risk-adjusted Expected Readmissions in future years for other hospitals as well. Hence, the lower bound of the incentive compatible region shifts downward as the Risk-adjusted Expected Readmissions decreases. In the all-conditions-applied readmissions penalty analysis, reducing the minimum Floor Adjustment Factor does not increase the number of incentivized hospitals. This suggests that the 0.97 Floor Adjustment Factor captures all of the potentially incentivized hospitals in the sample.

If hospitals were to maximize patient margin net income instead of operating margin net income, i.e., patients are viewed to be more profitable, only 3.8 percent of the hospitals are incentivized to reduce readmissions and 42.6 percent prefer to not reduce readmissions despite incurring penalty. If the objectives were to maximize net revenue (inpatient revenue HRRP penalty) which represents the highest possible patient profitability, none of the hospitals are incentivized to reduce readmissions and all hospitals incurring penalty (45.4 percent) maximize their net revenue by maintaining their current readmission performance. A larger financial return on readmitted patients translates to less ability for the current HRRP penalty to incentivize hospitals to reduce readmissions.

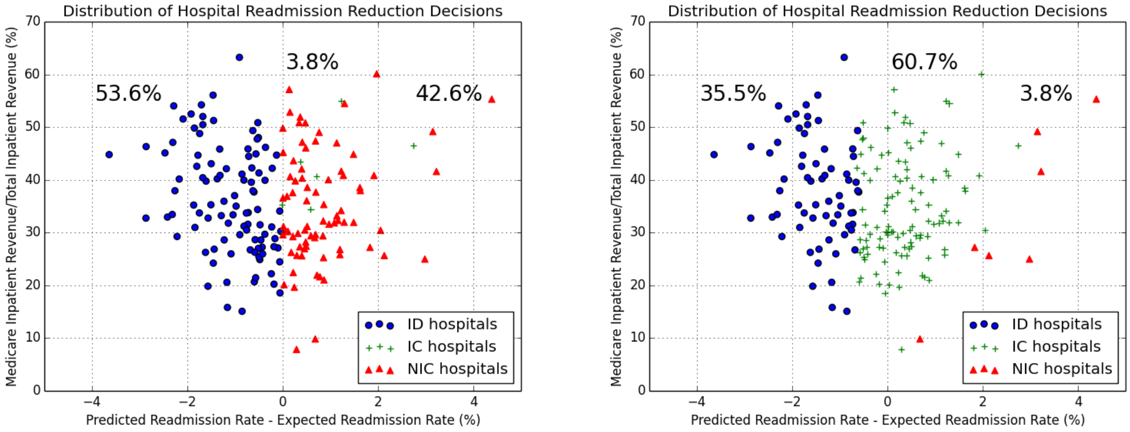
Under the all-conditions-applicable readmissions penalty and minimum Floor Adjustment Factor of 0.97, 60.7 percent of the patient margin net income maximizers and 24.0 percent of the revenue maximizers are incentivized to reduce readmissions, but both are short from operating margin net income maximizers (75.4 percent). As we decrease the



(a) Without competition

(b) With competition

Figure 2.5. Competition effect on incentive compatibility of 183 sample California hospitals: 0.97 minimum Floor Adjustment Factor, 3 applicable conditions, operating margin net income maximizers

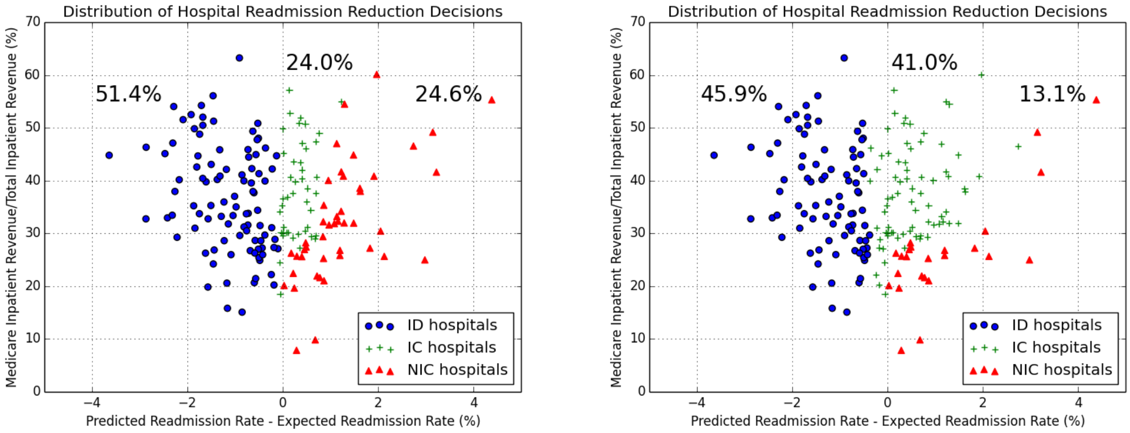


(a) Three applicable conditions

(b) All conditions applied

Figure 2.6. All conditions applied readmissions penalty effect on incentive compatibility of 183 sample California hospitals: 0.97 minimum Floor Adjustment Factor, patient margin net income maximizers

minimum Floor Adjustment Factor, the percentage of incentive compatible and non-incentive compatible hospitals increases and decreases respectively for both cases.



(a) Minimum Floor Adjustment Factor: 0.97 (b) Minimum Floor Adjustment Factor: 0.92

Figure 2.7. Minimum Floor Adjustment Factor effect on incentive compatibility of 183 sample California hospitals: All conditions applied penalty, net revenue maximizers

Figures 2.5, 2.6, 2.7 provide a graphical view of individual hospitals incentive compatibility. Indifferent (ID), Incentive Compatible (IC), and Non-Incentive Compatible (NIC) hospitals are identified by a circle, cross, and triangle respectively for various model parameters. The vertical axes represent the fraction of total inpatient revenue contributed by Medicare beneficiaries. The horizontal axes represent the difference between the Predicted and Expected Readmission Rates which are weighted averages of the Risk-adjusted Predicted and Expected Readmissions. The higher the value is the larger the gap between the Predicted and Expected Readmission Rate is which indicates a poorer readmission performance. Each data point on the plots represents one of the 183 sample California hospitals.

Figure 2.5 visualizes the effect of strategic gaming between hospitals. Figure 2.5a shows when the hospitals do not consider others reduction effort. In this case, hospitals that are performing better than their target (those with a negative predicted minus expected readmission rate) will not reduce their readmissions. They neglect a decrease in their future target readmission rates due to others reduction effort. However, if hospitals consider other hospitals reduction effort and their effect on their own expected readmission rate in future years, they will react by anticipating lower expected readmission rates. This is shown in Figure 2.5b where hospitals with a negative but close to zero (predicted readmission rate − expected readmission rate) value become incentive compatible. If hospitals are maximizing operating margin net income as in here, the competition effect induced by the relative performance measure structure financially incentivizes an additional 21.3 percent of hospitals, or increase from 39.9 percent to 61.2 percent, to reduce readmissions.

Figure 2.6 shows the effect of increasing the number of applied conditions from the initial three to all conditions when hospitals are maximizing patient margin net income under the current HRRP structure (Floor Adjustment Factor of 0.98 in FY 2014 and 0.97 in FY 2015 and beyond). AS noted in Subsection 2.3.2, all conditions applied penalty enables HRRP to incentivize hospitals that did not have an incentive compatible region before and significantly improves the HRRP effectiveness. Notably, the fraction of incentive compatible hospitals increases from 3.8 percent to 60.7 percent.

Even with the penalty applied to all conditions and a minimum Floor Adjustment Factor of 0.97, the program may not be fully effective as some net revenue maximizing hospitals are not incentivized because they are outside of the incentive compatible region (Figure 2.7a). While hospitals performing worse but close to their target are incentivized to reduce readmissions, ironically, those that perform far worse than the target are not incentivized to reduce. If the later were to reduce readmissions, loss in revenue due to reduced patient volume outweighs the savings in HRRP penalty. This effect becomes more substantial as the fraction of revenue from Medicare beneficiaries decreases (Figure 2.7a).

As discussed in Subsection 2.3.3, reducing the minimum Floor Adjustment Factor may incentivize hospitals outside of the incentive compatible region. Figure 2.7b shows the case with the minimum Floor Adjustment Factor reduced to 0.92 so that the Floor Adjustment Factor decreases by 0.01 each year in the time period until 2020. We observe 17 percent more hospitals are incentivized to reduce. However, the worst performers in terms of readmissions, hospitals on the far right of the plot, are still not incentivized to reduce readmissions. This pattern is observed in Figure 2.6b as well. Under the current HRRP penalty structure, an extremely low Floor Adjustment Factor will be required to incentivize them. This shows that the current penalty structure fails to some degree in protecting patients with the worst quality of care.

It is important to notice that, in general, the hospitals with a small fraction of revenue from Medicare beneficiaries, notably below 30 percent, are not incentivized to reduce readmissions (Figures 2.5, 2.6, 2.7). Since the revenue contribution from Medicare beneficiaries is small the revenue at risk for penalty is negligible compared to the total revenue. Policy makers need to be aware of the significance of this effect, especially if they expect to broaden the effectiveness of HRRP among non-Medicare heavy hospitals.

2.4. Discussion

In this manuscript we linked readmissions performance to hospital financials under the current HRRP penalty structure through a financial model of hospital admissions and identified the opposing effects of reducing readmissions on hospital net income: profit loss due to reduced patient volume and savings by avoiding the HRRP penalty. Our model predicts that not all hospitals are financially incentivized to reduce readmissions and for others there exists a specific window of readmission rate where hospitals are incentivized to reduce.

Our model demonstrates hospitals lack incentives to reduce readmissions for two reasons: the readmission rate is outside of the incentive compatible region or the incentive compatible region does not exist. For hospitals outside the incentive compatible region a potential solution is to reduce the minimum Floor Adjustment Factor which effectively widens the incentive compatible region. By increasing the maximal penalty under the HRRP, hospitals also stand to lose more revenue. For certain hospitals which are resource poor, such as safety-net hospitals, increasing the maximal penalty may have unintended consequences which would need to be addressed before implementing this policy change.

For hospitals without an incentive compatible region, increasing the number of conditions to which the readmission penalty is applied would create incentive compatible regions. Moreover, for these hospitals, reducing the minimum Floor Adjustment Factor will not have an impact on their incentive compatibility as there is no incentive compatible region to widen. Policymakers should note that this would affect the hospitals even more disproportionately as the hospitals incentivized even before the reduction in the minimum Floor Adjustment Factor will be punished more severely by the reduction while the unincentivized hospitals will not be affected whatsoever.

Our results support the CMS's decision to expand applicable conditions for FY 2015 to include patients admitted for an acute exacerbation of chronic obstructive pulmonary disease, elective total hip arthroplasty, and total knee arthroplasty (CMSb). CMS is currently considering expanding the number of applicable conditions to all-cause (all condition) readmission. In support of this effort, the National Quality Forum has already approved a "Hospital-Wide All-Cause Unplanned Readmission Measure" for this purpose (National Quality Forum). By expanding to an all-cause readmission metric, CMS will further incentivize hospitals to develop more robust efforts to prevent readmission among all their inpatients as opposed to concentrating on a few conditions and to make sure hospitals are engaged to reducing readmission rates.

We find that the worst performers in terms of readmissions, are still not incentivized to reduce readmissions even under extremely low Floor Adjustment Factors and all conditions applied readmissions. Policymakers should be aware of the potential shortage of the current penalty structure and combine an objective measure to ensure that all hospitals are financially incentivezed to reduce readmissions.

On a positive note, we find that the strategic gaming between hospitals induced by the relative performance measure in the penalty structure has a substantial effect in financially incentivizing even more hospitals to reduce readmissions.

In addition to the limitations associated with the model assumptions, we acknowledge the limitation that we do not consider the cost of reducing readmissions. Depending on the degree of reduction, the cost may vary widely from high cost hospital-wide operational changes to cheaper and easily installable toolkits. We leave this component out of the model and focus on the direct consequence of readmissions on the hospitals financials. Incorporating any cost associated with the effort of reducing readmissions will disincentivize hospitals from such activity and reduce the effectiveness of the Hospital Readmissions Reduction Program.

2.5. Conclusion

Our financial model of hospital readmissions suggests that only hospitals within a window of readmission rate that is unique to each hospital and depends on its own characteristics are financially incentivized to reduce readmissions as the Hospital Readmissions Reduction Program intends. Especially, the program fails to incentivize hospitals (1) with small fraction of Revenue contributed by Medicare beneficiaries, (2) with small fraction of Medicare revenue contributed by the applicable conditions, (3) with Risk-Adjusted Predicted Readmissions much larger than their Risk-Adjusted Expected Readmissions, i.e., the worst performers in terms of readmissions, and (4) and hospitals with sufficiently large operating margins, i.e., highly profitable hospitals. Within the current penalty structure, increasing the penalty amount by expanding the penalty applied conditions may solve some problems with incentive misappropriation while decreasing the floor adjustment factor may solve others. But neither will solve the first.

References

G. Allon and A. Bassamboo. Cheap talk in operations: Role of intentional vagueness. In *Consumer-Driven Demand and Operations Management Models*, pages 3–36. Springer, 2009.

G. Allon, A. Bassamboo, and I. Gurvich. We will be right with you: Managing customer expectations with vague promises and cheap talk. *Operations Research*, 59(6):1382–1394, 2011.

G. Allon, S. Deo, and W. Lin. The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research*, 61(3):544–562, 2013.

F. Aminzadeh and W. B. Dalziel. Older adults in the emergency department: a systematic review of patterns of use, adverse outcomes, and effectiveness of interventions. *Annals of emergency medicine*, 39(3):238–247, 2002.

O. K. Asamoah, S. J. Weiss, A. A. Ernst, M. Richards, and D. P. Sklar. A novel diversion protocol dramatically reduces diversion hours. *The American Journal of Emergency Medicine*, 26 (6):670–675, 2008.

R. J. Batt and C. Terwiesch. Doctors under load: An empirical study of state-dependent service times in emergency care. *Working Paper, The Wharton School*, 2012.

C. F. Baum. Probexog-tobexog: Stata modules to test exogeneity in probit/tobit. Statistical Software Components, Boston College Department of Economics, Dec. 2007. URL `http://ideas.repec.org/c/boc/bocode/s401102.html`.

R. A. Berenson, R. A. Paulus, and N. S. Kalman. Medicare's readmissions-reduction programa positive alternative. *New England Journal of Medicine*, 366(15):1364–1366, 2012.

L. G. Burke, N. Joyce, W. E. Baker, P. D. Biddinger, K. Dyer, F. D. Friedman, J. Imperato, A. King, T. M. Maciejko, M. D. Pearlmutter, A. Sayah, R. Zane, and S. Epstein. The effect of an ambulance diversion ban on emergency department length of stay and ambulance turnaround time. *Annals of Emergency Medicine*, 2013.

G. A. Caplan, A. J. Williams, B. Daly, and K. Abraham. A randomized, controlled trial of comprehensive geriatric assessment and multidisciplinary intervention after discharge of elderly from the emergency department –the DEED II study. *Journal of the American Geriatrics Society*, 52(9):1417–1423, 2004.

L. Cappellari and S. P. Jenkins. Multivariate probit regression using simulated maximum likelihood. *The Stata Journal*, 3(3):278–294, 2003.

Centers for Medicare and Medicaid Services. Official hospital compare data: Hospital readmission reduction, a. URL `https://data.medicare.gov/Hospital-Compare/Hospital-Readmission-Reduction/9n3s-kdb3`. Accessed: 2013-05-25.

Centers for Medicare and Medicaid Services. FY 2015 IPPS Final Rule: Hospital Readmissions Reduction Program Supplemental Data File, b. URL `http://cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html`. Accessed: 2014-09-17.

Centers for Medicare and Medicaid Services. Instructions for replicating your excess readmisson ration results, c. URL `http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1219069855273`. Accessed: 2013-05-25.

Centers for Medicare and Medicaid Services. Acute inpatient pps: Readmissions reduction program, 2013. URL `http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html`.

H. Centers for Medicare and Medicaid Services et al. Medicare program; hospital inpatient prospective payment systems for acute care hospitals and the long-term care hospital prospective payment system and fiscal year 2013 rates; hospitals' resident caps for graduate medical education payment purposes; quality reporting requirements for specific providers and for ambulatory surgical centers. final rule. *Federal register*, 77(170):53257, 2012.

S. Deo and I. Gurvich. Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science*, 57(7):1300–1319, 2011.

H. Do and M. Shunko. Pareto improving coordination policies in queueing systems: Application to flow control in emergency medical services. *Available at SSRN 2351965*, 2013.

N. Duan, W. G. Manning, C. N. Morris, and J. P. Newhouse. Choosing between the sample-selection model and the multi-part model. *Journal of Business & Economic Statistics*, 2 (3):283–289, 1984.

M. Eckstein and L. S. Chan. The effect of emergency department crowding on paramedic ambulance availability. *Annals of emergency medicine*, 43(1):100–105, 2004.

D. Fatovich, Y. Nagree, and P. Sprivulis. Access block causes emergency department overcrowding and ambulance diversion in Perth, Western Australia. *Emergency Medicine Journal*, 22(5):351–354, 2005.

R. Hagtvedt, M. Ferguson, P. Griffin, G. T. Jones, and P. Keskinocak. Cooperative strategies to reduce ambulance diversion. In *Proc. 2009 Winter Simulation Conf. (WSC), Austin, TX*, pages 1861–1874. IEEE, 2009.

N. R. Hoot and D. Aronsky. Systematic review of emergency department crowding: causes, effects, and solutions. *Annals of emergency Medicine*, 52(2):126–136, 2008.

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

K. E. Joynt and A. K. Jha. Who has higher readmission rates for heart failure, and why? implications for efforts to improve care using financial incentives. *Circulation: Cardiovascular Quality and Outcomes*, 4(1):53–59, 2011.

K. E. Joynt and A. K. Jha. Thirty-day readmissionstruth and consequences. *New England Journal of Medicine*, 366(15):1366–1369, 2012.

D. S. Kc and C. Terwiesch. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):1486–1498, 2009.

R. J. Lagoe, J. C. Kohlbrenner, L. D. Hall, M. Roizen, P. A. Nadle, and R. C. Hunt. Reducing ambulance diversion: A multihospital approach. *Prehospital Emergency Care*, 7(1):99–108, 2003.

M. L. McCarthy, A. D. Shore, G. Li, J. New, J. J. Scheulen, N. Tang, R. Collela, and G. D. Kelen. Likelihood of reroute during ambulance diversion periods in central maryland. *Prehospital Emergency Care*, 11(4):408–415, 2007.

K. J. McConnell, C. F. Richards, M. Daya, S. L. Bernell, C. C. Weathers, and R. A. Lowe. Effect of increased icu capacity on emergency department length of stay and ambulance diversion. *Annals of emergency medicine*, 45(5):471–478, 2005.

N. Mihal and R. Moilanen. When emergency rooms close: Ambulance diversion in the West San Fernando Valley. 2005.

National Quality Forum. Hospital-Wide All-Cause Unplanned Readmission Measure. URL http://www.qualityforum.org/Projects/NQF_All-Cause_Readmissions_Project.aspx. Accessed: 2014-09-17.

P. B. Patel, R. W. Derlet, D. R. Vinson, M. Williams, and J. Wills. Ambulance diversion reduction: the Sacramento solution. *American Journal of Emergency Medicine*, 24(2): 206–213, 2006.

J. C. Pham, R. Patel, M. G. Millin, T. D. Kirsch, and A. Chanmugam. The effects of ambulance diversion: a comprehensive review. *Academic Emergency Medicine*, 13(11):1220–1227, 2006.

S. R. Pitts, J. M. Pines, M. T. Handrigan, and A. L. Kellermann. National trends in emergency department occupancy, 2001 to 2008: Effect of inpatient admissions versus emergency department practice intensity. *Annals of Emergency Medicine*, 60:679–686, 2012.

A. Ramirez-Nafarrate, A. Baykal Hafizoglu, E. S. Gel, and J. W. Fowler. Optimal control policies for ambulance diversion. *European Journal of Operational Research*, 2013.

J. Rau. Medicare to penalize 2,217 hospitals for excess readmissions, 2012. URL http://www.kaiserhealthnews.org/stories/2012/august/13/medicare-hospitals-readmissions-penalties.aspx.

J. Rau. Armed with bigger fines, medicare to punish 2,225 hospitals for excess readmissions, 2013. URL http://www.kaiserhealthnews.org/Stories/2013/August/02/readmission-penalties-medicare-hospitals-year-two.aspx.

S. Schneider, F. Zwemer, A. Doniger, R. Dick, T. Czapranski, and E. Davis. Rochester, New York a decade of emergency department overcrowding. *Academic Emergency Medicine*, 8 (11):1044–1050, 2001.

M. J. Schull, L. J. Morrison, M. Vermeulen, and D. A. Redelmeier. Emergency department overcrowding and ambulance transport delays for patients with chest pain. *Canadian Medical Association Journal*, 168(3):277–283, 2003.

Y. Shen and R. Y. Hsia. Association between ambulance diversion and survival among patients with acute myocardial infarction. *JAMA: The Journal of the American Medical Association*, 305(23):2440, 2011.

R. P. Shenoi, L. Ma, J. Jones, M. Frost, M. Seo, and C. E. Begley. Ambulance diversion as a proxy for emergency department crowding: The effect on pediatric mortality in a metropolitan area. *Academic Emergency Medicine*, 16(2):116–123, 2009.

R. J. Smith and R. W. Blundell. An exogeneity test for a simultaneous equation tobit model with an application to labor supply. *Econometrica: Journal of the Econometric Society*, pages 679–685, 1986.

State of California Office of Statewide Health Planning & Development. Annual financial data. URL `http://www.oshpd.ca.gov/HID/Products/Hospitals/AnnFinanData/CmplteDataSet/index.asp`. Accessed: 2013-05-24.

B. C. Sun, S. A. Mohanty, R. Weiss, R. Tadeo, M. Hasbrouck, W. Koenig, C. Meyer, and S. Asch. Effects of hospital closures and hospital characteristics on emergency department ambulance diversion, Los Angeles County, 1998 to 2004. *Annals of Emergency Medicine*, 47(4):309–316, 2006.

G. M. Vilke, E. M. Castillo, M. A. Metz, L. U. RAY, P. A. Murrin, R. Lev, and T. C. Chan. Community trial to decrease ambulance diversion hours: The San Diego county patient destination trial. *Annals of Emergency Medicine*, 44(4):295–303, 2004.

J. M. Wooldridge. Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. *Working Paper, Michigan State University*, 2012.