Hospital Readmissions Reduction Program:
An Economic and Operational Analysis

Dennis J. Zhang, Itai Gurvich, Jan A. Van Mieghem, Eric Park
Kellogg School of Management, Northwestern University, Evanston, IL 60208,

Robert S. Young, Mark V. Williams
Feinberg School of Medicine, Northwestern University, Chicago, IL 60611,

May 31, 2014

The Hospital Readmissions Reduction Program (HRRP), a part of the US Patient Protection and Affordable Care Act, requires the Centers for Medicare and Medicaid Services to penalize hospitals with excess readmissions. We take an economic and operational (patient flow) perspective to analyze the effectiveness of this policy in encouraging hospitals to reduce readmissions. We introduce a single-hospital model to capture the dependence of a hospital's readmission-reduction decision on various hospital characteristics. We derive comparative statics that predict how changes in hospital characteristics impact the hospital's readmission-reduction decision. We then proceed to develop a game-theoretic model that captures the competition between hospitals introduced by the HRRP policy's benchmarking mechanism. We provide bounds that apply to any equilibrium of the game and show that the comparative statics derived from the single-hospital model remain valid after the introduction of competition. Importantly, the comparison of the single-hospitals and multi-hospital models shows that, while the competition among hospitals often encourages more hospitals to reduce readmissions, it can only increase the number of "worst offenders," which are hospitals that prefer paying penalties over reducing readmissions in any equilibrium. We calibrate our model with a dataset of hospitals in California which allows us to quantify the results and insights derived from the model. Last, we validate our model with recent hospitals' performance data collected since the policy was implemented.

Key words: Healthcare Operations, Public Policy.

History: This paper was first submitted on May 31st, 2014.

1. Introduction

According to the Medicare Payment Advisory commission (MedPAC) (MedPAC 2007), nearly a fifth of Medicare beneficiaries that are discharged from a hospital are readmitted within 30 days. Re-hospitalization of a patient shortly after the initial discharge is often viewed as a sign of poor quality of care (Ashton et al. 1997 and Gwadry-Sridhar et al. 2004). Past research has shown that hospital readmissions are often costly (Jencks et al. 2009 and MedPAC 2007) and avoidable through simple process changes (Hansen et al. 2013). The Centers for Medicare and Medicaid
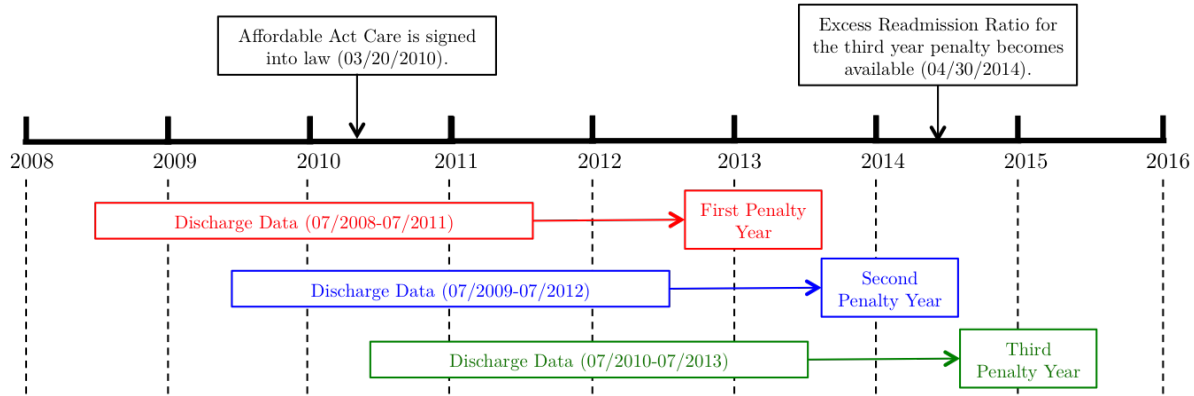
Figure 1: Timeline of the Hospital Readmissions Reduction Program

Services (CMS) estimated that a 20% reduction in hospital readmission rates could save 5 billion dollars by the end of fiscal year 2013 (Mor et al. 2010).

In response to the increasing costs associated with readmissions, the Hospital Readmissions Reduction Program (HRRP) was implemented by CMS on October 1, 2012. The program penalizes medicare payments to hospitals with high 30-day readmission rates for acute myocardial infarction (AMI), heart failure (HF), and pneumonia (PN). Two additional diseases will be added to the policy starting 2015. Using historical data, the CMS determines, for each hospital in the Inpatient Prospective Payment System (IPPS), whether its readmission rates are higher than they should be given the hospital's case mix. The CMS model determines the targets by benchmarking hospitals against peers with similar case mix.

Figure 1 gives a detailed view of the policy's timeline. For fiscal year 2013, CMS uses benchmarking data from July 2008 to July 2011 and, under current legislation, hospitals with higher-than-expected readmission rates have their total Medicare reimbursement for fiscal year 2013 cut by up to 1%. This maximum penalty cap is expected to increase to 2% in 2014 and to 3% in 2015.

Two common criticisms of the HRRP policy are: (i) hospitals are not the appropriate entities to be held accountable for readmissions, since some causes of readmissions are outside the control of hospitals. Only a small fraction of readmissions is claimed to be preventable by measures that hospitals directly control (van Walraven et al. 2011); and (ii) the readmission rate of a hospital is not a good proxy for its quality of care. There is empirical evidence that people who have severe illness or come from a disadvantaged socioeconomic status are at particularly high risk for readmission (Joynt et al. 2011).

Supporters of the HRRP policy point to the large number of patients whose discharge is fraught with poor communication, ineffective medication management, and inadequate hand-offs to the

primary care physicians or nursing homes. A report by the Medicare Payment Advisory Commission supports the HRRP policy in estimating a small but significant decrease in national rates of readmission for all causes from 15.6% in 2009 to 15.3% after the introduction of the HRRP (HealthCare.gov 2011).

This paper does not argue with the appropriateness of readmission as a quality-of-care metric. We also do not consider mechanism design questions. Rather, we take HRRP as a given government program and analyze its effectiveness in reducing readmissions. We take an economic and operational perspective to ask a simple question: assuming that hospitals are self-interested operating-margin maximizers and are strategically forward-looking, does the HRRP policy provide economic incentives for a hospital to reduce its readmissions? What are the characteristics of hospitals that prefer paying penalties over reducing readmissions? And how does the HRRP benchmarking (and the competition it induces) affect who these hospitals are?

Readmission-reduction decisions present hospitals with trade-offs between several cost and revenue drivers: (i) The reduction of the penalty due to readmission improvements: 2, 217 hospitals nationwide incurred more than $300 million HRRP penalties in the fiscal year 2013 (Fontanarosa and McNutt (2013)). Many hospitals incurred hundreds of thousands of dollars in penalties, while the "worst-offenders" incurred millions of dollars in penalties. These amounts could be tripled by 2015, when the maximum penalty cap is expected to increase to 3%. (ii) Contribution loss due to readmission reductions: taking a pure revenue perspective readmissions generate revenues. Specifically, if a non-negligible portion of the hospital's patients are covered under a pay-per-case insurance scheme, readmissions may account for a significant proportion of the hospital's contribution margin. (iii) Process-improvement cost: Reducing readmissions may involve costly long-term process changes.

CMS determines the expected readmission rate for each hospital using discharge-level data for IPPS hospitals from the previous three years. A logistic Hierarchical Generalized Linear Model (HGLM) is used to determine the national average performance conditioning on the case mix of the particular hospital – we refer to this conditional average as the *CMS expected readmission rate* for that hospital. If the hospital's predicted readmission rate, based on the hospital's actual performance, for the next year is greater than its CMS expected readmission rate, the hospital incurs a penalty up to the maximum penalty cap – currently 1%. For fiscal year 2014, the CMS expected readmission rate for each hospital is based on data from July 1st, 2009 to June 30th, 2012. This penalty mechanism inevitably introduces game theoretical elements into each hospital's decision making since one hospital's future penalty is determined not only by its own actions, but also by the performance of all other (similar) hospitals.

We use analytic modeling to study the impact of the HRRP policy on a hospital's long-term readmission reduction efforts. We develop a theoretical model which captures the patient flow from readmissions, the essential financial driving forces in each hospital's decision, and the game-theoretical nature of the policy design. Our stylized operational and financial model of the individual hospital (see Section 2) captures the three financial considerations mentioned above: the savings in penalty, the loss in contribution, and the cost of reducing readmissions. We allow for a flexible specification of the cost of reducing readmissions and capture the possible reduced contribution of readmitted patients vs. first-time patients.

One may take initially the view that hospitals are "naive" and do not take into account how the CMS-targets move with decisions made by other hospitals. Our single-hospital model then allows to capture characteristics of hospitals that are incentivized to reduce their readmission rates in response to the policy.

If hospitals do, however, take into account the decisions of their peers, one would want to make sure that the insights obtained from the single-hospital model persist when hospitals indeed consider this strategic interaction. We then proceed to introduce a static game where hospitals determine their readmission reduction efforts while taking other hospitals' actions into consideration. We show that existence and uniqueness of pure strategy equilibrium is not guaranteed but we are able to identify global bounds that apply to all equilibria of the game and provide meaningful insights into its outcomes. Specifically, we derive a lower bound on the number of hospitals that are not incentivized by HRRP–this, in particular, provides a bound on the policy's effectiveness. We further use the bound to show that, unexpectedly, competition through benchmarking can only increase the number of hospitals that prefer incurring penalties over reducing readmissions.

In Section 5 we apply our model to hospitals in California. We calibrate our model using the data, and report findings drawn from the dataset focusing on those insights pertaining to non-incentivized hospitals. We find that under various parameter choices (regardless of whether hospitals are short-sighted or forward looking) the percentage of hospitals that pay penalties but do not reduce readmissions in **any** equilibrium is non-negligible for all monitored diseases. Namely, there is a set of hospitals that are not incentivized by HRRP to improve their readmissions. Several other observations related to non-incentivized hospitals arise from our simulation study:

1. The effectiveness of the HRRP policy depends on various characteristics of a hospital: A hospital in an urban area, with greater competition and higher probability of patients being readmitted to a different hospital, has a greater financial incentive to reduce its readmission. Second, the current HRRP policy is not effective in inducing hospitals with poor performance since, for these hospitals, the cost of reducing readmissions is greater than the savings in penalties. Third, hospitals with low percentage of medicare revenue are less likely to reduce their readmissions.

Patients served by these hospitals are in relative disadvantage under the current structure of the policy. Forth, the higher the contribution margin of a hospital, the smaller the likelihood of the hospital to reduce its readmissions under the HRRP policy. This suggests that a better regulated payment system may be helpful in incentivizing hospitals to reduce readmissions.

2. The cost of process changes to reduce readmission plays an important role in the hospitals' responses to the HRRP policy. Consequently, research projects promoting simple (i.e, not costly) readmission reduction programs, such as BOOST (Hansen et al. 2013), can enhance the effectiveness of the HRRP.

3. The Medicare Payment Advisory Commission emphasizes the importance of the competition introduced by the HRRP benchmarking procedure (Glass et al. (2012)). It is therefore important to understand the hospitals' decisions when they consider the strategic interactions among them. We find that, while the competition induced by HRRP may indeed incentivize more hospitals to reduce readmissions relative to the individual hospital (no benchmarking) model, it can only increase the number of hospitals that prefer paying penalties to reducing readmissions.

The policy is still relatively new, and it is premature to draw definite conclusions about its long-run effectiveness. Nevertheless, in Section 5, we validate our model predictions about those non-incentivized hospitals by comparing the simulation results to the actual changes in hospitals readmission ratios between 2013 and 2015. This initial empirical evidence supports the predictive power of our model. Specifically: the hospitals that our model identifies as those that prefer paying penalties over reducing readmissions coincide with the actual hospitals that did not reduce readmissions and continue incurring penalties.

We conclude this introduction with a brief literature review. Much of the academic work in the medical literature focuses on understanding the causes of readmissions and proposing effective quality improvement programs to help hospitals reduce readmissions (Dharmarajan et al. 2013, Krumholz et al. 1997 and Stewart et al. 1999). Using 2003-2004 Medicare data, Jencks et. al. (Jencks et al. 2009) report the most frequent diagnoses for 30-day readmissions for 10 common conditions. Using national Medicare data from 2006 to 2008, Karen et al. (Joynt et al. 2011) examine 30-day readmissions for AMI, HF, and PN, and show that medicare patients from a poorer socioeconomic background have particularly high risk of being readmitted. The literature also demonstrates that simple quality-improvement programs – such as coaching the caregivers of chronically ill or older patients (Coleman et al. 2006), properly planning the discharge process (Naylor et al. 1999, Hansen et al. 2013), and conducting a nurse-directed multidisciplinary intervention (Rich et al. 1995) – can effectively reduce readmissions.

Since the introduction of the US Patient Protection and Affordable Care Act (ACA), in particular the HRRP policy, the medical literature studied the structure of this policy and its medical

effectiveness. Vaduganathan et al. 2013 question the validity of considering 30-day readmissions as a measure of one hospital's readmission conditions in the policy. Srivastava and Keren 2013 point out that the current policy does not cover pediatric hospitals and suggest the readmission quality measure for hospitalized children should focus on different conditions; Vashi et al. 2013 suggest that approximately 18% of hospitals discharges were followed by at least 1 hospital-based acute care encounter within 30 days, which suggests that 30-day readmissions do not necessarily reflect the quality of care in the hospital.

Here we take the HRRP as given and ask whether the HRRP policy financially incentivize hospitals to reduce their readmissions? Our research is, in turn, also related to the stream of literature in health economics, which studies moral hazard in hospitals' and physicians' behavior (Chiappori et al. 1998, Propper and Van Reenen 2010) and analyzes the effect of policy interventions (Cutler and Gruber 1996, Card et al. 2008).

2. Model

In this section, we introduce a stylized hospital-level patient flow model and link it to the contribution margin of a hospital, which sets the foundation for analyzing individual hospital's behavior in Section 3 and hospitals' joint equilibrium in Section 4.

2.1. Hospital Flow Model

Each hospital faces an exogenous arrival of patients for each disease i and each insurance type j, with a rate λ_{ij}^e. The HRRP distinguishes between Medicare patients, and patients with private insurance, Medicaid, military insurance and other insurance types. [1]

For every disease type i and insurance type j, the readmission rate is denoted as r_{ij}. A patient requiring readmission could either return to the hospital from which he/she was originally discharged, or visit a different hospital. We refer to the probability that a patient is readmitted to a different hospital as the *hospital-level readmission divergence probability* and denote it by d_h. Patients that were classified as belonging to group ij upon first visit may, upon readmission belong to a different disease group i'. The HRRP policy counts these towards the readmission in group ij. We call the probability that a patient is readmitted to a different disease group as *disease-level readmission divergence probability*, denoted by d_d. A flow diagram for patients with disease i and payment j in a hospital is shown in Figure 2, where $\lambda_{ij}^{d_h}$ is the rate of the readmitted patients coming from other hospitals, and $\lambda_{ij}^{d_d}$ is the rate of the patients readmitted form diseases other than i. Finally, λ_{ij} is the total throughput of patients in group ij.

[1] While some research claims that demand is not truly exogenous as patients can be induced to visit the hospitals (Acton 1975), recent econometric studies have shown that hospitals can hardly increase or decrease the incoming rate of patients (Dranove and Wehner 1994).
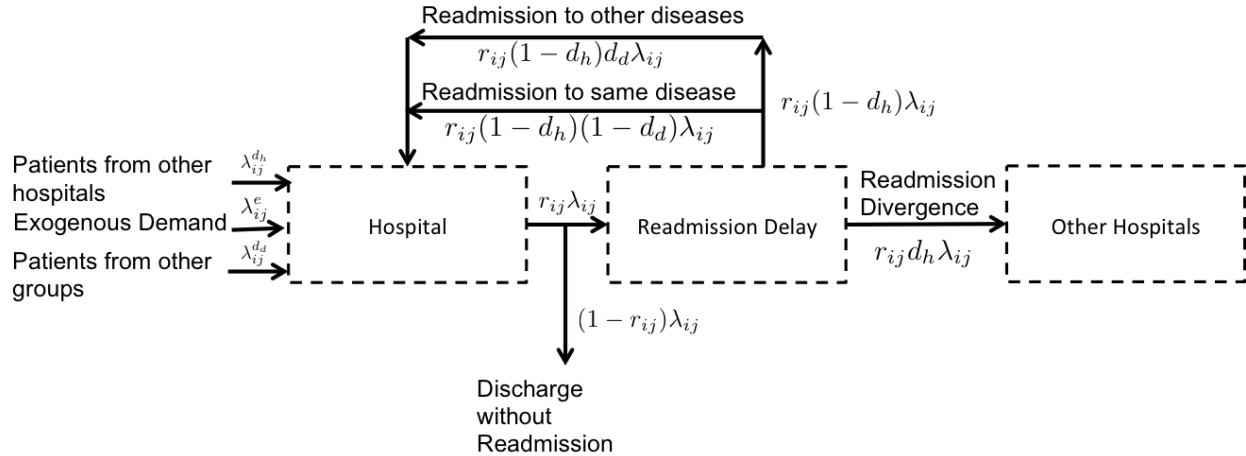
Figure 2: Operational flow of readmissions in a hospital

Each hospital receives a payment p_{ij} for treating a patient of type ij. Readmitted patients may be less profitable (Zook et al. 1980). We refer to the percentage that the contribution decreases per readmission as the *readmission loss factor* and we denote it by l. Thus, the contribution for the kth readmission is $l^k p_{ij}$.

Letting $\lambda_{ij}^a = \lambda_{ij}^e + \lambda_{ij}^{d_h} + \lambda_{ij}^{d_d}$ the total incoming rate for patient group ij, direct calculation yields that the combined arrival rate (exogenous patients plus readmitted patients) for disease ij satisfies:

$$\lambda_{ij} = \lambda_{ij}^a + r_{ij}(1-d_h)(1-d_d)\lambda_{ij}$$

so that

$$\lambda_{ij} = \frac{\lambda_{ij}^a}{1 - r_{ij}(1-d_d)(1-d_h)} \tag{1}$$

where λ_{ij}^a is the total throughput of the hospital. Adding the subscript h to denote a specific hospital, hospital h's revenue is:

$$\Pi_{ijh}^R(0) = \lambda_{ijh}^a p_{ijh}$$
$$\Pi_{ijh}^R(r_{ijh}) = \Pi_{ijh}^R(0)\frac{1}{1 - (1-l)(1-d_h)(1-d_d)r_{ijh}} \tag{2}$$
$$\Pi_h^R(r_h) = \sum_{ij}\Pi_{ijh}^R(r_{ijh}).$$

where, note, $\Pi_{ijh}(0)$ is the revenue from patients in group ij for hospital h if $r_{ijh} = 0$. We define *contribution margin*, Π_j^C, as the difference between the hospitals revenue and all the hospitalization variable labor and supply cost. We assume that the contribution margin is the same for one disease across all hospitals and is a fraction of C_m (for $C_m \leq 1$) of the total revenue, i.e,

$$\Pi_h^C(r_h) = C_m \Pi_h^R(r_h), \tag{3}$$

where the argument r_h captures the dependence of this contribution margin on the hospital's readmission rates.

2.2. Penalty Structure

In October 2012, CMS started penalizing Medicare payments to hospitals based on their excess readmission ratio for monitored diseases. The excess readmission of a hospital for each monitored disease is measured as a ratio of its *risk-adjusted predicted* readmission rate (r_{ijh} for patient with condition i and hospital h) and its *risk-adjusted expected* readmission rate (r_{ijh}^e for patient with condition i and hospital h). Here, "risk-adjusted" means that the estimated readmission rate for a given hospital is adjusted to the case mix of that hospital. Therefore, hospitals with more severe patients in their case mix, have higher risk-adjusted expected and predicted readmission rates. The risk adjustment prevents the discrimination of hospitals with more difficult patients.

CMS computes the expected and predicted readmission rates for every hospital by applying the HGLM model to discharge-level data. The detailed model is described in Appendix A. For each hospital h, the risk-adjusted predicted readmission rate, r_{ijh}, predicts the readmission rate of disease ij in hospital h (for $j = Med$, i.e, Medicare patients) in the following year conditional on its case mix remaining unchanged. Roughly speaking, the risk-adjusted expected readmission rate, r_{ijh}^e, is the readmission rate of the average hospital with the same case mix as hospital h. We will refer to these numbers as the CMS-predicted and CMS-expected readmission rates.

If the excess readmission ratio, r_{ijh}/r_{ijh}^e, for hospital h is greater than 1, CMS will impose a penalty equal to the sum of base operating diagnosis-related group (DRG) payments from Medicare patients with disease i multiplied by $\frac{r_{ih}}{r_{ih}^e} - 1$. However, the sum of all penalties is capped by a P_{cap} fraction of the total hospital's medicare revenue for the monitored diseases. Formally, the total penalty for hospital h and all monitored diseases is given by:

$$\mathbb{P}_h(r_h, r_h^e) = \min\left\{ \sum_{i,j=Med} \max\left(\frac{r_{ijh}}{r_{ijh}^e} - 1, 0 \right) \Pi_{ijh}^R, \; P_{cap} \sum_{i,j=Med} \Pi_{ijh}^R \right\} \tag{4}$$

3. Single-Hospital Model

To gain some structural insights, let us first focus on the case where there is a single group of patients (a single disease and a single insurance plan) so that we can drop the subscripts ij. Without loss of generality and to simplify the algebra, assume that the readmission divergence probability $d_h = 0$ and $d_d = 0$, and that the readmission loss factor is $l = 0$. The revenue, the contribution margin, and the penalty for hospital h with actual readmission rate r and risk-adjusted CMS-expected readmission ratio $\frac{r}{r^e}$ are then rewritten as:

$$\Pi_h^R(r) = \Pi_h^R(0)\frac{1}{1-r} = \lambda_h^a p_h \frac{1}{1-r},$$

$$\Pi_h^P(r) = C_m \Pi_h^R(r),$$

$$\mathbb{P}_h(r, r^e) = \phi_h^{med} \Pi_h^R(r) \min\left(\max\left(\frac{r}{r^e} - 1, 0\right), P_{cap}\right).$$

where ϕ_h^{med} is the percentage of the hospital h revenue that comes from Medicare patients.

We assume that hospitals are operating-margin maximizers. While nonprofit hospitals should not incentivize their management group to maximize any form of profit, past studies have shown that these hospitals behave as profit maximizers in a competitive market (Deneffe and Masson 2002). Also, we consider the *operating margin* as the contribution margin minus the penalty and the readmission reduction cost.

Reducing readmission may involve process changes (Naylor et al. 1999) or increases in staffing (Stewart et al. 1999). These are long-term commitments. Consequently, we assume that if a hospital reduces its readmission from r_0 to r, then an annual readmission-management cost $C(r_0, r)$ is added to the hospitals cost in each subsequent year. The function $C(\cdot, \cdot)$ is assumed to be continuous and bounded with a second derivative that satisfies $\left|\frac{\partial^2}{\partial x \partial x} C(y, x)\right| \leq \eta \frac{1}{(1-x)^3}$ for some constant η and $\forall x, r \in (0, 1)$. This technical assumption is satisfied, in particular, by functions of the form

$$C(y, x) = C_h^v(y - x)^\alpha + g(y),$$

for any function g and any $\alpha \geq 0$.

When a hospital with current readmission rate r_{h0} and CMS-expected readmission rate r_h^e decides to reduce readmissions to r_h, its operating margins for the subsequent year is given by:

$$
\begin{aligned}
R(r_{h0}, r_h, r_h^e) &= \Pi_h^P(r_h) - \mathbb{P}_h(r_h, r_h^e) - C(r_{h0}, r_h) \\
&= C_m \Pi_h^R(0) \frac{1}{1 - r_h}\left(1 - \phi_h^{med} \frac{1}{C_m} \min\left(\max\left(\frac{r}{r^e} - 1, 0\right), P_{cap}\right)\right) - C(r_{h0}, r_h).
\end{aligned}
\tag{5}
$$

The hospital then chooses a maximizer

$$r_h^*(r_{h0}, r_h^e) \in \arg\max_{x \leq r_{h0}} R(r_{h0}, x, r_h^e) \tag{6}$$

An implicit assumption here is that hospitals do not deliberately increase readmission rates. One could argue that this is reasonable for ethical reasons but it is also consistent with the spirit of our analysis that focuses on best case outcomes of the policy and tries to identify hospitals that "fall outside" of the policies effectiveness boundaries.

Our first result is that a hospital will have financial incentive to reduce its readmissions only when its readmission rate is contained in a certain interval, the width of which depends on the hospital's cost structure and other hospital parameters. This dependence provides some useful insights.
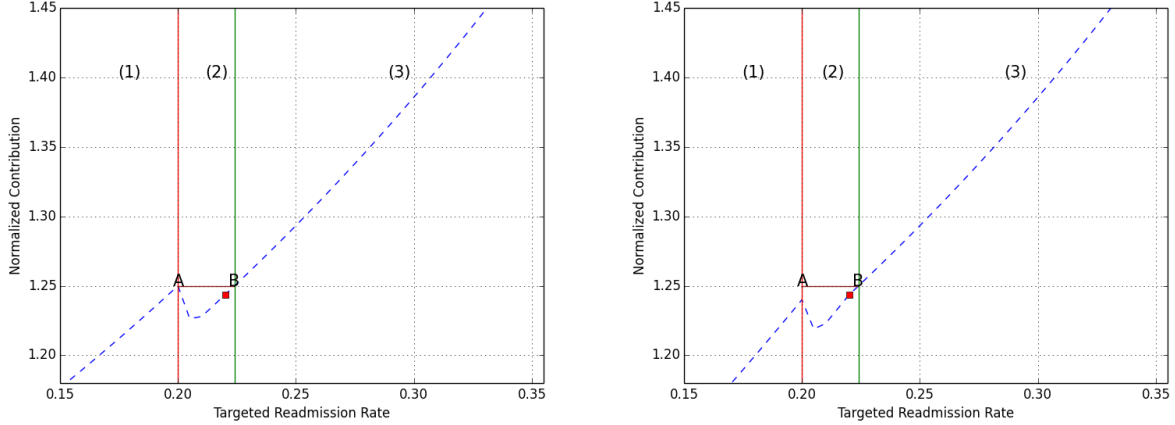
Figure 3: (LHS) No readmission reduction cost $C(r_0, r) \equiv 0$ (RHS) Linear cost $C(r_0, r) = C_v \times (r - r_0)$ (For $r_e = 20\%$ and $P_{cap} = 3\%$)

PROPOSITION 1. *The optimal decision for hospital h with original readmission rate r_{h0} is either to remain at its current readmission rate or to set the readmission rate at the expected readmission rate r_h^e:*

$$r_h^*(r_{h0}, r_h^e) = \begin{cases} r_h^e & \text{if } r_{h0} \in [r_h^e, f(r_{h0}, r_h^e)], \\ r_{h0} & \text{otherwise.} \end{cases} \tag{7}$$

where $f(r_{h0}, r_h^e)$ is the unique solution to the equation:

$$R(f(r_{h0}, r_h^e), r_h^e, r_h^e) = R(f(r_{h0}, r_h^e), f(r_{h0}, r_h^e), r_h^e) \tag{8}$$

and $R(\cdot, \cdot, \cdot)$ is as in Equation 5.

Figure 3 depicts hospital h's operating margin as a function of its readmission rate with different reduction-cost functions for given parameters when $r_h^e = 0.2$. The red vertical line, where A is, indicates the position of the expected readmission rate, r_h^e of the hospital, while the box shows the current readmission rate. The green vertical line, where B is, corresponds to $f(r_{h0}, r_h^e)$ and is, by definition, the readmission rate (greater than r_h^e) that generates the same contribution as setting r_h to r_h^e. As discussed above, a hospital has financial incentive to reduce its readmissions if and only if its current/predicted readmission rate falls in the region $[A, B]$. We define this region as the *policy effective region* – hospitals whose parameters place them in this region will optimally reduce readmissions in response to the HRRP penalties.

In the graph of Figure 3 there are three parameter regions:

Region (1) (Program-Indifferent Region, $[0, A]$). A hospital in this region has its original readmission rate r_0 smaller than its expected readmission r_h^e (here 0.2). Then its contribution is strictly increasing with its readmission rates, indicating that the optimal decision for the hospital is to not reduce its readmissions. We call these hospitals program-indifferent (PI) hospitals.

Region (2) (Program-Effective Region, $[A, B]$). If the hospital's original readmission rate r_{h0} is greater than r_h^e and the operating margin at current readmission rate is lower than that at r_h^e, then the hospital's optimal decision is to reduce its readmission rate to r_h^e (recall that we only allow hospital to reduce readmissions). In this case, the savings in penalty by reducing the readmission rate to r_h^e outweigh the loss of contribution. These hospitals are denoted as program-effective (PE) hospitals.

Region (3) (Non-Program-Effective Region, $[B, 1]$). In this area the margin loss by reducing readmissions is greater than the savings in penalties. In this case, the optimal strategy for the hospital is to take no action, and remain at the current readmission rate. We call these hospitals non-program-effective (NPE) hospitals.

For a hospital h that currently receives penalties $(r_{h0} > r_h^e)$, there are two measures that affect its decisions in reducing readmissions. The first is $r_{h0} - r_h^e$. For large values of $r_{h0} - r_h^e$ the hospital falls in Region (2). Therefore, the hospital is less likely to reduce its readmissions from a financial standpoint. The second is $f(r_{h0}, r_h^e) - r_h^e$, the width of Region (2). The wider Region (2) is, the more likely it will include r_{h0} and, consequently, the hospital will be incentivized to reduce its readmissions. In the corollary below, we summarize comparative statics linking $f(r_{h0}, r_h^e) - r_h^e$ with the primitives of a hospital $(l, d_d, d_h, \lambda^a, \phi)$.

COROLLARY 1. *For given primitives, the width of Region (2) $(f(r_{h0}, r_h^e) - r_h^e)$ for a hospital h is weekly increasing in its percentage of medicare patients (P_{med}), the hospital divergence probability (d_h), and the readmission loss factor (l), and weakly decreasing in the cost of reducing readmission $(C(y, x))$ and the contribution margin (C_m).*

In the following sections, we revisit these comparative statics, and prove that these comparative statics still hold in a game-theoretic setting with other hospitals' decision's affecting hospital h's expected readmission rate.

4. Game-Theoretic Model

Building on the model of the single hospital, we construct a game-theoretic model to describe hospitals' joint decisions assuming hospitals consider one year into the future. Our main analytical result is that, despite the difficulty of the game, the hospitals that prefer paying penalties to reducing readmissions in the single-hospital setting still do not want to reduce readmissions in the game-theoretic setting. Moreover, the comparative statics results in single-hospital setting remains valid in the game setting.

4.1. Single-year game

Recall that hospitals are benchmarked against their peers under the HRRP policy. This creates a game, in which each hospital maximizes its total operating margin by deciding how much it wants to reduce the current readmission rate at the beginning of the game. In looking one year into the future, the hospital will be taking into account how its decision (and those of its peers) affect its CMS-expected target r^e for the next year.

We will impose the following assumptions: (1) $\lambda_{ij}^{d_h} \equiv 0$. This assumption implies that a hospital's readmission reduction decision will affect other hospitals' decisions mostly through the expected readmission and not through the effects on the throughput. The hospital-level divergence probability is typically low, such as 7.5% in Corrigan and Martin (1992). Our experience at Northwestern Memorial Hospital is that this divergence rate is below 30%. For concreteness, consider two hospitals A and B. A reduction of 5% in readmissions (a very significant reduction) in Hospital A leads to a reduction in the divergence from hospital A to B of at most $0.3 * 0.05 = 0.015$ (or 1.5%) change in the input to hospital B. In reality there are multiple (i.e, more than two) hospitals located within a given geographic are so this number should be even smaller.

(2) $d_d = 0$, and in turn $\lambda_{ij}^d \equiv 0$. This assumption assumes that the probability that a patient in group ij to be readmitted to group $i'j'$ is zero. As Pearson et al. (2002) documented in their experiment, the commonest agreed reason for readmission is a relapse or complication of the initial illness. Among all the subjects in their experiment, less than 14% are identified with new disease problems.

CMS uses the discharge data at the beginning of the game to evaluate the penalty for each of the hospitals. The maximization problem for hospital h with r_{h0} is to choose r_{h1} which maximizes $R(r_{h0}, r_{h1}, r_{h1}^e)$ stated in Equation 5, where r_{h1}^e is the new expected readmission rate of the hospital after CMS observes the readmission reductions from each hospital and re-estimates the parameters at the beginning of the game.

Each hospital knows all other participating hospitals' current readmission rate and expected readmission rate (r_{h0}, r_{h0}^e). This information is made publicly available by CMS (http://www.medicare.gov/hospitalcompare/Data/30-day-measures.html). To calculate the exact expected readmission rate, the hospital has to acquire the patient-level discharge data from all other hospitals, and re-estimate the HGLM model used by CMS. This, as acknowledged by CMS (see FAQ in www.qualitynet.org), is a difficult undertaking for an individual hospital: First, the hospital does not have the patient-level discharge data from all other hospitals, and it has to devote substantial resources to get access to a subset of this data. Second, CMS changes its methods of calculating expected readmissions every year, and it is impossible to predict the changes. For example, from 2011 to 2012, CMS added the readmission cases from VA hospitals into the estimation model.

Accordingly, we assume that each hospital h (instead of precisely predicting it) infers its future expected readmission rate, denoted as $\bar{r}^e_{h1}$, from the existing data and other hospitals' actions according to a certain updating function, $\bar{r}^e_{h1} = g_h(\vec{r}_1, \vec{r}^e_0)$. The hospital h then makes its maximization decision based on its estimated expected readmission rate, $\bar{r}^e_1$. The only assumptions on the updating function g_h is that it is weakly increasing in r_{i1} for any hospital i. [2]

The dynamics of the game are as follows:

(0) Period 0:

a. r_{h0} stands for the current readmission rate at hospital h and r^e_{h0} stands for the current year's CMS-expected readmission rate. These are given.

b. Each hospital h makes a single decision: its targeted value of readmission for next year r_{h1}.

(1) Period 1: Hospital h incurs penalty based on its choice r_{h1} and the CMS-expected readmission rate r^e_{h1}.

In other words, the hospital's operating margin is set in period 0 and cannot be altered. The hospital is making a readmission-adjustment decision to maximize operating margin in year 1.

With the proper definition of updating mechanism, the formal definition of the static game with H hospitals is given below:

DEFINITION 1. Let $\vec{r}_0 = \{r_{10}, r_{20}, ..., r_{H0}\}$ be the initial predicted readmission rates, and $\vec{r}^e_0 = \{r^e_{10}, r^e_{20}, ..., r^e_{H0}\}$ be the initial expected readmission rates of the H hospitals. Hospital h's strategy space is $r_{h1} \in [0, r_{h0}]$. The payoff function for hospital h is $R(r_{h0}, r_{h1}, g_h(\vec{r}_1, \vec{r}^e_0))$ defined in Equation 5.

A Nash Equilibrium in pure strategies is a readmission vector $\vec{r}^*_1$ such that $r^*_{h1} \in \arg\max_{r_{h1} \in [0, r_{h0}]} R(r_{h0}, r_{h1}, g_h(\vec{r}_1, \vec{r}^e_0))$ $\forall r^*_{h1} \in \vec{r}^*_1$. A mixed-strategies Nash Equilibrium is $\pi = \{\pi_1, \pi_2, ..., \pi_H\}$ where π_h is a probability distribution with support $[0, r_{h0}]$.

The best response for hospital h given the other hospitals' decisions r_{-h1} is:

$$BR_h(r^e_{h0}, r_{h0}, r_{-h1}) = \begin{cases} g_h(\vec{r}_1, r^e_{h0}) & \text{if } r_{h0} \in [g_h(\vec{r}_1, r^e_{h0}), f_h(r_{h0}, g_h(\vec{r}_1, r^e_{h0}))], \\ r_{h0} & \text{otherwise,} \end{cases} \quad (9)$$

where f is the function defined in Proposition 1 to characterize the PE region of a hospital. Here we define the *no action* strategy for hospital h as $BR_h = r_{h0}$.

The effect of one hospital's decision on the actions of its peer is non-trivial. If one hospital decides to reduce its readmission rates, its decision effectively lowers the expected readmission rate for other hospitals, and decreases other hospitals' payoffs monotonically. Therefore, the action of each hospital exerts negative externality on other hospitals' contribution margins. However, a

---

[2] To the extent that the true CMS computations are monotone in the appropriate sense, our results continue to hold even if hospitals overcome the challenges associated with precise prediction of the CMS targets.

hospital's readmission reduction effort may affect other hospitals' readmission reduction effort in opposite directions. A reduction decision by hospital h lowers the expected readmission rates for other hospitals. However, if the expected readmission rate is lowered substantially, some hospitals may find themselves in Region (2) of Figure 2 and prefer incurring penalties over reducing readmissions. Therefore, it is unclear if a hospital's reduction effort would increase or decrease the total readmissions.

This game has a continuous payoff function and a compact strategy set and therefore has at least one mixed-strategies Nash Equilibrium (Glicksberg 1952). We cannot, however, guarantee the existence or the uniqueness of pure-strategy Nash Equilibria for this game.

LEMMA 1. *There exists at least one mixed-strategies Nash Equilibria in the single-year game for any continuous updating functions. For a specific updating function, the game may not have a pure Nash Equilibrium, or it may have multiple equilibria.*

In the absence of a uniqueness result, we turn to bounds. We bound the set of possible mixed-strategy Nash Equilibria, and provide bounds on the number of hospitals that are incentivized to reduce readmission rates in the game under *any* equilibrium. First, we provide a lower bound on the number of hospitals that prefer incurring penalties to reducing their readmission rates in this game. These are the hospitals on which the policy is not effective. Let $(\vec{r_0}, \vec{r_0^e})$ denote the initial expected and predicted readmission rates of hospitals in the game. We say that the hospital h is a strongly non-policy-effective (SNPE) hospital if it satisfies the condition:

$$r_{h0} > f_h(r_{h0}, g_h(\vec{r_0}, r_{h0}^e)) \tag{10}$$

where f_h is defined in Equation 8. The term strongly non-policy-effected is supported by the following proposition that shows that for any SNPE hospital, reducing to expected readmission rates is a dominated strategy.

PROPOSITION 2. *For any equilibrium π in the game:*

$$\forall h \in SNPE, \ \pi_h(r_{h0}) = 1. \tag{11}$$

In other words, the number of SNPE hospitals provides a lower bound on the number of NPE hospitals – those that incur penalties but assign probability 1 to no action strategy in any equilibrium of the game.

By definition, a hospital is SNPE if, considering its current CMS expected readmission rates, and its current readmission rate, reducing readmissions is a sub-optimal decision. Thus, the number of NPE hospitals in the single-hospital setting of Section 3 serves as a lower bound for the number

of hospitals that are SNPE in the game setting: competition introduced by the HRRP policy can only increase the number of NPE hospitals. In other words, the "worst offenders" are insensitive to the benchmarking.

We next develop an algorithm that allows us to derive an upper bound on the number of hospitals which are incentivized to reduce readmissions, PE hospitals, in this game. The algorithm is briefly summarized as follows:

0. Start with H hospitals with original predicted readmission rates $\vec{r_0} = \{r_{1,0}, r_{2,0}, ..., r_{H,0}\}$ and expected readmission rate $\vec{r_0^e} = \{r_{1,0}^e, r_{2,0}^e, ...r_{H,0}^e\}$.

1. Identify all SNPE hospitals. Set $n = 0$.

2. Update the readmission rate vector according to following:

$$r_{h,n+1} = \begin{cases} g_h(\vec{r_n}, \vec{r_n^e}) & \text{if } r_{h,n} > g_h(\vec{r_n}, \vec{r_n^e}), h \notin SNPE, \\ r_{h,n} & otherwise. \end{cases} \tag{12}$$

In words, for one hospital h that has initial readmission rate $r_{h,n}$ higher than its expected readmission rate $\bar{r}_{h,n}^e$ at step n, let it reduce to its expected readmission rate. Set $n \leftarrow n+1$.

3. If there are no hospitals that have reduced readmission rates in 2, terminate the algorithm and set $N = n$. Otherwise, go back to 2.

Let the terminal readmission vector of the algorithm be $\vec{r}_N$. A hospital h is an SPE hospital if:

$$r_{h,N} < r_{h0} \tag{13}$$

We say that a hospital is strongly PE (SPE) if it reduces readmissions in some stage of the algorithm. The following proposition proves that the number of SPE hospitals represents an upper bound on the number of hospitals that are PE in some equilibrium – i.e, hospitals that have positive probability to reduce their readmissions in any equilibrium. Combining this upper bound with the previous lower bound, we get a bound of the effectiveness of the HRRP policy in this model.

PROPOSITION 3. *Under any equilibrium, the number of PE hospitals is bounded above by the number of SPE hospitals:*

$$\forall \pi, \; \sum_{h=1}^{N} 1_{\{r_{h,\pi} < r_{h0}\}} \leq \sum_{h=1}^{N} 1_{\{h \in SPE\}} \tag{14}$$

Moreover, the set of SPE hospitals is mutually exclusive from the set of SNPE hospitals.

Last, we define a hospital to be strongly program-indifferent (SPI) if its current readmission rate will be lower than its expected readmission rate in any equilibrium. The set of SPI hospitals is simply the difference of the set of all hospitals and the sets of SNPE and SPE hospitals.

4.2. Dynamic game

We extend the single-stage model to an n-stage game where at each stage, hospitals play the one-stage game described in the previous section. This represents the scenario where hospitals may update their readmission rates at each period. To capture the plans to increase the penalty cap in 2014 and 2015, we allow the penalty cap to increase from one period to the next, i.e, we have P_{cap}^1 for the first year, P_{cap}^2 for the second and etc. Let $P_{cap}^{max} = \max_{l=1,\dots,n} P_{cap}^l$ be the maximum penalty cap in the horizon. Allowing hospitals to make readmission reduction decision every period also enables us to calibrate our model with real data and make predictions about the policy's outcomes in the long run.

Our concept of equilibrium in this dynamic setting is sub-game perfect Nash Equilibrium. Since there is no unique Nash Equilibrium in the single-stage game for any specific updating functions, the existence and uniqueness of a Nash equilibrium is not guaranteed in the multi-stage game. As before, we are able to develop bounds. In an equilibrium of the dynamic game (if it exists) a hospital is SPE if it reduces its readmission in some year (recall that we do not allow hospitals to increase readmission rates). It is SNPE if it does not reduce readmissions during the game horizon.

PROPOSITION 4. *The set of SNPE hospitals in a one-stage static game with $P_{cap} = P_{cap}^{max}$ is a subset of the hospitals that do not reduce their readmission rates in the multi-stage dynamic game. In turn, the number of these SNPE hospitals serves as a lower bound on the number of NPE hospitals in the multi-year game. Also, the number of SPE hospitals in a single-stage game with P_{cap}^{max} is greater than the number of PE hospitals that decide to reduce readmissions in any year in any equilibrium of the multi-stage game.*

The intuition is as follows: The number of SNPE hospitals in the single-stage game defined above is the hospitals that do not want to reduce readmissions even if the penalty cap is the maximum penalty cap forever. In particular, these hospitals will not be incentivized to reduce readmissions if the penalty gradually approaches the maximum penalty cap (from below). Similarly, the SPE bound in the proposition is derived with the maximum penalty cap. It is intuitively clear that the number of hospitals that are incentivized to reduce readmissions when penalty cap gradually increases to the maximum penalty cap should be bounded above the number of SPE hospitals with the maximum penalty cap.

5. Data and Simulation

In this section, we use hospital data from California in fiscal year 2013 to identify initial conditions of each hospital for our game-theoretical model. We use the data to better understand the value of competition that the HRRP policy introduces for the effectiveness of the policy and to propose

Variable	Mean	Std. Dev.	Min.	Max.
CMS_no	50321.886	225.372	50002	50764
Available_bed_occupancy	0.597	0.138	0.075	0.976
Revenue_fraction_from_medicare	0.362	0.108	0.079	0.792
Expected_readmission	21.586	1.475	17.136	26.545
Predicted_readmission	21.393	2.263	16.193	30.931
AMI_no_discharge	157.467	115.374	25	599
AMI_predicted_readmission	20.221	3.067	12.3	35.8
AMI_expected_readmission	20.337	2.16	15.2	29.4
HF_no_discharges	363.554	209.231	31	1714
HF_predicted_readmission	24.461	2.459	18	32.2
HF_expected_readmission	24.681	1.331	20.7	28.5
PN_no_discharges	305.799	163.521	43	1146
PN_predicted_readmission	18.64	2.226	14	27.7
PN_expected_readmission	18.827	1.453	14.7	23.5

Table 1: Summary statistics of hospital-level data in California

model-based predictions to the effect of the HRRP on hospitals' readmission reduction efforts. In Section 5.1, we describe the datasets. Simulation methods, necessary assumptions, and results are reported in Section 5.2. We validate our model predictions of SNPE hospitals with hospitals' readmission reduction efforts in practice in Section 5.3.

5.1. Data

We combine two datasets. First, the financial and operational data for 434 hospitals in California is obtained from the Office of Statewide Health Planning and Development's (OSHPD) *oshpd.ca.gov*, a state government organization that provides and ensures accessible healthcare in California. For fiscal 2013, the OSHPD dataset documents the fraction of each hospital's revenue derived from Medicare patients; see Table 1.

The CMS website lists the CMS-expected and predicted readmission rates for each IPPS hospital for fiscal year 2013. These readmission rates are estimated from the hospitals' discharge data from July 2008 to June 2011. Among 434 California hospitals in the OSHPD dataset, there are 312 IPPS hospitals that can be matched with the CMS data. The HRRP policy only targets IPPS hospitals, and therefore we can successfully match all HRRP-targeted hospitals in this data set. According to the selection rule of the HRRP, the hospitals with small number of readmissions in the monitored diseases, namely AMI, HF, and PN, are not considered in the penalty evaluation. Out of these hospitals that have CMS-expected and CMS-predicted readmission rates, 9 of them are missing the financial data. After removing these hospitals in the dataset, we have 186 hospitals for AMI, 250 hospitals for HF, and 249 hospitals for PN. The data is summarized in Table 1:

In our model, we assumed the payment per patient depends only on the disease and insurance type. This assumption is valid for insurance programs with a pay-per-case payment structure, such
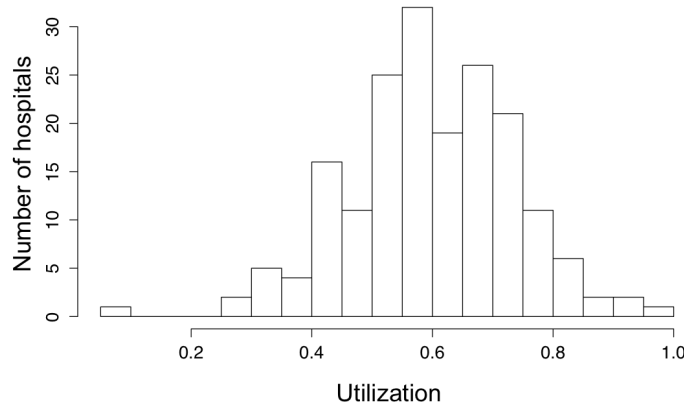
Figure 4: Histogram of Hospital Utilization

as Medicare and Medicaid. Private insurance programs may adopt different payment structures where the payment depends also, for example, on the length of stay (pay-per-diem) and the quality of treatment (pay-per-performance). Small changes to readmission rate should not, however, affect significantly the payment given that most hospitals in our dataset are not fully utilized: most hospitals in our dataset have utilization rates smaller than 85%; see Figure 4. Small changes to the readmission rates should not then have a big impact on the utilization of physicians, and in turn on the service quality of patients (Kc and Terwiesch 2009). There is also empirical evidence that readmission-reduction programs do not increase the length of stay of patients (Hansen et al. 2013), and therefore, changing readmissions is unlikely to change the average payments under pay-for-performance or pay-per-diem schemes.

5.2. Simulation and Results

We consider two cases: (i) hospitals can make decentralized decisions for each disease, and when they reduce readmissions for each disease, they only consider the cost and benefit related to that disease's processes, (ii) hospitals make centralized decisions and evaluate revenue and penalties aggregated from all three monitor diseases. In both cases, hospitals are modeled to maximize their undiscounted operating margin from 2013 to 2020.

As we proved in Section 2, the single-hospital setting, where hospitals are myopic and do not consider the interaction with other hospitals through the HRRP benchmarking, provides a lower bound on the number of SNPE hospitals. Similarly, the number of SPE hospitals is an upper bound on the number of hospitals that have positive probability to reduce their readmission rates in at least one equilibrium. Since these two sets are mutually exclusive by construction, the remaining hospitals do not incur penalties and are not incentivized to reduce readmissions in any equilibrium.

In order to numerically compute these bounds for the two models, we have to specify cost functions and updating functions:

(1) **Costs:** We assume that cost functions are of the form:

$$C_h(r, x) = C_h^v (r - x)^\alpha + C_h^s \frac{1}{r} \tag{15}$$

where $\alpha \in [0, 3]$, $C_h^v = \Pi_h(0)C_v$ and $C_h^s = \Pi_h(0)C_s$. The first term in the cost structure, $C_v(r - x)^\alpha$ represents the variable cost of reducing readmission rates from r to x. When $\alpha \in (0, 1)$, the cost structure is concave, representing economies of scale in reducing readmission rates. The cost is convex when $\alpha > 1$ representing a marginally increasing difficulty in reducing readmissions. The second term captures dependence of the cost on the initial starting point. For hospitals that have initially low readmission rates reducing further might be expensive – some small amount of readmission might be unavoidable and not easily further reduced by process changes.

In the following simulation we set $C_v = C_s = 0.001$, representing a cost of 0.1% of its total revenue for a hospital to reduce its readmissions by 1%.

(2) **Updating functions** $g(\cdot, \cdot)$**:** Hospital h predicts $\bar{r}_{h1}^e$ as the average of all readmission rates. In other words $\bar{r}_{h1}^e$ is the average readmission rates of $\{r_{h1}, \forall h\}$ weighted by the number of patients in each hospital. If all H participating hospitals have the same number of patients, then r_{h1}^e is simply $\frac{1}{H} \sum_k r_{h1}$. We write $\bar{r}_{h1}^e = g(\vec{r_1})$, where $\vec{r_1} = \{r_{11}, r_{21}, ..., r_{H1}\}$ is vector representing the new readmission rates of all hospitals.

5.2.1. Decentralized Model For a large teaching hospital like Northwestern Memorial it is reasonable to assume that different diseases are managed by different teams and readmission reduction decisions are made "locally" at the disease level. In this subsection we assume that process changes to reduce readmissions are made at the disease level. This is implemented by assuming that the each disease takes the maximum penalty cap as its own penalty cap. In this decentralized setting, the numerical study reduces to three single-disease models.

According to the latest version of the HRRP policy, the maximum cap penalty is 2% for 2014, and 3% thereafter. There are four parameters in the numerical study corresponding to the relevant parameters in the model of Section 2 (1) C_m, the contribution margin, (2) $(1-l)(1-d_h)$, the product of the inverse loss factor and inverse hospital divergence rate, (3) the cost function parameter α. In our base case, we assume that the contribution margin is 40%, the product is 1 ($l = d_h = 0\%$), and there is no cost $C_v = C_s = 0$. We later report results for a broader set of parameters; see Tables 2-4.

In Figure 5, we provide a graphic view of PI, PE, and NPE hospitals for each disease for the single-hospital setting (left-hand side) and the dynamic game setting (right-hand side) with the

bounds based on Strongly PE (SPE) and Strongly NPE (SNPE) hospitals. In each panel of Figure 5, the horizontal axis is the difference between a hospital's predicted readmission rate and its expected readmission rate for that disease. The vertical axis shows the percentage of the contribution margin of a hospital from Medicare patients. Each point on the graph represents a hospital. Circles, crosses, and triangles represent PI, PE, and NPE hospitals in the single-hospital setting and SPI, SPE, SNPE for the game setting.

As evident in Figure 5, for a hospital with a CMS-predicted readmission rate greater than its expected readmission rate, the lower the percentage of its revenue from Medicare patients, the lower the financial incentive to reduce readmissions. The loss in margin from reducing the readmission rate is greater for a hospital with higher readmission rate. If the sum of the readmission reduction cost and the loss of contribution outweighs the savings from penalties, the hospital is no longer financially incentivized to reduce its readmission rate. Importantly, even assuming that there is no process cost to reduce readmissions, with a maximum penalty of 0.03, about 10% of hospitals still prefer incurring penalties over reducing readmissions.

Further, by comparing the number of SPE hospitals in the single-hospital and the dynamic game setting, we quantify the maximum *value of competition* that the HRRP policy creates by benchmarking hospitals against their peers. Hospitals that are SPE hospitals in the dynamic game (but are PI in the single-hospital setting) are those that may reduce readmission rates due to the fact that the other hospitals' reduction effort changes the CMS-expected readmission rate in the subsequent years. As we show in this numerical study, when hospitals in this dataset consider the strategic interactions between them, there are at most 1 in 10 hospitals that become incentivized but would have not been such without the competition induced by the HRRP benchmarking. This result is robust to the updating function $g(\cdot)$. Intuitively, the SPE hospitals (see 5) are those that have an initial readmission rate that is relatively close to the expected readmission rate. Their readmission improvements are, thus, relatively small and bare little effect on the expected readmission rates of other hospitals.

Tables 2 and 3 depict the number of PI, SPE, and SNPE hospitals for a broad set of parameters. We consider the product of inverse readmission loss factor and inverse hospital-level divergence rate between 0% and 40%, the percentage of contribution margin of hospitals between 40% and 75%, and the cost of reducing readmissions being zero, concave $\alpha = 0.5$, or convex, $\alpha = 2$. For values $\alpha \geq 0$, the cost coefficients are $C_v = C_s = 0.001$.

Tables 2, 3, and 4 confirm that, as the readmission loss factor increases, hospitals are more inclined to reduce their readmissions. Note that by CMS's measure of readmission rates, a patient who is readmitted to a different hospital still contributes to the readmissions of the hospital from which the patient was initially discharged. Consequently, increasing the divergence probability
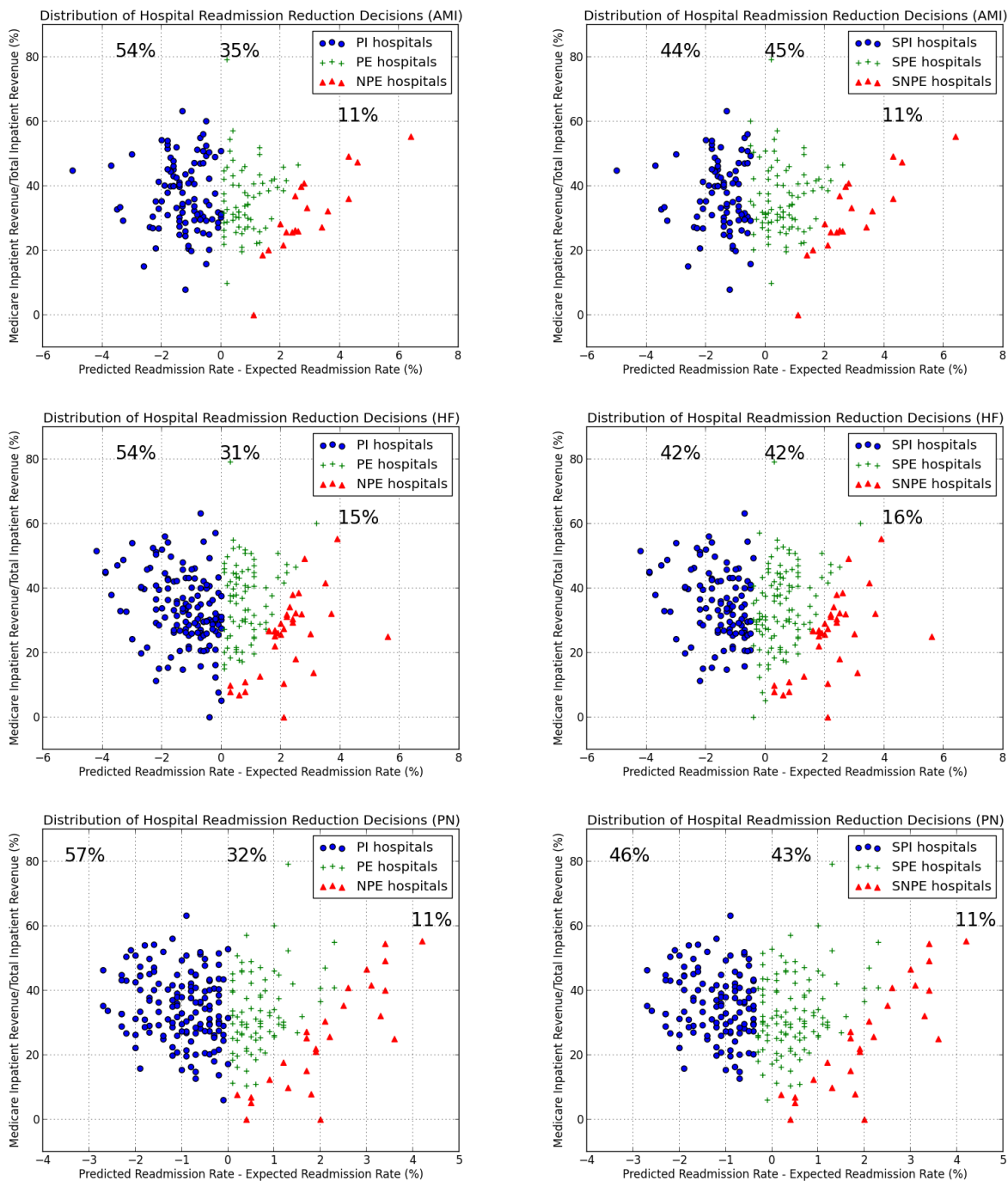
Figure 5: Model predictions on the hospitals' readmission reduction decisions in the base case. The single-hospital setting is on the left-hand side. Different rows correspond to different monitored diseases.

among hospitals is effectively equivalent to reducing the contribution from readmitted patients on average, and in turn decreasing the number of SNPE hospitals. Moreover, as the cost of reducing readmissions increases, the number of hospitals that are SNPE dramatically increased. This

		No Cost			Concave Cost			Convex Cost		
$(1-l) \times (1-d_h)$	Contribution Margin	SPI	SPE	SNPE	SPI	SPE	SNPE	SPI	SPE	SNPE
100%	40%	54%	35%	10%	54%	17%	27%	54%	18%	27%
	75%	54%	23%	22%	54%	5%	39%	54%	5%	39%
80%	40%	54%	38%	7%	54%	23%	22%	54%	23%	22%
	75%	54%	29%	16%	54%	12%	32%	54%	12%	32%
60%	40%	54%	43%	2%	54%	26%	18%	54%	26%	19%
	75%	54%	35%	10%	54%	15%	30%	54%	15%	30%

Table 2: Number of strongly program-indifferent (SPI), strongly program-effective (SPE) and strongly non-program effective (SNPE) hospitals for different parameters for the Californian hospital data set with 40% percentage of contribution margin for **Acute Myocardial Infarction (AMI)**

		No Cost			Concave Cost			Convex Cost		
$(1-l) \times (1-d_h)$	Contribution Margin	SPI	SPE	SNPE	SPI	SPE	SNPE	SPI	SPE	SNPE
100%	40%	57%	32%	10%	57%	13%	29%	57%	13%	29%
	75%	57%	23%	19%	57%	6%	36%	57%	6%	36%
80%	40%	57%	37%	5%	57%	17%	25%	57%	17%	25%
	75%	57%	28%	14%	57%	8%	34%	57%	8%	34%
60%	40%	57%	40%	2%	57%	21%	21%	57%	21%	21%
	75%	57%	32%	10%	57%	12%	30%	57%	12%	30%

Table 3: Number of strongly program-indifferent (SPI), strongly program-effective (SPE) and strongly non-program effective (SNPE) hospitals for different parameters for the Californian hospital data set with 40% percentage of contribution margin for **Heart Failure (HF)**

		No Cost			Concave Cost			Convex Cost		
$(1-l) \times (1-d_h)$	Contribution Margin	SPI	SPE	SNPE	SPI	SPE	SNPE	SPI	SPE	SNPE
100%	40%	54%	31%	14%	54%	16%	29%	54%	16%	29%
	75%	54%	18%	27%	54%	8%	36%	54%	8%	36%
80%	40%	54%	37%	8%	54%	20%	24%	54%	20%	24%
	75%	54%	22%	22%	54%	12%	33%	54%	12%	33%
60%	40%	54%	42%	3%	54%	26%	18%	54%	26%	18%
	75%	54%	31%	14%	54%	15%	30%	54%	15%	30%

Table 4: Number of strongly program-indifferent (SPI), strongly program-effective (SPE) and strongly non-program effective (SNPE) hospitals for different parameters for the Californian hospital data set with 40% percentage of contribution margin for **Pneumonia (PN)**

suggests that investment in readmission reduction technology is crucial to encourage hospitals to reduce readmissions. Last, we observe that the higher the percentage of contribution margin that a hospital has, the more contribution it generates from each patient admission so that the opportunity cost associated with reducing readmissions is larger. This implies that the percentage of contribution margin of a hospital has an inverse relation with its inclination to reduce readmissions.

5.2.2. **Centralized Model** In this centralized model, we assume that hospitals are making centralized readmission reduction decisions. They evaluate the penalties from all diseases and the cost of reducing readmissions in all diseases, and make a centralized decision where they may consider the trade-offs between investments in the various diseases. In this model, the maximum penalty cap represents the maximum percentage of penalties that a hospital pays in aggregation of all monitored diseases. Since here it is only the aggregate (across diseases) penalty that is capped, hospitals may be penalized more for readmissions. In this decision model, the definition of maximum penalty is consistent with the definition used by CMS.

Computing the optimal decision (or data-based bounds) for this centralized model is challenging since the payments from each disease vary dramatically from geographic location to another and from one year to the next (in the decentralized model such changes can be normalized). However, the intersection of the individual-disease SNPE hospitals provides a lower bound for the number of SNPE hospitals in the centralized model: the number of hospitals that receive penalties but do not reduce readmissions **in any disease** is bounded by the number of hospitals in the intersection of the sets of SNPE hospitals for three monitored diseases. Moreover, the union of the sets of SPE hospitals for three monitored diseases is an upper bound for the number of hospitals that decide to reduce readmissions when they make centralized decisions. Table 5 summarizes the results for the same broad set of parameters that have been considered above.

		No Cost			Concave Cost			Convex Cost		
$(1-l) \times (1-d_h)$	Contribution Margin	SPI	SPE	SNPE	SPI	SPE	SNPE	SPI	SPE	SNPE
100%	40%	34%	62%	2%	56%	33%	9%	56%	34%	9%
	75%	47%	46%	6%	70%	16%	13%	70%	16%	13%
80%	40%	32%	66%	1%	50%	42%	7%	50%	42%	7%
	75%	40%	55%	4%	63%	24%	12%	63%	24%	12%
60%	40%	29%	70%	0%	47%	46%	6%	48%	45%	6%
	75%	34%	62%	2%	58%	31%	10%	58%	31%	10%

Table 5: Number of strongly program-indifferent (SPI), strongly program-effective (SPE) and strongly non-program effective (SNPE) hospitals for different parameters for the Californian hospital data set with 40% percentage of contribution margin for **Centralized Decisions**

Importantly, also in the centralized model–and even when the cost of reducing readmissions is 0–there are still hospitals that pay penalties instead of reducing readmissions. Moreover, this portion has increased substantially if the cost of reducing readmissions is increasing. As the set of SNPE is based on the intersection of disease-level SNPE, all the comparative statics remain the same as in the decentralized case.

5.3. Model Validation

Recall from the HRRP timeline in Figure 1 that HRRP was signed into law along with the Affordable Act Care on March 20, 2010. The first penalty was charged in fiscal year 2013, which was based on the discharge data from July 2008 to July 2011. Since the hospitals became aware of the policy after 2010, the first penalties, all affected by hospital actions after HRRP was signed into law, are those levied in 2015, as they depend on discharge data from July 2010 to July 2013. The hospitals' penalties for 2015 became publicly available on April 30, 2014–a month before the date of this paper. For each of the hospitals, we use the difference between its readmission ratios in fiscal years 2013 and 2015 as a proxy for its readmission reduction efforts from 2010 to 2013. For example, if a hospital's excess readmission ratio in 2013 is larger than that in 2015, it suggests that this hospital did reduce its readmission rates from 2010 to 2013.

Our model is stylized and designed to capture financial trade-offs and identify bounds rather than predict exact outcomes. The key insights we derive from our model is the characterization of SNPE hospitals–those that prefer paying penalties over reducing readmission rates in the long run. The key observation is that the actual SNPE hospitals coincide with those predicted by our model.

	PN	HF	AMI
Number of SNPE hospitals (Model)	8	5	12
Number of SNPE hospitals not paying penalties for 2015	0 (0%)	0 (0%)	0 (0%)
Number of SNPE hospitals reducing readmissions in 2010-2013	2 (25%)	1 (20%)	3 (25%)

Table 6: Number of SNPE hospitals according to the prediction of the model that actually reduce readmissions or do not pay penalties by fiscal year 2015.

Table 6 takes the SNPE hospitals identified by our model and displays their actual readmission reduction. The first row reports the number of hospitals that are classified as SNPE hospitals by our model (used here with the baseline parameters; see Section 5.2) for each of the three monitored diseases. Since we assume, for this section, that hospitals' readmission reduction efforts are costless, the set of SNPE hospitals is a subset of the set of SNPE hospitals under any positive reduction costs. In other words, assuming costless reduction efforts is a more stringent test of the predictive power of our model. There are, for example, 8 SNPE hospitals for the disease PN. Having identified these 8 hospitals we look for their actual performance in the data and find that, indeed, all of them incurred penalties–this is reported in the second row. In this metric our model-based prediction of the SNPE hospitals does capture an lower bound on the hospitals that prefer paying penalties over reducing readmissions. The third row reports how many of these 8 hospitals actually did reduce readmissions in 2010-2013. Most SNPE hospitals did not reduce readmissions in the data. Among
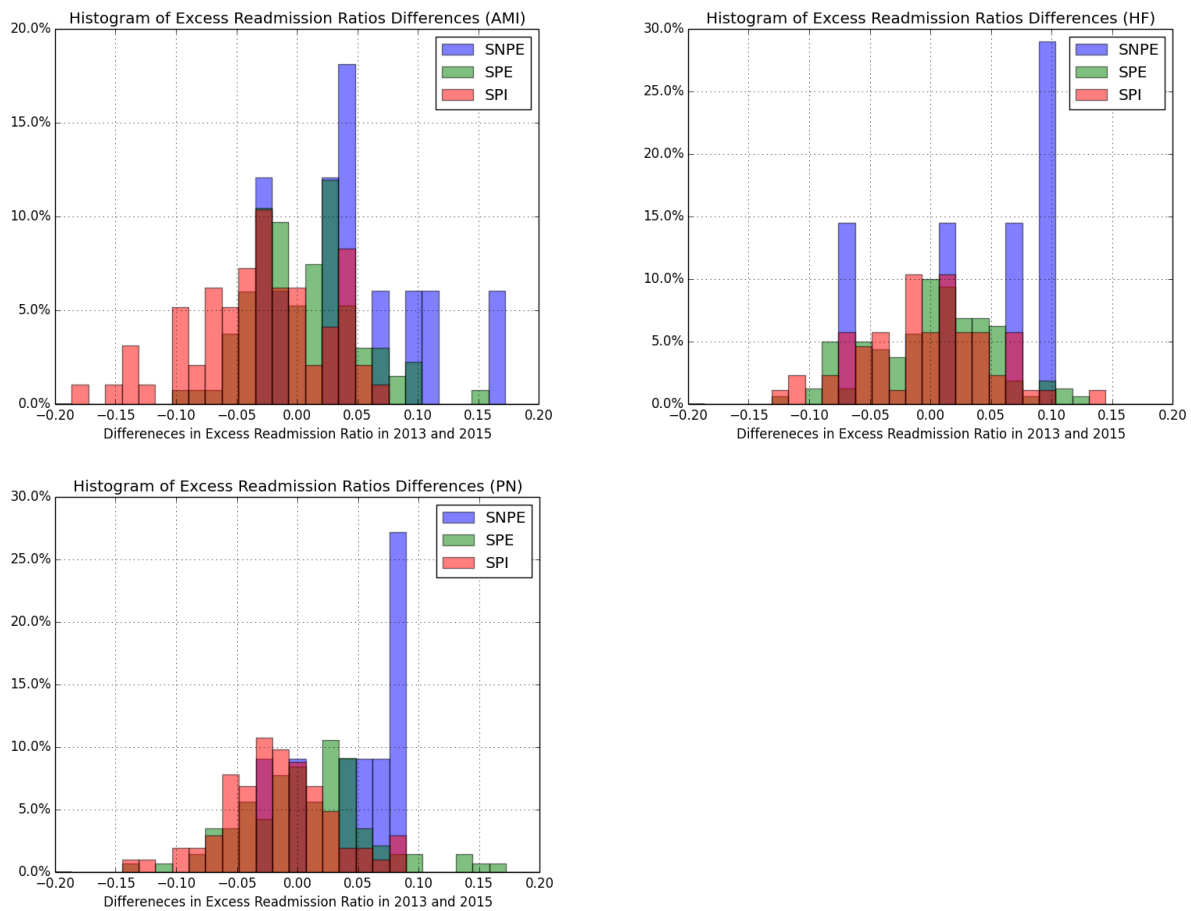
Figure 6: Excess readmission ratios for 2013 and 2015 for different sets of hospitals and different disease groups.

those that do reduce readmission (a total of 6 across the three diseases) none reduces it by more than 5%.

To further support the "separating" power of the model, we compare the distributions of actual readmission reduction efforts for SNPE, SPE, and SPI hospitals (classified by the model). Figure 6 displays the histogram of the differences of excess readmission ratios in 2013 and 2015 (reflecting hospitals' readmission reduction efforts in 2010-2013) for the different sets of hospitals (i.e. SNPE, NPE, PI) and different diseases. For a specific disease, blue, green, and red histograms correspond to the SNPE, SPE, and SPI hospitals respectively. Clearly, the distribution of SNPE hospitals is more skewed to the right while the distribution of the PE hospitals are more skewed to the left. This suggests that SNPE hospitals exert less effort in reducing readmissions compared to SPE hospitals in practice, which suggests that the model prediction is consistent with the hospital readmission reduction efforts in practice.

Last, we find that a significant number of SPI hospitals did reduce their readmissions even though

they did not pay any penalties. This suggests that, for the hospitals with better performance, factors other than financial incentives, perhaps such as ethics and reputation, play an important role in process improvements and readmission reduction decisions. This is not surprising. Our model is stylized and designed to capture financial trade-offs and identify bounds rather than predict exact outcomes. There is an opportunity here to explore other incentives as more data becomes available on hospital readmission-reduction efforts in response to HRRP.

6. Discussion

There are several noteworthy insights from our single-hospital model, game-theoretical model and the simulation results. The first five correspond to comparative statics. Whereas some of the comparative statics are intuitively expected, the study assigns magnitudes to them and draws corresponding conclusions. The latter items correspond to the value of competition and its relation to the single-hospital setting.

First, the HRRP policy is relatively more effective for hospitals that have higher readmission divergence probability. Such hospitals are typically located at more economically developed areas. Patients who live in urban areas, have access to more hospitals and better healthcare services relative to patients in the rural areas. The HRRP policy may magnify this healthcare quality gap as hospitals in rural areas will be less incentivized to reduce readmissions.

Second, despite its effectiveness of reducing readmission rates in general, the HRRP policy is less effective for hospitals with initial poor performance. These are hospitals with their readmission rates much greater than their CMS-expected readmission rates. The larger this difference is, the lower the financial incentive for a hospital to reduce readmissions. The HRRP policy may fail to induce improvement at these. Part of this ineffectiveness may stem from the presence of a low penalty cap that, in essence, protects these hospitals. However, even if the maximum penalty cap is exceedingly high, e.g. 8%, the worst performing 5% hospitals are still not incentivized to reduce their readmissions. In Figure 7, we included the number of SNPE, SPE, and PI hospitals for each of the monitored diseases applying the decentralized game from Section 5.2.1. We increase P_{cap} from 1% in 2013 to 5% in 2020. The simulation here is based on a loss factor is 0%, readmission divergence probability is 0%, the average percentage of contribution margin for hospitals is 40% and there is a linear cost of reducing readmissions.

The number of SNPE hospitals is, as expected, decreasing with the maximum penalty cap. However, the effect diminishes as the maximum penalty cap increases. In particular, changing P_{cap} from 3% to 9%, already captures more than 90% of the reduction in number of SNPE hospitals. More importantly, the set of SNPE hospitals stabilizes at 5% for HF and PN when the maximum
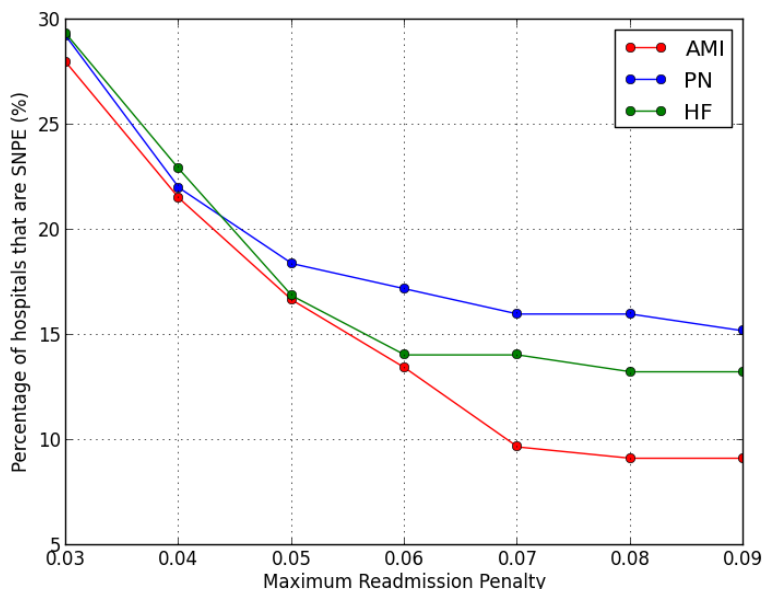
Figure 7: Equilibrium Behavior of Hospitals under Different Maximum Penalty Caps ($\alpha = 1$, $C_v = 0.01$, $l = 0.2$, $d_h = 0.15$, percentage of contribution margin=40%)

penalty cap is increasing from 3% to 9%. Yet, more than 5% of hospitals that receive penalties are not financially incentivized to reduce readmissions when the maximum penalty cap is as high as 9%. To incentivize these hospitals, increasing the penalty cap is not sufficient. CMS could instead consider a nonlinear penalty structure to incentivize worse performers by, for example, classifying hospitals based on performance and benchmarking hospitals with similar performance amongst themselves.

Third, hospitals with a low percentage of medicare patients are less likely to reduce their readmissions under the current policy. As the penalty is proportional to the contribution of a hospital from medicare patients (but revenue reduction due to reduced readmission applies to all patients), the penalty is less influential if a hospital's earnings from medicate patients are small relative to its total earnings. This property is inherent in the current government healthcare payment system, and may be difficult to change. One could utilize the penalties collected to reward hospitals with good performance. Absolute rewards (rather than rewards proportional to the revenue of a hospital from Medicare patients) may increase the impact of the current HRRP policy over these hospitals.

Fourth, hospitals with higher contribution margins are less likely to reduce readmissions in response to the HRRP policy. The higher the contribution margin, the more contribution a hospital earns from each readmission. Therefore, a higher contribution margin for a hospital implies that reducing readmissions is more costly for the hospitals which reduces its incentive to reduce readmissions in equilibrium. If treatment costs do not differ dramatically across hospitals, controlling

Medicare payments to hospitals (towards having more uniform contribution margins) can help to incentivize hospitals to reduce readmissions under the HRRP policy.

Fifth, the HRRP policy is very effective in incentivizing hospitals that receive penalties to reduce their readmissions if the cost associated with reducing readmissions is negligible. From Table 2, Table 3 and Table 4, under base case parameters, the numbers of hospitals paying penalties but not reducing readmissions are dramatically reduced by more than a half for all diseases if we move from convex cost to no cost. This points to the value of research projects (such as project BOOST) that aim to identify effective but cheap process changes to reduce readmissions.

Sixth, the HRRP policy creates a game between hospitals. As we proved in Section 2, even in the base case, the game induced by the HRRP policy is not guaranteed to have a pure-strategy Nash Equilibrium. Moreover, the game may have multiple pure-strategy and mixed-strategy Nash Equilibria. This uncertainty in outcomes reduces the visibility into the future for policy makers and hampers pro-active decision making by hospitals. In contrast, our findings based on the California data suggest that the level of uncertainty may be limited as the single-hospital setting captures the essential effects of the policy.

Finally, we have provided an low upper bound for the value of competition introduced by the HRRP policy. When hospitals consider the strategic interactions between them, there are at most one in ten hospitals that will become incentivized for AMI, HF, and PN respectively.

7. Conclusion

October 1, 2012 marked the nationwide initiation of the Hospital Readmissions Reduction Program (HRRP), an effort by the Centers for Medicare and Medicaid Services (CMS) to reduce the frequency of re-hospitalization of Medicare patients. According to CMS, approximately two thirds of U.S. hospitals will incur penalties of up to 1% of their reimbursement for Medicare patients, adding up to \$300 million in 2013 alone.

Our analysis started from considering the response of the individual hospital's in isolation ignoring the effects of the interactions with other hospitals through the benchmarking procedure of the CMS. We then proceeded to incorporate the dependence created by this policy and generated valid bounds on the effectiveness of the policy and connect the insights in the single-hospital setting to the game-theoretical setting. We applied our model to data of California hospitals, and quantified the effectiveness of the policy. We validate our model using hospitals' readmission reduction efforts during the first phase of the policy implementation from 2010 to 2012 (corresponding to penalties in fiscal year 2013 to 2015). The empirical evidence gives initial support to the predictive power of our model in identifying the SNPE hospitals.

This study can have policy implications to the extent that it provides a way to study the effectiveness of HRRP. It explicitly captures the game-theoretic aspect of this policy and relates

the individual optimization (the single-hospital setting) to the game setting. Our models can be used to evaluate the effectiveness of potential changes to the HRRP policy in the future.

The success of HRRP may be affected by various issues. In this paper we take hospitals as operating-margin maximizers. Readmission-reduction consideration may include other factors, such as the professional ethics of giving good medical services and the peer pressure from nearby hospitals given the increased visibility of performance metrics. Such considerations may increase its effectiveness relative to what our models can predict. As time passes, data will reveal the hospitals' readmission reduction decisions. Once such data is available it is our hope that our models can serve as a starting point for structural estimation on readmission reduction costs, hospitals incentives towards readmission reduction, and other factors that affect the policy outcomes.

Predicting the effectiveness of policy regulations on the individual decision making of hospitals is challenging. We hope that our paper serves as a starting point towards more elaborate analysis on the economic and financial consequences of the HRRP policy.

Appendix

A.   CMS's Estimation of readmission rates

CMS computes r_h and r_h^e for every hospital using patient level discharge and readmission data as follows: Let Y_{ilk} denote a binary variable indicating whether a discharge l of disease i in hospital k is associated with a readmission. For each discharge represented by Y_{ilk}, CMS collects the corresponding patient case covariates, denoted as Z_{ilk}. The logistic hierarchical generalized linear model is used to estimate the average and hospital-level intercepts to predict the readmission probability for each discharge:

$$log(P(Y_{ilk} = 1)) = \alpha_{ik} + \beta_i Z_{ijk}$$
$$\alpha_{ik} = \mu_i + \omega_{ik} \quad \omega_k \in N(0, \tau^2)$$

(16)

where, for each disease i, α_{ik} is the hospital-level intercepts for hospital k, μ_i is the average intercept, and β_i is the coefficient of case mix covariates.

With hospital-level and average intercepts as well as the coefficient of case mix covariates, CMS calculates the risk-adjusted predicted and the expected readmission rate for each hospital k by taking the average of the predicted readmission probabilities for all discharges of that hospital:

$$r_{ik}^e = \frac{1}{N_{ki}} \sum_{j=1}^{N_{ki}} \frac{1}{1 + e^{-\mu_i - \beta_i Z_{ilk}}}$$
$$r_{ik}^p = \frac{1}{N_{ki}} \sum_{j=1}^{N_{ki}} \frac{1}{1 + e^{-\alpha_{ik} - \beta_i Z_{ilk}}}$$

(17)

where n_{ik} is the number of Medicare discharge cases with disease i in hospital k.

B.   Proof of Proposition 1

Recall the maximization problem in Equation 5:

$$x^* = \arg\max_{x \le r} R(r, x, r_e) = \arg\max_{x \le r} \Pi_h(r)(1 - P_h(x, r_e)) - C(r, x) \tag{18}$$

where $\Pi_h(x)$, $P_h(x, r^e)$, and $C(r, x)$ are defined as:

$$
\begin{aligned}
\Pi_h(x) &= \Pi_h(0)\frac{1}{1-x} \\
P_h(x, r^e) &= \frac{P_{med}}{C_m} \min(\max(\frac{x}{r^e} - 1, 0), P_{cap}) \\
C(r, x) &= C_v(r^\alpha - x^\alpha)
\end{aligned}
\tag{19}
$$

For $x \in [0, r_e]$, $\Pi_k'(x) = -\frac{1}{(1-x)^2} > 0$, $C'(r, x) < 0$, and $P(x, r^e) = 0$. This means that a hospital's revenue is an increasing function of its readmission rate if the readmission rate is less than r_e. Therefore, the optimal solution for hospitals in this region is r_e.

Denote $x_m = \inf\{x : P_h(x, r_e) = P_{cap}\}$. Similarly, if $r > x_m$, for $x \in [x_m, r]$, $R(r, x, r_e)$ is increasing in x, and $C(r, x)$ is decreasing in x. Therefore, the optimal readmission rate in the region $[x_m, r]$ is r.

Last, let us consider the case where $x \in [r_e, x_m]$:

$$R'(r, x, r_e) = \frac{C_m}{(1-x)^2}\Pi_h(0)\Big[\frac{\frac{P_{med}}{C_m}(1 - r_e) - r_e}{r_e}\Big] - \frac{dC(r, x)}{dx} \tag{20}$$

where $\alpha \in 0, 1$

Since $\frac{dC(r, x)}{dx} < 0$, if $\frac{P_{med}}{C_m} > \frac{r_e}{1 - r_e}$, $\frac{dR(r, x, r_e)}{dx} > 0$ $\forall x \in [r_e, x_m]$, and in turn the optimal choice is x_m. If $\frac{P_{med}}{C_m} < \frac{r_e}{1 - r_e}$, since $|\frac{dC^2(r, x)}{dx^2}| \le \frac{1}{(1-x)^3}$ $\forall x \in (0, 1)$, we guarantee that $\frac{dR(r, x, r_e)}{dx}$ has the same sigh $\forall x \in (r_e, x_m)$. Therefore, we guarantee that the optimal choice is between r_e and x_m.

Overall we conclude that a hospital's optimal decision is either reducing to the expected readmission rate r_e or staying at the current readmission rate r.

C.   Proof of Corollary 1

Based Equation 8, $f(r_{h0}, r_h^e)$ is defined as:

$$R(f(r_{h0}, r_h^e), r_h^e, r_h^e) = R(f(r_{h0}, r_h^e), f(r_{h0}, r_h^e), r_h^e) \tag{21}$$

which is equivalent as:

$$\Pi_h^P(r^e) - \Pi_h^P(f(r_{h0}, r_h^e)) + P_h(f(r_{h0}, r_h^e), r^e) - C(f(r_{h0}, r_h^e), r^e) = 0 \tag{22}$$

Therefore, if ϕ^{med} increases, $P_h(f(r_{h0}, r_h^e), r^e)$ increases, and $f(r_{h0}, r_h^e)$ increases for a fixed r_h^e to balance Equation 22. Similarly, if d_h or l increases, $\Pi_h^P(r^e) - \Pi_h^P(f(r_{h0}, r_h^e)) + P_h(f(r_{h0}, r_h^e), r^e)$ increases, and $f(r_{h0}, r_h^e)$ increases for a fixed r_h^e.

If C_m increases, $\Pi_h^P(r^e) - \Pi_h^P(f(r_{h0}, r_h^e))$ decreases. Therefore, $P_h(f(r_{h0}, r_h^e), r^e) - C(f(r_{h0}, r_h^e), r^e)$ increases, which means $f(r_{h0}, r_h^e)$ decreases for a fixed r_h^e. Similarly, if $C(f(r_{h0}, r_h^e), r^e)$ increases, then $\Pi_h^P(r^e) - \Pi_h^P(f(r_{h0}, r_h^e)) + P_h(f(r_{h0}, r_h^e), r^e)$ increases, which means $f(r_{h0}, r_h^e)$ decreases for a fixed r_h^e.

D. Proof of Lemma 1

Notice that the payoff function described in Equation 5, is continuous in the hospital's readmission rate. Moreover, the strategy set for each hospital $[0, r_{h0}]$ is convex. According to Glicksberg's Theorem (Glicksberg 1952), this game has at least one Nash Equilibrium in mixed strategies.

To ease our following proofs, we have considered a concrete updating function. The updating functions we consider here is that hospital h calculates $\bar{r}_{h1}^e$ as the average of all readmission rates. In other words $\bar{r}_{h1}^e$ is the average readmission rates of $\{r_{h1}, \forall h\}$ weighted by the number of patients in each hospital. If all H participating hospitals have the same number of patients, then r_{h1}^e is simply $\frac{1}{H}\sum_k r_{h1}$. We denote this updating mechanism as $\bar{r}_{h1}^e = g_1(\vec{r_1})$, where $\vec{r_1} = \{r_{11}, r_{21}, ..., r_{H1}\}$ is vector representing the new readmission rates of all hospitals.

To prove that the game may not have pure Nash Equilibria, consider the following counter example:

There are 3 hospitals with initial readmission rates $\vec{r_0} = \{0.24, 0.244, 0.249\}$. Assume that all hospitals have all revenue from medicare patients ($\forall h \in \{1,2,3\}$, $P_{med,h} = 1$). Also assume that there is no cost to reduce readmissions ($\alpha = 0$), and the maximum penalty is 1% ($P_{cap} = 1\%$). Let the updating mechanism be $g_1(\vec{r_0})$, with all hospitals having same number of patients. In order words, $g_1(\vec{r_0}) = \frac{r_{01}+r_{02}+r_{03}}{3}$.

According to Proposition 1, each hospital only has two best-response actions: reducing to the average ($g_1(\vec{r_0})$) or staying at current readmission rates (r_{h0}). Therefore, we only have four possible pure-strategy Nash Equilibria: (1) hospital 2 does not reduce, and hospital 3 reduces: $\{0.24, 0.244, 0.242\}$, (2) both hospital 2 and hospital 3 reduce: $\{0.24, 0.24, 0.24\}$, (3) neither hospital 2 nor hospital 3 reduces: $\{0.24, 0.244, 0.249\}$. In (1), hospital 2 is better off to reduce to 0.241, indicating that (1) is not an equilibrium. In (2), hospital 3 is better off to stay at 0.249. In (3), hospital 3 gets more revenue if it reduces to the average readmission rates 0.242. Therefore, there is no pure-strategy Nash Equilibrium in the game described above. A similar example can be constructed with $g_2(\vec{r_0}, r_{h0}^e)$.

To show that the game may have multiple equilibria, consider the following example:

Consider the game discussed above but now with $\vec{r_0} = \{0.24, 0.251, 0.252\}$. There are four candidate pure-strategy Nash Equilibria: (1) hospital 2 does not reduce, and hospital 3 reduces: $\{0.24, 0.251, 0.2455\}$, (2) hospital 2 reduces while hospital 3 does not reduce: $\{0.24, 0.246, 0.252\}$, (3) both hospital 2 and hospital 3 reduce: $\{0.24, 0.24, 0.24\}$, (4) neither hospital 2 nor hospital 3 reduce: $\{0.24, 0.251, 0.252\}$. It is easy to verify using Equation 5 that both situation (1) and (2) are pure-strategy Nash Equilibria. Hence, the game could have multiple pure-strategy Nash Equilibria.

E. Proof of Proposition 2

By definition, a hospital h is a SNPE hospital if it satisfies:

$$r_{h0} > f_h(r_{h0}, g_i(\vec{r_0})) \tag{23}$$

By the definition of $f_h(r_{h0}, g_i(\vec{r_0}, r_{h0}^e))$ in Equation 8, $f_h(r_{h0}, g_i(\vec{r_0}, r_{h0}^e)) > r_{h0}^e$, indicating that SNPE hospitals incur penalties. From Equation 8, $f_h(r_{h0}, g_i(\vec{r_0}, r_{h0}^e))$ is increasing in $g_i(\vec{r_0}, r_{h0}^e)$. From the definition of updating functions, it is clear that $g_i(\vec{r'}, r_{h0}^e) \geq g_i(\vec{r}, r_{h0}^e)$ if $\vec{r'} \geq \vec{r}$ for $i = \{1, 2, ..., H\}$. Moreover, since we

restrict a hospital h's strategy set to $[0, r_{h0}]$, at any equilibrium of the game, the corresponding readmission vector is less than or equal to initial readmission vector, $\vec{r_1} \leq \vec{r_0}$.

Therefore, for any equilibrium π with the equilibrium readmission vector $\vec{r_{\pi 1}}$, we have that:

$$h \in SNPE => r_{h0} > f_h(r_{h0}, g_h(\vec{r_0}, \vec{r_0^e})) \geq f_h(r_{h0}, g_h(\vec{r_{\pi 1}}, \vec{r_0^e})) \tag{24}$$

Therefore, for hospitals that satisfy Equation 10, reducing readmissions is a strictly dominated strategy, and at any possible equilibrium, SNPE hospitals do not reduce readmissions. Formally:

$$\forall \pi, \ \forall h \in SNPE, \ \pi_h(r_{h0}) = 1, \tag{25}$$

which concludes the proof.

F.   Proof of Proposition 3

By step 2 of the algorithm, $r_{h,d} \neq r_{h,d-1}$ if $r_{h,1} > g_h(\vec{r}_{d-1}, r^e_{h,d-1})$ and $h \notin SNPE$. This means that if $h \in SNPE$, $r_{h,d} = r_{h,d-1}$ $\forall d$. Therefore, by construction, the set of SNPE hospitals and the set of SPE hospitals are mutually exclusive.

To show that the number of SPE hospitals provide an upper bound on the number of hospitals that reduce readmissions in any equilibrium, let us consider the readmission vector that the algorithm generates, denoted as $\vec{r_T}$. For any equilibrium π with readmission vector $\vec{r_\pi}$, following Proposition 3, the SNPE hospitals will not reduce readmissions. Fix one equilibrium π, for any hospital h that reduces readmissions in this equilibrium, it must satisfy:

$$\begin{aligned} h \notin SNPE \\ r_{h0} > g_h(\vec{r_\pi}, \vec{r_0^e}) \end{aligned} \tag{26}$$

By definition, $g_h(\vec{r_\pi}, r^e_{h0})$ is increasing in $\vec{r_\pi}$. Therefore, if $\vec{r_T} \leq \vec{r_\pi}$ $\forall \pi$, then $\forall \pi$ $r_{h0} > g_h(\vec{r_\pi}, r^e_{h0}) => r_{h0} > g_h(\vec{r_T}, r^e_{h0})$, which would imply that the number of hospitals that reduce readmissions in any equilibrium is bounded by the number of SPE hospitals generated by the algorithm.

Suppose that $\exists \pi$ $s.t.$ $\vec{r_T} > \vec{r_\pi}$, then it means that $\exists h$ such that $r_{h,T} > r_{h,\pi}$ and in turn $r_{h0} > g_h(\vec{r_T}, r^e_{h0})$. By the termination condition of the algorithm, we know that

$$r_{h0} \notin [g_i(\vec{r_T}, r^e_{h0}), f(r_{h0}, g_i(\vec{r_T}, \vec{r_0^e}))] \tag{27}$$

with $f(r_{h0}, g_h(\vec{r_T}, r^e_0))$ defined in Equation 8. Since $f(r_{h0}, g_h(\vec{r}, r^e_{h0}))$ is increasing in $g_h(\vec{r}, r^e_{h0})$, we have:

$$\vec{r_T} \geq \vec{r_\pi} => g_h(\vec{r_T}, r^e_{h0}) \geq g_h(\vec{r_\pi}, r^e_{h0}) => f(r_{h0}, g_h(\vec{r_T}, \vec{r_0^e})) \geq f(r_{h0}, g_h(\vec{r_\pi}, \vec{r_0^e})) \tag{28}$$

Therefore,

$$r_{h0} > f(r_{h0}, g_h(\vec{r_T}, \vec{r_0^e})) => r_{h0} > f(r_{h0}, g_h(\vec{r_\pi}, \vec{r_0^e})) => r_{h0} \notin [g_h(\vec{r_\pi}, \vec{r_0^e}), f(r_{h0}, g_h(\vec{r_\pi}, \vec{r_0^e}))] \tag{29}$$

Hence it is not optimal for the hospital h to reduce its readmissions. We reach a contradiction, and $\nexists \pi$ $\vec{r_T} \geq \vec{r_\pi}$.

G.   Proof of Proposition 4

Let us first prove that the set of SNPE hospitals in the single-year game where $P_{cap} = P_{cap}^{max}$ remains a valid lower bound. Rewrite equation 5 as:

$$R(P_{cap}, r_{h0}, r_h, r_h^e) = \Pi_h^P(r_h) - P_h(r_h, r_h^e, P_{cap}) - C(r_{h0}, r_h) \tag{30}$$

where $P_h(r_h, r_h^e, P_{cap})$ is the penalty under the cap P_{cap}. It is easy to see, by the definition of the penalty function, that $P_h(r, r_h^e, P_{cap})$ is increasing in P_{cap}. Therefore, if $r_1 > r_2 \geq r_h^e$ and $P_{cap}^1 > P_{cap}^2$:

$$R(P_{cap}^1, r_{h0}, r_1, r_h^e) - R(P_{cap}^1, r_{h0}, r_2, r_h^e) > 0$$

$$=> R(P_{cap}^2, r_{h0}, r_1, r_h^e) - R(P_{cap}^2, r_{h0}, r_2, r_h^e) > 0 \tag{31}$$

In other words, $R(P_{cap}, r_{h0}, r, r_h^e)$ is super-modular in (P_{cap}, r).

Denote the set of SNPE hospitals under P_{cap}^{max} as $SNPE^{max}$. Suppose that there exists a Sub-game Perfect Nash Equilibrium (SGPNE) π in the game (with the readmission vector, in the last period, given by $r_{h,\pi}$) such that a hospital h satisfies that $h \in SNPE^{max}$ and $r_{h,\pi} < r_{h0}$. Since h is an $SNPE^{max}$ hospital, it holds, in particular, that

$$R(P_{cap}^{max}, r_{h0}, r_{h0}, g_i(r_{h,\pi}, r_{h0}^e)) > R(P_{cap}^{max}, r_{h0}, r, g_i(r_{h,\pi}, r_{h0}^e)) \ \forall r < r_{h0} \tag{32}$$

which implies that for all $P_{cap} < P_{cap}^{max}$,

$$R(P_{cap}, r_{h0}, r_{h0}, g_i(r_{h,\pi}, r_{h0}^e)) > R(P_{cap}, r_{h0}, r, g_i(r_{h,\pi}, r_{h0}^e)) \ \forall r < r_{h0} \tag{33}$$

Moreover, it is clear to see that $R(P_{cap}^1, r_{h0}, r_1, r_h^e) - R(P_{cap}^1, x, r_1, r_h^e) > 0$ if $x < r_{h0}$ since the action set under x ($[0, x]$) is a subset of the action set under r_{h0} ($[0, r_{h0}]$).

Therefore, letting $\vec{r_{\pi}^t}$ be the readmission vector at the end of stage t. The total operating margin the hospital h collects in this SGPNE is less than the total operating margin collected under the strategy of doing nothing:

$$\sum_{t=1}^{N} R(P_{cap}^t, r_{h,\pi}^{t-1}, r_{\pi}^t, g_i(r_{\pi}^t, r_{h0}^e)) < \sum_{t=1}^{N} R(P_{cap}^t, r_{h0}, r_{\pi}^t, g_i(r_{\pi}^t, r_{h0}^e))$$

$$< \sum_{t=1}^{N} R(P_{cap}^t, r_{h0}, r_{h,0}, g_i(r_{\pi}^t, r_{h0}^e)) \tag{34}$$

Note that the first inequality follows the restriction of the action to reduction in readmissions ($r_{\pi}^t \leq r_{h0} \ \forall t$). This shows that the hospital h's no-action strategy is optimal, and in turn the SGPNE is not a valid equilibrium.

We next prove that the set SPE hospitals under the maximum penalty cap is also a valid upper bound on the number of hospitals that reduce readmissions relative to r_{h0} in any equilibrium. Following our strategy in the proof of Proposition 4, we must prove that for any SGPNE π, and at any period t, $\vec{r_{\pi}^t} \geq \vec{r_T}$ as generated by the algorithm with $P_{cap} = P_{cap}^{max}$. Since r_{π}^t is monotone decreasing in t, it suffices to show that $r_{\pi} = r_{\pi}^N \geq \vec{r_T}$. We can restrict attention to non-SNPE hospitals since we have shown above that the SNPE hospitals do not reduce readmission in any equilibrium π.

Suppose that there exists an equilibrium π such that $\vec{r}_\pi^N < \vec{r}_T$, then $\exists h$ such that $r_{h,T} > \vec{r}_{h,\pi,N}$ and, in particular, $r_{h0} > g_i(\vec{r}_T, r_{h0}^e)$. Proposition 4 was concerned with a single stage. Here, this single stage will be the last stage of the game. The initial condition of stage N is $r_{h,\pi}^{N-1}$. If we can now show that $r_{h,\pi}^{N-1} \notin [\vec{r}_T, f(r_{h0}, g_i(\vec{r}_T, r_{h0}^e))]$ (which is the analogue of Equation 27) then we can follow the proof in Proposition 4 to reach a contradiction.

To prove $r_{h,\pi}^{N-1} \notin [g_i(\vec{r}_T, r_{h0}^e), f(r_{h0}, g_i(\vec{r}_T, r_{h0}^e))]$ we use induction. For $n = 1$, this relation holds trivially since (recall) $R(P_{cap}, r_{h0}, r, r_h^e)$ is super-modular in (P_{cap}, r). Now, let us assume this relation holds for all $k \leq n-1$ and show it prove that $r_{h,\pi}^n \notin [g_i(\vec{r}_T, r_{h0}^e), f(r_{h0}, g_i(\vec{r}_T, r_{h0}^e))]$.

Suppose not, then:

$$r_{h,\pi}^n = g_i(\vec{r_\pi}^n, r_{h0}^e) < f(r_{h0}, g_i(\vec{r}_T, r_{h0}^e)) \tag{35}$$

Since $r_{h,\pi}^{n-1} \notin [g_i(\vec{r}_T, r_{h0}^e), f(r_{h0}, g_i(\vec{r}_T, r_{h0}^e))]$, there exist a set of hospitals, denoted as H, such that $r_{h,\pi}^{n-1} > f(r_{h0}, g_i(\vec{r}_T, r_{h0}^e))$, and they reduce to $g_i(\vec{r_\pi}^n, r_{h0}^e)$. By definition, both $g_1(\vec{r}, r_{h0}^e)$ and $g_2(\vec{r}, r_{h0}^e)$ are contraction mapping of $\vec{r}$ component-wise, in other words:

$$|g_i(\vec{r}, r_{h0}^e)) - g_i(\vec{r'}, r_{h0}^e))| < |r - r'| \tag{36}$$

we know that $\sum_{h \in H} (r_{h,\pi}^n - r_{h,\pi}^{n-1}) > \sum_{h \in H} [g_i(\vec{r_\pi}^n, r_{h0}^e) - g_i(\vec{r_\pi}^{n-1}, r_{h0}^e)]$. However, based on the optimality condition of the game, $\sum_{h \in H} r_{h,\pi}^n = \sum_{h \in H} g_i(\vec{r_\pi}^n, r_{h0}^e)$. This is a contraction, and therefore $r_{h,\pi}^n \notin [g_i(\vec{r}_T, r_{h0}^e), f(r_{h0}, g_i(\vec{r}_T, r_{h0}^e))]$. This concludes the proof.

References

Jan Paul Acton. Nonmonetary factors in the demand for medical services: some empirical evidence. *The Journal of Political Economy*, pages 595–614, 1975.

Carol M Ashton, Deborah J Del Junco, Julianne Souchek, Nelda P Wray, and Carol L Mansyur. The association between the quality of inpatient care and early readmission: a meta-analysis of the evidence. *Medical care*, pages 1044–1059, 1997.

David Card, Carlos Dobkin, and Nicole Maestas. The impact of nearly universal insurance coverage on health care utilization and health: evidence from medicare. *American Economic Review*, 98(5):2242–2258, 2008.

Pierre-André Chiappori, Franck Durand, and Pierre-Yves Geoffard. Moral hazard and the demand for physician services: first lessons from a french natural experiment. *European economic review*, 42(3-5): 499–511, 1998.

Eric A Coleman, Carla Parry, Sandra Chalmers, and Sung-Joon Min. The care transitions intervention: results of a randomized controlled trial. *Archives of internal medicine*, 166(17):1822, 2006.

Janet M Corrigan and James B Martin. Identification of factors associated with hospital readmission and development of a predictive model. *Health Services Research*, 27(1):81, 1992.

David M Cutler and Jonathan Gruber. Does public insurance crowd out private insurance? *The Quarterly Journal of Economics*, 111(2):391–430, 1996.

Daniel Deneffe and Robert T Masson. What do not-for-profit hospitals maximize? *International Journal of Industrial Organization*, 20(4):461–492, 2002.

Kumar Dharmarajan, Angela F Hsieh, Zhenqiu Lin, Héctor Bueno, Joseph S Ross, Leora I Horwitz, José Augusto Barreto-Filho, Nancy Kim, Susannah M Bernheim, Lisa G Suter, et al. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia readmissions. *JAMA*, 309(4):355–363, 2013.

David Dranove and Paul Wehner. Physician-induced demand for childbirths. *Journal of Health Economics*, 13(1):61–73, 1994.

Phil B Fontanarosa and Robert A McNutt. Revisiting hospital readmissions. *JAMA*, 309(4):398–400, 2013.

D Glass, C Lisk, and J Stensland. Refining the hospital readmissions reduction program. *Washington, DC: Medicare Payment Advisory Commission*, 2012.

Irving L Glicksberg. A further generalization of the kakutani fixed point theorem, with application to nash equilibrium points. *Proceedings of the American Mathematical Society*, 3(1):170–174, 1952.

Femida H Gwadry-Sridhar, Virginia Flintoft, Douglas S Lee, Hui Lee, and Gordon H Guyatt. A systematic review and meta-analysis of studies comparing readmission rates and mortality rates in patients with heart failure. *Archives of Internal Medicine*, 164(21):2315, 2004.

Luke O Hansen, Jeffrey L Greenwald, Tina Budnitz, Eric Howell, Lakshmi Halasyamani, Greg Maynard, Arpana Vidyarthi, Eric A Coleman, and Mark V Williams. Project BOOST: Effectiveness of a multi-hospital effort to reduce rehospitalization. *Journal of Hospital Medicine*, 8(8):421–427, 2013.

HealthCare.gov. *Report to Congress: national strategy for quality improvement in health care.* Health-Care.gov, 2011.

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

Karen E Joynt, E John Orav, and Ashish K Jha. Thirty-day readmission rates for medicare beneficiaries by race and site of care. *JAMA: The Journal of the American Medical Association*, 305(7):675–681, 2011.

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

Harlan M Krumholz, Eugene M Parent, Nora Tu, Viola Vaccarino, Yun Wang, Martha J Radford, and John Hennen. Readmission after hospitalization for congestive heart failure among medicare beneficiaries. *Archives of Internal Medicine*, 157(1):99, 1997.

MedPAC. *Report to the Congress: Medicare payment policy.* Medicare Payment Advisory Commission, 2007.

Vincent Mor, Orna Intrator, Zhanlian Feng, and David C Grabowski. The revolving door of rehospitalization from skilled nursing facilities. *Health Affairs*, 29(1):57–64, 2010.

Mary D Naylor, Dorothy Brooten, Roberta Campbell, Barbara S Jacobsen, Mathy D Mezey, Mark V Pauly, and J Sanford Schwartz. Comprehensive discharge planning and home follow-up of hospitalized elders. *JAMA: the journal of the American Medical Association*, 281(7):613–620, 1999.

Ben Pearson, Robert Skelly, Derek Wileman, and Tahir Masud. Unplanned readmission to hospital: a comparison of the views of general practitioners and hospital staff. *Age and ageing*, 31(2):141–143, 2002.

Carol Propper and John Van Reenen. Can pay regulation kill? panel data evidence on the effect of labor markets on hospital performance. *Journal of Political Economy*, 118(2):222–273, 2010.

Michael W Rich, Valerie Beckham, Carol Wittenberg, Charles L Leven, Kenneth E Freedland, and Robert M Carney. A multidisciplinary intervention to prevent the readmission of elderly patients with congestive heart failure. *New England Journal of Medicine*, 333(18):1190–1195, 1995.

Rajendu Srivastava and Ron Keren. Pediatric readmissions as a hospital quality measure. *JAMA*, 309(4): 396–398, 2013.

Simon Stewart, John E Marley, and John D Horowitz. Effects of a multidisciplinary, home-based intervention on planned readmissions and survival among patients with chronic congestive heart failure: a randomised controlled study. *The Lancet*, 354(9184):1077–1083, 1999.

Muthiah Vaduganathan, Robert O Bonow, and Mihai Gheorghiade. Thirty-day readmissions: The clock is ticking. *JAMA*, 309(4):345–346, 2013.

Carl van Walraven, Carol Bennett, Alison Jennings, Peter C Austin, and Alan J Forster. Proportion of hospital readmissions deemed avoidable: a systematic review. *Canadian Medical Association Journal*, 183(7):E391–E402, 2011.

Anita A Vashi, Justin P Fox, Brendan G Carr, Gail D'Onofrio, Jesse M Pines, Joseph S Ross, and Cary P Gross. Use of hospital-based acute care among patients recently discharged from the hospital. *JAMA*, 309(4):364–371, 2013.

Christopher J Zook, Sheila Flanigan Savickis, and Francis D Moore. Repeated hospitalization for the same disease: a multiplier of national health costs. *The Milbank Memorial Fund Quarterly. Health and Society*, pages 454–471, 1980.