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; revised Nov 23, 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 inherent in HRRP’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-hospital and multi-hospital models shows that, while 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 to quantify the results and insights derived from the model. We draw policy recommendations building on our study of the subtle interaction between various drivers of the policy effectiveness, such as localizing the benchmarking process. Last, we validate our model with recent hospitals’ performance data collected since the policy was implemented.

Key words: Healthcare Operations, Public Policy.

1. Introduction

According to the Medicare Payment Advisory commission (MedPAC) (Gerhardt et al. 2013), 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 Services

1
Figure 1: Timeline of the Hospital Readmissions Reduction Program

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

The Hospital Readmissions Reduction Program (HRRP) was implemented by CMS on October 1, 2012, as a response to the increasing costs of readmissions. The program penalizes Medicare payments to hospitals with high 30-day readmission rates for acute myocardial infarction (AMI), heart failure (HF), and pneumonia (PN). Chronic obstructive pulmonary disease (COPD) and hip/knee arthroplasty (THA/TKA) 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 expected given the hospital’s case mix. The CMS model determines the targets by benchmarking hospitals against their peers.

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 HRRP are: (i) hospitals are not the appropriate entities to be held accountable for readmissions, since some causes of readmission are outside the control of hospitals. Only a small fraction of readmissions may 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).

1 Hospitals that serve Medicare patients and are under the Medicare payment system are called IPPS hospitals.
Supporters of HRRP argue that the purpose of HRRP is not to directly affect quality of care but rather make the hospital accountable also for post-discharge processes\(^2\). Others point to the large number of patients whose discharge is fraught with poor communication, ineffective medication management and inadequate hand-offs to primary care physicians or nursing homes. A report by the Medicare Payment Advisory Commission supports HRRP in estimating a small but significant decrease in national rates of readmission for all diseases from 15.6% in 2009 to 15.3% after the introduction of HRRP (HealthCare.gov 2011). A recent study by CMS shows that readmission rates fell during 2012 in more than 239 out of the 309 hospital referral regions (HRR) (Gerhardt et al. 2013)

This paper does not argue 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 adopt 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 HRRP provide economic incentives for a hospital to reduce its readmissions? What are the characteristics of hospitals that prefer paying penalties over reducing readmissions (worst offenders)? And how does the HRRP benchmarking (and the competition it induces) affect who is a worst offender?

Readmission-reduction decisions present hospitals with trade-offs between cost and revenue drivers: (i) The reduction of the penalty due to readmission improvements: 2,217 hospitals nationwide cumulatively 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: If a non-negligible portion of a hospital’s patients are covered under a pay-per-case insurance scheme, readmissions may account for a non-negligible proportion of the hospital’s contribution margin. (iii) Process-improvement cost: Reducing readmissions may involve costly process changes.

CMS determines the expected readmission rate for each hospital using discharge-level data for IPPS hospitals from the previous three years. Per monitored disease, 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 and that disease. If the hospital’s predicted readmission rate, based on the hospital’s actual performance, for the next year is greater than its CMS-expected

\(^2\)http://blogs.sph.harvard.edu/ashish-jha/the-30-day-readmission-rate-not-a-quality-measure-but-an-accountability-measure/
readmission rate, the hospital incurs a penalty up to the maximal cap – currently 1% of overall Medicare payments to the hospital. 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 hospitals’ decision making as one hospital’s penalty is determined not only by its own actions, but also by the performance of all other (similar) hospitals.

We use analytic modeling, data analysis and simulation to study the impact of HRRP on a hospitals’ readmission reduction efforts. We develop a theoretical model that captures the patient flow from readmissions, the financial drivers in a hospital’s decision, and the game-theoretical nature of the policy. 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 readmission-reduction cost. We allow for a flexible specification of the process-improvement cost, which could capture other incentives related to readmission reductions, such as back-fill opportunities, reputation effects.

We take initially the view that hospitals are non-strategic and do not take into account how the CMS-targets are affected by decisions made by other hospitals. Our single-hospital model captures the characteristics of hospitals that are incentivized to reduce their readmission rates in response to HRRP. If hospitals do, however, take into account the decisions of their peers, one would want to assess the effect of the strategic interaction on hospital response. Consequently, we introduce a game where hospitals determine their readmission reduction efforts, taking other hospitals’ actions into account. We show that pure-strategy equilibria need not exist and, even if they exist, need not be unique. We are able, however, to identify bounds that apply to all equilibria of the game. Our lower bound on the number of hospitals that are not incentivized by HRRP captures the limits of the policy effectiveness.

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, particularly, on the set of strongly non-incentivized hospitals. In brief, the following are observations arise from our simulation study: **Hospital characteristics:** The effectiveness of HRRP depends on various hospital characteristics: 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 version of HRRP is not effective in inducing hospitals with poor performance (worst offenders) 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 at a relative disadvantage under the current structure of the policy. Forth, the higher the contribution margin ratio of a hospital, the
smaller the likelihood of the hospital to reduce its readmissions under HRRP. This suggests that a better regulated payment system may be helpful in incentivizing hospitals to reduce readmissions.

**Technology:** The cost of process improvement plays an important role in the hospitals’ responses to HRRP. 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.

Some of our findings and policy implications in Section 6 draw on the subtle ways in which the different drivers mentioned above interact with each other. For example, the fraction of Medicare patients in a given disease affects which diseases the government should target for quality improvement programs; see Section 6.4.

**Competition and “readmission dispersion”**. The Medicare Payment Advisory Commission emphasizes the importance of the competition introduced by the HRRP benchmarking procedure (Glass et al. 2012), and rejects the idea of having a fixed target for each hospital (MedPAC 2013). It is therefore important to understand the hospitals’ decisions when they consider the strategic interactions among them. We find that the competition induced by HRRP may indeed incentivize more hospitals to reduce readmissions relative to the individual hospital (no benchmarking) model. At the same time, we find, that competition increases the number of worst-offenders, hospitals that have high readmission rate yet are not incentivized by HRRP to improve these. The number of these hospitals is expected to increase as the set of monitored diseases is expanded in 2015 and beyond.

Importantly, the effectiveness of the policy depends on the dispersion in readmission rates over the set of hospitals. The higher the dispersion is, the less effective the policy is. Therefore we propose an alternative benchmarking mechanism of the policy and link the mechanism with existing recommendations from CMS (MedPAC 2013); see Section 6.2.

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 by comparing the simulation results to the actual changes in hospitals predicted readmission ratios between 2013 and 2015. This initial empirical evidence supports the power of our model in identifying those worst-offenders.

We conclude this introduction with a brief literature review. Much of the medical literature focuses on the causes of readmission and on hospital-level process improvement programs for readmission reduction (Dharmarajan et al. 2013, Krumholz et al. 1997 and Stewart et al. 1999). Using 2003-2004 Medicare data, 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, Joynt et al. (2011) examine 30-day readmissions for AMI, HF, and PN, and show that Medicare patients from a poor socioeconomic background have particularly high risk of being readmitted.
The literature also demonstrates that simple process-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, Hu et al. 2014), keeping patients longer in the in-patient unit (Bartel et al. 2014), using machine learning techniques (Bayati et al. 2014), and conducting a nurse-directed multidisciplinary intervention (Rich et al. 1995) – can effectively reduce readmissions. From a queuing perspective, readmissions are retrials to a queue. The operations management literature offers various insights into the dynamic management of queues with retrials (De Véricourt and Zhou 2005, Ren and Zhou 2008, Aksin et al. 2007) and can serve as a basis to study how process improvement within the hospital can affect readmissions. We abstract away from such questions in this paper.

Since the introduction of the US Patient Protection and Affordable Care Act (ACA), in particular its HRRP component, the medical literature studied the structure of the program and its 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 proposes, for pediatric hospitals, to focus on other monitored conditions; Vashi et al. (2013) estimate 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. Vest et al. (2010) suggests that the readmission reduction policy should carefully distinguish between preventable and non-preventable readmissions. Furthermore, van Walraven et al. (2011) conducts a survey of literature on preventable readmissions and concludes, similarly that it is unclear which readmission is avoidable, and some care is needed in this regard.

Here we take HRRP as given and ask whether HRRP financially incentivizes 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 hospital-level patient flow model and link it to the contribution margin of a hospital. This sets the foundations for analyzing individual hospital behavior in Section 3 and hospitals’ joint equilibria in Section 4.


2.1. Hospital flow model

A hospital faces an exogenous arrival of patients for each disease $i$, e.g., AMI, HF, and each insurance type $j$, e.g., Medicare, Medicaid, Private insurance, with a rate $\lambda_{ij}^E$. The HRRP distinguishes between Medicare, Medicaid, private insurance, military insurance and other insurance types. For every disease type $i$ and insurance type $j$, the readmission rate is denoted by $r_{ij}$. A patient requiring readmission could either return to the hospital from which 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 divergence probability and denote it by $d_h$. A hospital-flow diagram for patients with disease $i$ and payment $j$ is shown in Figure 2, where $\lambda_{ij}^{d_h}$ is the rate of the readmitted patients coming from other hospitals. Finally, $\lambda_{ij}$ is the total throughput of patients in group $ij$.

A hospital receives a payment $p_{ij}$ for treating a patient of type $ij$. Readmitted patients may receive a different payment. For example, Jencks et al. (2009) shows that the weighted payment index for initial admission is 1.41 while only 1.35 for 30-day readmission. We denote by $l$ the readmission adjustment factor which is the percentage difference in payment per readmission. Thus, the revenue for the $k^{th}$ readmission is $l^kp_{ij}$.

Let $\lambda_{ij} = \lambda_{ij}^E + \lambda_{ij}^{d_h}$ be the total incoming rate for patient group $ij$, the combined arrival rate (exogenous patients plus readmitted patients) for disease $ij$ satisfies: $\lambda_{ij} = \lambda_{ij}^a + r_{ij}(1-d_h)\lambda_{ij}$, so that

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

While some research claims that demand is not truly exogenous as patients can be induced to visit the hospitals (Acton 1975), econometric studies show that hospitals can only minimally alter their incoming rate of patients (Dranove and Wehner 1994).
where $\lambda_{ij}$ is the total throughput of the hospital. Adding the subscript $h$ to denote the specific hospital, hospital $h$’s revenue per disease and in total is computed as follows:

$$
\Pi_{ijh}^R(0) = \lambda_{ijh} p_{ijh}
$$

$$
\Pi_{ijh}^R(r_{ijh}) = \Pi_{ijh}^R(0) \frac{1}{1 - l(1 - d_h)r_{ijh}}
$$

$$
\Pi_h^R(r_h) = \sum_{ij} \Pi_{ijh}^R(r_{ijh}).
$$

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

$$
\Pi_h^C(r_h) = C_m \Pi_h^R(r_h).
$$

**Remark 1.** For clarity and tractability of the model, we do not specifically model the disease-level divergence of readmissions. In Appendix C, with a two-disease model and real data, we show that our main result and its implications are robust towards this simplification.

### 2.2. Penalty structure

On October 2012, CMS started penalizing Medicare payments to hospitals based on their excess readmission ratio for three monitored diseases. We let $D$ denote the set of all diseases and $M \subset D$ denote the set of monitored diseases. The excess readmissions of a hospital for each monitored disease $i$ is measured by the ratio of its risk-adjusted predicted readmission rate ($r_{ijh}$ for disease $i$ and hospital $h$) and its risk-adjusted expected readmission rate ($r_{ijh}^e$ for disease $i$ and hospital $h$). The term “risk-adjusted” refers to the fact that the estimated readmission rate for a hospital is adjusted to the risk profile of its patients. Therefore, the more severe patients a hospital has, the higher the risk-adjusted expected and predicted readmission rates for that hospital. This risk adjustment prevents the discrimination of hospitals with more severe patients.

CMS computes the expected and predicted readmission rates for each hospital and each disease by applying the HGLM model to discharge-level data; see Appendix A. For each hospital $h$, the risk-adjusted predicted readmission rate, $r_{ijh}$, predicts the readmission rate of disease-insurance pair $(i, \text{Med})$ in hospital $h$ for the following year, conditional on its case mix remaining unchanged. In our model, we treat a hospital’s current readmission rate as its predicted readmission rates. The risk-adjusted expected readmission rate $r_{ijh}^e$ is, roughly speaking, the readmission rate of the national average hospital with the same case mix as hospital $h$.

If the excess readmission ratio, computed as $\frac{r_{ijh}}{r_{ijh}^e}$, for hospital $h$ is greater than 1, the payment for excess readmissions in disease $i$ for hospital $h$ is defined as the excess readmission ratio for
disease \( i \) minus 1 multiplied by the revenue from Medicare patients in disease \( i \). In other words, the payment for excess readmissions in disease \( i \) is 
\[
\max \left( \frac{r_{ijh}}{r^c_{ijh}} - 1, 0 \right) \Pi_{ijh}(r_{ijh}),
\]
where \( j = \text{Med}. \)

CMS computes the aggregate payments for excess readmissions for hospital \( h \) as the sum of payments across monitored diseases,
\[
\sum_{i \in M, j = \text{Med}} \max \left( \frac{r_{ijh}}{r^c_{ijh}} - 1, 0 \right) \Pi_{ijh}(r_{ijh}).
\]

CMS then defines the readmission penalty ratio as the aggregate payments for excess readmissions divided by the Medicare revenue over all diseases,
\[
\sum_{i \in D, j = \text{Med}} \Pi_{ijh}(r_{ijh}).
\]
The readmission penalty ratio is capped at the maximum penalty cap, denoted as \( P_{\text{cap}}. \) The absolute amount of the penalty is the Medicare revenue across all diseases multiplied by the readmission penalty ratio:
\[
P_h(r_h, r^c_h) = \min \left( \sum_{i \in M, j = \text{Med}} \max \left( \frac{r_{ijh}}{r^c_{ijh}} - 1, 0 \right) \Pi_{ijh}(r_{ijh}), P_{\text{cap}} \right).
\]

Notice that the first summation is only over monitored diseases while the second one is over all diseases and captures the hospital’s total Medicare revenue.

### 3. Single-Hospital and Single-Disease Model

To gain structural insights, let us first suppose that there is a single monitored disease \((M = D = \{i\})\) and a single insurance type, which is Medicare \((j = \text{Med})\). In this setting, we can drop the subscripts \(ij\). In Section 5, we show how to combine multiple single-disease models to reflect the penalty across multiple monitored diseases dictated in the policy.

#### 3.1. Contribution

The revenue, the contribution margin, and the penalty for hospital \( h \) with actual readmission rate \( r \) and risk-adjusted CMS-expected readmission ratio \( r^c \) are then given by:
\[
\Pi_h^R(r) = \Pi_h^R(0) \frac{1}{1 - l(1 - d_h)r_h},
\]
\[
\Pi_h^P(r) = C_m \Pi_h^R(r),
\]
\[
\mathbb{P}_h(r, r^c) = \phi_h \Pi_h^R(r) \min \left( \max \left( \frac{r}{r^c} - 1, 0 \right), P_{\text{cap}} \right),
\]
where \( \phi_h \) is the percentage of hospital \( h \)’s revenue that comes from Medicare. In the special case that \( d_h = l = 0 \), we have the simpler expression
\[
\Pi_h^R(r) = \Pi_h^R(0) \frac{1}{1 - r} = \lambda_h p_h \frac{1}{1 - r}.
\]
3.2. Cost of process improvements

Reducing readmissions may require process changes (Naylor et al. 1999) and/or increases in staffing (Stewart et al. 1999). These are long-term commitments. We assume that if a hospital reduces its readmission from \( r_0 \) to \( r \), an annual readmission-management cost \( C(r_0, r) \) is added to the hospitals operating costs in each subsequent year. The function \( C(y, x) \) is assumed to be continuous, non-negative when \( x \leq y \), and has a second derivative that satisfies \( \frac{\partial^2}{\partial x \partial x} C(y, x) \leq \eta \frac{1 - x}{(1 - x)^3} \) for some constant \( \eta \) and \( \forall x, r \in (0, 1) \). This technical assumption is satisfied, in particular, by any function of the form

\[
C_h(y, x) = C^v_h(y - x)^\alpha + g(y),
\]

where \( g \) is any bounded function and \( \alpha \geq 0 \). The term \( C^v_h(r - x)^\alpha \) represents the variable cost of reducing readmission rates from \( r \) to \( x \). When \( \alpha \in (0, 1) \), this cost is concave, representing economies of scale in reducing readmissions. It is convex if \( \alpha > 1 \) representing a marginally increasing difficulty in reducing readmissions. The second term, \( g(y) \), captures dependence of the cost on the initial starting point. Some small amount of readmission might be unavoidable and, for hospitals that have initially low readmission rates, further reductions might be expensive.

In principle, hospitals could discriminate patients based on characteristics that are not accounted for in CMS’s risk adjustments but that do affect readmissions. For example, hospitals could choose to treat only patients with relatively high socioeconomic status and relatively simple conditions. As readmission rates are negatively correlated with patients’ socioeconomic status (Joynt et al. 2011) and health-condition complexity (Joynt and Jha 2013), this could decrease the hospital’s readmission without changing the target set by CMS. Such “gaming” (e.g., costless readmission reductions), no doubt, compromise the effectiveness of HRRP. Our purpose in this paper is to identify the fundamental limits (and drivers) of HRRPs effectiveness even in the absence of such gaming.

3.3. Structure of the optimal policy

A hospital \( h \) with current readmission rate \( r_{h0} \) and CMS-expected readmission rate \( r^e_h \), that decides to reduce its readmission to \( r_{h1} \) has operating margin

\[
R(r_{h0}, r_{h1}, r^e_h) = \Pi^h_{r_{h1}} - \Pi_h(r_{h1}, r^e_h) - C(r_{h0}, r_{h1})
\]

\[
= C_m \Pi^R_h(0) \frac{1}{1 - r_{h1}} \left( 1 - \phi^\text{med}_m \right) \frac{1}{C_m} \min \left( \max \left( \frac{r_{h1}}{r^e_h} - 1, 0 \right), P_{\text{cap}} \right) - C(r_{h0}, r_{h1}), \tag{8}
\]

for the subsequent year. Thus, the hospital solves the maximization problem

\[
r^*_h(r_{h0}, r^e_h) \in \arg \max_{x \leq r_{h0}} R(r_{h0}, x, r^e_h) \tag{9}
\]
We assume here that hospitals are operating-margin maximizers. Whereas, legally speaking, nonprofit hospitals should not incentivize their management group to maximize any form of profit, past studies have shown that these hospitals do behave as profit maximizers in a competitive market (Deneffe and Masson 2002). We also assume in this formulation that hospitals do not deliberately increase readmission rates. This is grounded in ethical reasons but is also consistent with the spirit of our analysis that focuses on best case outcomes of the policy and seeks to identify hospitals that “fall outside” of the policies effectiveness boundaries.

The following characterizes the optimal solution: a hospital has financial incentive to reduce its readmissions only if its readmission rate is contained in an interval, the width of which depends on the hospital’s cost structure and other parameters.

**Proposition 1.** The optimal decision for hospital \( h \) with current readmission rate \( r_{h0} \) is either to remain at its current readmission rate or to reduce its readmission rate to the CMS-expected readmission rate \( r_{e_h} \):

\[
r_{h1}^{*}(r_{h0}, r_{e_h}) = \begin{cases} r_{h}^{e} & \text{if } r_{h0} \in [r_{h}^{e}, f(r_{h0}, r_{h}^{e})], \\ r_{h0} & \text{otherwise,} \end{cases}
\]

where \( f(r_{h0}, r_{e_h}) \) is the maximal solution to the equation:

\[
R(f(r_{h0}, r_{h}^{e}), r_{e_h}) = R(f(r_{h0}, r_{h}^{e}), f(r_{h0}, r_{e_h}), r_{e_h})
\]

and \( R(\cdot, \cdot, \cdot) \) is as in Equation 8.

Notice that \( f(r_{h0}, r_{h}^{e}) \) is always greater than \( r_{e_h} \) since the contribution function is continuous. The left panel of Figure 3 depicts hospital \( h \)’s operating margin as a function of its targeted readmission rate, \( r_{h1} \), with \( d_h = l = 0, r_{h0}^{e} = 0.2 \), and no readmission reduction costs. The red vertical line indicates the position of the expected readmission rate, \( r_{h}^{e} \) and the red square denotes the initial readmission rate \( r_{h0} \). The green vertical line 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 \( r_{h1} \) to \( r_{h}^{e} \). A hospital has financial incentive to reduce its readmissions if and only if its current readmission rate falls in the region \([A, B]\). We define this region as the **policy effective region** – hospitals that fall in this interval act optimally by reducing readmissions in response to HRRP penalties.

There are three parameter regions in Figure 3:

Region (1) (Program-Indifferent Region, \([0, A]\)). A hospital is in this region if its original readmission rate \( r_{h0} \) is smaller than its CMS-expected readmission \( r_{h}^{e} \). Its operating margin is strictly increasing with its readmission rates, indicating that the optimal decision for the hospital is to stay at current readmission rate. 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 savings in penalties from reducing the readmission rate outweigh the loss of contribution. The hospital’s optimal decision is to reduce its readmission rate to \( r_{eh} \) (recall that hospitals in our model can only reduce readmissions). We refer to these hospitals 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. 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.

There are two metrics that affect a hospital’s decision to reduce readmissions, assuming the hospital has \( r_{h0} > r_{eh} \). The first is the magnitude of the excess rate \( r_{h0} - r_{eh} \). If this excess rate is large, the hospital falls in Region (3) and it is less financially beneficial for the hospital to reduce readmissions. The second metric is the width, \( f(r_{h0}, r_{eh}) - r_{eh} \), of Region (2). The wider this region is, the more likely it is to cover \( r_{h0} \), making it beneficial for the hospital to reduce readmissions.

Figure 3 shows how the width of Region (2), \( f(r_{h0}, r_{eh}) - r_{eh} \), changes with other parameters (i.e. \( \phi_{med}^m, C_m, d_h, \) and \( l \)). These observations lead to the following corollary that summarizes the comparative statics linking \( f(r_{h0}, r_{eh}) - r_{eh} \) to the primitives of a hospital (\( l, d_d, d_h, \lambda_a, \phi_{med} \)).

**Corollary 1.** The width of Region (2) \( f(r_{h0}, r_{eh}) - r_{eh} \) for a hospital \( h \) is weakly increasing in the percentage \( \phi_{med}^m \) of Medicare patients and the hospital divergence probability \( d_h \); and weakly decreasing in the readmission-reduction cost coefficient \( C_h \), the contribution margin ratio \( C_m \) and the adjustment factor \( l \).

These comparative statistics are inherent to the HRRP’s design. First, the fact that the hospital divergence probability and the inverse of adjustment factor have the same effect is expected. By CMS’s rules a readmitted patient contributes to the readmissions of the hospital from which the
patient was initially discharged. Consequently, increasing the divergence probability is effectively
equivalent to reducing the contribution from readmitted patients and has the same effect.

Second, since the penalty is proportional to the contribution of a hospital from Medicare patients
but the revenue reduction due to reduced readmission applies to all patients, hospitals with a low
percentage of Medicare patients are less affected by HRRP.

Last, as expected, when the margin $C_m$ is high the opportunity cost associated with reducing
readmissions is larger. The percentage of contribution margin of a hospital has an inverse relation-
ship with its inclination to reduce readmissions.

We will revisit these comparative statics after introducing an expanded model where hospitals' actions impose externalities on other hospitals through the CMS-expected readmission rate. Proposition 1 will be useful because the bounds that we derive in the next section are based on this single-hospital/single-disease model.

4. Game-Theoretic Model

Building on the single-hospital model of the previous section, we now construct a multi-player model to describe hospitals’ joint decisions assuming they are forward looking. In this multi-player setting hospitals impose externalities on each other through the calculation of the CMS-expected readmission rate.

The main result of this section is that, while strategic interactions between hospitals can increase
the number of PE hospitals, it can only increase the number of non-incentivized hospitals: hospitals that prefer paying penalties to reducing readmissions in the single-hospital setting do not reduce readmissions also in the multi-player game-theoretic setting. We start with a single-year game that is expanded to a multi-period game in Section 4.3.

4.1. Single-year game

There is a set of $H$ hospitals. Each hospital maximizes its operating margin by determining, at the beginning of the game, its reduction (possibly 0) from the current readmission rate. Looking one year into the future, a hospital takes into account how its decision (and those of its peers) affect its CMS-expected target $r^e$ for the next year.

We assume for now that $\lambda_{ij} \equiv 0$. That is, that the divergence throughput is 0. This is a rather strong assumption, but our main results, we argue, are not sensitive to this assumption; see Section 4.2.

Hospital $h$ with initial rate $r_{h0}$ makes a reduction decision at time 0. The penalty is paid at the end of the year against the expected readmission rate, $r^e_{h1}$, that CMS computes then based on the actions of all hospitals. In other words, hospital $h$ chooses a target readmission rate $r_{h1}$ so as to maximize $R(r_{h0}, r_{h1}, r^e_{h1})$ (see Equation 8).

The dynamics of the game are as follows:
(0) Period 0:

a. Let $r_{h0}$ denote the current readmission rate at hospital $h$ and $r_{h0}^e$ denote the current year’s CMS-expected readmission rate. These are given.

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

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

Each hospital knows the other hospitals’ current readmission rate and the CMS-expected readmission rate $r_{h0}$ and $r_{h0}^e$. This information is publicly available from CMS (http://www.Medicare.gov/hospitalcompare/Data/30-day-measures.html).

One challenge in the above is that hospitals cannot precisely predict $r_{h1}^e$ at the beginning of the year. For this, a hospital would need to acquire the patient-level discharge data of 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 the individual hospital as the hospital does not have the patient-level discharge data for all other hospitals and getting access to such data is costly. Moreover, CMS may change frequently its CMS-expected-readmissions calculations making it difficult for the hospital to predict in advance what formulas will be used.  

Instead of precisely predicting it, we assume that hospital $h$ estimates its future expected readmission rate from existing data and other hospitals’ actions according to an updating function, $g(\cdot, \cdot)$, where

$$\bar{r}_{h1}^e = g_h(\vec{r}_1, \vec{r}_0)$$

is a proxy for its true CMS-expected readmission rate. The hospital chooses $r_{h1}$ to maximize $R(r_{h0}, r_{h1}, \bar{r}_{h1}^e)$. The only property of $g_h$ that we use in our analysis is it is weakly increasing in $r_{h1}$ for any hospital $h$.

We are now ready for the formal definition of the static game with $H$ hospitals:

**Definition 1.** Let $\vec{r}_0 = \{r_{10}, r_{20}, \ldots, r_{H0}\}$ be the initial readmission rates and $\vec{r}_0^e = \{r_{10}^e, r_{20}^e, \ldots, r_{H0}^e\}$ 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 $R(r_{h0}, r_{h1}, g_h(\vec{r}_1, \vec{r}_0^e))$ defined in Equation 8.

A Nash Equilibrium in pure strategies is a readmission vector $\vec{r}_1^*$ such that $r_{h1}^* \in \arg \max_{r_{h1} \in [0, r_{h0}]} R(r_{h0}, r_{h1}, g_h(\vec{r}_1^*, \vec{r}_0))$ for every hospital $h$. A mixed-strategies Nash Equilibrium is $\pi = \{\pi_1, \pi_2, \ldots, \pi_H\}$ where $\pi_h$ is a probability distribution with support $[0, r_{h0}]$.

---

4 For example, from 2011 to 2012, CMS added the readmission cases from VA hospitals into the estimation model.

5 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.
Let $\mathbf{r} - h$ denote the decisions of hospital $h$'s peers. Hospital $h$'s best response is given by:

$$BR_h(r_h^c, r_{h0}, r_{-h1}) = \begin{cases} g_h(r_1, r_{h0}^c) & \text{if } r_{h0} \in [g_h(r_1, r_{h0}^c), f_h(r_{h0}, g_h(r_1, r_{h0}^c))], \\ r_{h0} & \text{otherwise,} \end{cases} \quad (13)$$

where $f$ is as in Equation 11 and characterizes the PE region of a hospital. The no action strategy for hospital $h$ is $BR_h = r_{h0}$. No-action is, in particular, the optimal strategy for any hospital $h$ that has an empty PE region ($f(r_{h0}, r_h^c) = r_h^c$). For such an hospital it is not financially beneficial for the hospital to reduce readmissions regardless of its current readmission rates and the maximum penalty.

It is apriori unclear how a individual hospital’s decision affects the decisions of other hospitals. If one hospital decides to reduce its readmission rates, it effectively lowers the expected readmission rate for other hospitals, and decreases other hospitals’ payoffs monotonically. This action of the hospital thus exerts negative externality on other hospitals’ contribution margins to which they may respond by reducing readmission in order to save on penalties. An opposite effect is, however, also possible. A reduction decision by hospital $h$ lowers the expected readmission rates for other hospitals. If the expected readmission rate is lowered substantially, some hospitals may find themselves in Region (3) of Figure 2 and prefer incurring penalties over reducing readmissions.

Since payoff function satisfies semi-continuity and the strategy set is compact, the existence of a Nash Equilibrium in mixed strategies is guaranteed (Dasgupta and Maskin 1986). Existence of pure-strategy Nash Equilibria (let alone uniqueness) is not, however, guaranteed.

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

In the absence of a uniqueness result, we turn to bounds. We first establish a lower bound on the number of hospitals that, in any equilibrium, prefer incurring penalties to reducing readmissions. These are hospitals on which the policy is not effective. Let $(\mathbf{r}_0^c, \mathbf{r}_0)$ denote the initial expected and the current readmission rates of hospitals in the game. We say that hospital $h$ is strongly non-program-effective (SNPE) hospital if it satisfies the condition:

$$r_{h0} > f_h(r_{h0}, g_h(\mathbf{r}_0, r_{h0})) \quad (14)$$

where $f_h$ is defined in Equation 11. The term strongly non-policy-effective is motivated by the following proposition showing that, for an SNPE hospital, reducing readmissions is a dominated strategy.
Proposition 2. For any equilibrium \( \pi \), and any SNPE hospital \( h \), \( \pi_h(r_{ho}) = 1 \). In particular, the number of SNPE hospitals provides a lower bound on the number of NPE hospitals – those that incur penalties but assign probability 1 to the no-action strategy in any equilibrium.

By definition, a hospital is SNPE if, considering its current CMS-expected readmission rates, and its current readmission, reducing readmissions is a sub-optimal decision. In effect, SNPE hospitals ignore the actions of their peers and make decisions following the single hospital model. Thus, even though HRRP may increase the overall number of hospitals that reduce readmissions, the competition it introduces can only increase the number of NPE hospitals: since readmission reduction by other hospitals can only further decrease the CMS-expected readmission rate, a hospital that is already NPE under \( r^e_{0} \), will find reducing readmissions even less appealing with the new (lower) targets. In other words, the “worst offenders” are indifferent to the benchmarking. Put differently, HRRP is ineffective in incentivizing worst offenders through benchmarking.

We turn to Program Effective (PE) hospitals, those that do respond to HRRP by reducing readmissions in equilibrium. The following is an algorithm whose output is an upper bound on the number of such hospitals. 0. Start with \( H \) hospitals with initial readmission rates \( \vec{r}_0 = \{r_{1,0}, r_{2,0}, ..., r_{H,0}\} \) and expected readmission rate \( \vec{r}^e_0 = \{r^e_{1,0}, r^e_{2,0}, ..., r^e_{H,0}\} \).

1. Identify all SNPE hospitals. Set \( n = 0 \).

2. Update the readmission rate vector as follows:

\[
    r_{h,n+1} = \begin{cases} 
        g_h(\vec{r}_n, \vec{r}^e_n) & \text{if } r_{h,n} > g_h(\vec{r}_n, \vec{r}^e_n), \ h \notin \text{SNPE}, \\
        r_{h,n} & \text{otherwise.} 
    \end{cases} 
\]

In words, any hospital \( h \) with initial readmission rate \( r_{h,n} \) at step \( n \) greater than its expected readmission rate \( \vec{r}^e_{h,n} \), reduces to its CMS-expected readmission rate. Set \( n \leftarrow n + 1 \).

3. If there are no hospitals that reduced readmission in step 2, terminate the algorithm and set \( N = n \). Otherwise, go back to step 2.

Let the terminal readmission vector of the algorithm be \( \vec{r}_N \). We say that \( h \) is an SPE hospital if:

\[
    r_{h,N} < r_{ho} \tag{16}
\]

We further say that \( h \) is strongly PE (SPE) if it reduces readmissions in some stage of the algorithm. In any equilibrium \( \pi \) (mixed or pure), the number of SPE hospitals is an upper bound on the number of PE hospitals.

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

\[
    \forall \pi, \sum_{h=1}^{H} 1_{\{E[r_{h,\pi}] < r_{ho}\}} \leq \sum_{h=1}^{N} 1_{\{h \in \text{SPE}\}} \tag{17}
\]

where \( E[r_{h,\pi}] \) is the expected readmission under the (possibly mixed) equilibrium \( \pi \). Moreover, the set of SPE hospitals is mutually exclusive from the set of SNPE hospitals.
Together, the SPE upper bound in Proposition 3 and the SNPE lower bound in Proposition 2, provide a measure of HRRP’s effectiveness. The remaining hospitals (those that are neither SPE or SNPE) are those that, under any equilibrium $\pi$, have a readmission rate that is (with probability 1 in a mixed equilibrium) lower than its CMS-expected readmission rate. We refer to these as Strongly Program Indifferent (SPI). Thus,

$$\text{SPI} = \{1, \ldots, H\} \setminus \{\text{SNPE} \cup \text{SPE}\}.$$  

4.2. Special case: a two-hospital game

The bounds we derived above provide a mechanism to derive insights into a game in which the existence or uniqueness of pure-strategy equilibria is not guaranteed. Two and three hospitals games are more tractable and could strengthen the confidence in the insights derived through the bounds.

Consider then a symmetric model with two hospitals and a single disease. The two hospitals have the same patient volume, same patient mix, with only Medicare patients and with a 100% adjustment factor ($l = 1$). The two hospitals may have, however, different initial readmission rates $r_{i0}$ for hospital $i \in \{1, 2\}$. The CMS-expected readmission rate, in this case, is simply the average $r_e^1 = r_e^2 = \frac{r_{10} + r_{20}}{2}$.

Label hospitals such that $r_{10} < r_{20}$. In this case $r_{10} < r^e_1$, hospital 1 does not incur any penalties and has no incentive to reduce readmissions. Hospital 2 decides between reducing readmission to $r_{10}$ (in which case $(r_{10}, r_{10})$ is the equilibrium) and remaining at $r_{20}$. Under the former strategy the hospital’s payoff is $\frac{1}{1-r_{10}}$ and it is $\frac{1}{1-r_{20}} \min\{\max\{r_{20}/r_{10} - 1, 0\}, P_{\text{cap}}\}$ under the latter. Thus, the hospital will reduce its readmissions if

$$\frac{1}{1-r_{10}} > \frac{1}{1-r_{20}} \left(1 - \min\left(\max\left(\frac{r_{20}}{r_{10}} - 1, 0\right), P_{\text{cap}}\right)\right).$$

The hospital will be indifferent between the options if $r_{20} = r_{10} + P_{\text{cap}}(1 - r_{10})$.

**Corollary 2.** A 2-hospital game has a unique Pareto-dominant pure-strategy equilibrium provided that $r_{20} \neq P_{\text{cap}} + r_{10}(1 - P_{\text{cap}})$.

The equilibrium underscores two drivers of HRRP effectiveness:

1. The first (and more obvious) driver is the maximum penalty cap $P_{\text{cap}}$. The larger $P_{\text{cap}}$ is, the more beneficial it is for hospital 2 to reduce readmissions since it incurs more penalties.

2. The readmission dispersion, which we define as $\mathcal{V}(\bar{r}_0) = \frac{\sum_{h \in S} r_{h0} - \bar{r}_0}{|S|}$ where $\bar{r}_0 = \frac{\sum_{h} r_{h0}}{H}$ and $S = \{i | r_{h0} > \bar{r}_0\}$. In the 2-hospital game $\mathcal{V}(r_{10}, r_{20}) = r_{20} - \frac{r_{10} + r_{20}}{2} = \frac{r_{20} - r_{10}}{2}$ (assuming $r_2 > r_1$) and it is easy to show that there is a threshold $d$ such that when $\mathcal{V}(r_{10}, r_{20}) > d$, hospital 2 reduces its readmission and it does not reduce otherwise (or is indifferent if the dispersion equals the
Figure 4: Trade-off of Maximum Penalty Cap and Readmission Distance (2-Hospital Model)

threshold. In other words, the greater this dispersion, the further away hospital 2 is from its expected readmission rate and hence, more likely to fall in region (3) of Figure 3.

This simple game also uncovers a relationship between the two drivers that is illustrated in Figure 4. In this numerical example $r_{10} = 0.2$, we vary $V(\vec{r})$ (by varying $r_{20}$ and $P_{cap}$) and compute the (unique Pareto-dominant equilibrium) for each pair of values. What we clearly see is that, the greater the readmission dispersion, the larger the cap that is required to incentivize the hospital 2 to reduce its readmissions. Since the cap is a clear policy variable, an obvious tool to incentivize hospital 2 is to increase this cap. But, in fact, one may also be able to control the dispersion. The policy could benchmark hospitals against similar peers to decrease the average dispersion. We revisit this policy recommendation in Section 6.

We conclude this section by re-visiting the issue of the divergence throughput $\lambda_{ij}^{dh}$ which, recall, we have assumed that $\lambda_{ij}^{dh} \equiv 0$ for our derivation of the SPE and SNPE bounds. The dynamics of the two-hospital game with positive divergence helps explain why setting $\lambda_{ij}^{dh} = 0$ may be in fact valid in a game with many players. To this end, assume that $d_i > 0$ so that $\lambda_{ij}^{dh} > 0$. Then (see Section 2.1),

$$\lambda_1^{dh} = \frac{r_2 d_h}{1 - r_2} \lambda_2^d = \frac{r_2 d_h}{1 - r_2} (\lambda_1^d + \lambda_2^d) \quad \text{and} \quad \lambda_2^{dh} = \frac{r_1 d_h}{1 - r_1} (\lambda_1^d + \lambda_2^d),$$

where $\lambda_i^d$, $i \in \{1, 2\}$ is the exogenous arrival rate to hospital $i$. It turns out, there is still one unique Pareto-dominate pure-strategy equilibrium.

The intuition is as follows: With the positive divergence, Hospital 1’s action can possibly change Hospital 2’s payoff (by reducing the number of diverted patients), and vice versa. However, the effect of Hospital i’s action on its own payoff remains the same after introducing the positive divergent throughput. In other words, a strategy set that does not have profitable unilateral deviation in the original game cannot have profitable unilateral deviation when there is positive divergent
throughput. This shows that any equilibrium in the original case remains an equilibrium in this new setting. Moreover, any strategy set that has a profitable unilateral deviation in the original setting also has a profitable unilateral deviation with the positive divergent throughput. Therefore, the set of equilibria do not change when we consider the positive divergent throughput. In other words, our model is robust towards the simplifying assumption that $\lambda_{d_{ij}} \equiv 0$.

In Appendix B, we also study a three-hospital game. The three-hospital game has a multiplicity of Pareto-efficient equilibria but we are, nevertheless, able to compute all pure-strategy Pareto-efficient equilibria in the three-hospital game. This serves to compare SNPE and NPE hospitals in equilibrium—a measure of the quality of the bounds—and to verify the robustness of the trade-off in Figure 4 to the multiplicity of equilibria.

4.3. Multi-year game

We next consider an n-period 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. Allowing hospitals to make readmission reduction decisions every period, and allowing CMS to change the penalty every period (as is planned for 2014 and 2015), allows us to calibrate our model with real data and make predictions about the long run outcomes of HRRP.

Let $P_{cap}^l$ denote the cap in the $l^{th}$ period. Let $P_{cap}^{max} = \max_{l=1,...,n} P_{cap}^l$ be the maximum penalty cap in the time horizon. Our concept of equilibrium is sub-game perfect Nash Equilibrium. Since there is no unique Nash Equilibrium in the single-stage game, the uniqueness of a Nash equilibrium is not guaranteed in the multi-stage game and we turn, as before, to bounds.

A hospital is SPE if, in some equilibrium and some year, it reduces its readmission in with positive probability. The hospital is SNPE if at any (possibly mixed) equilibrium the probability that it reduces readmissions during the game is 0. The following allows us to apply the bounds from the single-stage to the multi-year game.

**Proposition 4.** The set of SNPE hospitals in a one-year game with $P_{cap} = P_{cap}^{max}$ is a subset of the SNPE hospitals in the multi-year game and hence the number of the former is a lower bound on the number of the latter. Also, the number of SPE hospitals in the one-year game with $P_{cap}^{max}$ is an upper bound on the number of SPE hospitals in the multi-year game.

5. Simulation and Model Validation

We use hospital data from the state of California for fiscal year 2013 to better understand the drivers HRRP’s effectiveness and propose model-based predictions. 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 with actual hospitals’ readmission reduction efforts in Section 5.3.
5.1. Data

We combine three datasets:

First, we obtain financial and operational data for 434 hospitals in California 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.

Second, the CMS website lists the CMS-expected and predicted readmission rates for each hospital for fiscal year 2013. CMS computes these from the hospitals’ discharge data for July 2008 to June 2011. Out of 434 California hospitals in the OSHPD dataset, 312 are IPPS hospitals that can be matched with the CMS data. HRRP, recall, targets only IPPS hospitals. Financial data is missing for 9 of the IPPS hospitals. After removing these hospitals from the dataset, the number of hospitals per monitored disease are 186, 250 and 249 for AMI, HF, PN respectively. The differences in the hospital count across diseases are due to the selection rule of HRRP according to which a hospital with small number of readmissions in a monitored disease is not considered in the penalty evaluation for that disease. The data is summarized in Table 1:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMS Identifier</td>
<td>50002</td>
<td>50764</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Available Bed Occupancy Rate</td>
<td>0.597</td>
<td>0.138</td>
<td>0.075</td>
<td>0.976</td>
</tr>
<tr>
<td>Fraction of Total Revenue from Medicare</td>
<td>0.362</td>
<td>0.108</td>
<td>0.079</td>
<td>0.792</td>
</tr>
<tr>
<td>Average Expected Readmission Rate</td>
<td>21.586</td>
<td>1.475</td>
<td>17.136</td>
<td>26.545</td>
</tr>
<tr>
<td>Average Predicted Readmission Rate</td>
<td>21.393</td>
<td>2.263</td>
<td>16.193</td>
<td>30.931</td>
</tr>
<tr>
<td>Number of AMI Discharges</td>
<td>157.467</td>
<td>115.374</td>
<td>25</td>
<td>599</td>
</tr>
<tr>
<td>Predicted Readmission Rate for AMI</td>
<td>20.221</td>
<td>3.067</td>
<td>12.3</td>
<td>35.8</td>
</tr>
<tr>
<td>Expected Readmission Rate for AMI</td>
<td>20.337</td>
<td>2.16</td>
<td>15.2</td>
<td>29.4</td>
</tr>
<tr>
<td>Number of HF Discharges</td>
<td>363.554</td>
<td>209.231</td>
<td>31</td>
<td>1714</td>
</tr>
<tr>
<td>Predicted Readmission Rate for HF</td>
<td>24.461</td>
<td>2.459</td>
<td>18</td>
<td>32.2</td>
</tr>
<tr>
<td>Expected Readmission Rate for HF</td>
<td>24.681</td>
<td>1.331</td>
<td>20.7</td>
<td>28.5</td>
</tr>
<tr>
<td>Number of PN Discharges</td>
<td>305.799</td>
<td>163.521</td>
<td>43</td>
<td>1146</td>
</tr>
<tr>
<td>Predicted Readmission Rate for PN</td>
<td>18.64</td>
<td>2.226</td>
<td>14</td>
<td>27.7</td>
</tr>
<tr>
<td>Expected Readmission Rate for PN</td>
<td>18.827</td>
<td>1.453</td>
<td>14.7</td>
<td>23.5</td>
</tr>
</tbody>
</table>

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

Third, the Affordable Care Act requires CMS to report, per hospital, the hundred DRGs with the highest Medicare payments. For each hospital, we extract from the data the fraction of a hospital’s Medicare revenue generated by each HRRP monitored disease.

In our model, we assumed that payment per patient \( p_{ij} \) depends only on the disease \( i \) and insurance type \( j \). Medicare and Medicaid do have a pay-per-case payment structure. Private insurance
programs may adopt different payment structures where the payment may depend, for example, on the length of stay (pay-per-diem) and the quality of treatment (pay-per-performance). In hospitals that are not fully utilized, small changes to readmission rates should not affect the quality of care (Kc and Terwiesch 2009) or the length of stay (Freeman et al. 2014). Empirical evidence suggests that readmission-reduction programs do not increase the length of stay of patients (Hansen et al. 2013). Given that most hospitals in our dataset are not fully utilized (see Figure 5) small changes to the readmission rates should not affect the payments from pay-per-diem or pay-per-performance insurance types.

![Figure 5: Histogram of Hospital Utilization](image)

5.2. Simulation and results

We next apply the bounds developed in Section 4. To numerically compute these bounds for the two models, we must specify process-improvement costs and updating functions:

(1) **Process-improvement costs**: We assume that

\[ C_h(r, x) = C_v(r - x)^\alpha + C_s \frac{1}{r} \]

where \( \alpha \in [1, 2] \), \( C_v = \Pi_h(0)C_v \) and \( C_s = \Pi_h(0)C_s \). In the simulation we set \( C_v = C_s = 0.001 \), corresponding to a process-improvement cost of 0.1% of the hospital’s revenue from a disease in return for a 1% reduction in readmissions for that disease. We also test the sensitivity to the cost parameters by assuming that there is no cost of reducing readmissions. For a hospital with $1 billion in revenue, this means a process improvement cost of $1 million to reduce readmissions for all diseases by 1%.

(2) **Updating functions** \( g(\cdot, \cdot) \): The hospital’s prediction of the CMS-expected readmission rate \( \bar{r}_{hi} \) (see Equation (12)) is computed as the average readmission rates of \( \{r_{hi}, \forall h\} \) weighted by the number of patients in each hospital. If all \( H \) participating hospitals have the same number of patients, then \( \bar{r}_{hi} \) is simply \( \frac{1}{H} \sum_k r_{hi} \).
5.2.1. Multiple-Disease Decentralized Model: For a large teaching hospital like Northwestern Memorial Hospital, it is reasonable to assume that readmission reduction decisions are made “locally” at the disease level and that the hospital only acts by assigning a penalty cap to each disease so as to meet its overall targets. In this section, we demonstrate how to combine multiple single-disease models discussed in Section 4 to reflect the joint penalty applied across all monitored diseases in the policy.

Notice first that, per the definition of the policy, each monitored-disease’s excess readmissions induce penalties on all diseases (monitored or not). Denoting by $\Pi_{\text{all}}$ the total Medicare revenue of the hospital across all diseases and by $\Pi_i$ the Medicare revenue from disease $i$, the penalty induced by excess readmissions in disease $i$ is

$$\Pi_{\text{all}} \times \min \left( \max \left( \frac{r_{p,i}}{r_{e,i}} - 1, 0 \right), \frac{\Pi_i}{\Pi_{\text{all}}} P_{\text{cap}}^i \right)$$

where $r_{p,i}$ and $r_{e,i}$ are the predicted and expected readmission rates for disease $i$ and $P_{\text{cap}}^i$ is the cap assigned to disease $d$ by the hospital.

By setting the disease-level caps so that $\sum_i P_{\text{cap}}^i = P_{\text{cap}}$, the hospital guarantees that

$$\sum_{i \in \mathcal{M}} \Pi_{\text{all}} \times \min \left( \max \left( \frac{r_{p,i}}{r_{e,i}} - 1, 0 \right), \frac{\Pi_i}{\Pi_{\text{all}}} P_{\text{cap}}^i \right) \leq \Pi_{\text{all}} P_{\text{cap}}.$$

In other words, the total penalty across monitored diseases does not exceed the global penalty cap $\Pi_{\text{all}} P_{\text{cap}}$.

In particular, we assume that the hospital allocates the penalty proportionally to the relative Medicare of each disease:

$$P_{\text{cap}}^i = P_{\text{cap}} \frac{\Pi_i}{\sum_{j \in \mathcal{M}} \Pi_j}.$$

Then the penalty can be re-written as

$$\Pi_i \times \min \left( \max \left( \frac{r_p}{r_e} - 1, 0 \right), \frac{P_{\text{cap}}}{\sum_{i \in \mathcal{M}} \Pi_i} \right),$$

which reduces to Equation 7 in our single-disease model. Furthermore, the penalty cap can be interpreted as saying that disease $i$ has an effective maximum penalty cap,

$$\frac{P_{\text{cap}}}{\sum_{j \in \mathcal{M}} \Pi_j} \frac{\Pi_i}{\Pi_{\text{all}}}$$

If, for example, $P_{\text{cap}} = 3\%$, a hospital with total Medicare revenue of $1$ million will pay at most $30000$ in penalties. If 20% of its Medicare revenue ($200,000$) is attributed to monitored diseases, the effective penalty cap of monitored diseases is 15% since $15\% \times 200,000 = 30,000$. According to the latest version of HRRP, the penalty cap is 2% for 2014 and 3% thereafter. For most hospitals
in our data set the percentage of Medicare revenue from monitored diseases is between 5% and 40%, which results in an effective penalty cap that is between 5% and 60%.

If all diseases are monitored then the induced penalty of disease $i$ is

$$\Pi_i \times \min \left( \max \left( \frac{r_{p,i}}{r_{e,i}} - 1, 0 \right), P_{\text{cap}} \right).$$

In other words, the effective penalty cap for disease $i$ is simply the penalty cap dictated in the policy.

5.2.2. Applying model to data: With this decentralized view, the numerical study reduces to that of three single-disease models. We vary four parameters in the numerical study (1) $C_m$, the contribution margin ratio, (2) $l(1-d_h)$, the product of the readmission adjustment factor and inverse hospital divergence rate, (3) the cost function parameter $\alpha$.

Notice that our bounds (e.g., the percentage of SNPE hospitals) do not depend on the relative magnitudes of parameters between hospitals. For example, the percentage of SNPE hospitals for PN is 2% when contribution margin ratio is 40% and 5% when contribution margin ratio is 75%. This means that the percentage of SNPE hospitals for PN is between 2% and 5% for any combination of contribution margin ratios as long as each hospital’s ratio is between 40% and 75%.

Tables 2, 3 and 4 report the number of SPI, SPE, and SNPE hospitals for a broad set of parameters. The No-cost column corresponds to $C_v = C_s = 0$. Otherwise, we set $C_v = C_s = 0.001$ and use $\alpha = 1$ for linear cost and $\alpha = 2$ for convex cost.

<table>
<thead>
<tr>
<th>$l \times (1-d_h)$</th>
<th>Contribution Margin Ratio</th>
<th>No Cost</th>
<th>Linear Cost</th>
<th>Convex Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SPI</td>
<td>SPE</td>
<td>SNPE</td>
</tr>
<tr>
<td>100 %</td>
<td>40%</td>
<td>19%</td>
<td>81%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>23%</td>
<td>75%</td>
<td>2%</td>
</tr>
<tr>
<td>80 %</td>
<td>40%</td>
<td>19%</td>
<td>81%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>19%</td>
<td>79%</td>
<td>1%</td>
</tr>
<tr>
<td>60 %</td>
<td>40%</td>
<td>19%</td>
<td>81%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>19%</td>
<td>81%</td>
<td>0%</td>
</tr>
</tbody>
</table>

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 for disease AMI.

Tables 2-4 confirm the comparative statistics we found for the single-hospital model in Proposition 1. First, as the readmission adjustment factor increases, hospitals are more inclined to reduce their readmissions. Second, as the cost of reducing readmissions increases, the number of SNPE hospitals dramatically increases. Last, we observe that the higher the contribution margin ratio, $C_m$, the larger the number of SNPE hospitals.
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 for disease PN

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 for disease HF

We observe that the number of SPNE hospitals is rather small in the no-cost scenario. Even if the cost is positive and linear, we find that the number of SNPE hospitals barely changes with increases to the maximum penalty cap. In other words, HRRP seems to be rather effective and the current choice of the cap does not seem to compromise its performance.

Both the effectiveness and insensitivity to the maximum cap are driven to a large extent by the high effective penalty cap with under the current set of monitored diseases. Indeed, as the monitored diseases count for only a small fraction of total Medicare revenue in our data, the effective penalty cap $P_{\text{cap}} / (\sum_{d \in \mathcal{M}} \Pi_d / \Pi_a)$ rarely binds. This makes region (2) of Figure 3 wide so that most hospitals fall within this region and strictly gain by reducing readmissions. The fact that, with the current set of monitored diseases, the cap is not binding is also confirmed by national data from 2013: with a maximum cap of 1%, only 8% of hospitals in the U.S. paid the maximum penalty.\(^6\)

All this is likely to change as the set of monitored diseases is expanded starting from 2015. We re-examine this issue when discussing policy implications in Section 6.

\(^6\)http://www.advisory.com/daily-briefing/2014/10/07/khn-the-states-facing-the-most-readmission-penalties
5.3. Model validation

HRRP was signed into law along with the Affordable Care Act on March 20, 2010; see Figure 1. The first penalty was charged in fiscal year 2013 based on discharge data from July 2008 to July 2011. As the policy was advertised to hospitals in 2010, the first penalties affected by hospital actions in response to HRRP are those levied in 2015, computed based on discharge data from July 2010 to July 2013. The 2015 penalties were publicized on April 30, 2014. For each of the hospitals, we use the difference between its CMS-predicted readmissions rates in fiscal years 2013 and 2015 as a proxy for its readmission reduction efforts from 2010 to 2013. For example, if a hospital’s CMS-predicted readmission rate in 2015 is smaller than that of 2013 we take this as indication that the hospital did reduce its readmission rates from 2010 to 2013.

Our model and analysis are focused on the SNPE hospitals and we seek to validate our identification of these. Figure 6 displays the histogram of the change in predicted readmission rates between 2013 and 2015 (reflecting hospitals’ readmission reduction efforts in 2010-2013) for SNPE, NPE, and PI hospitals for the three monitored diseases assuming, for the application of the model with the base parameters ($C_m = 0.4$, $l = d_h = 0\%$ and $C_v = C_s = 0$).

For AMI and PN the distribution of SNPE hospitals is more skewed to the right (these have increased their excess readmissions ratio) while the distribution of the SPE 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. In the case of HF, the SNPE hospitals have hardly changed their predicted readmission rate. One of these hospitals did reduce readmissions but a mere 0.004.

Notice that a significant number of hospitals that our model categorizes as SPI (policy indifferent) did reduce their readmissions. This is not surprising, as reducing readmissions may bestow other financial benefits on the hospitals besides reducing the penalties such as reputation effects or the ability to backfill beds currently occupied by readmissions with higher-margin patients. These hospitals are, in any case, not the target hospitals for the policy. We are concerned with the effect of the policy on hospitals that do not have these “exogenous” incentives.

We next “project” these histograms to a table focusing only on SNPE hospitals. Table 5 displays the actual readmission reduction of the hospitals our model identifies as SNPE. The first row reports the number of these for each of the three monitored diseases. Note that, assuming zero readmission-reduction cost, we are forcing our model to overestimate the number of SNPE hospitals. Nevertheless, the matching for PN and AMI is perfect: hospitals that our model identifies as SNPE incur penalties but do not reduce readmissions. For HF we have a a hospital that we identify as SNPE but that does reduce readmissions in the data. This, recall from the corresponding histogram in Figure 6 is the hospital with 0.04 reduction so that this “miss” is in fact tiny.
Table 5: 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>
<thead>
<tr>
<th></th>
<th>PN</th>
<th>HF</th>
<th>AMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of SNPE hospitals (Model)</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Number of SNPE hospitals not paying penalties for 2015</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Number of SNPE hospitals reducing readmissions in 2010-2013</td>
<td>0 (0%)</td>
<td>1 (50%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

6. Policy Implications

Through our models and simulation results we have identified multiple drivers of policy effectiveness. The findings have implications to policy design that we discuss below by relating recommendations to the drivers they are intended to address.

6.1. The set of monitored diseases

In order to incentivize hospitals to reduce readmissions for diseases outside the current set of monitored conditions, CMS is expanding this set. The diseases COPD and TKA/THA will be
added to the set of monitored diseases already next year (2015). In order to assess the effect of adding diseases we consider here the extreme scenario that all diseases are monitored. In this case, the effective maximum penalty cap (see Section 5.2.1) is then equal to the maximum penalty cap $P_{\text{cap}}$. Tables 6, 7 and 8 report the results in this case.

<table>
<thead>
<tr>
<th>$l \times (1 - d_h)$</th>
<th>Contribution Margin Ratio</th>
<th>No Cost</th>
<th>Linear Cost</th>
<th>Convex Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPI</td>
<td>SPE</td>
<td>SNPE</td>
<td>SPI</td>
</tr>
<tr>
<td>100 %</td>
<td>40%</td>
<td>44%</td>
<td>10%</td>
<td>50%</td>
</tr>
<tr>
<td>75%</td>
<td>51%</td>
<td>27%</td>
<td>22%</td>
<td>52%</td>
</tr>
<tr>
<td>80 %</td>
<td>42%</td>
<td>51%</td>
<td>7%</td>
<td>46%</td>
</tr>
<tr>
<td>75%</td>
<td>50%</td>
<td>34%</td>
<td>16%</td>
<td>50%</td>
</tr>
<tr>
<td>60 %</td>
<td>31%</td>
<td>67%</td>
<td>2%</td>
<td>41%</td>
</tr>
<tr>
<td>75%</td>
<td>46%</td>
<td>44%</td>
<td>10%</td>
<td>50%</td>
</tr>
</tbody>
</table>

Table 6: 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 for disease AMI

<table>
<thead>
<tr>
<th>$l \times (1 - d_h)$</th>
<th>Contribution Margin Ratio</th>
<th>No Cost</th>
<th>Linear Cost</th>
<th>Convex Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPI</td>
<td>SPE</td>
<td>SNPE</td>
<td>SPI</td>
</tr>
<tr>
<td>100 %</td>
<td>48%</td>
<td>42%</td>
<td>10%</td>
<td>49%</td>
</tr>
<tr>
<td>75%</td>
<td>56%</td>
<td>26%</td>
<td>18%</td>
<td>56%</td>
</tr>
<tr>
<td>80 %</td>
<td>39%</td>
<td>56%</td>
<td>5%</td>
<td>46%</td>
</tr>
<tr>
<td>75%</td>
<td>50%</td>
<td>37%</td>
<td>13%</td>
<td>55%</td>
</tr>
<tr>
<td>60 %</td>
<td>32%</td>
<td>66%</td>
<td>2%</td>
<td>39%</td>
</tr>
<tr>
<td>75%</td>
<td>48%</td>
<td>42%</td>
<td>10%</td>
<td>49%</td>
</tr>
</tbody>
</table>

Table 7: 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 for disease PN

As the set of monitored diseases is expanded, the effective penalty cap per disease decreases, weakening the incentive to reduce readmissions as more hospitals now fall in region (3) of Figure 3 and are SNPE. Moreover, since the cap will be more frequently binding with the expanded set of monitored diseases, the effects of other drivers, e.g., dispersion in readmissions are magnified. This means that as more diseases are added, more fine-tuning of the policy (and hospital) parameters becomes imperative. Since CMS is constantly expanding its set of monitored diseases and the effect of other drives becomes more pronounced when the set of monitored diseases expands, we take the view that all diseases are monitored for the rest discussion in this section.

**Implication:** CMS plans to continue expanding the set of monitored diseases (MedPAC 2013). We show that increasing the number of monitored diseases may have the unintended consequence
Table 8: 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 for disease HF

of making the effective penalty cap smaller for monitored diseases. Hence, CMS should carefully increase the maximum penalty cap when it expands the set of monitored diseases.

6.2. Readmission dispersion

Recall (see Section 4.2) that to achieve a certain percentage of SNPE hospitals, the higher the readmission dispersion, the higher the penalty cap must be. This suggests that the policy may be more effective if hospitals are benchmarked against similar peers. To the extent that geographic proximity implies similar readmission rates, local benchmarking may provide a mechanism to increase the effectiveness of the policy.

To examine this, we study the dispersion at the Hospital Referral Region (HRR) level vs. the nationwide. Past study (Zhang et al. 2010) has shown hat hospitals within the same HRR has similar performance in various measures of quality of cares. Therefore, we hypothesize that benchmarking hospitals in the HRR level results in smaller readmission dispersion.

Clearly, we cannot re-simulate our model since benchmarking hospitals locally requires discharge-level data to re-run the logistical regression and compute risk-adjusted expected readmission rates. Instead, we compute the dispersion in risk-adjusted predicted readmission rates once for nationwide and once for each of HRR for all hospitals in 2013. Notice that, since we no longer need the financial data (percentage of revenue from Medicare patients in this simulation), we use all IPPS hospitals that are eligible for HRRP nationwide for this investigation.

Figure 7 displays the results. The horizontal axis of the histogram is the HRR-wise dispersion of hospital’s predicted readmission rates in 2013, while the vertical axis represents the number of hospitals. The red vertical line is the national dispersion. Evidently, for all diseases, most hospitals have lower HRR-wise variance. This suggests that benchmarking hospitals 'locally' (i.e. HRR-wise) may be more effective than the current national benchmarking. Local benchmarking also has the added benefit appeal of simplicity. Moreover, it may increase the fairness of the policy given that—as higher economic status reduces the likelihood of readmission—it currently penalizes unfairly hospitals in localities with low socioeconomic status.
Recall that benchmarking has a positive effect: it increases the number of PE hospitals by incentivizing PI hospitals in the single-hospital model to reduce their readmissions. It has, however, also a negative consequence: it increases the number of the NPE hospitals. The alternative benchmarking mechanism offered above retains the positive effect while minimizing the negative effect by reducing the number of worst-offenders. Moreover, in using local instead of national benchmarking, this alternative mechanism alleviates concerns with regards to the unfairness that may be introduced by HRRP.\textsuperscript{7,8}

**Implication:** CMS is considering to benchmark hospitals against similar peers to reduce the unfairness created by not adjusting for socioeconomic status (MedPAC 2013). We show that local benchmarking has other important benefits: it may decrease the number of NPE hospitals, and in turn increase the effectiveness of the policy.

\textsuperscript{7} http://www.manchin.senate.gov/public/index.cfm/press-releases?ID=e43f6f51-bec5-4be0-9c77-9d1096a121ff

\textsuperscript{8} http://www.charlestondaily-mail.com/article/20140620/DM0104/140629951

---

Figure 7: HRR-wise variance vs Nation-wise variance for all diseases in 2013
6.3. Penalty cap

HRRP is less effective on SNPE hospitals—those that have initial readmission rates that are significantly greater than their CMS-expected readmission rate. The greater the distance, the lower the financial incentive for a hospital to reduce readmissions. The penalty cap protects these hospitals from paying excessive penalties. A possible remedy is to increase the penalty cap. To assess the effectiveness of such action, we simulate our model with base parameters \( C_m = 0.4, l = d_h = 0\% \) and \( C_v = C_s = 0 \) and varying the maximum penalty cap between 3% and 100%.

![Figure 8: Equilibrium Behavior of Hospitals under Different Maximum Penalty Caps (\( \alpha = 1, C_v = 0.01, l = 0.8, d_h = 0.15, C_m = 40\% \))](image)

Figure 8 displays the percentage change in the number of SNPE hospitals as we increase the penalty cap for each of the three monitored diseases. For the California hospitals, most of the reduction in SNPE hospitals is achieved by increasing the cap to 10% for all three diseases. The effect diminishes as the maximum penalty cap increases. Once the penalty cap exceeds 20%, the SNPE hospitals are those with relatively small Medicare revenue in the monitored diseases. To incentivize these hospitals, increasing the penalty cap is not sufficient. CMS could instead consider expanding the set of monitored diseases, or extending the penalties to non-Medicare patients.

Moreover, we observe that cap-increase has a differential effect on diseases. Increasing the maximum penalty cap from 3% to 10%, reduces by 94% the number of SNPE hospitals for AMI, but only by 81% for PN. The main driver here is the percentage of Medicare patients for the existing SNPE hospitals: 31%, 26% and 28% respectively for AMI, PN and HF. Consequently, the disease with the highest percentage of revenue from Medicare patients among SNPE hospitals benefits the most from increases to the penalty cap.
Implication: We show that increasing the penalty may be more helpful to reduce readmissions for diseases, for which the NPE hospitals have higher percentage of Medicare revenue.

6.4. Process-improvement costs

As CMS notes, the major goal of designing the policy is that the penalty for not meeting reduction target is greater the incremental cost of reducing readmissions and the lost marginal profit from those readmissions (MedPAC 2013). Therefore, the costlier the process-improvement required to reduce readmissions, the less incentivized are hospitals to reduce these. To obtain more refined insights, we re-visit the linear cost case but vary the cost coefficient $C_v$ (it was set to 0.001 in Section 5). Figure 9 demonstrates the percentage of SNPE hospitals for each disease as a function of changes to the cost parameter relative to the base cost of $C_v = 0.001$.

Making the readmissions-reduction costless reduces the number of SNPE hospitals by 9%, 19%, and 29% for PN, HF, AMI respectively. The effect of cost changes differs by disease: the biggest reduction in the number of SNPE is for AMI and the smaller for PN. As in the cap this heterogeneity is driven by the differences, across diseases, in the average percentage of revenue from Medicare patients among SNPE hospitals.

![Equilibrium Behavior of Hospitals under Different Variable Costs (α = 1, $P_{cap} = 0.03$, $l = 0.8$, $d_h = 0.15$, $C_m = 40\%$)]

Facing a decision between investments in readmission reduction toolboxes (such as Boost Hansen et al. (2013)) the government may want to target diseases where the average percentage of revenue from Medicare patients is highest.
Implication: We show that diseases, for which the NPE hospitals have higher percentage of Medicare revenue, benefit more from lowering the readmission reduction cost. Therefore, the government could prioritize research on diseases that have a higher percentage of Medicare patients within the NPE hospitals.

6.5. Hospital characteristics and incentives
First, HRRP has relatively greater influence on hospitals that have higher divergence probability. Such hospitals are typically located at more developed and dense urban areas. Residents of these areas already have access to more hospitals and better healthcare relative to patients in the rural areas. HRRP may then magnify the health-care-access gap as hospitals in rural areas will be less incentivized to reduce readmissions.

Second, the policy is less effective for hospitals with a low fraction of Medicare revenue. This limitation is inherent to the government payment system, and may be difficult to change. One could, however, utilize the penalties collected to reward hospitals with good performance. Absolute rewards (rather than those proportional to the revenue of a hospital from Medicare patients) may increase the effect that HRRP has on these hospitals.

Third, hospitals with higher contribution margin ratios are less likely to reduce readmissions in response to HRRP. Therefore, payment programs that lower the contribution margin ratio of “worst offenders”, such as pay-for-performance program, can improve the effectiveness of the policy.

Implication: We show that hospitals with lower divergence probability, low fraction of Medicare revenue, or higher contribution margin ratio are less likely to be incentivized by HRRP.

7. Conclusions
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 incur penalties of up to 1% of their reimbursement for Medicare patients, adding up to $280 million, with an average $125,000 penalty per hospital in 2013.

The success of HRRP may be affected by various issues. In this paper we take the view that hospitals are operating-margin maximizers. Readmission-reduction decisions may be affected by other factors, such as the medical ethics and peer pressure that is enhanced by the increased visibility of hospital performance metrics that accompany the implementation of HRRP. Such considerations may increase the policy’s effectiveness and mitigate some of the challenges we identify in this paper.

Predicting the effectiveness of policy regulations on individual decision makers is challenging. With time, however, data will become available documenting actual hospital actions in response to HRRP. Once such data is available it is our hope that our model 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.

Appendix

A. CMS’s Estimation of readmission rates

CMS computes \( r_h \) and \( r_{eh} \) for every hospital using patient level discharge and readmission data as follows:

Let \( Y_{ilk} \) be a binary variable indicating whether discharge \( l \) of disease \( i \) in hospital \( k \) is associated with a readmission (either to the same hospital or to another hospital). For each discharge CMS collects the corresponding patient case covariates, denoted by \( Z_{ilk} \) for discharge \( l \) in disease \( i \) and hospital \( k \). The logistic hierarchical generalized linear model is used to estimate the average and individual-hospital intercepts to predict the readmission probability for each discharge:

\[
\log(P(Y_{ilk} = 1)) = \alpha_{ik} + \beta_i'Z_{ijk}
\]

where, for each disease \( i \), \( \alpha_{ik} \) is the hospital-level intercept for hospital \( k \), \( \mu_i \) is the average intercept, and \( \beta_i \) 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:

\[
l_{ik}^e = \frac{1}{N_{ki}} \sum_{j=1}^{N_{ki}} \frac{1}{1 + e^{-\mu_i - \beta_i Z_{ilk}}}
\]

\[
l_{ik}^p = \frac{1}{N_{ki}} \sum_{j=1}^{N_{ki}} \frac{1}{1 + e^{-\alpha_{ik} - \beta_i Z_{ilk}}}
\]

where \( N_{ki} \) is the number of Medicare discharge cases with disease \( i \) in hospital \( k \).

B. Three-hospital Game

We consider a symmetric three-hospital model. In contrast to the two-hospital game, there may be here multiple pure strategy equilibria. Suppose, for example, that \( \vec{r} = (0.2, 0.24, 0.24) \). Then, we have the two pure-strategy equilibria. The first \((0.2, 0.22, 0.24)\), i.e., hospital 2 reduces to \( r_e = 0.22 \), while hospital 3 does not reduce. The second is \((0.2, 0.24, 0.22)\), i.e, hospital 2 does not reduce, but hospital 3 reduces to \( r_e = 0.22 \).

To compute all pure-strategy equilibria in the game, we design a tatonnement algorithm (Cheng and Wellman 1998). In principle, we search through all sequences of best-response plays. The number of possible sequences of best responses exponentially increases since each period there are two possible outcomes (i.e., either Hospital 2 plays first or Hospital 3 plays first). However, for 3 hospitals, we can use branch-and-cut techniques (Padberg and Rinaldi 1991) to avoid searching through certain sequences and reduce the execution time.

Figure 10 displays the results of a numerical study based on this algorithm. We vary the readmission dispersion \( a \). For each value of \( a \), we draw 100 random samples \(((x_1(n), x_2(n), x_3(n)); n = 1, \ldots, 100)\) from
three independent uniform \([0,1]\) random variables. The initial readmission vector in the \(n^{th}\) simulation is set to \([0.2 + a \cdot x_1(n), 0.2 + a \cdot x_2(n), 0.2 + a \cdot x_3(n)]\). For each realization we compute the average (across pure-strategy equilibria) number of NPE hospitals, and the average number of SNPE hospitals. We then average these across the 100 realizations.

Figure 10: Number of SNPE Hospitals v.s. Number of NPE Hospitals (3-Hospital Model)

Since the number of SNPE hospitals is a lower bound on the number of NPE hospitals the blue curve should indeed be above the green curve. The average number of SNPE hospitals is relatively close to the average expected number of NPE hospitals in all cases supporting the use of SNPE hospitals to evaluate the policy effectiveness.

Next we vary both \(P_{\text{cap}}\) and the readmission dispersion. Given a readmission dispersion \(x\) and \(P_{\text{cap}}\), we generate the 100 readmission vectors \(\vec{r}_0 = [0.2 - 2x, 0.2 + u \cdot x, 0.2 + (1 - u) \cdot x]\) where \(u\) is drawn from a uniform distribution on \([0,1]\). For each realization we compute the average number (across equilibria) of SNPE hospitals and the average number of NPE hospitals. We then average across realizations to obtain a point for each pair \((x, P_{\text{cap}})\).

Figure 11 displays the results for three value different maximum penalty caps. We see, again, that to achieve to target a given average number of NPE hospitals, the penalty cap has to be increased as the dispersion increases.

C. Disease-level divergence

For the clarity and simplicity of our model, we assume that the readmission divergence probability between different diseases is 0. In practice, disease-level divergence is quite common. As Jencks et al. (2009) suggests, more than half of the patients are readmitted with different diseases for the three monitored diseases. In this section, we propose a two-disease model and investigate, under what conditions, our bound on the number of NPE hospitals remains valid.
There is a single hospital $h$ with two diseases: disease 1 and disease 2. Disease 1 is the only monitored disease under HRRP with exogenous arrival rate $\lambda_1$. The original readmission rate of disease 1 is $r_0$, and the expected readmission rate for disease 1 is $r^e$. The average payments of one admission of disease 1 and 2 are $p_1$ and $p_2$ respectively. The disease-level divergence rate from $d_1$ to $d_2$ is denoted as $d_d$. Without loss of generosity, suppose that the contribution margin ratio is 100%, and the percentage of revenue from Medicare patients is 100%.

If $d_d = 0$, the operating margin of the hospital for readmission $r_1$ of disease 1 is:

$$R_0(r_0, r_1, r^e) = p_1 \frac{\lambda_1}{1 - r_1} \left(1 - \min \left(\max \left(\frac{r_1}{r^e} - 1, 0\right), P_{cap}\right)\right) - C(r_0, r_1)$$

If $d_d > 0$, the operating margin of the hospital for readmission $r_1$ of disease 1 is:

$$R_1(r_0, r_1, r^e) = p_1 \frac{\lambda_1}{1 - (1 - d_d)r_1} \left(1 - \min \left(\max \left(\frac{r_1}{r^e} - 1, 0\right), P_{cap}\right)\right) - C(r_0, r_1) + p_2 \frac{\lambda_1 d_d}{1 - (1 - d_d)r_1}.$$

Obviously, if $p_1 \approx p_2$ and $d_d > 0$, the contribution from each readmission remains the same, while the penalty becomes less. Suppose that $p_1 \approx p_2$, a hospital which does not reduce readmissions when $d_d = 0$, will not reduce readmissions if $d_d > 0$.

Jencks et al. (2009) carefully examines the readmissions of more than 2.9 million patients and conclude that the average payment index, for the three monitored diseases, is 1.41 while it is 1.35 for the 30-day readmissions of the monitored diseases. Their analysis also suggests that the average length of stay for 30-day readmission (of the monitored disease) is 0.6 (13.2%) days longer. Combining these two observations, it is evident that the average payments of each initial admission and its 30-day readmission are comparable. In other words, $p_1 \approx p_2$. Therefore, our characterization of SNPE hospitals remain valid when we incorporate the disease-level divergence. Hence, our implications based on SNPE hospitals are robust towards the assumption that the disease-level divergence is 0.
D. Proof of Proposition 1

Recall the maximization problem in Equation 8:

$$x^* = \arg \max_{x \leq r} R(x, r, r_e) = \arg \max_{x \leq r} \Pi_h(r) (1 - \mathbb{P}_h(x, r_e)) - C(r, x)$$

where $\Pi_h(x)$, $\mathbb{P}_h(x, r_e)$, and $C(r, x)$ are defined as:

$$\Pi_h(x) = \Pi_h(0) \frac{1}{1-x},$$

$$\mathbb{P}_h(x, r_e) = \phi_{med} \min(\max(\frac{x}{r_e} - 1, 0), P_{cap}),$$

$$C(r, x) = C_v(r^\alpha - x^\alpha).$$

Notice that $\Pi'_h(x) = -\frac{1}{(1-x)^2} > 0$ and that, by assumption, $C'(r, x) < 0$, and $\mathbb{P}_h(x, r_e) = 0$. This means that a hospital’s revenue is an increasing function of its readmission rate for values less than $r_e$. Therefore, the hospital’s optimal solution in this region is $r_e$.

Let

$$x_m = \inf \{x : \mathbb{P}_h(x, r_e) = P_{cap}\}.$$ 

If $r > x_m$, for $x \in [x_m, r]$, $R(r, x, r_e)$ is strictly increasing in $x$, and $C(r, x)$ is decreasing in $x$. Therefore, the optimal readmission rate in the region $[x_m, r]$ is $r$. Finally, for $x \in [r_e, x_m]$:

$$R'(r, x, r_e) = \frac{C_m}{(1-x)^2} \Pi_h(0) \left[ \frac{\phi_{med}}{C_m} \left( 1 - r_e \right) - \frac{r_e}{r_e} \right] - \frac{dC(r, x)}{dx}. \tag{22}$$

Since, by assumption, $\frac{dC(x, r)}{dx} < 0$ then $\frac{dR(x, r, r_e)}{dx} > 0 \forall x \in [r_e, x_m]$, if $\frac{\phi_{med}}{C_m} > \frac{r_e}{1-r_e}$, so that the optimal choice is $x_m$. If, on the other hand, $\frac{\phi_{med}}{C_m} < \frac{r_e}{1-r_e}$, then since $\left| \frac{dC(x, r)}{dx} \right| \leq \frac{1}{(1-x)^2} \forall x \in (0, 1)$, it must be the case that $\frac{dR(x, r, r_e)}{dx}$ has the same sign $\forall x \in (r_e, x_m)$. Therefore, the optimal choice must be between $r_e$ and $x_m$.

We have proved, then, that for all values $x$ of initial readmissions, a hospital’s optimal decision is either to reduce to the expected readmission rate $r_e$ or to remain at the current readmission rate $x$.

E. Proof of Corollary 1

By Equation (11), $f(r_{h0}, r^*_h)$ is defined as:

$$R(f(r_{h0}, r^*_h), r^*_h, r^*_h) = R(f(r_{h0}, r^*_h), f(r_{h0}, r^*_h), r^*_h), \tag{23}$$

or equivalently,

$$\Pi'_h(r^*) - \Pi'_h(f(r_{h0}, r^*_h)) + \mathbb{P}_h(f(r_{h0}, r^*_h), r^*) - C(f(r_{h0}, r^*_h), r^*) = 0 \tag{24}$$

Recall that $\mathbb{P}_h(x, r^*) = \frac{\phi_{med}}{C_m} \min(\max(\frac{x}{r_e} - 1, 0), P_{cap})$. Thus, as $\phi_{med}$ increases, $\mathbb{P}_h(f(r_{h0}, r^*_h), r^*)$ increases.

In turn, Equation (24) guarantees that $f(r_{h0}, r^*_h)$ must increase (for a fixed $r^*_h$). Similarly, if $d_h$ increases or $l$ decreases, $\Pi'_h(r^*) - \Pi'_h(f(r_{h0}, r^*_h)) + P_h(f(r_{h0}, r^*_h), r^*)$ increases, and $f(r_{h0}, r^*_h)$ increases for a fixed $r^*_h$.

If $C_m$ increases, $\Pi'_h(r^*) - \Pi'_h(f(r_{h0}, r^*_h))$ decreases. Therefore, $\mathbb{P}_h(f(r_{h0}, r^*_h), r^*) - C(f(r_{h0}, r^*_h), r^*)$ increases, which means—using again (24)—that $f(r_{h0}, r^*_h)$ decreases for a fixed $r^*_h$. Finally, if there are two cost functions $C(\cdot, \cdot)$ and $\tilde{C}(\cdot, \cdot)$ such that $\tilde{C}(x, r) \geq C(x, r)$ for all $x, r$ then, $\Pi'_h(r^*) - \Pi'_h(f(r_{h0}, r^*_h)) + \mathbb{P}_h(f(r_{h0}, r^*_h), r^*)$ is higher when the cost function is $\tilde{C}$ relative to $C$ which implies, in particular, that $f(r_{h0}, r^*_h)$ decreases for a fixed $r^*_h$. 


F. Proof of Lemma 1

Notice that the payoff function described in Equation 8, is continuous in the hospital’s readmission rate except at \( r_{h0} \), when there is a fixed cost to implement readmission reduction programs. Moreover, the strategy set for each hospital \([0, r_{h0}]\) is convex. According to a variant of Glicksberg’s Theorem (Dasgupta and Maskin 1986), this game has at least one Nash Equilibrium in mixed strategies.

For counter examples we use the updating function where \( \bar{r}_{h1}^e \) is the average readmission rates of \( \{r_{h1}, \forall h\} \) weighted by the number of patients in each hospital. If all \( H \) hospitals have the same number of patients, \( r_{h1}^e \) reduces to \( \frac{1}{H} \sum_k r_{h1} \). We denote this updating mechanism by \( \bar{r}_{h1}^e = g(r_{h1}^e) \), where \( r_{h1}^e = \{r_{11}, r_{21}, \ldots, r_{H1}\} \).

The first example shows that the game may not have pure-strategy Nash Equilibria:

There are 3 hospitals with initial readmission rates \( r_0^e = \{0.24, 0.244, 0.249\} \). 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 \( g_1(r_0^e) \), with all hospitals having same number of patients. In order words, \( g(r_0^e) = (r_{01} + r_{02} + r_{03})/3 \).

By Proposition 1, each hospital chooses between reducing to the average \( g(r_0^e) \) or remaining at current readmission rates \( r_{h0} \). Therefore, we only have four candidates for pure-strategy Nash Equilibria: (1) hospital 2 does not reduce, and hospital 3 reduces: \( \{0.24, 0.244, 0.242\} \), (2) both hospitals 2 and 3 reduce: \( \{0.24, 0.24, 0.24\} \), (3) neither hospital 2 nor hospital 3 reduce: \( \{0.24, 0.244, 0.249\} \). In (1), hospital 2 is better off reducing to 0.241, indicating that (1) is not an equilibrium. In (2), hospital 3 is better off staying at 0.249. In (3), hospital 3 increases its revenue by reducing to the average (expected) readmission rate 0.242.

Therefore, there is no pure-strategy Nash Equilibrium in the game described above.

The following example shows that there may be multiple pure-strategy Nash equilibria:

Consider the same game but with \( r_0^e = \{0.2, 0.24, 0.24\} \). There are four candidate pure-strategy Nash Equilibria: (1) hospital 2 does not reduce, and hospital 3 reduces: \( \{0.2, 0.22, 0.24\} \), (2) hospital 2 reduces while hospital 3 does not reduce: \( \{0.2, 0.24, 0.22\} \), (3) both hospitals 2 and 3 reduce: \( \{0.2, 0.2, 0.2\} \), (4) neither hospital 2 nor hospital 3 reduce: \( \{0.2, 0.24, 0.24\} \). Using Equation 8 it is easily verified that both (1) and (2) are pure-strategy Nash Equilibria.

G. Proof of Proposition 2

By definition, a hospital \( h \) is SNPE hospital if

\[
r_{h0} > f_h(r_{h0}, g_h(r_{h0}^e, r_{h0}^e)) \tag{25}\]

By the definition of \( f_h(r_{h0}, g_h(r_{h0}^e, r_{h0}^e)) \) (see Equation 11), \( f_h(r_{h0}, g_h(r_{h0}^e, r_{h0}^e)) > r_{h0}^e \). Thus, in particular, \( r_{h0} > f_h(r_{h0}, g_h(r_{h0}^e, r_{h0}^e)) > r_{h0}^e \) indicating that SNPE hospitals have readmissions that exceed their CMS-expected rates and hence pay penalties.

By Equation 11, \( f_h(x, y) \) is increasing in \( y \). By the monotonicity of \( g_h \), \( g_h(r_i, r_{h0}) \geq g_h(r_{i'}, r_{h0}) \) if \( r_i \geq r_{i'} \) for \( i = \{1, 2, \ldots, H\} \). Moreover, since hospital \( h \)'s strategy set is \([0, r_{h0}]\), it must be that, at any equilibrium of the game, the equilibrium readmission vector is less than or equal to the initial readmission vector, i.e., \( \bar{r}_i \leq r_{i0} \).

Therefore, given any equilibrium \( \pi \) and a readmission vector \( r_{\pi 1}^e \) that has a positive probability under \( \pi \), we have

\[
h \in SNPE \implies r_{h0} > f_h(r_{h0}, g_h(r_{h0}^e, r_{h0}^e)) \geq f_h(r_{h0}, g_h(r_{\pi 1}^e, r_{h0}^e)) \tag{26}\]
Therefore, for SNPE hospitals reducing readmissions is a strictly dominated strategy, and they do not reduce readmission at any equilibrium, i.e.,

\[ \forall \pi, \forall h \in \text{SNPE}, \pi_h(r_{h0}) = 1. \]  

(27)

\[ \blacksquare \]

H. Proof of Proposition 3

In step 2 of the algorithm, \( r_{h,n} \neq r_{h,n-1} \) if \( r_{h,n-1} > g_h(r^e_{n-1}, r^e_{h,n-1}) \) and \( h \notin \text{SNPE} \). This means that if \( h \in SNPE, r_{h,n} = r_{h,n-1} \forall n \). 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 is an upper bound on the number of hospitals that reduce readmissions in some equilibrium with positive probability, let us consider the readmission vector \( r_N^e \), that the algorithm generates. By Proposition 3, at any equilibrium \( \pi \) SNPE hospitals do not reduce readmissions. Fix one such equilibrium \( \pi \). Then any hospital \( h \) that reduces readmissions in this equilibrium must satisfy:

\[ h \notin \text{SNPE}, \quad \text{and} \quad r_{h0} > g_h(r^e_x, r^e_0). \]

By the assumed monotonicity of \( g_h \), if we could show that \( r_N^e \leq r^e_x \) for all equilibrium \( \pi \), then we would in particular have that a hospital \( h \) with \( r_{h0} > g_h(r^e_x, r^e_{h0}) \) also has \( r_{h0} > g_h(r_N^e, r^e_{h0}) \). In turn, the number of hospitals that reduce readmissions in some equilibrium is bounded by the number of SPE hospitals generated by the algorithm.

It remains then to prove that \( r_N^e \leq r^e_x \) for any equilibrium \( \pi \). Suppose, to reach a contradiction, that \( \exists \pi \ s.t. \ r_N^e > r^e_x \). Then there must exist \( h \) such that \( r_{h,N} > r_{h,x} \) and \( r_{h0} > g_h(r^e_x, r^e_{h0}) \). Indeed, we claim that if every hospital that has \( r_{h,N} > r_{h,x} \) also has \( r_{h0} \leq g_h(r_N^e, r^e_{h0}) \) then \( r^e_x \) could not be an equilibrium.

To see this let \( H \) be the set of hospitals with \( r_{h,N} > r_{h,x} \). Assume that \( \forall h \in H, r_{h0} \leq g_h(r_N^e, r^e_{h0}) \). Since \( r_{N} > r_{x}, r_{h0} \geq g_h(r^e_x, r^e_{h0}) \forall h \in H \). So it must be that \( \sum_{h \in H} r_{h0} - g_h(r^e_x, r^e_{h0}) > \sum_{h \in H} g_h(r^e_N, r^e_{h0}) - g_h(r^e_x, r^e_{h0}) \), which is a contradiction to the assumption that \( r_{h0} \leq g_h(r_N^e, r^e_{h0}) \) for all \( h \in H \).

Pick then \( h \) that has \( r_{h,N} > r_{h,x} \) and \( r_{h0} > g_h(r^e_{N}, r^e_{h0}) \). By the termination condition of the algorithm, no hospitals (in particular \( h \)) have incentive to reduce and hence

\[ r_{h0} \notin [g_h(r^e_{N}, r^e_{h0}), f(r_{h0}, g_h(r^e_x, r^e_{h0}))] \]  

(28)

with \( f(r_{h0}, g_h(r^e_N, r^e_{h0})) \) defined in Equation 11. Since \( f(r_{h0}, g_h(r^e_N, r^e_{h0})) \) is increasing in \( g_h(r^e_N, r^e_{h0}) \), we have:

\[ r_N^e \geq r^e_x \Rightarrow g_h(r^e_N, r^e_{h0}) \geq g_h(r^e_x, r^e_{h0}) \Rightarrow f(r_{h0}, g_h(r_N^e, r^e_{h0})) \geq f(r_{h0}, g_h(r^e_x, r^e_{h0})) \]  

(29)

Therefore,

\[ r_{h0} > f(r_{h0}, g_h(r^e_{N}, r^e_{h0})) \Rightarrow r_{h0} > f(r_{h0}, g_h(r^e_x, r^e_{h0})) \Rightarrow r_{h0} \notin [g_h(r^e_x, r^e_{h0}), f(r_{h0}, g_h(r^e_x, r^e_{h0}))]. \]  

(30)

Hence it is not optimal for the hospital \( h \) to reduce its readmissions. We reach a contradiction to the assumption that \( r_N^e > r^e_x \).

\[ \blacksquare \]
I. Proof of Proposition 4

We first prove that the set of SNPE hospitals in the single-year game where \( P_{\text{cap}} = P_{\text{cap}}^{\text{max}} \) is a lower bound on the number of NPE hospitals in the multi-year game. We rewrite equation 8 as:

\[
R(P_{\text{cap}}, r_{h0}, r_h, r^*_h) = \Pi^r_h(r_h) - P_h(r_h, r^*_h, P_{\text{cap}}) - C(r_{h0}, r_h),
\]

where \( \Pi^r_h(r_h) \) is the penalty under \( P_{\text{cap}} \). By construction, the penalty function \( \Pi^r_h(r_h, P_{\text{cap}}) \) is increasing in \( P_{\text{cap}} \). Therefore, if \( r_1 > r_2 \geq r^*_h \) and \( P_{\text{cap}}^1 > P_{\text{cap}}^2 \):

\[
R(P_{\text{cap}}^1, r_{h0}, r_1, r^*_h) - R(P_{\text{cap}}^1, r_{h0}, r_2, r^*_h) > 0 \Rightarrow R(P_{\text{cap}}^2, r_{h0}, r_1, r^*_h) - R(P_{\text{cap}}^2, r_{h0}, r_2, r^*_h) > 0
\]

In other words, \( R(P_{\text{cap}}, r_{h0}, r, r^*_h) \) is super-modular in \( (P_{\text{cap}}, r) \).

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

\[
R(P_{\text{cap}}^{\text{max}}, r_{h0}, r_{h0}, g_h(r^T_{h,\pi}, r^*_h)) > R(P_{\text{cap}}^{\text{max}}, r_{h0}, r, g_h(r^T_{h,\pi}, r^*_h)) \quad \forall r < r_{h0},
\]

which implies, by the argued supermodularity, that for all \( P_{\text{cap}} < P_{\text{cap}}^{\text{max}} \),

\[
R(P_{\text{cap}}, r_{h0}, r_{h0}, g_h(r_{h,\pi}, r^*_h)) > R(P_{\text{cap}}, r_{h0}, r, g_h(r_{h,\pi}, r^*_h)) \quad \forall r < r_{h0}.
\]

Moreover, \( R(P_{\text{cap}}^1, r_{h0}, r_1, r^*_h) - R(P_{\text{cap}}^1, r, r^*_h) > 0 \) 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, with \( r^{-1}_t \), being the readmission vector at the end of stage \( t \), the total operating margin of hospital \( h \) in this sub-game perfect Nash Equilibrium (SGPNE) is less than the total operating margin collected under the no-action strategy:

\[
\sum_{t=1}^{T} R(P_{\text{cap}}^t, r_{h,\pi}, r^*_t, g_h(r^*_t, r_{h0})) < \sum_{t=1}^{T} R(P_{\text{cap}}^t, r_{h,\pi}, r^*_t, g_h(r^*_t, r_{h0}))
\]

The first inequality above follows from the restriction to reductions in readmissions (\( r^*_t \leq r_{h0} \forall t \)). This then shows that hospital \( h \)'s no-action strategy is optimal, and in turn the SGPNE is not a valid equilibrium.

We next prove that the set of SPE hospitals under the maximum penalty cap is an upper bound on the number of hospitals that reduce readmissions relative to \( r_{h0} \) in any equilibrium of the multi-year game. Following our strategy in the proof of Proposition 4, we must prove that for any SGPNE \( \pi \), and at any period \( t \), \( r^*_t \geq r^*_N \) as generated by the algorithm with \( P_{\text{cap}} = P_{\text{cap}}^{\text{max}} \). Since \( r^*_t \) is monotone decreasing in \( t \) (because readmissions can only be decreased), it suffices to show that \( r^*_t = r^*_t \geq r^*_N \) (where \( N \) is the terminal step of the algorithm). We can restrict attention to non-SNPE hospitals since we have shown above that the SNPE hospitals do not reduce readmission in any equilibrium \( \pi \).

Suppose that there exists an equilibrium \( \pi \) such that \( r^*_t < r^*_N \), then \( \exists h \) such that \( r_{h,N} > r^*_t \) and, in particular, \( r_{h0} > g_h(r^*_N, r^*_0) \). The initial condition of stage \( T \) is \( r^*_{h,\pi} \). If we can now show that \( r^*_{h,\pi} \notin \)
[\vec{r}_N, f(r_{h_0}, g_h(r_N, r_{h_0}^e))] (which is the analogue of Equation 28) then we can follow the proof in Proposition 4 to reach a contradiction.

To prove that \( r_{h_0, \pi}^{T-1} \notin [g_h(r_N, r_{h_0}^e), f(r_{h_0}, g_h(r_N, r_{h_0}^e))] \) we use induction. For \( n = 1 \), this relationship holds trivially since (recall) \( R(P_{cap}, r_{h_0}, r, r_k^e) \) is super-modular in \( (P_{cap}, r) \). Now, let us assume this relation holds for all \( k + T - 1 \) and show it proves that \( r_{h_0, \pi}^{T-1} \notin [g_h(r_N, r_{h_0}^e), f(r_{h_0}, g_h(r_N, r_{h_0}^e))] \). If this were not the case then:

\[
r_{h_0, \pi} = g_h(r_N, r_{h_0}^e) < f(r_{h_0}, g_h(r_N, r_{h_0}^e)) \tag{36}
\]

Since \( r_{h_0, \pi}^{T-1} \notin [g_h(r_N, r_{h_0}^e), f(r_{h_0}, g_h(r_N, r_{h_0}^e))] \), there exist a set of hospitals, \( \mathcal{H} \), such that each \( h \) in this set has \( r_{h_0, \pi}^{T-1} > f(r_{h_0}, g_h(r_N, r_{h_0}^e)) \), and reduce readmissions to \( g_h(r_N, r_{h_0}^e) \) in equilibrium. By definition, both \( g_1(\vec{r}, r_{h_0}^e) \) and \( g_2(\vec{r}, r_{h_0}^e) \) are contraction mapping of \( \vec{r} \) component-wise, in other words:

\[
|g_h(\vec{r}, r_{h_0}^e)) - g_h(\vec{r}', r_{h_0}^e)| < |r - r'|
\tag{37}
\]

we know that \( \sum_{h \in \mathcal{H}} (r_{h_0}^e - r_{h_0, \pi}^{T-1}) > \sum_{h \in \mathcal{H}} [g_h(r_N, r_{h_0}^e) - g_h(r_N, r_{h_0}^e)] \). However, based on the optimality condition of the game, \( \sum_{h \in \mathcal{H}} r_{h_0, \pi} = \sum_{h \in \mathcal{H}} g_h(r_N, r_{h_0}^e) \). This is a contraction, and therefore \( r_{h_0, \pi} \notin [g_h(r_N, r_{h_0}^e), f(r_{h_0}, g_h(r_N, r_{h_0}^e))] \). This concludes the proof.

\section{Proof of Corollary 2}

If Hospital 2 never reduces its readmissions beyond Hospital 1’s current readmission rate \( r_{10} \), then Hospital 1’s dominant strategy is simply staying at its current readmission rate \( r_{10} \). Therefore, the equilibrium becomes a static analysis of Hospital 2’s optimal decision when the expected readmission rate is \( \frac{r_{20} + r_{22}}{2} \) where \( r_{21} \) is the readmission decision of Hospital 2.

The objective function of Hospital 2 in this case is:

\[
\frac{1}{1 - r_{21}} \left( 1 - \min \left( \max \left( \frac{2r_{21}}{r_{21} + r_{10}}, 0 \right), 1 - r_{21} \right) \right) P_{cap}
\]

It can be easily seen that Hospital 2’s optimal decision is simply to reduce readmissions to \( r_{10} \) if \( r_{20} < r_{10} + P_{cap}(1 - r_{10}) \), and not reduce if \( r_{20} > r_{10} + P_{cap}(1 - r_{10}) \). Therefore, if Hospital 2 never reduces its readmissions beyond \( r_{10} \), there is a unique pure-strategy equilibrium \( (r_1 = (r_{10}, r_{10}) \text{ or } r_1 = (r_{10}, r_{20})) \).

If \( r_{20} > r_{10} + P_{cap}(1 - r_{10}) \), reducing beyond \( r_{10} \) is a strictly dominated strategy for Hospital 2. If \( r_{20} < r_{10} + P_{cap}(1 - r_{10}) \), Hospital 2 could potentially reduce to \( r' \) such that \( r' < r_{10} \). This is only an equilibrium if and only if the best response of Hospital 1 is also to reduce to \( r' \). In this case, the equilibrium is \( (r', r') \), which is strictly Pareto dominated by the equilibrium \( (r_{10}, r_{10}) \).

Hence, there is a unique Pareto-dominant pure-strategy equilibrium.

\section{References}


