Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0000-0000 | EISSN 0000-0000 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000


# Hospital Readmissions Reduction Program: An Economic and Operational Analysis

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

Eric Park
Sauder School of Business, University of British Columbia, Vancouver, Canada

Robert S. Young, Mark V. Williams
College of Medicine, University of Kentucky, Lexington, KY 40508,

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 (CMS) 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 develop a game-theoretic model that captures the competition among hospitals inherent in HRRP's benchmarking mechanism. We show that this competition can be counter-productive: it increases the number of non-incentivized hospitals, which prefer paying penalties over reducing readmissions in *any* equilibrium. We calibrate our model with a dataset of more than 3,000 hospitals in the United States and show that under the current policy, and for a large set of parameters, 4% to 13% of the hospitals remain non-incentivized to reduce readmissions. We also validate our model against the actual performance of hospitals in the three years since the introduction of the policy. We draw several policy recommendations to improve this policy's outcome. For example, localizing the benchmarking process – comparing hospitals against similar peers – improves the performance of the policy.

*Key words*: Healthcare Operations, Public Policy.

*History*: This paper was first submitted on May 31, 2014, revised on Nov 23, 2014 and Apr 4, 2015, and accepted on May 28, 2015

## 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

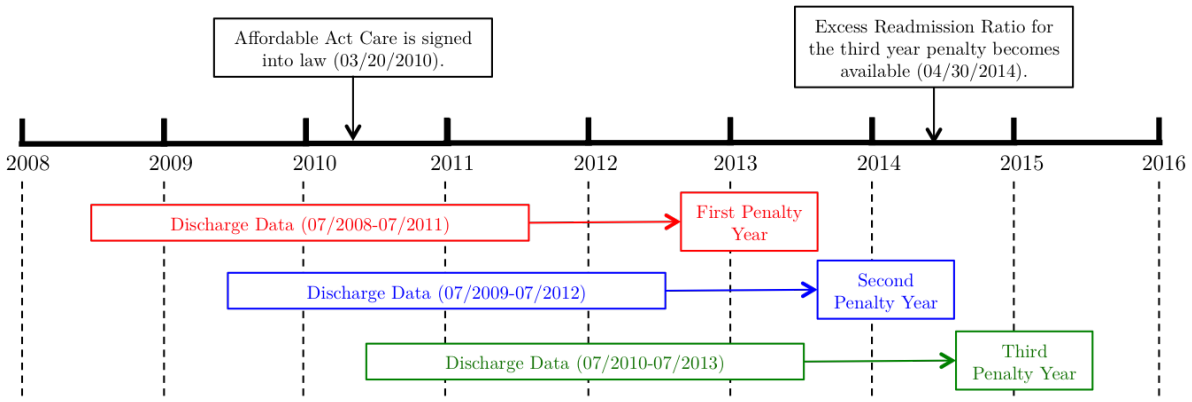


Figure 1: Timeline of the Hospital Readmissions Reduction Program

(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, in 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)[1] 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* increased 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 severely ill patients or those that come from a disadvantaged socioeconomic status are at particularly high risk for readmissions (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 the program is not to directly affect quality of care but rather to make hospitals accountable also for post-discharge processes.[2] They point to the large number of patients whose discharges are 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 during 2012 readmission rates fell in 239 out of the 309 hospital referral regions (HRR) (Gerhardt et al. 2013)

Unlike the above research focusing on whether readmission is a good quality-of-care metric, this paper takes HRRP as a given government program and analyzes 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 in HRRP penalties in the fiscal year 2013 (Fontanarosa and McNutt 2013). Many hospitals incurred hundreds of thousands of dollars in penalties, while the worst-performing hospitals incurred millions of dollars in penalties. These amounts could be tripled by the end of 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 of all 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, for the next year, the hospital's predicted readmission rate, based on its actual performance, is greater than its CMS-expected readmission

[2] http://blogs.sph.harvard.edu/ashish-jha/the-30-day-readmission-rate-not-a-quality-measure-but-an-accountability-measure/

rate, it incurs a penalty up to the maximal cap – 1% of overall Medicare payments to the hospital in 2013. 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 and 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 nation-wide. We calibrate our model using the data from Medicare Cost dataset, and report findings focusing, particularly, on the set of hospitals that prefer paying penalties to reducing readmissions. The following observations arise from our simulation study:

**Effectiveness of the policy:** By applying our model to data we show that, across 3,000 hospitals in the US, there is a non-negligible fraction of hospitals (between 4%-13% for reasonable parameter choices) that the policy would not incentivize to reduce readmissions – these would rather incur penalties than reduce readmissions. Moreover, the number of such hospitals could increase as CMS is expanding the list of monitored diseases. For example, in 2015, CMS is adding COPD and TKA/THA to the current list of three monitored diseases: AMI, PN, and HF. CMS is advocating monitoring as many diseases as possible (MedPAC 2013). As more diseases are added,

the number of non-incentivized hospitals might increase significantly, reaching 30% if all diseases are monitored.

**Subtle effects of benchmarking**: 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 in light of 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 non-incentivized hospitals, hospitals that have high readmission rate yet are not incentivized by HRRP to improve their readmissions. The number of these hospitals is expected to increase as the set of monitored diseases is expanded in 2015 and beyond.

In other words, our model shows that, while competition among hospitals often encourages more hospitals to reduce readmissions, *competition can only increase the number of non-incentivized hospitals,* which are hospitals that prefer paying penalties over reducing readmissions in any equilibrium.

**Readmission dispersion**: Importantly, the effectiveness of the benchmarking depends on how readmission rates vary across hospitals (we call this readmission dispersion). The higher the dispersion is, the **less** effective the policy is. We propose an alternative benchmarking mechanism of the policy and link the mechanism with existing CMS recommendations (MedPAC 2013); see Section 6.2. We also apply our proposed benchmarking to data and show that it may reduce by half the number of non-incentivized hospitals for all three monitored diseases in 2013.

**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 readmissions. 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 of a hospital, the smaller the likelihood of the hospital to reduce its readmissions under HRRP. This suggests that regulating payment system, such as decreasing the probability of up-coding, could potentially reduce the contribution margin and increase the effectiveness of the policy.

**Technology:** The cost of reducing readmissions by improving processes plays an important role in hospitals' decisions on whether to reduce readmissions. Reducing the costs of readmission reduction programs can effectively reduce the number of non-incentivized hospitals. 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 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 6.3.

In Section 5, we also validate our model predictions by comparing the simulation results to the actual changes in hospitals predicted readmission ratios between 2013 and 2015. It is premature to draw definite conclusions about its long-run effectiveness. Nevertheless, this initial empirical evidence supports the ability of our model to identify those non-incentivized hospitals. Specifically, we show that the set of hospitals that paid penalties in 2013 and would still pay penalties in 2015 (i.e., non-incentivized hospitals in practice) has on average 70% overlap with the set of non-incentivized hospitals identified by our model. This shows that our model performs fairly well in identifying non-incentivized hospitals within the considered parameter spaces.

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 a 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 *e*xogenous arrival of patients for each disease $i$, e.g., AMI, HF, and PN, and each insurance type $j$, e.g., Medicare, Medicaid, Private insurance, with a rate $\lambda_{ij}^{\mathcal{E}}$.[3] 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 $ij$.

[3] 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).

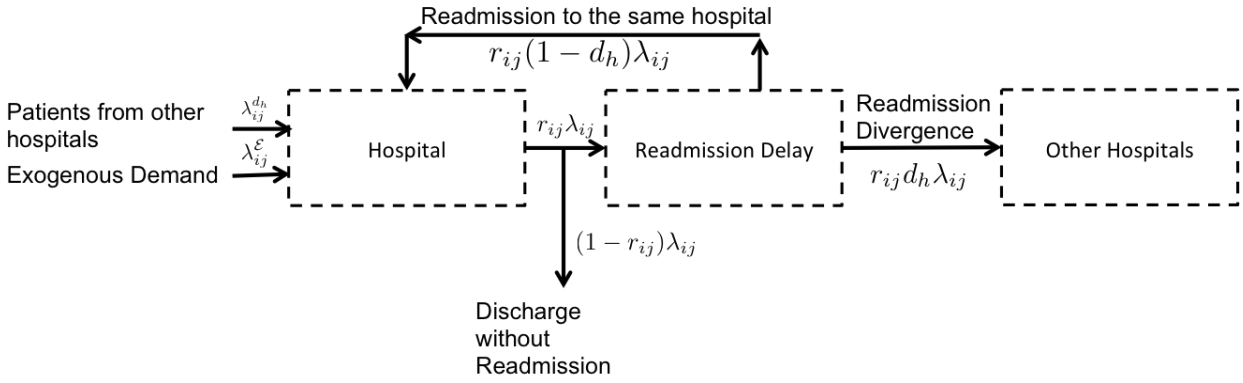


Figure 2: Operational flow of readmissions in a hospital

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 $l^k p_{ij}$.

Let $\lambda_{ij}^a = \lambda_{ij}^{\mathcal{E}} + \lambda_{ij}^{d_h}$ 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

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

where $\lambda_{ij}^a$ 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:

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

where $\Pi_{ijh}^R(0)$ 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 for a given disease $i$ 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,

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

REMARK 1. For clarity and tractability of the model we do not explicitly model the disease-level divergence. In Appendix C, we use a two-disease model and real data to show that this assumption should not affect the main implications of our model.

### 2.2. Penalty structure

On October 2012, CMS started penalizing Medicare payments to hospitals based on their excess readmission ratios for three monitored diseases. We let $\mathcal{D}$ denote the set of all diseases and $\mathcal{M} \subseteq \mathcal{D}$ 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 ($r^e_{ijh}$ for disease $i$ and hospital $h$). The term "risk-adjusted" refers to the fact that the estimated readmission rates (expected and predicted) for a hospital are 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 rate. The risk-adjusted *expected* readmission rate $r^e_{ijh}$ 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 $\frac{r_{ijh}}{r^e_{ijh}}$, for hospital $h$ is greater than 1, the penalty 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 $i$: $max\left(\frac{r_{ijh}}{r^e_{ijh}} - 1, 0\right)\Pi^R_{ijh}(r_{ijh})$, where $j = \text{Med}$.

CMS computes the *aggregate payments for excess readmissions* for hospital $h$ as the sum of payments across monitored diseases,

$$\sum_{i \in \mathcal{M}, j = \text{Med}} \max\left(\frac{r_{ijh}}{r^e_{ijh}} - 1, 0\right)\Pi^R_{ijh}(r_{ijh}).$$

CMS then defines the readmission penalty ratio as the aggregate penalty for excess readmissions divided by the Medicare revenue over all diseases, $\sum_{i \in \mathcal{D}, j = \text{Med}} \Pi^R_{ijh}(r_{ijh})$. The readmission penalty ratio is capped at the maximum penalty cap, denoted as $P_{cap}$. The absolute amount of the penalty

is the Medicare revenue across all diseases multiplied by the minimum of the readmission penalty ratio and the cap:

$$\mathbb{P}_h(r_h, r_h^e) = \left( \sum_{i\in\mathcal{D}, j=\text{Med}} \Pi_{ijh}^R(r_{ijh}) \right) * \min \left( \frac{\sum_{i\in\mathcal{M}, j\in\text{Med}} \max\{\frac{r_{ijh}}{r_{ijh}^e} - 1, 0\} \Pi_{ijh}^R(r_{ijh})}{\sum_{i\in\mathcal{D}, j=\text{Med}} \Pi_{ijh}^R(r_{ijh})}, P_{cap} \right)$$

which can be rewritten as

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

For example, if $P_{cap} = 3\%$, a hospital with total Medicare revenue of \$1 million will pay at most \$30,000 in penalties (i.e., $P_{cap} \sum_{i\in\mathcal{D}, j=Med} \Pi_{ijh}^R(r_{ijh}) = 0.03 * 1000000 = 30000$). If there is only one monitored disease (i.e., $\mathcal{M} = \{1\}$) that accounts for 20% of hospital's Medicare revenue ($\Pi_{1,\text{Med},h}^R(r_{1,\text{Med},h}) = 200000$), and the excess readmission ratio is 1.1 ($\max\left(\frac{r_{1,\text{Med},h}}{r_{1,\text{Med},h}^e} - 1, 0\right) = 1.1$), the penalty will be \$20,000 since the penalty is less than the cap. If the readmission ratio is greater than 1.15, the penalty will be greater than \$30,000 ($\max\left(\frac{r_{1,\text{Med},h}}{r_{1,\text{Med},h}^e} - 1, 0\right) \Pi_{1,\text{Med},h}^R(r_{1,\text{Med},h}) > 0.15 * 200000 = 30000$). In this case, the penalty will be capped at \$30,000. Notice that the cap is applied to **all** Medicare revenue, while the penalty is proportional to revenue from monitored diseases only.[4]

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

To gain structural insights, let us first suppose that there is a single hospital with a single monitored disease ($\mathcal{M} = \mathcal{D} = \{1\}$) 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.

### 3.1. Contribution

In this single-hospital and single-disease model, the revenue, the contribution margin, and the penalty for hospital $h$ with actual readmission rate $r_h$ and risk-adjusted CMS-expected readmission ratio $\frac{r_h}{r_h^e}$ are then given by:

$$\Pi_h^R(r_h) = \Pi_h^R(0) \frac{1}{1 - l(1 - d_h) r_h}, \tag{5}$$

$$\Pi_h^P(r_h) = C_m \Pi_h^R(r_h), \tag{6}$$

$$\mathbb{P}_h(r_h, r_h^e) = \phi_h \Pi_h^R(r_h) \min\left( \max\left( \frac{r_h}{r_h^e} - 1, 0 \right), P_{cap} \right), \tag{7}$$

where $\phi_h$ is the percentage of hospital $h$'s revenue that comes from Medicare. In the special case that $d_h = 0$ and $l = 1$, we have the simpler expression

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

[4] For more details of the policy, refer to http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html

### 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 hospital's operating costs in each subsequent year. The function $C(y, x)$ is assumed to be continuous, non-negative when $x \le y$, and to have a second derivative that satisfies $\left|\frac{\partial^2}{\partial x \partial x} C(y,x)\right| \le \eta \frac{1}{(1-x)^3}$ for some constant $\eta$ and $\forall x, r \in (0,1)$. This technical assumption is satisfied, in particular, by any function of the form

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

where $g$ is a bounded function and $\alpha \ge 0$. The term $C_h^v (r-x)^\alpha$ 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 this cost on the initial readmission rate $r_0$. 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, compromises 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 $r_h^e$, that decides to reduce its readmission from $r_{h0}$ to $r_{h1}$ has *operating margin*

$$\begin{aligned} R(r_{h0}, r_{h1}, r_h^e) &= \Pi_h^P(r_{h1}) - \mathbb{P}_h(r_{h1}, r_h^e) - C(r_{h0}, r_{h1}) \\ &= C_m \Pi_h^R(0) \frac{1}{1-r_{h1}} \left(1 - \phi_h^{med} \frac{1}{C_m} \min\left(\max\left(\frac{r_{h1}}{r^e} - 1, 0\right), P_{cap}\right)\right) - C(r_{h0}, r_{h1}), \end{aligned} \tag{8}$$

for the subsequent year. The hospital solves the maximization problem

$$r_{h1}^*(r_{h0}, r_h^e) \in \arg\max_{x \le r_{h0}} R(r_{h0}, x, r_h^e) \tag{9}$$

We assume here that hospitals are operating-margin maximizers. Whereas nonprofit hospitals should not incentivize their management to maximize 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 parameters and the policy parameters (the penalty cap).

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 $r_h^e$:*

$$r_{h1}^*(r_{h0}, r_h^e) = \begin{cases} r_h^e & \text{if } r_{h0} \in [r_h^e, f(r_{h0}, r_h^e)], \\ r_{h0} & \text{otherwise}, \end{cases} \tag{10}$$

*where $f(r_{h0}, r_h^e) \geq r_h^e$ is the maximal solution to the equation:*

$$R(f(r_{h0}, r_h^e), r_h^e, r_h^e) = R(f(r_{h0}, r_h^e), f(r_{h0}, r_h^e), r_h^e), \tag{11}$$

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

The left panel of Figure 3 depicts hospital $h$'s operating margin as a function of its targeted readmission rate, $r_{h1}$, with $d_h = 0$, $l = 1$, $r_h^e = 0.2$, and no readmission reduction costs. The left vertical line (i.e. A) indicates the position of the expected readmission rate, $r_h^e$, and the square denotes the initial readmission rate $r_{h0}$. The green vertical line (i.e. B) 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 $(r_{h1}, r_h^e)$. 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 initial readmission rate $r_{h0}$ is greater than $r_h^e$ and the operating margin at $r_{h0}$ is lower than that at $r_h^e$, then the savings in

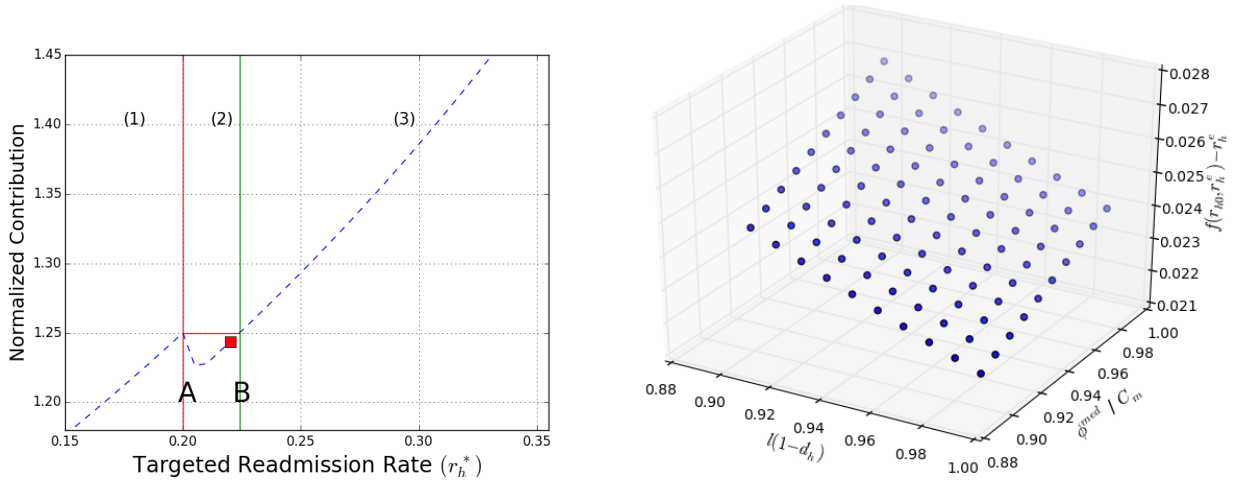


Figure 3: Left panel: No readmission reduction cost $C(r_0, r) \equiv 0$, Right panel: $f(r_{h0}, r_h^e) - r_h^e$ vs. $l(1-d_h)$ and $\frac{\phi_{med}}{C_m}$ (For $r_e = 20\%$ and $P_{cap} = 3\%$)

penalties from reducing the readmission rate outweigh the loss of contribution. The hospital's optimal decision is to reduce its readmission rate to $r_h^e$ (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 contribution 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 $r_{h0} > r_h^e$. The first is the magnitude of the excess $r_{h0} - r_h^e$. If this excess 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, $f(r_{h0}, r_h^e) - r_h^e$, 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. The right panel of Figure 3 shows how the width of Region (2), $f(r_{h0}, r_h^e) - r_h^e$, changes with other parameters (i.e. $\phi^{med}$, $C_m$, $d_h$, and $l$.). These observations lead to the following corollary that summarizes the comparative statics linking $f(r_{h0}, r_h^e) - r_h^e$ to the primitives of a hospital $(l, d_d, d_h, \lambda^a, \phi^{emd})$.

COROLLARY 1. *The width of Region (2) ($f(r_{h0}, r_h^e) - r_h^e$) is weakly increasing in the percentage of contribution from Medicare patients $\phi^{med}$ and the hospital divergence probability $d_h$. It is also weakly decreasing in the readmission-reduction cost coefficient $C_h^v$, the contribution margin ratio $C_m$, and the readmission adjustment factor $l$.*

These comparative statics 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.

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

Last, as expected, when the margin $C_m$ is high the opportunity cost associated with reducing readmissions is larger. The contribution margin of a hospital has an inverse relationship 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, which, recall, results from benchmarking among hospitals.

The main result of this section is that, while strategic interactions among hospitals can increase the number of PE hospitals, it can only increase the number of NPE 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 which is expanded to a multi-year 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 none) 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 that $\lambda_{ij}^{d_h} \equiv 0$. That is, that the divergence throughput is 0. This is seemingly a rather strong assumption, but our main results 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, $r_{h1}^e$, that CMS computes based on the actions of all hospitals. The dynamics of the game are as follows:

(0) Period 0:

a. Let $r_{h0}$ denote the current readmission rate at hospital $h$ and $r^e_{h0}$ 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 so as to maximize $R(r_{h0}, r_{h1}, r^e_{h1})$ (see Equation 8).

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

In making decisions, each hospital knows all other hospitals' initial readmission rate $r_{h0}$ and the CMS-expected readmission rate $r^e_0$. This information is publicly available from CMS (http://www.Medicare.gov/hospitalcompare/Data/30-day-measures.html).

Hospitals, however, cannot precisely predict $r^e_{h1}$ 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 it does not have the patient-level discharge data for all other hospitals. Getting access to such data is costly. Moreover, CMS may change its CMS-expected-readmissions calculations frequently making it difficult for hospitals to predict in advance what formulas will be used. For example, from 2011 to 2012, CMS added the readmission cases from VA hospitals into the estimation model.

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, $g_h(\cdot,\cdot)$, where

$$\bar{r}^e_{h1} = g_h(\vec{r}_1, \vec{r^e_0}) \tag{12}$$

is a proxy for its true CMS-expected readmission rate. The hospital chooses $r_{h1}$ to maximize $R(r_{h0}, r_{h1}, \bar{r}^e_{h1})$. The only property of $g_h$ that we use in our analysis is it is weakly increasing in $r_{\tilde{h}1}$ for any hospital $\tilde{h}$. [5]

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

Definition 1. Let $\vec{r_0} = \{r_{10}, r_{20}, ..., r_{H0}\}$ be the initial readmission rates and $\vec{r^e_0} = \{r^e_{10}, r^e_{20}, ..., r^e_{H0}\}$ be the initial expected readmission rates of the $H$ hospitals. Hospital $h$'s strategy space is $r_{h1} \in [0, r_{h0}]$. The payoff function for hospital $h$ is $R(r_{h0}, r_{h1}, g_h(\vec{r_1}, \vec{r^e_0}))$ defined in Equation 8.

A Nash Equilibrium in pure strategies is a readmission vector $\vec{r^*_1}$ such that $r^*_{h1} \in \arg\max_{r_{h1}\in[0,r_{h0}]} R(r_{h0}, r_{h1}, g_h((\vec{r^*}_{(-h)1}, r_{h1}), \vec{r^e_0}))$ for every hospital $h$, where $\vec{r}_{(-h)1}$ represents a vector of readmission rates for hospitals except the hospital $h$. A mixed-strategies Nash Equilibrium is $\pi = \{\pi_1, \pi_2, ..., \pi_H\}$ where $\pi_h$ is a probability distribution with support $[0, r_{h0}]$.

[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 and are able to perfectly predict the CMS targets.

Let $\vec{r}_{(-h)1}$ denote the decisions of hospital $h$'s peers. Hospital $h$'s best response is given by:

$$BR_h(r^e_{h0}, r_{h0}, r_{(-h)1}) = \begin{cases} g_h(\vec{r_1}, r^e_{h0}) \text{ if } r_{h0} \in [g_h(\vec{r_1}, r^e_{h0}), f_h(r_{h0}, g_h(\vec{r_1}, r^e_{h0}))], \\ r_{h0} \qquad\qquad \text{otherwise,} \end{cases} \tag{13}$$

where $f$ is as in Equation 11 and characterizes the PE region of a hospital. That is, given the actions of all other hospitals, hospital $h$ is solving a single-hospital optimization problem as in Section 3.3. 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 ($f(r_{h0}, r^e_h) = r^e_h$). For such a hospital it is financially inefficient for the hospital to reduce readmissions.

It is a-priori unclear how a hospital's decision affects the decisions of other hospitals. If one hospital decides to reduce its readmission rate, it effectively lowers the expected readmission rate for other hospitals, and decreases other hospitals' payoffs monotonically. This action 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 the payoff function is semi-continuous 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 specific updating functions, 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 $(\vec{r_0}, \vec{r^e_0})$ 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:

$$r_{h0} > f_h(r_{h0}, g_h(\vec{r_0}, r^e_{h0})) \tag{14}$$

where $f_h$ is defined in Equation 11. We also call SNPE hospitals "worst offenders."

A hospital is SNPE if, considering its current CMS-expected readmission rates, and its current readmission, reducing readmissions is a sub-optimal decision. 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 $h$, $\pi_h(r_{h0}) = 1$. 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.*

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 SNPE hospitals: since readmission reduction by other hospitals can only further decrease the CMS-expected readmission rate, a hospital that is SNPE under $r_0^e$, will find reducing readmissions even less beneficial with the new (lower) targets. Put differently, the worst offenders are indifferent to the benchmarking. In other words, 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 $\vec{\bar{r}_0^e} = \{r_{1,0}^e, r_{2,0}^e, ...r_{H,0}^e\}$.

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

2. Update the readmission rate vector as follows:

$$r_{h,n+1} = \begin{cases} g_h(\vec{r_n}, \vec{r_n^e}) & \text{if } r_{h,n} > g_h(\vec{r_n}, \vec{r_n^e}), h \notin \text{SNPE}, \\ r_{h,n} & \text{otherwise.} \end{cases} \tag{15}$$

Any hospital $h$ with a readmission rate $r_{h,n}$ at step $n$ that is greater than its expected readmission rate $\bar{r}_{h,n}^e$, reduces to its CMS-expected readmission rate. Set $n \leftarrow n+1$.

3. If there are no hospitals that reduce readmissions 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 strongly program-effective (SPE) hospital if:

$$r_{h,N} < r_{h0} \tag{16}$$

In other words, an SPE hospital reduces its 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:*

$$\forall \pi, \ \sum_{h=1}^{H} 1_{\{\mathbb{E}[r_{h,\pi}] < r_{h0}\}} \leq \sum_{h=1}^{H} 1_{\{h \in SPE\}} \tag{17}$$

*where $\mathbb{E}[r_{h,\pi}]$ 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 readmission rates that are (with probability 1 in a mixed equilibrium) lower than their CMS-expected readmission rates. 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 obtain insights into a game in which the existence or uniqueness of pure-strategy equilibria is not guaranteed. Two-hospitals and three-hospitals games are more tractable and support insights derived through the bounds.

Consider a symmetric model with two hospitals and a single disease. The two hospitals have the same patient volume, same patient mix, only Medicare patients and a 100% readmission 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 $r_1^e = r_2^e = \frac{r_{10}+r_{20}}{2}$.

Label hospitals such that $r_{10} < r_{20}$. In this case $r_{10} < r_1^e$, 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 $\frac{1}{1-r_{20}} \min\{\max\{r_{20}/r_{10} - 1, 0\}, P_{cap}\}$ under the latter. Thus, hospital 2 will reduce its readmissions if, and only if,

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

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

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

The equilibrium underscores two drivers of HRRP effectiveness:

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

(2) The second is **readmission dispersion**, which represents how disperse a set of hospitals' readmission rates are from their average. We formally define readmission dispersion as $\mathcal{V}(\vec{r_0}) = \frac{\sum_{h \in \mathcal{S}} r_{h0} - \bar{r}_0}{|\mathcal{S}|}$ where $\bar{r}_0 = \frac{\sum_h r_{ho}}{H}$ and $\mathcal{S} = \{i | r_{h0} > \bar{r}_0\}$. 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

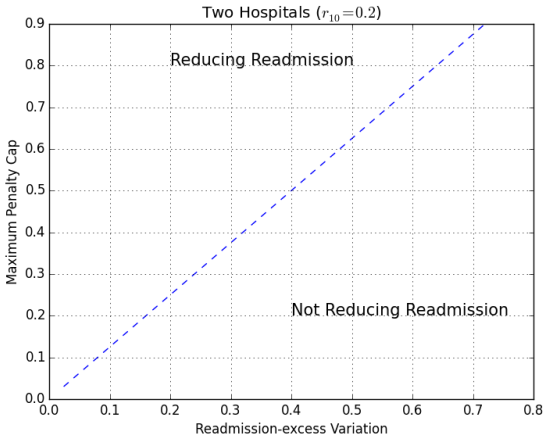


Figure 4: Trade-off of Maximum Penalty Cap and Readmission Dispersion (2-Hospital Model)

indifferent if the dispersion equals the 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 $\mathcal{V}(\vec{r})$ (by varying $r_{20}$ and $P_{cap}$) and compute the (unique Pareto-dominant equilibrium) for each pair of values. We clearly see 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 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 6.3.

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

$$\lambda_1^{d_h} = \frac{r_2 d_h}{1-r_2}\lambda_2^a = \frac{r_2 d_h}{1-r_2}(\lambda_2^{\mathcal{E}} + \lambda_2^{d_h}) \text{ and } \lambda_2^{d_h} = \frac{r_1 d_h}{1-r_1}(\lambda_1^{\mathcal{E}} + \lambda_1^{d_h}),$$

where $\lambda_i^{\mathcal{E}},\ i \in \{1,2\}$ is the exogenous arrival rate to hospital $i$. It turns out that there is still one unique Pareto-dominant 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 (no-divergence) game cannot have profitable unilateral deviation when there is positive divergence throughput. This shows that any equilibrium in the original game 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 positive divergence throughput. In other words, our model is robust towards the simplifying assumption that $\lambda_{ij}^{d_h} \equiv 0$.

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 equilbria.

### 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 impact of HRRP.

Let $P_{cap}^{l}$ denote the cap in the $l^{th}$ period. Let $P_{cap}^{max} = \max_{l=1,\ldots,n} P_{cap}^{l}$ 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 with positive probability. A 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 Medicare Cost and Hospital Compare datasets for fiscal year 2013 to better understand the drivers of 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 construct financial and operational data for 3,408 hospitals nationwide from Medicare Cost dataset provided by the CMS *www.cms.gov*. For fiscal year 2013, this data allows us to compute the fraction of in-patient revenue from Medicare patients; see Table 1.

Second, the CMS reports the CMS-expected and predicted readmission rates for each hospital for fiscal year 2013. CMS computes these from the hospitals' discharge data from July 2008 to June 2011. Out of 3,408 hospitals in the Medicare Cost dataset, 3,304 hospitals have CMS-expected and predicted readmission rates from fiscal year 2013 to 2015. After matching these two datasets, the number of hospitals per monitored disease are 2,041, 2,732 and 2,752 for AMI, HF, PN respectively. The differences in the hospital counts across diseases are due to the selection rule of HRRP, according to which a hospital with a small number of readmissions in a monitored diseases is not considered in the penalty evaluation for that disease. The data is summarized in Table 1.

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

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Fraction of Revenue from Medicare | 3,209 | 0.170 | 0.077 | 0.000 | 1.000 |
| Fraction of Medicare Revenue from Monitored Diseases | 3,317 | 0.375 | 0.188 | 0.000 | 1.000 |
| Number of AMI Discharges | 3,301 | 149.381 | 205.410 | 0 | 1,652 |
| Predicted Readmission Rate for AMI | 2,212 | 20.031 | 3.303 | 10.200 | 35.800 |
| Expected Readmission Rate for AMI | 2,212 | 19.933 | 2.377 | 14.300 | 30.700 |
| Number of HF Discharges | 3,301 | 354.767 | 353.426 | 0 | 3,667 |
| Predicated Readmission Rate for HF | 3,000 | 24.256 | 2.609 | 17.400 | 34.800 |
| Expected Readmission Rate for HF | 3,000 | 24.206 | 1.355 | 19.500 | 30.600 |
| Number of PN Discharges | 3,301 | 283.560 | 232.517 | 0 | 2,223 |
| Predicated Readmission Rate for PN | 3,012 | 18.222 | 2.330 | 11.600 | 31.000 |
| Expected Readmission Rate for PN | 3,012 | 18.166 | 1.429 | 12.800 | 25.900 |

Table 1: Summary statistics of hospital-level data

In our model, we assumed that average payment per patient $p_{ij}$ depends only on the disease $i$ and insurance type $j$ since Medicare and Medicaid systems do have a pay-per-case payment structure based on diagnosis group of patients. 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). According to past literature, small perturbation on readmission rates, and in turn on workload of the system, should not have large effects on the length of stay. For instance, Freeman et al. (2014) demonstrates that the workload has both positive direct effects and negative indirect effects on post-birth length of stay in a delivery unit, and in turn does not have significant effect on length of stay in general. Empirical evidence suggests that readmission-reduction programs do not increase the length of stay of patients (Hansen et al. 2013).

### 5.2. Simulation and results

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

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

$$C_h(r,x) = C_h^v(r-x)^\alpha + C_h^s\frac{1}{r} \tag{18}$$

where $\alpha \in [1,2]$, $C_h^v = \Pi_h(0)C_v$ and $C_h^s = \Pi_h(0)C_s$. 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. 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%. We also test the sensitivity to the cost parameters.

(2) **Updating functions $\vec{g}(\cdot,\cdot)$:** The hospital's prediction of the CMS-expected readmission rate $\bar{r}_{h1}^e$ (see Equation (12)) is computed as the average readmission rates of $\{r_{h1},\ \forall h\}$ weighted by the number of patients in each hospital. If all $H$ participating hospitals have the same number of patients, then $\bar{r}_{h1}^e$ is simply $\frac{1}{H}\sum_k r_{h1}$.

**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 aggregate 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 $\Pi_{all}$ 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

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

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

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

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

In other words, the total penalty across monitored diseases does not exceed the global penalty cap $\Pi_{all}P_{cap}$. We assume that the hospital assigns the penalty cap proportionally to the relative Medicare revenue of each disease:

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

Then the penalty can be re-written as

$$\Pi_i \times \min\left(\max\left(\frac{r_p}{r_e} - 1, 0\right), P_{cap}\left(\frac{\sum_{j\in\mathcal{M}}\Pi_j}{\Pi_{all}}\right)^{-1}\right),$$

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**,

$$P_{cap}\left(\frac{\sum_{j\in\mathcal{M}}\Pi_j}{\Pi_{all}}\right)^{-1}.$$

For example, if $P_{cap} = 3\%$, a hospital with total Medicare revenue of \$1 million will pay at most \$30,000 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 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 ($\mathcal{M} = \mathcal{D}$) then the induced penalty of disease $i$ is

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

so that the effective penalty cap for disease $i$ is simply the policy penalty cap $P_{cap}$.

**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) the contribution margin ratio $C_m$, (2) the product of the readmission adjustment factor and inverse hospital divergence rate $l(1-d_h)$, (3) the cost function parameter $\alpha$.

Whether a given hospital is SNPE or not does not depend on the parameters of other hospitals. Thus, assigning the *same* contribution margin for all hospitals and varying it from 40% to 75%

generated the same outcomes in terms of SNPE as allowing hospitals to have different contribution margins within the box $[40\%, 75\%]^H$.

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.

| | | No Cost | | | Linear Cost | | | Convex Cost | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $l \times (1 - d_h)$ | Contribution Margin | SPI | SPE | SNPE | SPI | SPE | SNPE | SPI | SPE | SNPE |
| 100 % | 40% | 15% | 81% | 4% | 22% | 71% | 7% | 29% | 62% | 9% |
| | 75% | 44% | 28% | 28% | 46% | 22% | 32% | 48% | 19% | 33% |
| 80 % | 40% | 7% | 92% | 1% | 13% | 84% | 3% | 22% | 73% | 5% |
| | 75% | 34% | 52% | 14% | 38% | 43% | 19% | 41% | 39% | 20% |
| 60 % | 40% | 2% | 98% | 0% | 5% | 94% | 1% | 15% | 82% | 3% |
| | 75% | 16% | 80% | 4% | 24% | 69% | 7% | 31% | 59% | 10% |

Table 2: Number of strongly program-indifferent (SPI), strongly program-effective (SPE) and strongly non-program effective (SNPE) hospitals for different parameters for disease AMI

| | | No Cost | | | Linear Cost | | | Convex Cost | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $l \times (1 - d_h)$ | Contribution Margin | SPI | SPE | SNPE | SPI | SPE | SNPE | SPI | SPE | SNPE |
| 100 % | 40% | 22% | 73% | 5% | 26% | 67% | 7% | 32% | 60% | 8% |
| | 75% | 45% | 30% | 25% | 47% | 24% | 29% | 50% | 21% | 29% |
| 80 % | 40% | 11% | 87% | 2% | 19% | 77% | 4% | 27% | 68% | 5% |
| | 75% | 36% | 51% | 13% | 41% | 41% | 18% | 43% | 38% | 19% |
| 60 % | 40% | 5% | 94% | 1% | 9% | 89% | 2% | 21% | 75% | 4% |
| | 75% | 22% | 73% | 5% | 28% | 64% | 8% | 34% | 57% | 9% |

Table 3: Number of strongly program-indifferent (SPI), strongly program-effective (SPE) and strongly non-program effective (SNPE) hospitals for different parameters for disease PN

| | | No Cost | | | Linear Cost | | | Convex Cost | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $l \times (1 - d_h)$ | Contribution Margin | SPI | SPE | SNPE | SPI | SPE | SNPE | SPI | SPE | SNPE |
| 100 % | 40% | 34% | 53% | 13% | 38% | 44% | 18% | 43% | 37% | 20% |
| | 75% | 51% | 8% | 41% | 51% | 6% | 43% | 52% | 5% | 43% |
| 80 % | 40% | 24% | 70% | 6% | 30% | 60% | 10% | 36% | 51% | 13% |
| | 75% | 44% | 27% | 29% | 47% | 19% | 34% | 49% | 16% | 35% |
| 60 % | 40% | 13% | 85% | 2% | 20% | 75% | 5% | 30% | 62% | 8% |
| | 75% | 33% | 54% | 13% | 38% | 43% | 19% | 43% | 36% | 21% |

Table 4: Number of strongly program-indifferent (SPI), strongly program-effective (SPE) and strongly non-program effective (SNPE) hospitals for different parameters for disease HF

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. Lastly, the higher the contribution margin ratio, $C_m$, the larger the number of SNPE hospitals.

We observe that the number of SNPE hospitals is significant even if there is no cost of reducing readmissions. In other words, there are *more than one hundred hospitals nationwide that are not incentivized* to reduce readmissions even if the cost of reducing readmission is only from the loss of throughput. If the cost is positive and linear, we find that the number of SNPE hospitals increases substantially. This shows that the HRRP has limited power in incentivizing worst-performing hospitals to reduce their readmissions.

There are two characteristics of worst offenders. First, some hospitals may be too far away from the penalty cap (i.e., too far away in Region (3)) and in turn are not incentivized to reduce. Second, some hospitals may have very high readmission rates and revenue percentage from Medicare patients so that reducing readmissions is always costly to them regardless of their positions.

The result is likely to become even worse when the set of monitored diseases is expanding starting from 2015. We re-examine this issue and propose recommendations to alleviate it when discussing policy implications in Section 6.

### 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 estimated 2015 penalties were publicized on April 30, 2014.

Our model and analysis are focused on the SNPE hospitals and we seek to validate our identification of these. First, 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 back-fill 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 perform validation study separately for each major states in the U.S. since contribution margins and divergence rates may differ across states. We define major states as states with more than 150 hospitals, which includes California, Florida, New York, Pennsylvania, and Texas. We

| State | Disease | Number of Hospitals | Recall | Precision |
|---|---|---|---|---|
| CA | AMI | 302 | 0.73 | 0.77 |
| CA | PN | 302 | 0.77 | 0.62 |
| CA | HF | 302 | 0.91 | 0.57 |
| FL | AMI | 165 | 0.86 | 0.71 |
| FL | PN | 165 | 0.83 | 0.75 |
| FL | HF | 165 | 0.95 | 0.60 |
| NY | AMI | 159 | 0.91 | 0.72 |
| NY | PN | 159 | 0.80 | 0.77 |
| NY | HF | 159 | 0.97 | 0.77 |
| PA | AMI | 155 | 0.98 | 0.77 |
| PA | PN | 155 | 0.94 | 0.67 |
| PA | HF | 155 | 1.0 | 0.66 |
| TX | AMI | 308 | 0.85 | 0.53 |
| TX | PN | 308 | 0.91 | 0.5 |
| TX | HF | 308 | 1.0 | 0.41 |
| Overall | AMI | 1089 | 0.85 | 0.69 |
| Overall | PN | 1089 | 0.85 | 0.63 |
| Overall | HF | 1089 | 0.96 | 0.57 |

Table 5: Recall and Precision of the identification of SNPE hospitals.

use parameter sets within the parameter space of original simulation to conduct this validation. Following traditions in the binary classification in machine learning literature (Forman 2003), we examine our results in two directions: **Recall:** The probability that a hospital that was penalized in 2013 and still pays penalty in 2015 is not a SPE hospital (i.e., the proportion of actual positives which are predicted positive); **Precision:** The probability that a hospital that is classified as SNPE hospital does pay penalty in 2015 (i.e., the proportion of predicted positives which are actual positive).

Table 5 shows that, overall, the recall and precision are around 70%. This shows that the set of hospitals that paid penalties in 2013 and will still pay penalties by the end of 2015 has roughly 70% overlap with the set of SNPE hospitals identified by our model. This demonstrates that our model performs remarkably well in identifying worst offenders within the parameter space we consider.

## 6. Policy Implication

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 are added

to the set of monitored diseases already this 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 $P_{cap}$. Tables 6, 7 and 8 report the results in this case.

| | | No Cost | | | Linear Cost | | | Convex Cost | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $l \times (1 - d_h)$ | Contribution Margin | SPI | SPE | SNPE | SPI | SPE | SNPE | SPI | SPE | SNPE |
| 100 % | 40% | 48% | 25% | 27% | 49% | 21% | 30% | 48% | 22% | 30% |
| | 75% | 51% | 8% | 41% | 50% | 7% | 43% | 50% | 8% | 42% |
| 80 % | 40% | 45% | 34% | 21% | 48% | 27% | 25% | 48% | 26% | 26% |
| | 75% | 50% | 15% | 35% | 50% | 13% | 37% | 50% | 14% | 36% |
| 60 % | 40% | 38% | 50% | 12% | 44% | 36% | 20% | 46% | 32% | 22% |
| | 75% | 49% | 24% | 27% | 50% | 19% | 31% | 50% | 20% | 30% |

Table 6: Number of strongly program-indifferent (SPI), strongly program-effective (SPE) and strongly non-program effective (SNPE) hospitals for different parameters for disease AMI

| | | No Cost | | | Linear Cost | | | Convex Cost | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $l \times (1 - d_h)$ | Contribution Margin | SPI | SPE | SNPE | SPI | SPE | SNPE | SPI | SPE | SNPE |
| 100 % | 40% | 50% | 26% | 24% | 50% | 23% | 27% | 50% | 23% | 27% |
| | 75% | 52% | 10% | 38% | 53% | 8% | 39% | 52% | 9% | 39% |
| 80 % | 40% | 47% | 35% | 18% | 50% | 27% | 23% | 50% | 27% | 23% |
| | 75% | 53% | 16% | 31% | 52% | 14% | 34% | 52% | 15% | 33% |
| 60 % | 40% | 37% | 53% | 10% | 46% | 37% | 17% | 48% | 33% | 19% |
| | 75% | 50% | 25% | 25% | 52% | 20% | 28% | 50% | 22% | 28% |

Table 7: Number of strongly program-indifferent (SPI), strongly program-effective (SPE) and strongly non-program effective (SNPE) hospitals for different parameters for disease PN

| | | No Cost | | | Linear Cost | | | Convex Cost | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $l \times (1 - d_h)$ | Contribution Margin | SPI | SPE | SNPE | SPI | SPE | SNPE | SPI | SPE | SNPE |
| 100 % | 40% | 50% | 18% | 32% | 51% | 14% | 35% | 51% | 15% | 34% |
| | 75% | 52% | 3% | 45% | 51% | 3% | 46% | 51% | 3% | 46% |
| 80 % | 40% | 47% | 28% | 25% | 49% | 22% | 29% | 49% | 22% | 29% |
| | 75% | 51% | 8% | 41% | 51% | 6% | 43% | 51% | 7% | 42% |
| 60 % | 40% | 43% | 40% | 17% | 47% | 30% | 23% | 48% | 28% | 24% |
| | 75% | 51% | 18% | 31% | 52% | 13% | 35% | 51% | 15% | 34% |

Table 8: Number of strongly program-indifferent (SPI), strongly program-effective (SPE) and strongly non-program effective (SNPE) hospitals for different parameters for disease HF

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 remainder of 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 of making the effective penalty cap smaller for monitored diseases, which may increase the number of SNPE hospitals. 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. A past study (Zhang et al. 2010) has shown hat hospitals within the same HRR have similar performance in various measures of quality of care. Therefore, we hypothesize that benchmarking hospitals in the HRR level will result in a smaller readmission dispersion.

We cannot rerun the hierarchical logistic regression to estimate the HRR-level fixed effect since this requires the patient-level data for all hospitals from CMS. Instead, we compute each hospital's HRR-level expected readmission rate as its national expected readmission rate multiplied by a HRR correction factor, which represents the ratio of average hospitals' performance in the nation and in the HRR. We calculate the correction factor for each disease and each HRR as the weighted average predicted readmission rate for hospitals in that HRR divided by the weighted average predicated rate for hospitals in the nation. We merge an HRR with less than 5 hospitals to an adjacent HRR.

With the new HRR-level expected readmission rates for each hospital, we re-compute the set of SNPE hospitals for each HRR. Table 9 shows the percentage of hospitals that are SNPE under the baseline parameter for both the nationwide benchmark and the HRR-wise benchmark. Indeed, we

| | AMI | PN | HF |
|---|---|---|---|
| Percentage of SNPE hospitals (Nationwide and 3 monitored disease) | 4% | 5% | 13% |
| Percentage of SNPE hospitals (HRR-wise and 3 monitored disease) | 1% | 2% | 9% |
| Percentage of SNPE hospitals (Nationwide and all disease monitored) | 27% | 24% | 32% |
| Percentage of SNPE hospitals (HRR-wise and all disease monitored) | 17% | 21% | 26% |

Table 9: Fraction of hospitals that are SNPE under different comparison policies ($C_m = 0.4$, $l \times (1 - d_h) = 100\%$ and $C_v = C_s = 0$).

decreased the number of SNPE hospitals by 75%, 60%, and 31% for AMI, PN, and HF respectively by comparing hospitals HRR-wise.

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 SNPE hospitals. The alternative local 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 unfairness concerns raised about HRRP such as the possibility of over-penalizing hospitals in geographic regions with low socioeconomic population.[6][7]

**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.

### 6.3. Penalty cap and Process-improvement costs

There are two important metrics that affect the incentives of all hospitals nationwide: the penalty cap and the cost of reducing readmissions.

**Penalty Cap:** HRRP is less effective on SNPE hospitals—those that have initial readmission rates that are significantly greater than their CMS-expected readmission rates. 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 5 displays the percentage change in the number of SNPE hospitals as we increase the penalty cap for each of the three monitored diseases. Most of the reduction in the number of SNPE

[6] http://www.manchin.senate.gov/public/index.cfm/press-releases?ID=e43f6f51-bee5-4be0-9c77-9d1096a121ff

[7] http://www.charlestondailymail.com/article/20140620/DM0104/140629951

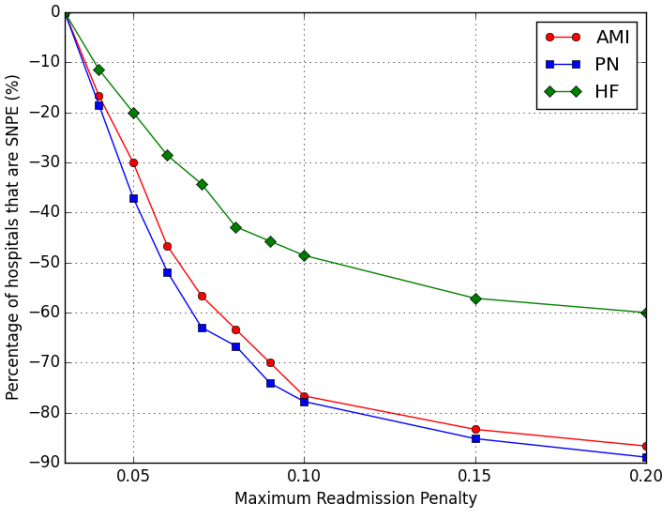


Figure 5: Equilibrium Behavior of Hospitals under Different Maximum Penalty Caps ($\alpha = 1$, $C_v = 0.01$, $l = 0.8$, $d_h = 0.15$, $C_m = 40\%$)

hospitals is achieved by increasing the cap to 10% for all three diseases. The effect diminishes as the maximum penalty cap increases.

Moreover, we observe that cap-increase has a differential effect on diseases. Increasing the maximum penalty cap from 3% to 10%, reduces by 80% the number of SNPE hospitals for AMI and PN, but only by 50% for HF. The main driver here is there are many hospitals that are much less incentivized to reduce for HF compared to AMI and PN. For instance, among the SNPE hospitals when the max penalty is 3%, there are 20% hospitals which are at least 3 percent away from their target, while only 14% and 17% such hospitals exist for PN and AMI. Moreover, there are 41% hospitals among SNPE hospitals for HF having percentage of revenue from Medicare patients less than 15%, while AMI and HF only have 29% and 36% such hospitals respectively among their SNPE hospitals.

**Process-improvement costs:** As CMS notes, the major goal of designing the policy is that the penalty for not meeting reduction targets exceeds than 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 6 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 10%, 11%, and 8% for AMI, PN, HF respectively. Similarly, diseases benefit differently from reducing process-

improvement costs depending on the number of hospitals that are close to the boundary between PE and NPE regions. The more such hospitals, the more benefit there will be to reductions in process improvement costs. Facing a decision between investments in readmission reduction toolboxes (such as Boost, see Hansen et al. (2013)) the government may want to selectively target diseases.

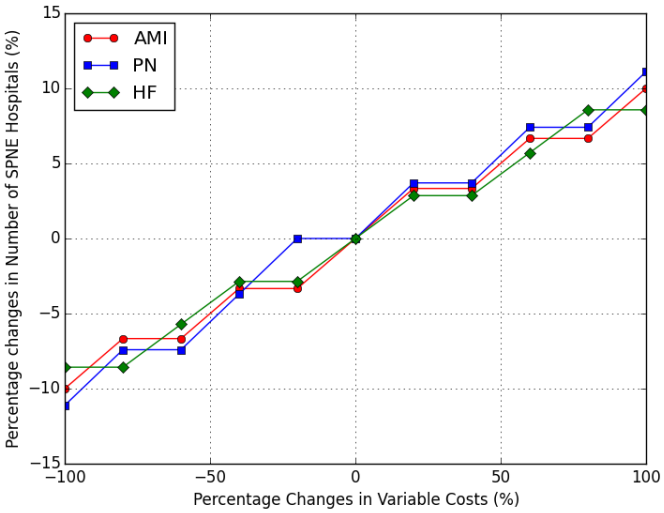


Figure 6: Equilibrium Behavior of Hospitals under Different Variable Costs ($\alpha = 1$, $P_{cap} = 0.03$, $l = 0.8$, $d_h = 0.15$, $C_m = 40\%$)

**Implication:** We show that increasing the penalty or reducing process-improvement costs is helpful in incentivizing more hospitals to reduce readmissions. However, different diseases benefit differently. The diseases which have fewer hospitals that are close to the PE-NPE boundary would benefit the most.

### 6.4. Hospital characteristics

First, HRRP has relatively greater influence on hospitals that have higher divergence probability. Under base parameters, the number of SNPE hospitals will decrease by 75%, 60%, and 50% for AMI, PN, and HF respectively if the divergent probability increases from 0% to 20%. Hospitals with higher divergence rates 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 would give those hospitals higher incentives to reduce readmissions, and in turn make the policy more effective. Since those hospitals which have higher readmission rates tend to have low quality of care, pay-for-performance programs can effectively achieves this.

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

## 7. Concluding Remarks

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 $300 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. We show that, while competition among hospitals introduced by HRRP often encourages more hospitals to reduce readmissions, it can only increase the number of non-incentivized hospitals, which are hospitals that prefer paying penalties over reducing readmissions in any equilibrium. Moreover, the unintended consequences of competition are likely to become worse as CMS follows its plan to expand the set of monitored diseases (MedPAC 2013). We propose to localize the benchmarking process in order to mitigate the negative effect of competition.

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 we hope that our model can serve as a starting point for estimating readmission reduction costs, hospitals incentives towards readmission reduction, and other factors that affect the policy outcomes. Moreover, HRRP's benchmarking is not unique among health care policies. Recently, CMS introduced the Bundled Payments for Care Improvement (BPCI) initiative in which parameters are also set through benchmarking against peers.[8] We hope that our findings (and our approach to the policy analysis) can be applied to evaluate these programs and to minimize the unintended consequences introduced by competition.

[8] http://innovation.cms.gov/initiatives/bundled-payments/

## References

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

Zeynep Aksin, Mor Armony, and Vijay Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688, 2007.

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

Ann P Bartel, Carri W Chan, and Song-Hee Hailey Kim. Should hospitals keep their patients longer? the role of inpatient and outpatient care in reducing readmissions. Technical report, National Bureau of Economic Research, 2014.

Mohsen Bayati, Mark Braverman, Michael Gillam, Karen M Mack, George Ruiz, Mark S Smith, and Eric Horvitz. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PloS one*, 9(10):e109264, 2014.

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

John Q Cheng and Michael P Wellman. The walras algorithm: A convergent distributed implementation of general equilibrium outcomes. *Computational Economics*, 12(1):1–24, 1998.

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

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

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

Partha Dasgupta and Eric Maskin. The existence of equilibrium in discontinuous economic games, i: Theory. *The Review of Economic Studies*, pages 1–26, 1986.

Francis De Véricourt and Yong-Pin Zhou. Managing response time in a call-routing problem with service failure. *Operations Research*, 53(6):968–981, 2005.

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

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

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

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

George Forman. An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research*, 3:1289–1305, 2003.

Michael Freeman, Nicos Savva, and Stefan Scholtes. Decomposing the effect of workload on patient outcomes: An empirical analysis of a maternity unit. 2014.

Geoffrey Gerhardt, A Yemane, P Hickman, A Oelschlaeger, E Rollins, and N Brennan. Data shows reduction in medicare hospital readmission rates during 2012. *Medicare & Medicaid Research Review*, 3(2):E1–E12, 2013.

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

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

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

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

Michael Hu, Bruce L Jacobs, Jeffrey S Montgomery, Chang He, Jun Ye, Yun Zhang, Julien Brathwaite, Todd M Morgan, Khaled S Hafez, Alon Z Weizer, et al. Sharpening the focus on causes and timing of readmission after radical cystectomy for bladder cancer. *Cancer*, 120(9):1409–1416, 2014.

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

Karen E Joynt and Ashish K Jha. Characteristics of hospitals receiving penalties under the hospital readmissions reduction program. *JAMA*, 309(4):342–343, 2013.

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

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

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

MedPAC. *Report to the Congress: Medicare and the health care delivery system.* MedPAC, 2013.

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

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

Manfred Padberg and Giovanni Rinaldi. A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems. *SIAM review*, 33(1):60–100, 1991.

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

Z Justin Ren and Yong-Pin Zhou. Call center outsourcing: Coordinating staffing level and service quality. *Management Science*, 54(2):369–383, 2008.

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

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

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

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

Carl van Walraven, Carol Bennett, Alison Jennings, Peter C Austin, and Alan J Forster. Proportion of hospital readmissions deemed avoidable: a systematic review. *Canadian Medical Association Journal*, pages cmaj–101860, 2011.

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

Joshua R Vest, Larry D Gamm, Brock A Oxford, Martha I Gonzalez, and Kevin M Slawson. Determinants of preventable readmissions in the united states: a systematic review. *Implementation Science*, 5(1): 88, 2010.

Yuting Zhang, Katherine Baicker, and Joseph P Newhouse. Geographic variation in the quality of prescribing. *New England Journal of Medicine*, 363(21):1985–1988, 2010.

## Appendix

## A. CMS's Estimation of readmission rates

CMS computes $r_h$ and $r_h^e$ for every hospital using patient level discharge and readmission data as follows: Let $Y_{ilk}$ 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:

$$\begin{aligned} \log(P(Y_{ilk}=1)) &= \alpha_{ik} + \beta_i' Z_{ijk} \\ \alpha_{ik} &= \mu_i + \omega_{ik} \quad \omega_k \in N(0,\tau^2) \end{aligned} \tag{19}$$

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:

$$\begin{aligned} r_{ik}^e &= \frac{1}{N_{ki}} \sum_{j=1}^{N_{ki}} \frac{1}{1+e^{-\mu_i - \beta_i Z_{ilk}}} \\ r_{ik}^p &= \frac{1}{N_{ki}} \sum_{j=1}^{N_{ki}} \frac{1}{1+e^{-\alpha_{ik} - \beta_i Z_{ilk}}} \end{aligned} \tag{20}$$

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 7 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 $[0.2 + a * x_1(n), 0.2 + a * x_2(n), 0.2 + a * x_3(n)]$. 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.

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

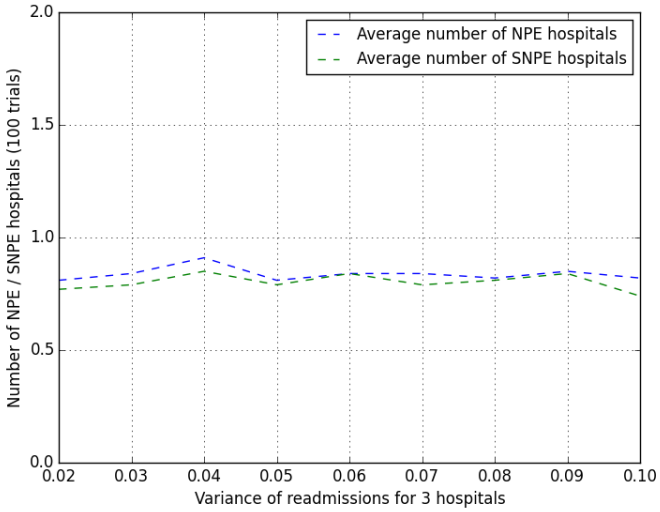


Figure 7: Number of SNPE Hospitals v.s. Number of NPE Hospitals (3-Hospital Model)

average expected number of NPE hospitals in all cases supporting the use of SNPE hospitals to evaluate the policy effectiveness.

Next we vary both $P_{cap}$ and the readmission dispersion. Given a readmission dispersion $x$ and $P_{cap}$, we generate the 100 readmission vectors $\vec{r}_0 = [0.2 - 2x, 0.2 + u * x, 0.2 + (1 - u) * x]$ 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 $(x, P_{cap})$.

Figure 8 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%.

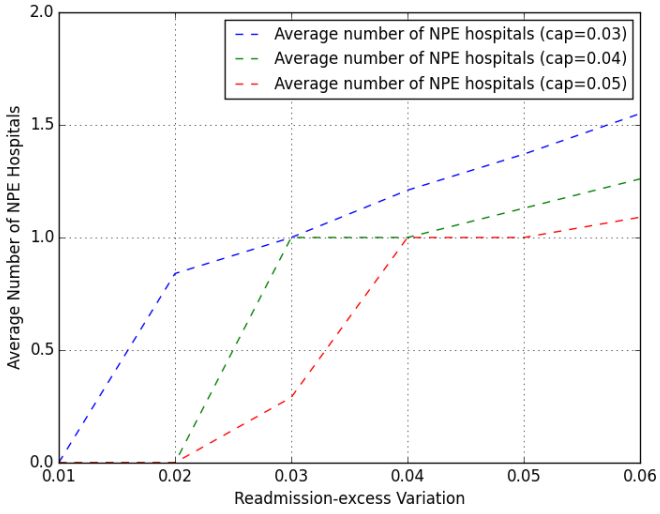


Figure 8

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:

$$x^* = \arg\max_{x \le r} R(r, x, r_e) = \arg\max_{x \le r} \Pi_h(r)(1 - \mathbb{P}_h(x, r_e)) - C(r, x) \tag{21}$$

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

$$\Pi_h(x) = \Pi_h(0)\frac{1}{1-x},$$
$$\mathbb{P}_h(x,r^e) = \frac{\phi_{med}}{C_m}\min(\max(\frac{x}{r^e}-1,0),P_{cap}),$$
$$C(r,x) = C_v(r^\alpha - x^\alpha).$$

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 $x \in [r_e, x_m]$:

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

Since, by assumption, $\frac{dC(r,x)}{dx} < 0$ then $\frac{dR(r,x,r_e)}{dx} > 0\ \forall x \in [r_e, x_m]$, if $\frac{P_{med}}{C_m} > \frac{r_e}{1-r_e}$, so that the optimal choice is $x_m$. If, on the other hand, $\frac{P_{med}}{C_m} < \frac{r_e}{1-r_e}$, then since $|\frac{dC^2(r,x)}{dx^2}| \le \frac{1}{(1-x)^3}\ \forall x \in (0,1)$, it must be the case that $\frac{dR(r,x,r_e)}{dx}$ 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), $f(r_{h0}, r_h^e)$ is defined as:

$$R(f(r_{h0},r_h^e),r_h^e,r_h^e) = R(f(r_{h0},r_h^e),f(r_{h0},r_h^e),r_h^e), \tag{23}$$

or equivalently,

$$\Pi_h^P(r^e) - \Pi_h^P(f(r_{h0},r_h^e)) + \mathbb{P}_h(f(r_{h0},r_h^e),r^e) - C(f(r_{h0},r_h^e),r^e) = 0 \tag{24}$$

Recall that $\mathbb{P}_h(x,r^e) = \frac{\phi_{med}}{C_m}\min(\max(\frac{x}{r^e}-1,0),P_{cap})$. Thus, as $\phi^{med}$ increases, $\mathbb{P}_h(f(r_{h0},r_h^e),r^e)$ increases. In turn, Equation (24) guarantees that $f(r_{h0},r_h^e)$ must increase (for a fixed $r_h^e$). Similarly, if $d_h$ increases or $l$ decreases, $\Pi_h^P(r^e) - \Pi_h^P(f(r_{h0},r_h^e)) + P_h(f(r_{h0},r_h^e),r^e)$ increases, and $f(r_{h0},r_h^e)$ increases for a fixed $r_h^e$.

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

## 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 $\bar{r}^e_{h1}$ 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, $r^e_{h1}$ reduces to $\frac{1}{H}\sum_k r_{h1}$. We denote this updating mechanism by $\bar{r}^e_{h1} = g(\vec{r_1})$, where $\vec{r_1} = \{r_{11}, r_{21}, ..., r_{H1}\}$.

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

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

By Proposition 1, each hospital chooses between reducing to the average ($g(\vec{r_0})$) 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 $\vec{r_0} = \{0.2, 0.24, 0.24\}$. 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

$$r_{h0} > f_h(r_{h0}, g_h(\vec{r_0}, r^e_{h0})). \tag{25}$$

By the definition of $f_h(r_{h0}, g_h(\vec{r_0}, r^e_{h0}))$ (see Equation 11), $f_h(r_{h0}, g_h(\vec{r_0}, r^e_{h0})) > r^e_{h0}$. Thus, in particular, $r_{h0} > f_h(r_{h0}, g_h(\vec{r_0}, r^e_{h0})) > r^e_{h0}$ 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 $g_h$ $g_h(\vec{r'}, r^e_{h0}) \geq g_h(\vec{r}, r^e_{h0})$ if $\vec{r'} \geq \vec{r}$ for $i = \{1, 2, ..., H\}$. 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., $\vec{r_1} \leq \vec{r_0}$.

Therefore, given any equilibrium $\pi$ and a readmission vector $\vec{r_{\pi 1}}$ that has a positive probability under $\pi$, we have

$$h \in \text{SNPE} => r_{h0} > f_h(r_{h0}, g_h(\vec{r_0}, \vec{r^e_0})) \geq f_h(r_{h0}, g_h(\vec{r_{\pi 1}}, \vec{r^e_0})) \tag{26}$$

Therefore, for SNPE hospitals reducing readmissions is a strictly dominated strategy, and they do not reduce readmission at any equilibrium, i.e.,

$$\forall \pi,\ \forall h \in \text{SNPE},\ \pi_h(r_{h0}) = 1. \tag{27}$$

■

## H. Proof of Proposition 3

In step 2 of the algorithm, $r_{h,n} \neq r_{h,n-1}$ if $r_{h,n-1} > g_h(\vec{r}_{n-1}, r^e_{h,n-1})$ and $h \notin SNPE$. 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 $\vec{r_N}$, 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:

$$h \notin SNPE, \text{ and } r_{h0} > g_h(\vec{r_\pi}, \vec{r^e_0}).$$

By the assumed monotonicity of $g_h$, **if** we could show that $\vec{r_N} \leq \vec{r_\pi}$ for all equilibrium $\pi$, then we would in particular have that a hospital $h$ with $r_{h0} > g_h(\vec{r_\pi}, r^e_{h0})$ also has $r_{h0} > g_h(\vec{r_N}, r^e_{h0})$. 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 $\vec{r_N} \leq \vec{r_\pi}$ for any equilibrium $\pi$. Suppose, to reach a contradiction, that $\exists \pi\ s.t.\ \vec{r_N} > \vec{r_\pi}$. Then there must exist $h$ such that $r_{h,N} > r_{h,\pi}$ and $r_{h0} > g_h(\vec{r_N}, r^e_{h0})$. Indeed, we claim that if every hospital that has $r_{h,N} > r_{h,\pi}$ also has $r_{h0} \leq g_h(\vec{r_N}, r^e_{h0})$ then $r_\pi$ could not be an equilibrium.

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

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

$$r_{h0} \notin [g_h(\vec{r_N}, r^e_{h0}), f(r_{h0}, g_h(\vec{r_T}, \vec{r^e_0}))] \tag{28}$$

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

$$\vec{r_N} \geq \vec{r_\pi} => g_h(\vec{r_N}, r^e_{h0}) \geq g_h(\vec{r_\pi}, r^e_{h0}) => f(r_{h0}, g_h(\vec{r_N}, \vec{r^e_0})) \geq f(r_{h0}, g_h(\vec{r_\pi}, \vec{r^e_0})) \tag{29}$$

Therefore,

$$r_{h0} > f(r_{h0}, g_h(\vec{r_N}, \vec{r^e_0})) => r_{h0} > f(r_{h0}, g_h(\vec{r_\pi}, \vec{r^e_0})) => r_{h0} \notin [g_h(\vec{r_\pi}, \vec{r^e_0}), f(r_{h0}, g_h(\vec{r_\pi}, \vec{r^e_0}))]. \tag{30}$$

Hence it is not optimal for the hospital $h$ to reduce its readmissions. We reach a contradiction to the assumption that $\vec{r_N} > \vec{r_\pi}$. ■

## I. Proof of Proposition 4

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

$$R(P_{cap}, r_{h0}, r_h, r_h^e) = \Pi_h^P(r_h) - \mathbb{P}_h(r_h, r_h^e, P_{cap}) - C(r_{h0}, r_h), \tag{31}$$

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

$$R(P_{cap}^1, r_{h0}, r_1, r_h^e) - R(P_{cap}^1, r_{h0}, r_2, r_h^e) > 0 \Rightarrow R(P_{cap}^2, r_{h0}, r_1, r_h^e) - R(P_{cap}^2, r_{h0}, r_2, r_h^e) > 0 \tag{32}$$

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

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

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

which implies, by the argued supermodularity, that for all $P_{cap} < P_{cap}^{max}$,

$$R(P_{cap}, r_{h0}, r_{h0}, g_h(r_{h,\pi}^T, r_{h0}^e)) > R(P_{cap}, r_{h0}, r, g_h(r_{h,\pi}^T, r_{h0}^e)) \ \forall r < r_{h0}. \tag{34}$$

Moreover, $R(P_{cap}^1, r_{h0}, r_1, r_h^e) - R(P_{cap}^1, x, r_1, r_h^e) > 0$ if $x < r_{h0}$ since the action set under $x$ ($[0, x]$) is a subset of the action set under $r_{h0}$ ($[0, r_{h0}]$). Therefore, with $\vec{r_\pi^t}$ 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:

$$\begin{aligned} \sum_{t=1}^{T} R(P_{cap}^t, r_{h,\pi}^{t-1}, r_\pi^t, g_h(r_\pi^t, r_{h0}^e)) &< \sum_{t=1}^{T} R(P_{cap}^t, r_{h0}, r_\pi^t, g_h(r_\pi^t, r_{h0}^e)) \\ &< \sum_{t=1}^{T} R(P_{cap}^t, r_{h0}, r_{h,0}, g_h(r_\pi^t, r_{h0}^e)). \end{aligned} \tag{35}$$

The first inequality above follows from the restriction to reductions in readmissions ($r_\pi^t \leq r_{h0} \ \forall t$). 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, $\vec{r_\pi^t} \geq \vec{r_N}$ as generated by the algorithm with $P_{cap} = P_{cap}^{max}$. Since $r_\pi^t$ is monotone decreasing in $t$ (because readmissions can only be decreased), it suffices to show that $r_\pi = r_\pi^T \geq \vec{r_N}$ (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 $\vec{r_\pi^T} < \vec{r_N}$, then $\exists h$ such that $r_{h,N} > \vec{r_{h,\pi}^T}$ and, in particular, $r_{h0} > g_h(\vec{r_N}, r_{h0}^e)$. The initial condition of stage $T$ is $r_{h,\pi}^{T-1}$. If we can now show that $r_{h,\pi}^{T-1} \notin$

$[\vec{r_N}, f(r_{h0}, g_h(\vec{r_N}, r^e_{h0}))]$ (which is the analogue of Equation 28) then we can follow the proof in Proposition 4 to reach a contradiction.

To prove that $r^{T-1}_{h,\pi} \notin [g_h(\vec{r_N}, r^e_{h0}), f(r_{h0}, g_h(\vec{r_N}, r^e_{h0}))]$ we use induction. For $n = 1$, this relationship holds trivially since (recall) $R(P_{cap}, r_{h0}, r, r^e_h)$ is super-modular in $(P_{cap}, r)$. Now, let us assume this relation holds for all $k \leq T-1$ and show it prove that $r^T_{h,\pi} \notin [g_h(\vec{r_N}, r^e_{h0}), f(r_{h0}, g_h(\vec{r_N}, r^e_{h0}))]$. If this were not the case then:

$$r^t_{h,\pi} = g_h(\vec{r_\pi}^{\,t}, r^e_{h0}) < f(r_{h0}, g_h(\vec{r_N}, r^e_{h0})) \tag{36}$$

Since $r^{T-1}_{h,\pi} \notin [g_h(\vec{r_N}, r^e_{h0}), f(r_{h0}, g_h(\vec{r_N}, r^e_{h0}))]$, there exist a set of hospitals, $\mathcal{H}$, such that each $h$ in this set has $r^{T-1}_{h,\pi} > f(r_{h0}, g_h(\vec{r_N}, r^e_{h0}))$, and reduce readmissions to to $g_h(\vec{r_\pi}^{\,T}, r^e_{h0})$ in equilibrium. By definition, both $g_1(\vec{r}, r^e_{h0})$ and $g_2(\vec{r}, r^e_{h0})$ are contraction mapping of $\vec{r}$ component-wise, in other words:

$$|g_h(\vec{r}, r^e_{h0})) - g_h(\vec{r'}, r^e_{h0}))| < |r - r'| \tag{37}$$

we know that $\sum_{h\in\mathcal{H}}(r^t_{h,\pi} - r^{t-1}_{h,\pi}) > \sum_{h\in\mathcal{H}}[g_h(\vec{r_\pi}^{\,t}, r^e_{h0}) - g_h(\vec{r_\pi}^{\,t-1}, r^e_{h0})]$. However, based on the optimality condition of the game, $\sum_{h\in\mathcal{H}} r^t_{h,\pi} = \sum_{h\in\mathcal{H}} g_h(\vec{r_\pi}^{\,t}, r^e_{h0})$. This is a contraction, and therefore $r^t_{h,\pi} \notin [g_h(\vec{r_N}, r^e_{h0}), f(r_{h0}, g_h(\vec{r_N}, r^e_{h0}))]$. This concludes the proof. ■

## J. 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 analysis of the equilibrium becomes a static analysis of Hospital 2's optimal decision when the expected readmission rate is $\frac{r_{10}+r_{21}}{2}$ where $r_{21}$ is the readmission decision of Hospital 2.

The objective function of Hospital 2 in this case is:

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

.

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 ($r_1 = (r_{10}, r_{10})$ 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. ■