Hospital Readmissions Reduction Program does not provide the right incentives: Issues and remedies

Kenan Arifoğlu
School of Management, University College London, 1 Canada Square, London E14 5AA, UK
k.arifoglu@ucl.ac.uk

Hang Ren
School of Business, George Mason University, 4400 University Dr, Fairfax, VA 22030, US
hren5@gmu.edu

Tolga Tezcan
London Business School, Regent's Park, London NW1 4SA, UK
ttezcan@london.edu

The Hospital Readmissions Reduction Program (HRRP) reduces Medicare payments to hospitals with higher-than-expected readmission rates where the expected readmission rate for each hospital is determined based on the readmission levels at other hospitals. Although similar relative-performance-based schemes are shown to lead to socially optimal outcomes in other settings (e.g., cost cutting efforts), HRRP differs from these schemes in three respects: (i) deviation from the targets are adjusted using a multiplier; (ii) the total financial penalty for a hospital with higher-than-expected readmission rate is capped; and (iii) hospitals with lower-than-expected readmission rates do not receive bonus payments. We study three regulatory schemes derived from HRRP to determine the impact of each feature, and use a principal-agent model to show that: (i) HRRP over-penalizes hospitals with excess readmissions because of the multiplier and its effect can be substantial; (ii) having a penalty cap can curtail the effect of financial incentives and result in a no-equilibrium outcome when the cap is too low; and (iii) not allowing bonus payments leads to many alternative symmetric equilibria, including one where hospitals exert no effort to reduce readmissions. These results show that HRRP does not provide the right incentives for hospitals to reduce readmissions. Next we show that a bundled payment type reimbursement method, which reimburses hospitals once for each episode of care (including readmissions), leads to socially optimal cost and readmissions reduction efforts. Finally we show that, when delays to accessing care are inevitable, the reimbursement schemes need to provide additional incentives for hospitals to invest sufficiently in capacity.

Key words: Readmissions, healthcare, regulation, information asymmetry, queues, yardstick regulation

1. Introduction

Hospital readmissions have drawn increased attention worldwide as medical professionals and healthcare regulators start to identify the substantial cost of excess readmissions and the potential to reduce healthcare costs and improve the quality of care by eliminating avoidable readmissions. The (unplanned) readmission rates are especially high for Medicare patients in the US; almost one-fifth of beneficiaries are rehospitalized within 30 days of discharge (Jencks et al. 2009) and

they account for over $17 billion in Medicare spending annually. (In Canada, UK and France, studies reported readmission rates between 6–9% for different patient cohorts, see Robinson (2010), Lanièce et al. (2008), Monette (2012).) Although it is not clear what proportion of these readmissions is avoidable (estimates range from 9–59%), a 10% reduction in readmissions would save over one billion dollars annually in the US (MedPAC 2013).

Regulators in different countries introduced payment reforms specifically to incentivize hospitals to reduce avoidable readmissions. The Centers for Medicare & Medicaid Services (CMS) implemented the HRRP in the US in 2012, which reduces the reimbursement to hospitals with excess 30-day readmissions (i.e., those hospitals whose readmission rates are higher than the risk-adjusted national average) for Medicare patients and it is expected to implement similar policies for other healthcare providers (Joynt et al. 2016). The National Health Service (NHS) in the UK implemented a similar policy but the benchmark is set by a local clinic review for each hospital. In Germany hospitals receive a single payment for each episode of treatment, including all potential readmissions following the initial (i.e., index) hospitalization (Kristensen et al. 2015). Despite worldwide attention on hospital readmissions, there is little understanding about the effectiveness of these payment reforms. The goal of this paper is to determine their impact on hospitals' actions and to guide their design using a stylized principal-agent framework.

Agency issues in healthcare regulation: One of the main challenges in healthcare regulation is the intrinsic information asymmetry between the regulator and the hospitals. This is because hospitals (and their employees) are more informed about the condition and the cost of treating patients and about the potential ways to improve their operations.[1] Yardstick regulation (or competition), which ties payments to a hospital with its relative performance compared to others, can help a regulator evade information asymmetry. Specifically, Shleifer (1985) shows that a (benevolent) regulator can induce firms (e.g. hospitals) to exert socially optimal cost-reduction efforts (i.e., maximize total social welfare) by observing the marginal and investment costs of multiple firms from accounting data and then setting the reimbursement level of each firm equal to the average cost of all the other firms (also see Laffont and Tirole (1993), Chp. 1.7). Although firms are assumed to be identical in Shleifer's original model, a risk-adjustment procedure that accounts for observable differences between firms is shown to restore socially optimal cost-reduction efforts if *all differences* can be captured in this procedure. In fact, the prospective payment system (PPS) introduced by CMS in early 1980s (and later adopted in other countries) employs this scheme,

[1] In the absence of information asymmetry regulation is straightforward. A fully informed regulator would easily determine the optimal operating parameters and penalize the hospitals that refuse to operate at these levels. Although most regulators have considerable bargaining power, e.g. the NHS in the UK is the single payer and Medicare and Medicaid revenues constitute over half of the revenues of a typical hospital in the US, such contracts have not been implemented in practice. See (Laffont and Tirole 1993, p. 40) for more details.

along with a risk-adjustment procedure, to determine reimbursement levels based on Diagnosis Related Groups (DRG). However, Shleifer's model, and so intrinsically the PPS, assumes that quality of service is exogenous. In situations where quality (e.g., hospital readmissions) depends on providers' actions, it is not clear how a yardstick competition-type payment scheme can be used to elicit socially optimal effort levels from providers in both cost-reduction and quality improvement, which are inextricably connected.

The goal of HRRP is to induce hospitals to invest in reducing readmissions by adding a yardstick competition type penalty term to the PPS based on hospitals' readmission rates. However, HRRP regulation has three additional unique features:[2] (i) *Multiplier effect:* Deviations from the readmission targets are adjusted using a multiplier. Medical practitioners are concerned that this feature, referred to as the 'multiplier effect', leads to excessive penalties (MedPAC 2013); (ii) *Capped penalty:* The financial penalty imposed on a hospital with higher-than-expected readmission rate is capped, currently at 3% of the total payments a hospital receives from CMS; (iii) *No bonus payments:* Hospitals with a lower-than-expected readmission rates do not receive bonus payments.

It is not difficult to see the rationale behind capping penalties and not allowing bonus payments. Capping penalties limits HRRP's financial impact on hospitals, and not allowing bonus payments reduces the financial burden of HRRP on CMS. The multiplier, on the other hand, amplifies the impact of the financial incentives that relative benchmarking provides (by making the excess readmission penalty amount larger than the reimbursement level) (MedPAC 2013). However, the precise impact of these provisions on hospitals' actions and hence on the equilibrium outcomes (and so on the long-run 'cost') under a yardstick competition type regulation is not known because the readmission targets depends on the actions of all the hospitals. In addition, CMS continues to use these provisions in other initiatives, for example, the Hospital-Acquired Condition Reduction Program does not allow bonus payments, and the Value-Based Purchasing program pays bonuses but the penalties are capped, see Bastani et al. (2016). Hence a thorough understanding of the impact of these three features is critical, not just for HRRP but for other healthcare payment reforms as well.

Analysis of HRRP: In this paper we use a principal-agent model (similar to Shleifer (1985)) to establish hospitals' equilibrium actions under HRRP. We proceed as follows:

- To better understand the impact of each individual element of HRRP, we first remove the cap and allow bonus payments to hospitals with lower-than-expected readmission rates. We show that hospitals over-invest (relative to socially optimal levels) in reducing readmissions because of the

[2] In addition, HRRP monitors multiple (but not all) conditions but imposes the penalty based on the total payments for all conditions. Although this is not the main focus of our paper we show that our main conclusions hold in this case as well, see Remark 1

multiplier effect. We also show that the excess penalty induced by the multiplier is unbounded. We demonstrate that the gap between the total cost of care under HRRP and that under social optimum might be as much as 30% using a numerical study whose parameters are calibrated using actual values from Medicare patient data for three monitored diseases. Interestingly, hospitals under-invest in reducing costs because the marginal benefit of reducing cost is lower than that in social optimum due to lower total demand (including readmissions) in equilibrium.

- Next, we show that, if the adjustments to payments (penalties as well as bonus payments) are capped, there might be no equilibrium (when the cap is too low), as some hospitals may choose to exert no effort in reducing readmissions.
- Finally, under HRRP–when bonus payments are not allowed and penalty is capped–we show that there are multiple (uncountably many to be more specific) equilibria. Although this set may contain the socially optimal readmission rate, a less than fully informed regulator cannot determine the optimal cost and readmission reduction effort levels.

More generally, we show that the theoretical support for the PPS to incentivize hospitals to pick socially optimal cost-cutting efforts *when readmission levels are assumed to be exogenous* (i.e., not affected by hospital actions) does not extend to HRRP when readmission levels are endogenous. Consequently, HRRP is unlikely to yield the desired (i.e., socially optimal) readmission levels in practice.

Our model relies on two important simplifying assumptions. First, we assume that hospitals have ample capacity. We make this assumption because HRRP solely focuses on hospitals' readmissions and does not consider capacity issues (and timely access to healthcare). However, readmission reduction efforts have a direct impact on the effective capacity of hospitals and this impact needs to be taken into account in capacity-constrained settings. We explore the impact of readmission reduction efforts on capacity issues later in the paper (see below). Second, we assume that each disease is monitored separately. Although HRRP uses the same mechanism in our model to calculate the financial penalty for each monitored disease, the cap is applied to the total penalty for all diseases, not separately to each disease. We present an extension to our basic model to demonstrate that our main result holds in this case as well.

Our paper is not the first to point out issues with the HRRP. However the main focus in this stream of literature has been on the insufficient risk adjustment (e.g., HRRP is criticized for not risk-adjusting readmission targets for socio-economic status of patients) used in determining target readmission rates–see, for example Barnett et al. (2015) and Joynt et al. (2016). The detrimental impact of insufficient risk adjustment in yardstick regulation when agents (e.g., hospitals) are heterogeneous has been long well known, see Armstrong and Sappington (2007). In the context of HRRP, Zhang et al. (2016) show (among other results) that hospitals whose readmission rates are

higher than a certain threshold will find it more beneficial to pay the penalty instead of exerting costly efforts to reduce readmissions. They demonstrate that more hospitals would reduce readmissions if the readmission targets were further adjusted using geographical locations of hospitals, indirectly accounting for socio-economic status of patients.

We believe that the more fundamental problem lies in the incentive mechanism HRRP uses. As described above, a relative performance-based reimbursement scheme that yields socially optimal efforts in a setting with homogeneous hospitals can be modified to account for heterogeneity using an accurate risk-adjustment procedure under certain assumptions (Shleifer 1985). However, if an incentive mechanism does not provide the right incentives in a setting with homogenous hospitals to begin with, a risk-adjustment procedure is not going to fix this. Therefore, our results imply that HRRP will not lead to socially optimal outcomes even if CMS improves the risk adjustment procedure and accounts for all the differences between hospitals (assuming it is at all possible).

Socially optimal regulation: We next show that hospitals exert socially optimal efforts to reduce costs and readmissions if bonus payments to hospitals with lower-than-expected readmission rates are allowed, and the multiplier and cap are removed from the HRRP. Surprisingly the resulting reimbursement scheme is similar to the well-known bundled payment scheme. Under this scheme the regulator reimburses hospitals for each episode of treatment once, including readmissions, instead of per-visit reimbursement with a separate financial penalty for excess readmissions (e.g., HRRP). Bundled payment systems have been used for certain conditions in Germany (Kristensen et al. 2015), the Netherlands (Struijs (2015)), and the US (Gruessner 2016). To the best of our knowledge, however, the fact that bundled payments can induce socially optimal cost and readmission reduction efforts has not been established in the literature.

Finally, we consider the impact of readmission reduction regulation on hospitals' capacity investment decisions and in turn on patients' welfare (until this point we assume patients do not experience significant delays to accessing care). Readmission rates have a direct impact on the effective capacity of hospitals since reducing readmissions can free up bottleneck resources. Hence reducing readmissions can also decrease delays to accessing care and increase hospital throughput, boosting hospital revenues as well as patient welfare. We propose a payment scheme that reimburses hospitals based on a bundled payment scheme alongside a yardstick competition type payment adjustment based on each hospital's performance in waiting times. Then we show that hospitals pick costs, readmission rates and capacities at socially optimal levels under this payment scheme. When capacity is limited, excessive delays may force patients to seek treatment through emergency care, and also may have a detrimental impact on their health. We show that the 'cost' of such delays has to be taken into account in determining the (bundled) payment amount. We also demonstrate

that the proposed schemes can be extended to account for heterogeneity and competition between hospitals, and we present additional modeling extensions.

One of the interesting theoretical findings of our paper is that the no-bonus and capped-penalty provisions have significantly different impact on hospital actions when the readmission target and reimbursement levels in HRRP payment scheme (without the multiplier) are set exogenously (at socially optimal levels) as opposed to endogenously (e.g., based on other hospitals' actions as done under HRRP and other payment systems). Specifically, we show that, when targets in the HRRP payment scheme without the multiplier are set exogenously at socially optimal levels, hospitals do choose socially optimal actions, and the capped-penalty (if the cap is not too low) and no-bonus provisions have *no impact* on hospital actions. This may have encouraged CMS to use these provisions in HRRP as well as in other payment schemes.[3] These results are in stark contrast to those under endogenous targets. When targets are set endogenously, these provisions greatly diminish the incentives provided by HRRP's yardstick-based scheme, leading to multiple equilibria. To the best of our knowledge, our study is the first to discover the dramatic impact of these provisions on hospital actions when used in conjunction with endogenous targets.

2. Literature review

Our paper contributes to three streams of literature: (i) analysis of HRRP and bundled payment schemes; (ii) yardstick competition; and (iii) operational impact of payment schemes in healthcare delivery.

HRRP and bundled payments: Hospital readmissions have attracted a lot of attention in the literature (see Burgess and Hockenberry (2014) for historical background and Kristensen et al. (2015) for an international perspective). Our paper is most related to research on consequences of different components of the HRRP scheme, specifically the impact of a cap and no-bonus clause. MedPAC (2013) argues that the current penalty multiplier is too high. Zhang et al. (2016) (see below for a detailed review) shows that the cap is too low to incentivize all hospitals, a conclusion supported also by Bastani et al. (2016) and Aswani et al. (2016). In this paper we characterize *the precise impact* of the multiplier, cap, and no-bonus provisions on hospital actions in equilibrium.

Zhang et al. (2016) is perhaps the most related to our paper in this stream. Although they do not consider incentive mechanism design questions or whether HRRP attains social optimum, they do analyze the impact of different features of HRRP using analytical modeling and data analysis. Taking the HRRP scheme as a given, they analyze the readmission reduction efforts of hospitals with different characteristics and show that certain hospitals may choose to pay the

[3] Also, when the penalty cap is too low, hospitals exert no effort in reducing readmissions. Thus, upon observing hospitals' actions, a regulator can progressively increase the cap if it is regarded too low.

penalty instead of reducing readmissions, thus demonstrating that HRRP is ineffective for these hospitals. They show that one of the main reasons behind this phenomenon is the fact that the readmission benchmarks for these hospitals are too low because HRRP does not incorporate the socio-economic status of the local population of a hospital in risk adjustment. Using hospital data from the US for fiscal year 2013, they show that benchmarking hospitals locally, which indirectly captures the impact of socio-economic status of the patients in their locality, improves the results. We instead focus on the incentives provided by HRRP in a principal-agent framework and assume that hospitals are identical to alienate the impact of risk adjustment. We show that HRRP will not lead to desired readmissions reduction efforts, even in this 'ideal' setting, hence demonstrating that HRRP is unlikely to be effective, even if all the issues surrounding risk-adjustment are addressed. Zhang et al. (2016) also study how different characteristics of hospitals, changing the penalty cap, and the procedure to determine target readmissions in HRRP, may impact readmission reduction efforts based on a structural model calibrated using the aforementioned data. We, on the other hand, precisely determine the impact of different provisions (i.e. the multiplier, cap and no-bonus) of HRRP on hospitals' actions, and then show that there are other payment schemes that lead to socially optimal readmission reduction efforts even when the capacity is limited.

There is an increasing interest in understanding the benefits and risks associated with bundled payments relative to other payment schemes–see Struijs (2015) and Porter and Kaplan (2016), more so in the US as CMS rolls out different bundled payment schemes–see Mechanic and Tompkins (2012). Our paper contributes to this literature in two ways. We show that (i) bundled payments provide the right incentives for hospitals to reduce readmissions, and (ii) in capacity-constrained settings, the cost of patients' inability to access healthcare in a timely manner needs to be taken into account while determining the reimbursement amounts (currently CMS uses different approaches, see Gupta and Mehrotra (2015), Baggot and Edeburn (2015)).

Yardstick competition: Yardstick competition is first analyzed in Shleifer (1985) in a setting where the main focus is on cost reduction. It was implemented in the prospective reimbursement schemes in healthcare (Pope 1989), in addition to many other industries where local monopolies arise naturally. Recent extensions most related to our research focus on models to capture quality improvement efforts while containing costs, see Ma (1994), Chalkley and Malcomson (1998), Tangerås (2009) and references therein, mostly focusing on adverse-selection issues. Savva et al. (2018) proposes a payment scheme based on yardstick regulation to incentivize hospitals to reduce waiting times. Part of the reimbursement system we propose in the capacity-constrained setting is based on that scheme. Our work contributes to this literature by establishing the equilibrium outcomes when bonus payments are not allowed and penalty is capped in yardstick competition, using a model inspired by Shleifer (1985).

Operational impact of payment schemes: A stream of literature focuses on the operational impact of different healthcare regulations by explicitly modeling hospital actions. Adida et al. (2016) compares fee-for-service with bundled payments in a model where hospitals choose the treatment intensity for heterogeneous patients. Zorc et al. (2017) considers contracting issues in a setting where patients are treated by a general practitioner and a specialist, and their decisions jointly affect patients' health status. Bastani et al. (2016) studies incentive schemes that have small reward and/or penalty terms to model the current practices of CMS. Andritsos and Tang (2018) compares fee-for-service and bundled payment schemes in a model with readmissions. Guo et al. (2019) extends this analysis by modeling the capacity of hospitals and its impact on waiting time of patients explicitly. Adida and Bravo (2019) studies reimbursement contracts between a managing organization, which provides basic care, and an external provider, which provides advanced care and which is reimbursed by the former. They show that a penalty-only contract (without bonus payments) can elicit system-wide or socially optimal effort levels for successful treatment, for example, by reducing readmissions. (See also So and Tang (2000), Ata et al. (2013), Gupta and Mehrotra (2015), Jiang et al. (2012), Bavafa et al. (2017) for studies that focus on regulation on other healthcare settings, and Batt et al. (2018a) and Chen and Savva (2018) for empirical analysis of the impact of HRRP on hospital actions.) Typically, papers in this stream assume that the reimbursement amounts are either exogenous or chosen optimally under varying assumptions about available information at regulator's disposal. Although Zhang et al. (2016) study endogenous readmission targets, they take the reimbursement levels as exogenous and establish a lower bound on the number of hospitals that would not exert any readmission reduction effort (primarily due to the fact that their initial readmission levels are too high, as explained above). To better explore the impact of payment systems used in practice and to capture the fact that the regulator has limited information about hospitals' cost parameters, we focus on PPS (and HRRP) where the reimbursement amount as well as the readmission target are endogenous. In addition, we propose socially optimal payment schemes that have the same information burden on the regulator as the PPS and HRRP, a research direction that is not explored in Zhang et al. (2016).

3. Model

We focus on the treatment of a single condition and consider a model (similar to Shleifer (1985), Savva et al. (2018), Tangerås (2009)) with three parties: (i) patients; (ii) hospitals/providers; and (iii) the regulator (see Remark 1 for an extension to multiple diseases). It is assumed that each hospital is a monopoly in its catchment area and provides treatment to a fixed population. With the objective of maximizing total welfare, the regulator sets the terms of the reimbursement scheme that dictates how providers are reimbursed for providing treatment to patients. Informed with

the terms of the reimbursement scheme, each provider chooses its marginal cost and readmission rate. In order to determine the effectiveness of a reimbursement scheme, we analyze the actions of providers in equilibrium of a (Stackelberg) game where the regulator moves first and each entity is self-interested. Specifically, we assume that the regulator commits to the reimbursement scheme and we establish the (pure-strategy) Nash equilibrium in a one-round game in which each player holds correct expectations about the other players' actions, with the aim of assessing the long-term impact of reimbursement schemes (see, Fudenberg and Tirole 1991, Osborne and Rubinstein 1994).

3.1. Interactions between patients, providers and regulator

Patients: We assume that patients arrive to each hospital at a fixed rate λ and that they need to be admitted to the hospital for treatment. (We consider a case with demand dependent on hospitals' actions in §5.2.) A patient's index hospitalization may be followed by a hospital readmission to (successfully) complete the treatment and, for notational simplicity, we assume that each patient can be readmitted (at most) once, an assumption we relax in §5.3.

Providers: We consider N providers and assume that each provider operates as a local monopoly in its catchment area (see §5.3 for an extension) and we consider two decisions by each provider: (i) readmission rate (or the probability that a patient is readmitted) r; and (ii) marginal cost of treating a patient c. More specifically, c denotes the cost of treating a patient each time the patient seeks treatment. We assume that the cost of treatment c is the same for index hospitalizations and readmissions (see §5.3 for an extension), and it does not depend on r. Each hospital has an initial constant marginal cost c_{max} and readmission level r_{max} and can reduce the marginal cost to c and readmission level to r by spending $R(r, c)$. This represents the fixed cost of all activities undertaken by a hospital to reduce the readmission rate to r and the marginal cost to c.[4]

In practice, hospitals can adopt various interventions aimed at reducing readmissions, such as patient education, better discharge planning, and better co-ordination of clinical intervention with community and social care providers (see Hansen et al. (2011) for more on readmission-reduction interventions). They can also improve the cost efficiency of treatment by purchasing new equipment that allows for more precise and faster diagnosis/treatment, employing more and better-qualified staff, improving staff training programs, and/or process re-engineering among others. These interventions are typically costly since they require investment in facilities, labor, staff/patient education, and so on.[5]

[4] There are alternative ways to model hospital actions that result in similar insights. For example hospital actions can be modeled using a fixed (investment) cost F and a variable cost c, which in turn jointly determine the readmission levels. Under mild conditions, such a model is equivalent to our model.

[5] We only consider the impact of hospital actions on cost of treatment and readmissions but not on, for example,

A hospital's objective, Π, hence can be written as

$$\Pi(r,c) = T - c(1+r)\lambda - R(r,c), \tag{1}$$

where T denotes the payment from the regulator to the provider (see below for details). To avoid technical subtleties, we assume that $r \in [r_{min}, r_{max}]$ and $c \in [c_{min}, c_{max}]$, for some $0 < r_{min} \leq r_{max} < 1$ and $0 \leq c_{min} < c_{max} < \infty$.

Regulator: The regulator's objective is to maximize the total social welfare S (i.e., total patient surplus minus total cost), given by

$$S(r,c) = V(\lambda) - c(1+r)\lambda - R(r,c), \tag{2}$$

subject to the constraint that providers break even, i.e., $\Pi(r,c) \geq 0$. Here $V(\lambda)$ denotes the total utility of patients from receiving treatment and the last two terms constitute the total cost of providing treatment. Notice that T is absent from this expression since it is a transfer payment from the regulator (financed typically indirectly by patients via taxes or insurance premiums) to the providers. Also because $V(\lambda)$ is a constant, the objective of the regulator can be viewed as minimizing the total cost of providing care (to a fixed population). We keep $V(\lambda)$ in our model because of the extensions we consider later in the paper where its value depends on hospitals' actions.

Assumptions and preliminaries: Throughout, we make the following standard technical assumptions that guarantee that the regulator and the providers have unique optimal actions that can be determined using first order conditions (FOCs). We assume that $R(r,c)$ is decreasing and jointly convex in r and c, and is twice differentiable. Also we assume that $R_{rr}R_{cc} > (R_{rc} + \lambda)^2$, i.e., R is sufficiently convex[6] (where R_x denotes partial derivative with respect to variable x), and $R_{cr} \geqslant 0$, i.e., readmission reduction is more costly when the treatment cost is low. We impose additional boundary conditions to ensure that optimal actions are interior. These conditions are presented in Appendix A for the sake of brevity. We use these assumptions throughout the paper without further mention.

Under these assumptions, the socially optimal (or first-best) readmission rate and treatment cost, denoted by (r^*, c^*), are unique and can be characterized using the FOCs.

other medical outcomes. We do this because HRRP does not take any other measures, besides readmission rates, into account in calculating penalties. However, CMS introduced Hospital Value-Based Purchasing and Hospital-Acquired Condition Reduction Programs that tie hospital payments to various additional medical outcomes. These programs should encourage hospitals to take actions that increase the prospect of good medical outcomes while reducing readmissions.

[6] This assumption is just a sufficient condition that ensures that objective functions of the regulator and hospitals are unimodal (i.e. have unique maximums). There are other sufficient conditions ensuring unimodality of all objective functions. For example, it can be shown that conditions in Lemma A8 of Appendix F are sufficient to ensure the regulator's objective function $S(r,c)$ defined in (2) to be unimodal.

Lemma 1 (First-best benchmark). *The regulator's objective in* (2) *has a unique maximizer* (r^*, c^*)*, where* $r^* \in (r_{min}, r_{max})$ *and* $c^* \in (c_{min}, c_{max})$ *satisfy FOCs*

$$\lambda c^* + R_r(r^*, c^*) = 0, \tag{3}$$

$$\lambda(1 + r^*) + R_c(r^*, c^*) = 0. \tag{4}$$

We note that the information asymmetry in this setting is embodied in the inability of the regulator to fully estimate the cost function R. A fully informed regulator could easily determine the first-best cost and readmission levels from (3) and (4) (assuming that the demand rate is observable, which is usually the case in practice), and then penalize those hospitals that refuse to operate at these levels. However, due to this information asymmetry, the regulator is compelled to use yardstick competition type schemes, e.g., the PPS and HRRP, that only utilize available information. Specifically, under these payment systems, the regulator only needs to observe the marginal treatment cost, readmission rate and investment cost of each hospital from accounting data. This information is already collected by CMS to determine the DRG reimbursement levels as well as the penalties under HRRP.

For the rest of the paper we use $h(r)$ to denote the solution of (4) in c for fixed r, that is

$$\lambda(1 + r) + R_c(r, h(r)) = 0. \tag{5}$$

By the assumptions above, h is well defined and is decreasing in r, for $r \in [r_{min}, r_{max})$, see Lemma A1 in Appendix B.

3.2. HRRP payment scheme

We use a model similar to Shleifer (1985) to model the HRRP reimbursement scheme. Because we focus on a single disease, the payment scheme we consider is slightly different from the payment scheme CMS uses in HRRP. We discuss the differences and how our model can be extended to incorporate these differences in Remark 1 below, after we explain our basic model.

First the regulator observes the marginal cost c_i, the readmission rate r_i, and the investment cost $R_i \equiv R(r_i, c_i)$ for each hospital and sets

$$\bar{c}_i = \frac{1}{N-1}\sum_{k \neq i} c_k, \quad \bar{r}_i = \frac{1}{N-1}\sum_{k \neq i} r_k, \text{ and } \bar{R}_i = \frac{1}{N-1}\sum_{k \neq i} R_k, \text{ for } i = 1, \ldots, N. \tag{6}$$

Then the transfer payment T_i to hospital i is given by

$$T_i = \bar{c}_i\lambda(1 + r_i) - \underbrace{\left(\min\left\{\frac{r_i - \bar{r}_i}{\bar{r}_i}, P_{cap}\right\}\right)^+ \bar{c}_i(1 + r_i)\lambda}_{\pi(r_i|\bar{r}_i, \bar{c}_i)} + \bar{R}_i, \tag{7}$$

where $x^+ = max\{x, 0\}$. We refer to this reimbursement scheme as 'HRRP reimbursement scheme' for the rest of the paper, or just HRRP for simplicity, with a slight abuse of terminology. The HRRP consists of three components: (i) a payment equal to $\bar{c}_i$ per patient; (ii) a penalty based on the hospital's relative performance on readmission rates (term π); and (iii) a lump-sum payment $\bar{R}_i$ to help recover the hospital's investment cost. Under this reimbursement scheme the hospital's i objective can be written, from (1) and (7), as

$$\Pi(r_i, c_i) = (\bar{c}_i - c_i)(1 + r_i)\lambda - \pi(r_i|\bar{r}_i, \bar{c}_i) + \bar{R}_i - R(r_i, c_i). \tag{8}$$

We note that CMS does not make lump-sum payments but makes per-discharge capital payments that cover costs for depreciation, interest, rent and tax related costs. However, the lump-sum payment $\bar{R}_i$ in our model can be instead made per-discharge by dividing the total payment amount by the demand, i.e., $\bar{R}_i/(1 + r_i)\lambda$. In addition, the readmission penalty in our model is based only on the total per-patient payments but not on the payments to cover the cost and readmission reduction efforts (i.e., $\bar{R}_i$,). This is in line with CMS's practice because the penalties under HRRP are only applied to operating payments (that cover labor and supply costs to treat a patient) but not to capital payments (CMS 2019).

Differences between PPS and HRRP: Shleifer (1985) models the PPS using (7) without the readmission penalty term π and shows that this scheme (referred to as PPS for simplicity from here on) leads to socially optimal cost-reduction efforts when the readmission rates are exogenous.[7] The PPS's capability to induce socially optimal efforts (when readmission levels are assumed to be exogenous) is based on the indirect cost competition it induces between hospitals that otherwise operate in different markets in that those hospitals that are more cost-efficient than the other hospitals, i.e., $c_i < \bar{c}_i$, will enjoy positive profits and others will suffer losses.

Under the PPS, hospitals have no incentive to reduce readmissions when readmission levels are assumed to be endogenous (in fact we show below that they do not exert any effort at all in equilibrium–see Proposition 3) and the novelty of HRRP is the introduction of the additional term π to the PPS reimbursement scheme in order to provide this missing incentive. This penalty term induces an indirect competition between hospitals in reducing readmissions (in a way similar to what the PPS does for cost-reduction efforts) by comparing provider i's readmission rate, r_i, to its expected (yardstick) readmission rate $\bar{r}_i$. Clearly, if hospital i's readmission rate is higher than $\bar{r}_i$, then it pays a penalty, proportional to how much worse its performance relative to its target is (i.e., $r_i/\bar{r}_i$) and to the total amount of per patient reimbursement it receives (i.e., $\bar{c}_i(1 + r_i)$).

[7] In Shleifer (1985)'s original model, patients bear the cost of treatment and the total demand is price sensitive. However his result can easily be extended to our model where we assume that insurance (e.g., Medicare) covers the cost of treatment, excluding a potential co-payment independent from hospital's readmission rate and marginal cost.

As discussed above, the HRRP readmission penalty mechanism has three features that the PPS does not share. Having presented the details of our model, we are in a good position to explain these differences in detail using the definition of this reimbursement schemes in our setting. (i) *Multiplier effect:* The deviation from the target is normalized using the average readmission rates, term $((1+r_i)/\bar{r}_i)\,(r_i-\bar{r}_i)$ in (7), in penalty calculations under HRRP, whereas under the PPS, the cost-reduction incentive is provided by the absolute deviation from the target, term $(\bar{c}_i - c_i)$ in (8). (ii) *No bonus payments:* HRRP is a penalty-only scheme (note that $\pi = 0$ if $r_i < \bar{r}_i$), and so hospitals with lower-than-expected readmission rates do not receive bonus payments, whereas under the PPS hospitals do enjoy profits if they are more cost efficient. (iii) *Capped penalty:* The penalty is capped, denoted by $P_{cap}(\geq 0)$ in (7), under HRRP, limiting the potential financial burden of the scheme on hospitals, but no such cap exists under the PPS on profits or losses. Hence, although the PPS is shown to lead to socially optimal cost-reduction efforts when readmission levels are assumed to be exogenous and HRRP uses a similar incentive mechanism, it is not clear what the impact of these additional features on hospitals' efforts on cost and readmission reduction is when readmission levels are assumed to be endogenous. We characterize their precise impact in the next section.

Remark 1 (Modeling assumptions). We note that the reimbursement scheme we use is simpler than the actual implementation of the HRRP[8] in that HRRP monitors multiple diseases (six as of 2017), imposes a penalty as a proportion of the total payments from CMS (including payments for unmonitored diseases), and the total penalty is capped–see, for example, Zhang et al. (2016). Our main result can be extended to a model that incorporates multiple diseases. The details are presented in Appendix D. Also, if the financial penalty is imposed on total payments (as opposed to payments for the monitored disease as in (7)), then this can be incorporated in our model by increasing the cap (P_{cap}), assuming that the total reimbursement for monitored diseases is a constant fraction of the total payments received from the regulator, see (4) in Zhang et al. (2016).

Using a single disease model allows us to derive clean insights from the analysis. Also by focusing on identical hospitals, we remove the impact of the HRRP's risk-adjustment procedure on hospital's actions, which is studied widely elsewhere. Finally, when hospitals are heterogeneous, Shleifer (1985) shows that the PPS yields socially optimal cost-reduction efforts (when the readmission rates are exogenous) with an accurate risk adjustment procedure and a similar result holds in our setting as well once the issues with incentive mechanism are addressed, see §5.3 for more details.

[8] We delineate the procedure CMS uses to determine reimbursement amounts under Inpatient Prospective Payment System and penalties under HRRP in Appendix I. In addition, we describe the connection between CMS' procedures and our model along with the implications of recent changes in HRRP for our results.

4. Analysis of HRRP

In this section we analyze the equilibrium outcomes under the HRRP reimbursement scheme using the principal-agent framework described above. To precisely determine the impact of each of the three provisions (i.e., the multiplier, no-bonus, and penalty cap) we analyze two additional reimbursement schemes obtained from HRRP by altering term π in (7). Specifically we establish the equilibrium under the following cases. (i) We allow bonus payments for hospitals with lower-than-expected readmissions and remove the cap on financial incentives, which enables us to isolate the impact of the multiplier effect. We refer to this scheme as HRRP-I. (ii) Next we re-introduce the cap but still allow hospitals to receive bonus payments. Combined with the analysis of HRRP-I, this enables us to study the impact of the cap. We refer to this scheme as HRRP-II. (iii) Finally we consider the original HRRP scheme defined in (7) where hospitals do not receive bonus payments and the penalty is capped.

4.1. Equilibrium under HRRP-I scheme

When there is no cap on financial incentives (i.e., setting $P_{cap} = \infty$ in (7)) and bonus payments are allowed in HRRP, the transfer payment to hospital i becomes

$$T_i = \bar{c}_i\lambda(1+r_i) - \underbrace{\frac{1}{\bar{r}_i}(r_i - \bar{r}_i)(1+r_i)\bar{c}_i\lambda}_{\pi^{\mathrm{I}}(r_i|\bar{r}_i,\bar{c}_i)} + \bar{R}_i. \tag{9}$$

The next proposition characterizes the equilibrium outcomes under this scheme, referred to as HRRP-I.

Proposition 1 (Equilibrium under HRRP-I). *Under the reimbursement scheme HRRP-I there exists at least one symmetric equilibrium $(\tilde{r}, \tilde{c})$, and all symmetric equilibria satisfy $\tilde{r} < r^*$, $\tilde{c} > c^*$. Furthermore, if there are only two hospitals, i.e., $N = 2$, there cannot be any asymmetric equilibrium.*

This result confirms the argued impact of the multiplier in the medical literature and shows that hospitals over-invest in readmission reduction efforts in symmetric equilibrium, since $\tilde{r} < r^*$ in all symmetric equilibria. In addition, over-investment in readmission reduction leads to under-investment in cost-reduction efforts (i.e., $\tilde{c} > c^*$), relative to first-best, because in equilibrium the total number of admissions (index admissions plus readmissions) for each hospital is lower than what it would have been under the first-best readmission levels, and so the total benefit of reducing marginal cost is lower.

To demonstrate the multiplier effect more precisely, consider the relative performance-based penalty under HRRP-I, term π^{I} in (9). For simplicity assume that $\bar{r}_i = r^*$. By (9) hospital i with

readmission rate r_i will incur a penalty equal to $\frac{(1+r_i)}{\bar{r}_i}(r_i - \bar{r}_i)\bar{c}_i\lambda$, assuming $r_i > \bar{r}_i$. However, the cost of excessive readmissions to the regulator is only $\bar{c}_i\lambda(r_i - \bar{r}_i)$ (see (2)) and so the readmission penalty is $\frac{(1+r_i)}{\bar{r}_i}$ times higher. We refer to this quantity as the multiplier from here on. We note that the multiplier is always greater than 3 if $r_{max} \leq 0.5$ (a range that covers the current readmission rates for most diseases), and, for any fixed $r_i/\bar{r}_i$, it is increasing in $\bar{r}_i$ and goes to *infinity* as $\bar{r}_i \to 0$. Therefore, penalties imposed in practice could be excessive if there is no cap (especially if $\bar{r}_i$ is low) because of the multiplier. In the presence of a penalty cap, most hospitals may have to pay the maximum penalty, even when they deviate slightly from their readmission target. We determine the precise impact of the multiplier on hospitals' actions and on treatment costs using parameters estimated from Medicare patient data for three monitored diseases in §4.1.1 below.

Remark 2 (Asymmetric equilibria). Our result does not rule out the possibility of asymmetric equilibria in general, which is common in such settings–see for example, Shleifer (1985). We do not focus on asymmetric equilibrium for three main reasons. First, it is possible to slightly alter the reimbursement scheme to rule out the possibility of asymmetric equilibrium by dividing the hospitals into two groups and setting the target readmission level of one group using the average performance of the hospitals in the other group and vice versa (proof is similar to that in EC4 of Savva et al. (2018) by using the fact that there are no asymmetric equilibrium outcomes with two hospitals). Second, in our model all hospitals are identical, hence the readmission reduction effort levels in an asymmetric equilibrium cannot be socially optimal. Third, we show that there exists no asymmetric equilibrium under HRRP if R satisfies an additional mild condition, see Proposition 3 below.

In the next two sections we establish the equilibrium outcomes under HRRP-II and HRRP, and the results depend on those under HRRP-I. We make the following assumption for the rest of the analysis for simplicity (we provide sufficient conditions for this assumption to hold in Appendix F).

Assumption 1. *Under HRRP-I there is a unique symmetric equilibrium* $(\tilde{r}, \tilde{c})$.

It is possible to extend the subsequent analysis to cases with multiple symmetric equilibria, see Appendix G.

4.1.1. Numerical examples Although Proposition 1 and the subsequent discussion show that the multiplier can have a substantial impact on financial incentives, its impact on ensuing equilibria is less clear. In this section we demonstrate the multiplier's impact on hospital actions and the total cost of care using a simple example. We use the following investment cost function, which allows us to determine first-best and equilibrium outcomes in closed forms,

$$R(r, c) = \frac{\tau_1\lambda}{r} + \frac{\tau_2\lambda}{c}, \tag{10}$$

for constants $\tau_1, \tau_2 > 0$. By Lemma 1 socially optimal actions satisfy

$$r^* = \sqrt{\tau_1/c^*} \text{ and } c^* = \sqrt{\tau_2/(1+r^*)} \tag{11}$$

and equilibrium actions under HRRP-I satisfy

$$\tilde{r} = \sqrt{\tilde{r}/(1+\tilde{r})}\sqrt{\tau_1/\tilde{c}} \text{ and } \tilde{c} = \sqrt{\tau_2/(1+\tilde{r})}. \tag{12}$$

(Results (12) follow from (A17) in the proof of Proposition 1 assuming FOCs are necessary and sufficient to determine a hospital's optimal choices.) By (11) and (12) (after some algebra), if $r^* \leq 0.5$

$$\tilde{r}/r^* \leq 70\%, \tag{13}$$

that is, the equilibrium readmission rates under HRRP-I is at least 30% less than socially optimal rates.

To demonstrate the impact of the multiplier more precisely, we present the results of a set of numerical experiments where we use the cost and readmission rate values for three HRRP monitored diseases: acute myocardial infarction (AMI), heart failure (HF) and pneumonia (PN) with complications and co-morbidities, from CMS data in 2013 to calibrate the parameters of our cost model as follows. First, for each disease, we assume that r^* and c^* are around 90% of the current respective national average rates and then determine τ_1 and τ_2 using (11). Finally we find $\tilde{r}$ and $\tilde{c}$ using (12). (For all numerical examples here, we verified that the convexity assumption $R_{rr}R_{cc} > (R_{rc}+\lambda)^2$ and all assumptions in Appendix A are satisfied and that there is a unique symmetric equilibrium by Lemma A9(ii) in Appendix F.) Socially optimal and equilibrium actions, along with the increase in the total cost of care when hospitals pick equilibrium actions instead of socially optimal actions, are presented in Table 1.

Disease	c^*	r^*	$\tilde{c}$	$\tilde{r}$	Increase in cost
AMI	5900	18%	6316.9	2.9%	30%
HF	5300	22%	5732.4	4.3%	27%
PN	5000	16%	5323.0	2.4%	32%

Table 1 First-best and equilibrium actions, and increase in cost due to deviation from first-best in three settings driven from average costs and readmission rates for HRRP monitored diseases.

Clearly the multiplier effect is substantial in these cases (much more than the bound in (13) estimates); $\tilde{r}$ is around 20% of r^* on average. In addition, the total cost of care increases by almost 30% on average because of suboptimal hospital actions in equilibrium. Interestingly the deviation in marginal cost is much lower; on average $\tilde{c}/c^* \approx 107\%$. This follows from the fact that the impact of readmission rates is limited in (11) and (12) and one can show that $\tilde{c}$ is at most 22.5% higher than c^* for $r^* \leq 0.5$.

4.2. Equilibrium under HRRP-II scheme

Next we consider a reimbursement scheme obtained from HRRP-I by imposing a cap, equal to $P_{cap} \geq 0$, on rewards and penalties. Consider the following reimbursement scheme, referred to as HRRP-II, with the transfer payment for hospital i equal to

$$T_i = \bar{c}_i\lambda(1+r_i) - \pi^{\text{II}}(r_i|\bar{r}_i,\bar{c}_i) + \bar{R}_i, \tag{14}$$

where

$$\pi^{\text{II}}(r_i|\bar{r}_i,\bar{c}_i) = \begin{cases} \min\left[\frac{1}{\bar{r}_i}(r_i-\bar{r}_i), P_{cap}\right](1+r_i)\bar{c}_i\lambda, \text{ if } r_i \geq \bar{r}_i, \\ -\min\left[\frac{1}{\bar{r}_i}(\bar{r}_i-r_i), P_{cap}\right](1+r_i)\bar{c}_i\lambda, \text{ if } r_i < \bar{r}_i. \end{cases}$$

Clearly this reimbursement scheme can also be derived from the HRRP scheme by adding capped bonus payments when $r_i < \bar{r}_i$ in (7). The next proposition characterizes the equilibrium outcomes.

Proposition 2 (Equilibrium under HRRP-II). *Suppose Assumption 1 holds. Under HRRP-II scheme:*

(i) There exists $\bar{P}_{cap} > 0$ such that $(\tilde{r},\tilde{c})$ is the unique symmetric equilibrium if $P_{cap} \geq \bar{P}_{cap}$.

(ii) No symmetric equilibrium exists if $0 < P_{cap} < \bar{P}_{cap}$.

(iii) $(r_{max}, h(r_{max}))$ is the unique symmetric equilibrium if $P_{cap} = 0$.

This proposition shows that, if the cap is low enough, it can have a drastic impact on hospitals' actions, leading to no symmetric equilibrium and nullifying the effect of readmission reduction financial incentive scheme. Interestingly there is a phase transition and the cap has no impact if it is larger than a certain value (i.e., the equilibrium is identical to the case with no cap).

To explain the intuition behind this result, and why $\tilde{r}$ may no longer be an equilibrium outcome, consider a scenario with just two hospitals. If hospital 2 chooses readmission level $\tilde{r}$, it might be more profitable for hospital 1 not to exert any readmission reduction effort if the cap is low enough, because the readmission penalty (term $\pi^{\text{II}}(r_i|\bar{r}_i,\bar{c}_i)$) is bounded. Observing this, hospital 2 then can re-choose a readmission level slightly below hospital 1 and obtain positive profits, but hospital 1 would lose money in that case, and would choose to change its action. Hence there might not exist a symmetric equilibrium–again we cannot rule out the possibility of asymmetric equilibria.

Finally, Proposition 2(iii) shows that, if the penalty cap is zero, hospitals will exert no readmission reduction effort in equilibrium. Note that in this case HRRP-II reimbursement scheme is equivalent to the PPS by (7) and (14) (because $\pi \equiv 0$ in (7) when $P_{cap} = 0$), thus verifying the necessity of introducing additional incentives to the PPS to reduce readmissions.

4.3. Equilibrium under HRRP scheme

Finally we establish the equilibrium outcomes under the HRRP scheme defined in (7), where hospitals with lower-than-expected readmission rates do not receive bonus payments and the penalty is capped.

We need to introduce additional terminology to specify the equilibrium outcomes in this case. First, let

$$r_e = r_{max}/(P_{cap}+1) \tag{15}$$

and $r_p = \max\{r_e, \tilde{r}\}$. Intuitively, r_e is a threshold on the readmission target beyond which the financial cap has no impact. Hence, a hospital's equilibrium actions will be different depending on their target being above or below this level. To specify the equilibrium outcomes we set

$$\mathcal{S}_p = \{(r, h(r)) : r \in [r_p, r_{max}]\} \tag{16}$$

and note that, if $P_{cap} > 0$, $r_p < r_{max}$ by Proposition 1 and so $\mathcal{S}_p$ is non-empty. Finally let $\mathcal{S} = \{(r, h(r)) : r \in [\tilde{r}, r_p]\}$ and $P_{max} = \frac{r_{max}}{\tilde{r}} - 1$. We have the following result.

Proposition 3 (Equilibrium under HRRP). *Suppose Assumption 1 holds. The following hold under the HRRP scheme:*

(i) For any $P_{cap} \geq 0$, there exists $\mathcal{S}_o \subset \mathcal{S}$ (depending on P_{cap}), such that any $(r, c) \in \mathcal{S}_o \cup \mathcal{S}_p$ is a symmetric equilibrium and there is no other symmetric equilibrium.

(ii) There exists $\bar{\bar{P}}_{cap} \in (0, P_{max})$ such that for any $(r, c) \in S_o \cup \mathcal{S}_p$, $r > r^*$ and $c < c^*$ for $P_{cap} < \bar{\bar{P}}_{cap}$.

(iii) If $P_{cap} \geq P_{max}$, then any

$$(r, c) \in \{(r, h(r)) : r \in [\tilde{r}, r_{max}]\} \, (= S \cup \mathcal{S}_p)$$

is a symmetric equilibrium.

(iv) There is no asymmetric equilibrium if $dR(r, h(r))/dr < 0$ for all $r \in [r_{min}, r_{max}]$ (even when Assumption 1 does not hold).

Proposition 3 shows that removing bonus payments has a non-trivial impact on hospitals' actions as any point in a set (namely, $\mathcal{S}_o \cup \mathcal{S}_p$) with uncountably many outcomes, which include no effort levels (i.e., (r_{max}, c_{max})) and might include the first-best outcomes. In addition, parts (i)–(iii) together establish that the equilibrium set shrinks as the penalty cap becomes smaller. More specifically, part (ii) shows that if P_{cap} is small enough, hospitals always underinvest in readmission reduction efforts, and part (iii) shows that, if P_{cap} is large enough, any $r \geq \tilde{r}$ can be an equilibrium. Note that part (iii) includes the case when there is no cap (i.e., $P_{cap} = \infty$) and therefore establishes the impact of removing bonus payments from HRRP-I. Part (iv) shows that no asymmetric equilibrium exists if the cost function satisfies a mild technical condition and this result does not require

Assumption 1. In addition, it can be shown that $dR(r,h(r))/dr < 0$ if $R_{cc} > R_c(\lambda + R_{cr})/R_r$, for $r \in [r_{min}, r_{max}]$, i.e., if R is "sufficiently" convex in c for all $r \in [r_{min}, r_{max}]$.

Propositions 1–3 together show that removing bonus payments has a very different impact on equilibrium than imposing caps. Specifically introducing a cap will never increase the set of potential equilibria, unlike removing bonus payments which increases the number of equilibrium points to uncountably many.

To demonstrate the impact of removing bonus payments and explain the intuition behind the proof Proposition 3, assume that $P_{cap} = \infty$ (part (iii) in Proposition 3) and again consider a case with two hospitals. If hospital 2 chooses a readmission level higher than $\tilde{r}$ then hospital 1 would not pick a lower readmission level because it will not receive a bonus payment for performing better (unlike the case under HRRP-I). Also hospital 1's cost is increasing in r (due to concavity of the hospital's objective function, and because $r \geq \tilde{r}$) hence it will not choose a readmission level higher than hospital 2. As a result, any $r \geq \tilde{r}$ will be an equilibrium. Proof of part (ii) of Proposition 3 follows from the fact that, if $P_{cap} < \infty$, it may be optimal for hospital 1 to exert no effort when hospital 2 picks a low enough readmission level, in a way similar to the case under HRRP-II. Hence, for low enough P_{cap}, only $r > r^*$ can be an equilibrium.

Remark 3 (Mixed-strategy equilibrium). Throughout this section we only focused on pure-strategy equilibrium, however, there exists at least one mixed-strategy equilibrium under HRRP-II and there may exist additional equilibria under HRRP payment schemes by Theorem 12.4 in Fudenberg and Tirole (1991). Nevertheless, it is difficult to anticipate how a mixed-strategy equilibrium can be implemented by the hospitals in the current context. First, determining mixed-strategy equilibrium proved to be elusive, even in our stylised model. Hence it is unlikely that hospitals can even identify these strategies in practice. Second, various interpretations of mixed equilibrium do not apply to rational hospitals making long-term investment decisions in cost and readmission reduction efforts, see Chapter 3.2 in Osborne and Rubinstein (1994) for a detailed discussion. The most relevant interpretation is based on the celebrated result of Harsanyi (1973), where mixed equilibrium strategies can be approximated by the set of pure-strategy equilibria for a disturbed game (e.g., updating the cost function by adding a small random component) of incomplete information in which the payoffs of each player are known to themselves but not their opponents. However, a mixed-strategy equilibrium will clearly not be socially optimal.

Our equilibrium analysis does not render a definitive conclusion about how hospitals would react to HRRP incentives in practice.[9] Therefore, we turn to the extant empirical research on HRRP

[9] Standard tools used in the literature to refine the equilibrium concept when multiple equilibria exist do not help in this case. For example, it is easy to show that all equilibria under HRRP is stable, see van Damme (1983).

to infer what equilibria might be observed in practice. First, Zuckerman et al. (2016), Batt et al. (2018b), and Wasfy et al. (2017) report that the average readmission rates of Medicare patients have decreased after the introduction of HRRP. However Desai et al. (2016), Chen and Grabowski (2019), and Wasfy et al. (2017) (among others) show that hospitals with the highest pre-HRRP readmission levels had the greatest improvement. In addition, Mellor et al. (2017) and Ziedan (2018) demonstrate (using different identification strategies) that there is a statistically significant decrease in readmission rates only for AMI patients for low-performing hospitals. They conclude that there is no statistically significant reduction for other hospitals (for AMI) nor for the other two monitored conditions (PN and HF) in all hospitals and find that the readmission rates for these conditions decreased as well, also see Samsky et al. (2019), and Ody et al. (2019). Although the conclusions of the empirical research on the impact of HRRP on hospital actions are somewhat mixed, findings in Ziedan (2018) and Mellor et al. (2017) provide additional support for the impact of no-bonus and capped-penalty provisions that our study identified on hospital actions. Their findings also indicate that hospitals may be settling in the no-effort equilibria for PN and HF.

In conclusion, *HRRP does not provide the right incentives* for hospitals to pick socially optimal readmission reduction levels. The equilibrium action of hospitals is not clear (even when one ignores all the additional complicating factors such as risk adjustment) and the eventual outcome probably depends on the initial readmission levels of the hospitals when the program started. In addition, findings based on empirical analysis of hospital readmissions data in Ziedan (2018), and Mellor et al. (2017) support our (theoretical) findings. In theory the regulator could use additional policy tools to enforce hospitals to settle in a desired output. However, the essential reason that the regulator has to use a yardstick competition type regulation in the first place is its inability to estimate hospitals' cost function. The very same information asymmetry hinders regulators' ability to determine the 'desired' levels for marginal cost and readmission rates. In the next section we show that a similar reimbursement scheme, which imposes no additional informational burden on the regulator, leads to first-best outcomes, unequivocally.

5. Socially optimal relative performance-based regulation

The purpose of this section is to show that it is possible to design reimbursement schemes, guided by our findings in the previous section, that lead to socially optimal outcomes. We do this under two settings. First in §5.1 we consider the same setting as §4 for a comparative analysis. Then in §5.2 we consider a setting where hospitals have limited capacity and delays to access healthcare are inevitable. Lastly in §5.3 we demonstrate how the proposed payment schemes can be modified to account for various additional extensions in the proposed model.

5.1. Modified HRRP

Consider the following payment system, referred to as m(odified)-HRRP; the regulator sets the transfer payment T_i to hospital i as

$$T_i = \bar{c}_i(1+r_i)\lambda + \underbrace{(\bar{r}_i - r_i)\,\bar{c}_i\lambda}_{\pi^m(r_i|\bar{r}_i,\,\bar{c}_i)} + \bar{R}_i, \tag{17}$$

where $\bar{c}_i$, $\bar{r}_i$, and $\bar{R}_i$ are defined as in (6). Reimbursement scheme m-HRRP is similar to HRRP-I (see (9)) in that it does not impose a cap on payment adjustments specified in term π^m, and does allow bonus payments for hospitals with lower-than-expected readmission rates. However, unlike HRRP-I, m-HRRP does not adjust the relative performance of each hospital by the multiplier (i.e. term $(1+r_i)/\bar{r}_i$). We highlight the fact that m-HRRP imposes the same informational burden on the regulator as HRRP. We next show that it restores first-best, and then discuss the implications of this result.

Proposition 4. *Under m-HRRP there exists a unique equilibrium and each hospital chooses the first-best readmission and cost levels* (r^*, c^*) *in this equilibrium.*

There are two different ways to interpret the financial incentives that m-HRRP provides. First interpretation is based on the observation that hospital i's objective can be written as

$$\Pi(r_i, c_i) = (\bar{c}_i - c_i)\,(1+r_i)\lambda + (\bar{r}_i - r_i)\,\bar{c}_i\lambda - R(r_i, c_i) + \bar{R}_i \tag{18}$$

by (1) and (17). It is clear from (18) that m-HRRP provides two different (but interrelated) incentives: i) to reduce costs (cost-efficient hospitals are rewarded owing to term $(\bar{c}_i - c_i)$); and ii) to reduce readmissions (hospitals with lower-than-expected readmission rates are rewarded owing to term $(\bar{r}_i - r_i)$). More generally, this scheme provides evidence that it is possible to design yardstick competition type reimbursement schemes to incentivize hospitals to take desired actions on multiple fronts, e.g., cost and readmission reduction.

Another (arguably more interesting) interpretation is based on the observation that hospital i's objective under m-HRRP can also be written (again by (1) and (17)) as

$$\Pi(r_i, c_i) = (1+\bar{r}_i)\bar{c}_i\lambda - (1+r_i)c_i\lambda - R(r_i, c_i) + \bar{R}_i. \tag{19}$$

To wit, hospital i receives a per patient payment equal to the average expected marginal cost of all other hospitals to *successfully* treat a patient (i.e., when a patient is permanently discharged from the hospital), equal to $(1+\bar{r}_i)\bar{c}_i$. Therefore, in addition to the investment cost R_i, the regulator only needs to estimate the cost of treating a patient, including the cost of potential readmissions.

In fact, this is the idea behind the well-known *bundled payment* scheme, which is widely used by regulators around the world, see Kristensen et al. (2015), Altman (2012).

Finally m-HRRP is equivalent to the reimbursement scheme that reimburses hospital i by $\bar{c}_i(1+\bar{r}_i)/(1+r_i)$ per admission, for index hospitalization or readmission, in addition to the transfer payment equal to $\bar{R}_i$. (It can easily be shown that the objective of hospital i is identical to (19) with this per-patient reimbursement.) Hence, if m-HRRP is implemented this way, hospitals are reimbursed per admission basis under m-HRRP (as is the case under the PPS) as opposed to per patient basis under bundled payments. Therefore m-HRRP can be viewed as a mechanism to alter the per admission reimbursement based on hospitals' relative performance in reducing readmissions, without imposing a separate financial penalty for excess readmissions. On a different note, if a patient is readmitted to a hospital different from the index-admission hospital (which is not captured in our model), the payment for the readmission hospital can easily be determined using this scheme.

Remark 4 (Impact of no-bonus and penalty cap without the multiplier effect). The reimbursement scheme m-HRRP is obtained by removing the multiplier from the HRRP-I scheme and we showed that the equilibrium outcome moves to r^* from $\tilde{r}$. Interestingly a similar result holds for HRRP (and HRRP-II) as well. To demonstrate, we obtain the following transfer payment by removing the multiplier from HRRP reimbursement scheme (see (7))

$$T_i = \bar{c}_i(1+r_i)\lambda - \left(\min\left\{\frac{r_i}{\bar{r}_i} - 1, P_{cap}\right\}\right)^+ \bar{r}_i\bar{c}_i\lambda + \bar{R}_i. \tag{20}$$

(Note that is equivalent to m-HRRP if we allow bonus payments and remove the cap.) Results of Proposition 3 are still valid after replacing $\tilde{r}$ with r^* in definitions of r_p, $\mathcal{S}$, and $\mathcal{S}_o$ under the reimbursement scheme defined in (20). In addition, if we allow bonus payments in (20), results of Proposition 2 hold as well, again after replacing $\tilde{r}$ with r^*. Together these results imply that the suboptimal outcomes under HRRP are not caused entirely by the multiplier but also by the no-bonus payment and capped-penalty provisions.

Remark 5. In Appendix E we establish the optimal hospital actions under m-HRRP with no-bonus and capped-penalty provisions (see (20)) and when the reimbursement level and the readmission target for a hospital are set (exogenously) at socially optimal levels (i.e, $\bar{c}_i = c^*$ and $\bar{r}_i = r^*$). We show that there is a fundamental difference between hospitals' optimal actions when the targets are set exogenously and endogenously. Specifically, we show that when targets are set exogenously at socially optimal levels in m-HRRP, hospitals still choose socially optimal actions, and imposing the capped-penalty (if the cap is not too low) and no-bonus provisions have *no impact* on hospital actions. These results are in stark contrast with those under endogenous targets; these provisions greatly diminish the incentives that the yardstick-based scheme m-HRRP utilizes; see Remark 4 and Proposition 3.

5.2. Limited capacity

So far we have studied the impact of reimbursement schemes on hospital decisions, focusing on the trade-off between cost efficiency and reduced readmissions, under the assumption that hospitals have ample capacity to treat all patients in a timely manner. However, hospitals typically need to operate under high utilization (Campbell 2017) resulting in excessive waiting times, even in developed OECD countries (Viberg et al. 2013). In addition, readmissions have a direct impact on the effective capacity of hospitals, since avoiding readmissions would free up bottleneck resources. This additional capacity can be utilized to treat more patients and to reduce delays to access healthcare, increasing total patient welfare.

Before we present a detailed model that captures the impact of delays to access care, we first demonstrate the interaction between a system's capacity and readmissions in a simple example. Assume that a hospital can be modeled as a G/G/1 queue (i.e., a single server queue with generally distributed service and interarrival times) with readmissions (referred to as retrials in queueing literature), i.e., a patient can be readmitted with probability r after receiving treatment for the first time (e.g., index hospitalization). Let $\tau(r)$ denote the throughput (i.e., the rate patients are treated successfully or the rate they leave the system) of this system as a function of the readmission rate, assuming the service rate (i.e., capacity), μ, and the arrival rate Λ are fixed. Then, assuming that readmitted patients have preemptive priority over newly admitted patients,

$$\tau(r) = \begin{cases} \Lambda, & \text{if } (1+r)\Lambda \le \mu, \\ \frac{\mu}{1+r}, & \text{if } (1+r)\Lambda > \mu. \end{cases}$$

The result implies that, if $(1+r)\Lambda > \mu$ then the throughput will be limited by the available capacity. This follows from the fact that, if this condition holds, then the server does not have enough capacity to treat all patients and its total throughput will be limited to $\frac{\mu}{1+r}$. If the regulator ignores the impact of readmissions on capacity and uses a reimbursement scheme similar to (19) for such systems, the equilibrium readmission rate can be large and the system might be overloaded, i.e., $(1+r)\Lambda > \mu$, especially when the nominal utilization (i.e., Λ/μ) is close to 100%. In this case the throughput will be lower than Λ and the total welfare will be less than $V(\Lambda)$, hence our results in §5.1 would not hold. Therefore, the regulator needs to use an additional mechanism to ensure hospitals invest sufficiently in capacity in its reimbursement scheme.

If the regulator is able to observe the capacity of each hospital, it is possible to augment the reimbursement scheme (19) by another relative performance-based scheme on hospitals' capacity, just as that for readmissions in (18). However, the regulator is unlikely to have the capability to determine the precise capacity of a hospital because of various reasons including, but not limited to: randomness in treatment times; the fact that resources are shared among different patient

groups; and complicated patient flows. Although throughput is typically observable, the regulator cannot infer capacity from this information because, in most cases, the (potential) demand is not observable. In contrast, the regulator can usually track the waiting times of patients, which reflects the capacity-demand balance. Therefore we propose a reimbursement scheme based on patient waiting times.

We next present a model similar to Savva et al. (2018) to capture the impact of waiting times on patients' and hospitals' choices and then show that a reimbursement scheme, obtained by adding a relative performance-based financial incentive term for waiting times to m-HRRP, achieves socially optimal outcomes.

5.2.1. Treatment model when capacity is limited

Effect of waiting times on patient behavior: To model the impact of waiting times on patients' decisions and their welfare, we assume that patients are delay-sensitive and let u denote their utility from the treatment. Patients are heterogeneous in their treatment utilities. We use Θ to denote the distribution of treatment utility across the population.[10] We assume that the treatment utility u is private information (i.e., only patients know their own treatment utility), but hospitals have accurate information about its distribution Θ. Let $W(\lambda, r, \mu)$ denote the expected waiting time[11] of patients with equilibrium arrival rate λ (which we define next), readmission rate r, and capacity μ. We assume that waiting cost is t per unit time and so all patients whose treatment utilities are larger than $tW(\lambda, r, \mu)$ seek service. Hence the equilibrium arrival rate, $\lambda(r, \mu)$, is the unique solution of

$$\lambda(r,\mu) = \Lambda\bar{\Theta}\left(tW(\lambda(r,\mu), r, \mu)\right), \tag{22}$$

where $\bar{\Theta}(x) := 1 - \Theta(x)$. For simplicity we assume that patients make joining decisions only at the time of index hospitalization and will always seek treatment when they need to be readmitted.

Parameter t can be interpreted in different ways–besides the way we presented above–depending on the severity of the disease under consideration. Viewing t as the patients' tolerance for delay might be more appropriate for less severe conditions. For time-sensitive conditions, t can be viewed as the rate a patient's condition deteriorates and, given a patient's 'utility' u, quantity u/t can be viewed as the time a patient needs be treated by via the (regular) care channel. If not treated

[10] We assume that Θ is differentiable and denote its derivative by θ. We also assume that θ is strictly positive everywhere in $[0, \infty)$.

[11] Throughout we assume that conditions stated in §3 continue to hold. We make the following technical assumptions: for any $\lambda \in (0, \mu)$,

$$W(\lambda, r, \mu) > W(0, r, \mu) \text{ and } \lim_{\mu \downarrow (1+r)\lambda} W(\lambda, r, \mu) = \infty, \tag{21}$$

which, for example, hold for $M/M/1$ queues with readmissions, see Guo et al. (2019).

within this time limit, the patient may have to be admitted to the emergency department, and in some situations may suffer from additional complications and health problems due to excessive delay. Finally, smaller values of t along with a suitable distribution Θ can be used when delay to accessing care is more tolerable.

Providers: We assume that each provider picks its capacity level μ in addition to its readmission rate r and marginal cost c and we use $R(r, \mu, c)$ to denote the cost associated with actions (r, μ, c). A hospitals' objective, Π, can be written as

$$\Pi(r, \mu, c) = T - c(1 + r)\lambda(\mu, r) - R(r, \mu, c), \tag{23}$$

in a way similar to (1), where T denotes the payment from the regulator to the provider. We assume for simplicity that $\mu \geq \mu_{min}$ for some $\mu_{min} > 0$.

Regulator: The regulator's objective is again to maximize the total social welfare but, in the current context, the cost of delays to access care has to be accounted for in total patient welfare. First, some of the patients will not be able to access care and we assume that the regulator incurs a penalty equal to c_e for each such patient. We use the cost parameter c_e as a catch-all cost parameter to account for the total cost of emergency care and the cost of additional complications caused by excessive delay. On the other hand, patients who eventually access care experience disutility from waiting, equal to $tW(\lambda, \mu, r)$ on average. Therefore the objective of the regulator can be written as follows

$$S(r, \mu, c) = \Lambda \int_{tW(\lambda,\mu,r)}^{\infty} (x - tW(\lambda, \mu, r))\, d\Theta(x) - c(1 + r)\lambda - R(r, \mu, c) - c_e(\Lambda - \lambda), \tag{24}$$

where we set $\lambda = \lambda(\mu, r)$ for notational simplicity. The regulator's objective S consists of three components: i) the total patient utility for those who eventually access care (the first term $\Lambda \int_{tW(\lambda,\mu,r)}^{\infty} x d\Theta(x)$ is the total treatment utility and $\Lambda \int_{tW(\lambda,\mu,r)}^{\infty} tW(\lambda, \mu, r) d\Theta(x)$ is the total disutility from waiting); ii) total cost of providing treatment (the second and third terms); and iii) cost of excessive delays (the last term). We assume that there is a unique optimal solution (r^*, μ^*, c^*) to the regulator's problem, referred to as socially optimal actions, and FOCs of $S(r, \mu, c)$ are necessary and sufficient to determine regulator's optimal actions.

Remark 6. The model we use to capture the impact of delays to access care is similar to that in Guo et al. (2019) except they assume that patients are homogeneous in their treatment utilities, hence Θ assigns probability 1 to a single point. Our results can be generalized to this case easily. Zorc et al. (2017) also use a similar model for the impact of delay on treatment utility but they do not model patient's 'joining' decisions. Our model is also similar to Savva et al. (2018) but they assume $c_e = 0$ and they do not consider (endogenous) readmissions in their model. Cost of

excessive delay, c_e, might be small in certain healthcare settings. For example, for some emergency room patients, there are more appropriate healthcare channels (Uscher-Pines et al. 2013). However, elective care patients are carefully screened by general practitioners, hence they typically need to receive treatment from specialists in a timely manner. Excessive delays to access care may deteriorate their health to a degree that they have to seek more costly emergency care.

Reimbursement scheme when capacity is limited: For each hospital i, let $\bar{c}_i$ and $\bar{R}_i$ be defined as in (6) and let

$$\bar{W}_i = \frac{1}{N-1}\sum_{j\neq i} W_j, \text{ and } \bar{\lambda}_i = \frac{1}{N-1}\sum_{j\neq i}\lambda_j \tag{25}$$

denote the average waiting time and the average arrival rate, respectively, at all hospitals excluding hospital i. Consider the reimbursement scheme, which we refer to as m-HRRPW(ait), where the transfer payment to hospital i is given by

$$T_i = c_e\lambda_i - t(W(r_i,\mu_i) - \bar{W}_i)\lambda_i + \bar{R}_i + (\bar{c}_i(1+\bar{r}_i) - c_e)\bar{\lambda}_i. \tag{26}$$

Before we discuss the interpretation of different components of this reimbursement scheme, we first show that it leads to socially optimal outcomes in equilibrium. For the next result we assume that FOCs of hospitals' objective functions are necessary and sufficient to determine their unique optimal actions.

Proposition 5. *Under the reimbursement scheme given in* (26)*, the unique symmetric equilibrium is for each hospital i to pick $r_i = r^*$, $\mu_i = \mu^*$, and $c_i = c^*$.*

The proposed reimbursement scheme (26) has three components: (i) payment for successful treatment (the first term); (ii) a yardstick competition type financial incentive term for waiting times (the second term); and (iii) transfer payment to make sure hospitals break even and do not collect rents (the last two terms). Jointly the first two components ensure that the hospitals exert socially optimal efforts to reduce readmissions and costs while ensuring that patients have timely access to care. Also, although Proposition 5 does not rule out the possibility of asymmetric equilibrium, it is possible to modify the reimbursement scheme (26) as outlined in Remark 2 to guarantee the uniqueness of the equilibrium outcomes, because we can rule out the possibility of asymmetric equilibrium when $N = 2$ in this setting as well (the proof is identical to that of the second part of Proposition 1).

The proposed reimbursement system imposes additional information burden on the regulator relative to the reimbursement scheme for the case with ample capacity. Specifically the regulator needs to observe average waiting times W_is and estimate the parameters t and c_e. The waiting

times in different healthcare systems have already being collected as one of the quality measures in healthcare (Viberg et al. 2013). Also, the cost of waiting t and cost of excessive delay c_e can be estimated from current system-wide patient flow data for time-sensitive conditions. In addition, if $c_e = \bar{c}_i(1 + \bar{r}_i)$ then the proposed reimbursement scheme reduces to the bundled payment scheme (with an additional incentive payment for waiting time performance). Hence, a bundled-payment type reimbursement scheme still induces socially optimal actions for conditions for which (average) cost of excessive delay is not significantly different from the cost of providing care through the 'regular' channel, when used with an incentive payment for waiting time performance.

5.3. Other extensions

The following additional features can be incorporated in our models and reimbursement schemes.

Heterogeneous hospitals: In our proposed reimbursement scheme, the regulator should be able to identify (at least pairs of) identical hospitals. In practice, however, hospitals may differ along multiple dimensions due to, for example, geographical and demographic factors. Nevertheless, if these factors are observable to the regulator and exogenous to hospitals, then the proposed scheme can be modified in a way similar to that in Shleifer (1985)§4, also see Savva et al. (2018). In fact CMS uses a hierarchical generalized linear model to estimate the target readmission rate for each hospital.

In our setting, reimbursement scheme (26) can be adjusted as follows. The regulator can use the information on the total costs, readmission rates, arrival and average waiting times of all other providers, along with the vector of observable characteristics to predict the target readmission rates, costs and waiting times for each provider. This new adjusted target can then be used in (26) in place of the 'bare' averages. As discussed in the introduction, the predictive accuracy of this estimation procedure is essential for a relative performance-based reimbursement mechanism to lead to desired outcomes when hospitals are heterogeneous, see Laffont and Tirole (1993) for a more extensive discussion.

Disutility of readmissions: We can extend our model to incorporate an explicit cost for patient readmissions, accounting for additional health risks due to increased time spent in hospital and for patient disutility. To illustrate, let K denote the cost of readmission to a patient, and so to the regulator. Then (26) can be modified by adding the following term

$$K\left(\bar{r}_i - r_i\right)\lambda_i. \tag{27}$$

Intuitively, term (27) makes hospitals internalize the total cost of readmissions to patients. It can be shown similar (proof is virtually identical) to Proposition 5 that, with this additional term, the reimbursement scheme (26) induces socially optimal actions.

Competition between hospitals: Typically the US hospitals in urban areas are not monopolies and compete for patients. If in addition patients use readmission rates to pick among competing hospitals, our model needs to be modified. This can be done in a similar way to Savva et al. (2018), which studies the impact of waiting times on patient choice in the presence of hospital competition. The basic idea is to benchmark the group of hospitals operating in the same catchment area using the performance of hospitals operating in a different but similar market.

Multiple Readmissions per Patient: A patient may be readmitted multiple times following the index hospitalization[12]. This could be incorporated in our reimbursement scheme m-HRRPW by redefining the cost of successful treatment (equal to $(1+\bar{r}_i)\bar{c}_i$ in our base case) as follows. Suppose that each patient can be admitted at most $M \geq 2$ times. Let r_i^j denote the proportion of patients whose treatment is successful after j visits and assume that the cost of treating a patient in jth admission is c_i^j at hospital i. For notational simplicity set $\mathbf{r}_i = (r_i^1, \ldots, r_i^M)$ and $\mathbf{c}_i = (c_i^1, \ldots, c_i^M)$. Now assume that each hospital determines its readmission levels $\mathbf{r}$, cost of treatment $\mathbf{c}$, and capacity μ and let $R(\mathbf{r}, \mu, \mathbf{c})$ denote the cost of these actions. (Note that there are $2M+1$ decision variables in this setting as opposed to three decision variables in the model in §5.2.1.)

Let

$$\bar{\bar{c}}_i \equiv \frac{1}{N-1} \sum_{\ell \neq i} \sum_{j=1}^{M} c_\ell^j r_\ell^j.$$

Term $\bar{\bar{c}}_i$ can be interpreted as the average cost of successfully treating a patient. We note that it reduces to $(1+\bar{r}_i)\bar{c}_i$ if a patient can only be readmitted once and the (average) cost of treatment is the same in each admission, i.e. $c_i^1 = c_i^2 = c_i$, as in our base model. Then the regulator can use the same reimbursement scheme in (26) except by changing the last term $((\bar{c}_i(1+\bar{r}_i) - c_e)\bar{\lambda}_i)$ to $(\bar{\bar{c}}_i - c_e)\bar{\lambda}_i$ to induce socially optimal actions.

Spillover between hospitals: A patient requiring readmission could either return to the hospital from which she was originally discharged, or visit a different hospital, see Zhang et al. (2016). The proposed payment scheme m-HRRP still elicits socially optimal actions from hospitals under symmetric demand and spillovers if hospitals have ample capacity. That is, assume that each patient whose index admission is at hospital i and requires readmission visit to hospital i with probability ρ and to one of the other hospitals j with probability $(1-\rho)/(N-1)$. Then we show in Appendix H that m-HRRP restores first-best.

When capacity is limited, a direct extension of m-HRRPW (see (26)) does not elicit socially optimal outcomes because of the interlink between hospitals' arrival rates. In particular, if a hospital increases its readmission rate, this would increase the arrival rate of all the hospitals operating

[12] Under HRRP, if a patient experiences multiple readmissions within 30 days of the index hospitalization, only the first readmission is included in the total number of readmissions used in penalty calculations.

in the same catchment area. This, in turn, would reduce other hospitals' waiting time performance, increasing the waiting time target for the hospital itself. Hence, if we use m-HRRPW without altering the waiting time benchmarks, it would introduce perverse incentives. Instead, we benchmark each hospital using the performance of hospitals operating in a different catchment area, an idea first proposed in Savva et al. (2018). With this modification, m-HRRPW also restores first-best, see Appendix H.

6. Conclusion and policy implications

The traditional PPS rewards hospitals with high readmission levels because it reimburses hospitals on a per admission basis. To eliminate this perverse incentive, CMS has introduced financial penalties to hospitals with higher-than-expected readmission rates for targeted conditions. Although readmissions for these conditions declined after the introduction of HRRP, the debate about this program's effectiveness and the effects of its financial incentive mechanism on hospital actions continues. Specifically, HRRP is criticized for over-penalizing hospitals because of the multiplier effect, and there is growing evidence in the literature on the negative impact of no-bonus and penalty cap provisions on hospitals' incentives (see for example Bastani et al. (2016) and Zhang et al. (2016)).

Analysis of HRRP: To understand the precise impact of these three features, we analyze the readmission reduction efforts of hospitals under reimbursement schemes derived from HRRP in a setting where customers do not experience significant delays to access care. First we remove the cap and allow bonus payments and show that HRRP does indeed over-penalize hospitals, which results in hospitals over-investing in readmission reduction efforts and under-investing in cost-reduction efforts (relative to social optimum). An even more troubling observation is that the multiplier goes to *infinity* (so too the potential penalties in the absence of a penalty cap) as readmission targets fall when the percentage deviation of a hospital from its target is held fixed. Next we introduce a cap on payment adjustments, while still allowing bonus payments for hospitals with lower-than-expected readmissions, and show that there might be no equilibrium if the cap is too low. We then show that, when bonus payments are not allowed (with or without a penalty cap), there are multiple equilibria and the impact of HRRP on hospital actions is unclear.

These results have important policy implications. Our results show that CMS should remove the multiplier from the HRRP scheme to better align its own incentives with hospitals. Potentially, CMS might have considered additional factors, which our model fails to capture, in using a multiplier. However, we did not find any supporting argument in the literature for using a multiplier. In fact, some industry observers suggest that this penalty multiplier was "simply a drafting error in the legislation" (MedPAC 2013). Although the multiplier is problematic, we also show that removing

the multiplier by itself is not enough to restore social optimum as long as the adjustments to payments are capped and bonus payments are not allowed.

The proponents of capping adjustments to payments argue that hospitals with low operating margins (such as teaching hospitals and those with a relatively greater share of low-income patients) will not be able to afford further reductions in CMS payments, see, for example, Eijkenaar et al. (2013), Barnett et al. (2015). However, such concerns should be addressed by improving the risk-adjustment procedure but not by curtailing the financial incentives provided to reduce readmissions. In fact we show that a low enough cap may completely eliminate the incentives to reduce readmissions. As a result, we suggest further increasing the cap (CMS has increased the cap from 1% to 3% from 2013 to 2015), in addition to the other changes in HRRP, in line with the results in Bastani et al. (2016), Zhang et al. (2016) and Aswani et al. (2016). While the issues with risk-adjustment procedure are being addressed,[13] CMS should at least commit to future increases in the cap so as to make it more profitable to reduce readmissions than to pay the maximum penalty, in the long run.

By using a penalty-only scheme perhaps CMS aimed to demonstrate immediate cost savings from the HRRP program because there was a consensus that avoidable readmissions were already at high levels. In addition, adding bonus payments would increase the overall cost of the program to CMS, at least in the short run. However, we show that a penalty-only yardstick competition scheme does not provide the right incentives for hospitals and long-term costs of suboptimal actions are likely to be much larger than the short-term savings. This result is driven by the fact that hospitals whose readmission levels are below the target do not invest in additional readmission reduction efforts due to lack of bonus payments. In turn, this results in higher target readmission rates for all other hospitals, limiting the impact of the regulation. In addition, not making bonus payments may have other unintended consequences because those hospitals that were already successful in reducing readmissions to below their targets are more likely to discover novel methods to further reduce readmissions in the future as more conditions are monitored under the HRRP program, and treatment methods advance with new techniques. Finally, instead of merely financially penalizing hospitals with excessive readmissions, it may be more effective to channel (at least some of) these 'savings' to readmission reduction efforts in hospitals with higher-than-expected readmission rates. This could reduce the readmission targets for all hospitals. Both the NHS and CMS have already initiated programs that implement this idea, see Kristensen and Sutton (2016), James (2013) for details.

[13] In November 2017, CMS announced that it will use a new stratified methodology to account for socio-economic status in risk adjustment starting in the calculation of financial penalties starting from 2019 (CMS 2017).

Improving HRRP: We show that hospitals exert socially optimal efforts in cost cutting and readmission reduction if bonus payments to hospitals with lower-than-expected readmission rates are allowed and the multiplier and cap are removed from the HRRP scheme. Interestingly the proposed payment system is similar to the well-known (prospective) bundled payment system, where the regulator reimburses hospitals once for the entire episode of care. This payment system does not rely on explicit financial penalties on excessive readmissions because hospitals that can treat patients in a more cost-effective way, through a combination of reduced readmissions and reduced marginal costs, would enjoy higher profits, leading to socially optimal actions, akin to results of Shleifer (1985). One of the challenges in implementing a bundled payment system is that DRG classifications used for the PPS, which has a separate code for each hospitalization, need to be expanded to cover the whole episode of care. However, bundled payments were successfully implemented in Germany in 2004 along with suitable DRGs (Kristensen et al. 2015).

Finally we show that, when capacity is limited and delays to accessing care might deteriorate a patient's condition, reimbursement schemes need to incentivize hospitals to install sufficient capacity in addition to reducing readmissions and costs. We propose a yardstick regulation scheme that introduces additional financial incentives to reduce delays using the average waiting time across all (other) hospitals as a benchmark. In addition, the reimbursement levels in this scheme are tied to the cost of treating patients who cannot access care in a timely manner due to long delays. Therefore the reimbursement amount should be determined using disease-specific cost information for each condition and an effective one-size-fits-all scheme (e.g., yardstick regulation) may not be appropriate when delays to accessing care are inevitable.

6.1. Limitations of our study

In this section we discuss the limitations of our study in addition to potential extensions of our models and directions for future research.

In the analysis of HRRP in §4 we assumed that hospitals are identical. In practice hospitals are heterogeneous and HRRP uses a risk-adjustment procedure to account for differences in hospitals. Considering identical hospitals enabled us to identify the incentives the multiplier, no-bonus, capped-penalty provisions provide. However it is not clear how the HRRP risk adjustment procedure affects the impact of the three provisions on hospital actions. Yet it is unlikely that a risk-adjustment procedure will ameliorate the negative impact of these provisions on hospital actions and we provided a payment scheme (m-HRRP) that yields socially optimal outcomes when used with a proper risk adjustment procedure in case hospitals are heteregenous.

Our equilibrium analysis of HRRP (Proposition 3) shows that there is a continuum of equilibria and our analysis does not yield which equilibrium is more likely. To gain more insight into how

hospitals are reacting to incentives provided by HRRP, data on current hospital actions can be used (see §4.3 for a summary of the current empirical research on HRRP). However, HRRP is relatively new and it is not clear how to use current data on hospital actions to determine equilibrium outcomes. Another interesting potential research direction is to compare how hospitals react to payment reforms with different provisions that CMS implemented. For example, the Value-Based Purchasing program pays bonuses but the penalties are capped. Such an analysis could help shed more light on the impact of some of the provisions HRRP uses.

Our results on HRRP-II and HRRP rely on the assumption that there is a unique symmetric equilibrium under HRRP-I, i.e., Assumption 1. Although we extended our analysis to multiple symmetric equilibrium (see Appendix G), our analysis does not cover the case with asymmetric equilibrium. Also we ignore the fact that there is uncertainty in readmission rates for the sake of analytical tractability. Finally, we assume that the investment cost function satisfies the conditions in Assumption A1. These conditions ensure that the regulator's and hospitals' objective functions have unique optimal solutions. We did not explore the impact of HRRP on hospital actions if the regulator and/or hospitals have multiple optimal actions. We leave these directions for future work.

The desired outcomes under yardstick-competition-based reimbursement systems can only be obtained if performance targets are risk-adjusted for factors outside hospitals' control. However, it can be difficult to select appropriate factors to be included in a risk-adjustment formula, an issue outside the scope of this paper. Moreover, some of the influential factors might be unobservable to the regulator. In this setting the adverse-selection issues have to be taken into account and the optimal regulation for coordinating cost-cutting efforts is a hybrid scheme between fee-for-service and yardstick-competition-based schemes (Laffont and Tirole 1993). However, readmissions are not studied explicitly in the current literature (to the best of our knowledge). Our results can guide the design of new reimbursement mechanisms when adverse-selection issues are present and readmission rates are determined by hospitals' actions.

Hospitals have additional 'low-powered' financial incentives to reduce readmissions. For example, readmission rates of hospitals are publicly reported by CMS and hospitals might be concerned about the negative impact that reporting high readmission rates on their reputation (hence on their future demand). Also competition between hospitals in close proximity might amplify this effect. Our model does not take hospital reputation and competition into account. However, reimbursement schemes that align the incentives of hospitals and the regulator are likely to yield better outcomes, even when these additional factors are present.

References

Adida, E. and F. Bravo (2019). Contracts for healthcare referral services: Coordination via outcome-based penalty contracts. *Management Science 65*(3), 1322–1341.

Adida, E., H. Mamani, and S. Nassiri (2016). Bundled payment vs. fee-for-service: Impact of payment scheme on performance. *Management Science 63*(5), 1606–1624.

Altman, S. H. (2012). The lessons of Medicare's prospective payment system show that the bundled payment program faces challenges. *Health Affairs 31*(9), 1923–1930.

Andritsos, D. A. and C. S. Tang (2018). Incentive programs for reducing readmissions when patient care is co-produced. *Production and Operations Management 27*(6), 999–1020.

Armstrong, M. and D. E. Sappington (2007). *Recent Developments in the Theory of Regulation*, Volume 3 of *Handbook of Industrial Organization*. Elsevier, Amsterdam, Netherlands.

Aswani, A., Z. M. Shen, and A. Siddiq (2016). Data-driven incentive design in the Medicare shared savings program. Technical report, UC Berkeley.

Ata, B., B. L. Killaly, T. L. Olsen, and R. P. Parker (2013). On hospice operations under Medicare reimbursement policies. *Management Science 59*(5), 1027–1044.

Baggot, D. and A. Edeburn (2015). Mandated bundled payments compel hospitals to rethink post-acute care. *Healthcare Financial Management 69*, 64–69.

Barnett, M., J. Hsu, and J. McWilliams (2015). Patient characteristics and differences in hospital readmission rates. *JAMA Intern. Med. 175*(11), 1803–1812.

Bastani, H., M. Bayati, M. Braverman, R. Gummadi, and R. Johari (2016). Analysis of Medicare pay-for-performance contracts. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2839143`, last accessed on December 03, 2017.

Batt, R. J., H. Bavafa, and M. Soltani (2018a). Quality improvement spillovers: Evidence from the hospital readmissions reduction program. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3132770`, last accessed on October 03, 2018.

Batt, R. J., H. Bavafa, and M. Soltani (2018b). Quality improvement spillovers: Evidence from the hospital readmissions reduction program. Technical report, University of Wisconsin.

Bavafa, H., S. Savin, and C. Terwiesch (2017). Redesigning primary care delivery: Customized office revisit intervals and e-visits. Technical report, University of Wisconsin.

Burgess, J. F. and J. M. Hockenberry (2014). Can all cause readmission policy improve quality or lower expenditures? A historical perspective on current initiatives. *Health Econ. Policy Law 9*(2), 193–213.

Campbell, D. (2017). NHS bosses sound alarm over hospitals already running at 99% capacity. *The Guardian,* `https://www.theguardian.com/society/2017/dec/07/nhs-bosses-sound-alarm-over-hospital-bed-shortages-patient-safety-concern`, last accessed on August 12, 2018.

Chalkley, M. and J. M. Malcomson (1998). Contracting for health services with unmonitored quality. *The Economic Journal 108*(449), 1093–1110.

Chen, C. and N. Savva (2018). Unintended consequences of hospital regulation: The case of the Hospital Readmissions Reduction Program. Technical report, London Business School.

Chen, M. and D. C. Grabowski (2019). Hospital readmissions reduction program: Intended and unintended effects. *Medical Care Research and Review 76*(5), 643–660. PMID: 29199504.

CMS (2017). New stratified methodology hospital-level impact file user guide. `https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Downloads/HRRP_StratMethod_ImpctFile_UG.PDF`, last accessed on August 12, 2018.

CMS (2019). Acute care hospital inpatient prospective payment system. `https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/AcutePaymtSysfctsht.pdf`, last accessed on August 8, 2019.

Desai, N. R., J. S. Ross, J. Y. Kwon, J. Herrin, K. Dharmarajan, S. M. Bernheim, H. M. Krumholz, and L. I. Horwitz (2016, 12). Association Between Hospital Penalty Status Under the Hospital Readmission Reduction Program and Readmission Rates for Target and Nontarget Conditions. *JAMA 316*(24), 2647–2656.

Eijkenaar, F., M. Emmert, M. Scheppach, and O. Schöffski (2013). Effects of pay for performance in health care: A systematic review of systematic reviews. *Health Policy 110*(2-3), 115–130.

Fudenberg, D. and J. Tirole (1991). *Game theory*. MIT Press, Cambridge, MA.

Gruessner, V. (2016). CMS includes rich history of healthcare bundled payments. `https://healthpayerintelligence.com/news/cms-includes-rich-history-of-healthcare-bundled-payments`, last accessed on October 03, 2018.

Guo, P., C. S. Tang, Y. Wang, and M. Zhao (2019). The impact of reimbursement policy on social welfare, revisit rate, and waiting time in a public healthcare system: Fee-for-service versus bundled payment. *Manufacturing & Service Operations Management 21*(1), 154–170.

Gupta, D. and M. Mehrotra (2015). Bundled payments for healthcare services: Proposer selection and information sharing. *Operations Research 63*(4), 772–788.

Hansen, L. O., R. S. Young, K. Hinami, A. Leung, and M. V. Williams (2011). Interventions to reduce 30-day rehospitalization: A systematic review. *Ann. Intern. Med. 155*(8), 520–528.

Harsanyi, J. C. (1973, Dec). Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory 2*(1), 1–23.

James, J. (2013). Medicare hospital readmissions reduction program. Health Affairs Health Policy Brief.

Jencks, S. F., M. V. Williams, and E. A. Coleman (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *N. Engl. J. Med. 360*(14), 1418–1428.

Jiang, H., Z. Pang, and S. Savin (2012). Performance-based contracts for outpatient medical services. *Manufacturing & Service Operations Management 14*(4), 654–669.

Joynt, K. E., J. Figueroa, J. Oray, and A. K. Jha (2016). Opinions on the hospital readmission reduction program: Results of a national survey of hospital leaders. *Am. J. Manag. Care. 22*(8), e287–e294.

Kristensen, S. R., M. Bech, and W. Quentin (2015). A roadmap for comparing readmission policies with application to Denmark, England, Germany and the United States. *Health Policy 119*(3), 264–273.

Kristensen, S. R. and M. Sutton (2016). Financial penalties for readmissions in the English NHS. 2016 Royal Economic Society Annual Conference.

Laffont, J. J. and J. Tirole (1993). *A Theory of Incentives in Procurement and Regulation*. MIT Press, Cambridge, MA.

Lanièce, I., P. Couturier, M. Dramé, G. Gavazzi, S. Lehman, D. Jolly, T. Voisin, P. O. Lang, N. Jovenin, J. B. Gauvain, et al. (2008). Incidence and main factors associated with early unplanned hospital readmission among french medical inpatients aged 75 and over admitted through emergency units. *Age Ageing 37*(4), 416–422.

Ma, C. A. (1994). Health care payment systems: Cost and quality incentives. *Journal of Economics & Management Strategy 3*(1), 93–112.

Mechanic, R. and C. Tompkins (2012). Lessons learned preparing for Medicare bundled payments. *N. Eng. J. Med. 367*(20), 1873–1875.

MedPAC (2013). Report to congress: Medicare and the health care delivery system. Chapter 4: Refining the hospital readmissions reduction program. `http://www.medpac.gov/docs/default-source/reports/jun13_ch04.pdf`, last accessed on August 16, 2019.

Mellor, J., M. Daly, and M. Smith (2017). Does it pay to penalize hospitals for excess readmissions? intended and unintended consequences of medicare's hospital readmissions reductions program. *Health Economics 8*(26), 1037–1051.

Monette, M. (2012, 09). Hospital readmission rates under the microscope. *Can. Med. Assoc. J. 184*(12), e651–e652.

Ody, C., L. Msall, L. Dafny, D. Grabowski, and D. Cutler (2019, 1). Decreases in readmissions credited to medicare's program to reduce hospital readmissions have been overstated. *Health Affairs (Project Hope) 38*(1), 36–43.

Osborne, M. J. and A. Rubinstein (1994). *A Course in Game Theory*. MIT Press Books. The MIT Press.

Pope, G. C. (1989). Hospital nonprice competition and Medicare reimbursement policy. *J. Health Econ. 8*(2), 147–172.

Porter, M. E. and R. S. Kaplan (2016). How to pay for health care. *Harvard Business Review 94*(7/8), 88–100.

Robinson, P. (2010). Hospitals readmissions and the 30 day threshold. `http://www.chks.co.uk/userfiles/files/CHKS%20Report%20Hospital%20readmissions.pdf` last accessed on October 03, 2018.

Samsky, M. D., A. P. Ambrosy, E. Youngson, L. Liang, P. Kaul, A. F. Hernandez, E. D. Peterson, and F. A. McAlister (2019, 05). Trends in Readmissions and Length of Stay for Patients Hospitalized With Heart Failure in Canada and the United States. *JAMA Cardiology 4*(5), 444–453.

Savva, N., T. Tezcan, and O. Yildiz (2018). Can yardstick competition reduce waiting times? To appear in *Management Science*.

Shleifer, A. (1985). A theory of yardstick competition. *The RAND Journal of Economics 16*(3), 319–327.

So, K. C. and C. S. Tang (2000). Modeling the impact of an outcome-oriented reimbursement policy on clinic, patients, and pharmaceutical firms. *Management Science 46*(7), 875–892.

Struijs, J. (2015). How bundled health care payments are working in the Netherlands. Harvard Business Review.

Tangerås, T. P. (2009). Yardstick competition and quality. *Journal of Economics & Management Strategy 18*(2), 589–613.

Uscher-Pines, L., J. Pines, A. Kellermann, E. Gillen, and A. Mehrotra (2013). Deciding to visit the emergency department for non-urgent conditions: A systematic review of the literature. *Am. J. Manag. Care 19*(1), 47–59.

van Damme, E. (1983). *Refinements of the Nash equilibrium concept*, Volume Lecture Notes in Vol. 219. Berlin/New York: Springer-Verlag.

Viberg, N., B. C. Forsberg, M. Borowitz, and R. Molin (2013). International comparisons of waiting times in health care – Limitations and prospects. *Health Policy 112*(1), 53 – 61.

Wasfy, J. H., C. M. Zigler, C. Choirat, Y. Wang, F. Dominici, and R. W. Yeh (2017). Readmission rates after passage of the hospital readmissions reduction program: A pre–post analysis. *Annals of Internal Medicine 166*, 324–331.

Zhang, D. J., I. Gurvich, J. A. Van Mieghem, E. Park, R. S. Young, and M. V. Williams (2016). Hospital Readmissions Reduction Program an economic and operational analysis. *Management Science 62*(11), 3351–3371.

Ziedan, E. (2018). The intended and unintended consequences of the hospital readmission reduction program. Technical report, Tulane University.

Zorc, S., S. E. Chick, and S. Hasija (2017). Outcomes-based reimbursement policies for chronic care pathways. Technical report, INSEAD.

Zuckerman, R. B., S. H. Sheingold, E. J. Orav, J. Ruhter, and A. M. Epstein (2016). Readmissions, observation, and the hospital readmissions reduction program. *N. Engl. J. Med. 374*(16), 1543–1551.