Healthcare Payment Policy Impact on Hospital Readmissions

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Manufacturing and Service Operations Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>Draft</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Service Operations</td>
</tr>
<tr>
<td>Keywords:</td>
<td>Health Care Management</td>
</tr>
</tbody>
</table>
Healthcare Payment Policy Impact on Hospital Readmissions

(Authors’ names blinded for peer review)

Problem Definition: In the burgeoning pay-for-performance reimbursement paradigm, readmissions are treated as non-value added quality failures. We examine if and how bundled payment reimbursement plans motivate healthcare providers to reduce readmissions. We consider how a large provider (hospital) and smaller provider (post-discharge) allocate their readmission reduction efforts.

Academic/Practical Relevance: Both readmissions and bundled payment schemes are relatively new and lie at the forefront of the healthcare industry. Relevant to payers designing bundled payment schemes considering readmissions, we develop a model of a multi-provider bundle involving an ill-behaved profit function. This requires developing new modeling and functional analysis techniques.

Methodology: Through analytical modeling and non-linear optimization we analyze individual providers and their behavior within the broader bundle, which enables us to differentiate readmission reduction behavior between large and small providers. We leverage functional analysis to determine properties of provider response to various care parameters that lead to insights into their behavior.

Results: While bundled payment plans incentivize at least cost-effective readmission reduction effort, competition for the bundled payment can cause providers to perform excessive effort beyond what is cost effective. These misalignments are exacerbated when combining bundled payments with the Hospital Readmissions Reduction Program. In general, the hospital is more motivated by the penalty, while the post-discharge provider is more motivated by the bonus (profit). Risk profiling and vertical integration may be effective in reducing the misalignment.

Managerial Implications: When bundled payments are too high, the smaller provider behaves more like a Fee-For-Service provider, which bundled payments are designed to avoid. Lower payments motivate the hospital to perform cost-effective readmission reduction whereas higher payments motivate the post-discharge provider to perform excessive effort. HRRP misaligns the hospitals incentives and should never be used in conjunction with bundled payments. Finally, vertical integration of healthcare providers and risk-adjusted bundled payment approaches can reduce the incentives for providers to deviate from cost-effective effort.

Key words: Hospital readmissions, bundled payment plans, readmission reduction effort

1. Introduction

Hospital readmissions are increasingly being associated with quality failures in care delivery. It is widely reported that nearly 20% of Medicare patients are readmitted within 30 days of discharge. This resulted in over $15 billion of readmission cost to the Medicare system in 2010 (Foster and Harkness 2010), which increased to $26 billion in 2014 (Rau 2014).
Researchers and policy makers believe that approximately 75% of Medicare readmissions are potentially preventable (Hansen et al. 2011, Foster and Harkness 2010). Meanwhile, traditional fee-for-service reimbursement provides payment for every service performed, which can actually encourage the practice of readmitting patients, due to the artificially increased demand (Jencks et al. 2009).

Recently, policy makers and payers began developing pay-for-performance reimbursement models that transfer the financial risk of readmissions to healthcare providers. Bundled payment plans are an increasingly prevalent pay-for-performance model currently being tested and developed by the Center for Medicare and Medicaid Services (CMS). Under bundled payment plans, a group of healthcare providers are paid a fixed payment for an entire episode of care. For example, an episode of care could include pre-operative care, the surgical procedure, and any post-operative care and recovery including readmissions. Each bundled payment plan is negotiated between the healthcare providers and Medicare.

A relatively unique feature of bundled payments is that they present both a penalty, if readmissions are high, and a bonus, if readmissions are reduced. If a readmission occurs, healthcare providers are not reimbursed an extra amount for the readmission cost. Conversely, if healthcare providers can reduce readmission (and other) costs, they split any extra amount up to the fixed payment as a bonus for achieving more cost efficient quality care (CMS 2015).

Bundled payments introduce interesting dynamics among healthcare providers. Specifically, providers seek to claim as much of the fixed bundled payment as possible. This creates a provider versus provider tension, since all providers in the bundled payment are competing for a larger piece of the same pie (the fixed payment). In many plans, a higher percentage of charged revenue leads to gaining a larger fraction of the fixed payment. If the total charges go over the bundled payment amount, however, providers may prefer a lower percentage of charged revenue since the “penalty” is split similarly among providers.

Another pay-for-performance model is the Hospital Readmission Reduction Program (HRRP), which was introduced in 2013 and penalizes hospitals up to 3% of their total Medicare reimbursement when they exceed readmission thresholds in specific categories (Zhang et al. 2016). The goal of pay-for-performance initiatives (related to readmissions) is to motivate healthcare providers to exert a readmission reduction effort level that maximizes readmission reduction relative to the cost of that effort level.
In this paper, we analyze if (and under what conditions), pay-for-performance reimbursement initiatives incentivize a cost-effective approach to readmission reduction. We focus our analysis on hospitals (including physician groups) and post-discharge follow-up providers to determine how they interact with each other and how their interactions impact the motivation to reduce readmissions. We specifically study: 1) Will bundled payment plans motivate all healthcare providers to be cost-effective in their care delivery with respect to reducing readmissions? 2) Will combining HRRP and bundled payment plans together better motivate all healthcare providers to cost effectively reduce readmissions? 3) How can bundled payment plans be designed to better motivate cost-effective readmission reduction?

In answering these questions, our research provides several important insights to bundled payment policy makers and providers. The impact of bundled payments on readmissions has not been studied, therefore these results also fill an important gap in the healthcare operations literature. First, we find that bundled payment plans achieve the CMS-desired outcome of motivating at least the cost-effective level of readmission reduction effort. However, sometimes providers will exceed this cost-effective effort level to increase their individual share of the bundled payment payout, causing them to behave more like Fee For Service (FFS) providers, which is what payers wanted to avoid. Second, this motivation to perform excess effort is particularly strong for the smaller post-discharge provider in the payment bundle. Figure 1 provides a visualization of insufficient and excessive effort, discussed later in the paper, as opposed to cost-effective effort as a function of the bundled payment price. Figure 1 is based on typical parameter values and analysis presented later in this paper. Third, vertical integration among healthcare providers eliminates the incentive for this excessive effort, leading to cost-effective readmission reduction effort by all healthcare providers. Interestingly, adding a readmission penalty, such as that leveraged through HRRP, to bundled payments can be counterproductive as it can result in over-motivation to perform excessive effort both with and without vertical integration. Finally, we propose a new approach to bundled payment plans involving risk-adjusted bundled payments that can mitigate some of the incentive to perform excess effort.

We proceed with a brief literature review in Section 2. Section 3 develops an individual healthcare provider contribution model and a cost-effective payer preferred model. Section 4 discusses the results and interactions of the individual healthcare providers and cost-effective models both analytically and numerically and compares them to a vertically
 integrative model. Finally, Section 5 presents a new bundled payment formulation with risk-adjusted payments that reduces excessive effort and more properly motivates cost-effective readmission reduction. We conclude in Section 6 with a summary of insights and implications for policy makers and future research.

2. Literature Review

Currently, most readmission reduction research focuses on one particular readmission avoidance tactic in a specific healthcare system. These readmission avoidance tactics range from pre-discharge initiatives that involve nurse staffing ratios, day of week influences, and risk prediction models run by the hospital (McHugh and Ma 2013, Kansagara et al. 2011, van Walraven et al. 2010) to post-discharge follow-up initiatives that involve phone calls and tele-medicine opportunities, optimal scheduling of follow-up appointments, dedicated post-discharge management teams, nursing facilities, home health services, and greater levels of patient education by their primary care provider (Helm et al. 2016, Costantino et al. 2013, Wallmann et al. 2013, Tao et al. 2012, Harrison et al. 2011, Watson et al. 2011, Cardozo and Steinberg 2010, Jack et al. 2009). All of this research illustrates that there are many ways to reduce readmissions in various situations, but there is not one perfect way to dramatically reduce readmissions in every case. In fact, in a meta-analysis of readmission reduction at the Mayo clinic (Leppin et al. 2014), doctors found that almost any readmission reduction initiative that focused on the patient’s individual needs was effective in reducing readmissions. Our paper does not attempt to select the best medical procedure or initiative for reducing readmissions, but our contribution is on studying the additional effort to achieve cost-effective readmission reduction.

Medicare and other payers have experimented with pay-for-performance concepts for many years to motivate improved healthcare quality at lower costs. Lee and Zenios (2012)
investigated Medicare’s End-Stage Renal Disease program for dialysis providers and found that the scheme proposed by Medicare would not provide the desired incentives. They introduced an improved design for incorporating full risk adjustment and pay-for-compliance into the payment system. Ata et al. (2013) used a dynamic model to analyze how Medicare’s reimbursement policy in hospice care may give incentives for selecting short-lived patients and may be causing an increasing number of hospice bankruptcies. Jiang et al. (2012) studied optimal contracts for scheduling with a principal-agent framework for outpatient medical services. Their results showed that popular plans used in practice cannot implement the first-best solution and propose a threshold-penalty approach to better coordinate the system for patient mix. Dai et al. (2016) used a strategic queuing framework to study imaging test orders and found that even without fee-for-service payment systems over-testing can occur due to distorted price signals. Aswani et al. (2019) studied contracts in the Medicare Shared Savings Program and found that introducing a performance-based subsidy to partially reimburse providers initial investment in efficiency initiatives could boost Medicare savings by 40% without compromising provider participation in the program. None of these papers have looked at the impact of bundled payments on readmissions, as we do in this paper.

Recently, several more pay-for-performance reimbursement initiatives have been developed to better motivate readmission reduction. One of these is the Hospital Readmission Reduction Program (HRRP), which has a growing amount of operations management literature. Empirical research shows that HRRP is reducing readmissions for monitored and non-monitored conditions (Batt et al. 2018, Chen and Savva 2018). However, Chen and Savva (2018) showed that observation unit usage is also increasing since observation unit stays are not counted as readmissions. Using a principal-agent model, Bastani et al. (2016) demonstrated that adding a financial bonus to HRRP could further incentivize hospitals to reduce readmissions. Zhang et al. (2016) used a game theory approach to determine how hospitals will react to HRRP penalties, assuming all hospitals maximize their contribution margins. They determined that many hospitals will simply accept the penalty as a cost of doing business and not make any extra effort to reduce readmissions. Conversely, our paper looks at the interaction between the hospital and other non-hospital healthcare providers to determine how they will interact to reduce readmissions. While HRRP penalties are included in our analysis, the focus of our paper is bundled payment reimbursement plans.
Bundled payments plans are another pay-for-performance reimbursement initiative developed by Medicare. Recent research on bundled payment plans has focused on what could be improved through better coordination (Miller et al. 2011, Rosen et al. 2013), where to draw the boundaries on episodes of care (Sood et al. 2011, Cutler and Ghosh 2012), and how to set the payment price and which patients or providers to exclude (Rosen et al. 2013, Gupta and Mehrotra 2015, Adida et al. 2016). Rana and Bozic (2015) looked at actual implementation of bundled payment plans that cover the time from admission to discharge in an orthopedic hospital and found savings of 7%-10% per episode of care. Most of these savings came from reducing the implant costs of artificial knees and hips. Gupta and Mehrotra (2015) focused on the mechanics of how bundled payment agreements are reached between healthcare providers and Medicare. They use a principal-agent model to determine how the healthcare providers should act and what mechanisms the overall payer (Medicare) should use to maximize the benefit to society. Adida et al. (2016) also considered the mechanics of bundled payment agreements by looking at patient selection and the providers utility and financial risk. They find that performance of bundled payments is sensitive to the bundled payment price and the provider’s risk aversion. However, none of these papers consider the impact on readmissions, which is a contribution and the focus of our paper.

Guo et al. (2019) compared bundled payment and fee-for-service reimbursement schemes on how they impact social welfare, revisit rates (readmissions), and patient waiting time. They found that bundled payments reduce readmissions, but that the size of the patient pool determines if waiting times will be reduced or increased. They use a combined Stackelberg game and queuing model with assumptions that higher service rates increase readmissions, while we are investigating how bundled payments actually motivate healthcare providers to perform additional effort to reduce readmissions. Andritsos and Tang (2018) looked at readmission reduction with respect to different payment plans, including fee-for-service, bundled payments, and other pay-for-performance initiatives. They use a health co-production model to investigate how readmissions are jointly controlled through the efforts of patients and the hospital. Conversely, our research does not use a co-production model because we investigate how various healthcare providers (hospitals and post-discharge providers) are motivated to perform additional effort to reduce readmissions under bundled payments and HRRP.
3. Bundled Payment Models for Readmission Reduction

In this section, we introduce several models for healthcare providers participating in a bundled payment plan. Our primary decision variable, \( x_i \), is the units of additional readmission reduction effort exerted by healthcare provider \( i \). By additional effort, we refer to initiatives specifically targeting readmission reduction that are not considered part of the general treatment for that episode of care. For example, a hospital could exert additional effort by keeping a patient longer or at higher levels of care (e.g. discretionary use of ICU), improving the nurse-to-patient ratios, hiring a full-time discharge planner, or improving hand-washing programs to reduce infections. Additional efforts of post-discharge providers could include increasing the frequency of patient visits, providing additional home care, using additional tele-medicine monitoring, or making frequent phone calls. For examples of specific additional effort initiatives to reduce readmissions (for both hospitals and post-discharge providers), see Project RED (Jack et al. 2013) and related reports from the US Department of Health & Human Services on reducing readmissions (Boutwell et al. 2016). For the rest of this paper, we will refer to \( x_i \) as readmission reduction effort or simply effort performed by provider \( i \).

Without loss of generality, we define a universal readmission reduction probability function, \( p(\cdot) \), which is based on the units of effort for any provider (see below). The activities required to move along the \( p(\cdot) \) curve may be different for each provider and hence have different costs. Therefore, we use a common \( p(\cdot) \) function for all providers, with differentiation coming from the cost of these efforts to reduce readmissions. This allows us to consider the impact of the total readmission reduction effort of all providers to reduce readmissions, and hence, we define it as a scalar function.

**Definition 1 (Probability of readmission \( p(\bar{x}) \), due to total effort, \( \bar{x} \)).**

Given readmission function \( p(\bar{x}) \), a unit of effort is defined as the amount of effort required for any provider to move from \( p(\bar{x}) \) to \( p(\bar{x} + 1) \). The cost to achieve this improvement, one unit of effort, is given by \( c_i(x_i + 1) - c_i(x_i) \), which could be different for each provider \( i \).

3.1. Contribution Margin Model

In this paper, we follow Zhang et al. (2016) in assuming for-profit and not-for-profit healthcare providers are contribution margin maximizers, since many authors including Weisbrod (1988) and Capps et al. (2017) show that not-for-profit healthcare providers behave very similarly to for-profit healthcare providers. In our model, we capture the impact of bundled
payment plans on readmission reduction efforts under this contribution margin maximizing goal. While there are a number of bundled payment plan implementations, we focus on one of the more common implementations in which care providers submit claims throughout the episode of care. Then at the end of an episode of care, each provider is allocated a fraction of the bonus or penalty amount equal to the fraction of their submitted claims relative to the total claims for the episode of care (CMS 2015).

For example, in their Bundled Payment Guide for Physicians, the Toward Accountable Care Consortium lists a strategic tip in dealing with bundled payment accounting as: “The other simpler option for smaller organizations is to have the payer pay fee-for-service to all providers and then perform a “true up” at the end of the episode to determine if additional payment is due to the provider, or vice versa.” (Consortium 2014).

The charged revenue for provider $i$, $r_i(x)$, and associated costs, $c_i(x)$, where $x$ is the vector of effort levels for each provider, are divided into three different categories: (1) General or standard services performed, $r_i^G, c_i^G$, which does not depend on readmission reduction effort; (2) Additional activities to avert readmissions $r_i^A(x_i), c_i^A(x_i)$, which depends on the readmission reduction effort of the provider, $x_i$. Throughout the paper, we assume $c_i^A(x_i)$ is linear in $x_i$, which simply means each unit of effort costs the same; (3) services performed during unscheduled readmissions, with revenue $\theta_i$, and costs, $\gamma_i$, which occur with probability $p(\bar{x})$, depends on the total readmission effort by all providers, $\bar{x} = \sum_i x_i$. We make the reasonable assumption that $p(\bar{x})$ is convex decreasing in total readmission reduction effort, $\bar{x}$, which means the first units of effort have the most impact on readmission reduction. In addition, we make the global assumption that effort costs for hospitals are greater than or equal to effort costs for post-discharge providers, since hospitals typically have higher overhead costs and even if one post-discharge provider has high costs there will be other post-discharge providers with lower costs that can be used within the bundled payment plan. Thus, $r^A_h \geq r^A_{pd}$ and $c^A_h \geq c^A_{pd}$.

Equation 1 formulates the expected contribution margin, $E[\pi_i(x)]$, for provider $i$ in a bundled payment plan with fixed bundled payment price, $\tau$. The bundled payment bonus or penalty (seen in lines 2 and 3 of Equation 1) is allocated based on provider $i$’s revenue percentage, which is the ratio of provider $i$’s reimbursable charges over all reimbursable charges for that episode of care. Line 2 shows the bonus/penalty when a readmission occurs,
with probability $p(x)$. Line 3 shows the bonus/penalty when a readmission does not occur, with probability $1 - p(x)$.

$$E[\pi_i(x)] = (r_i^G - c_i^G) + (r_i^A(x_i) - c_i^A(x_i)) + (p(x)\theta_i - p(x)\gamma_i)$$

$$+ p(x) \left[ \sum_{k=1}^{K} \left[ \frac{r_k^G + r_k^A(x_k)}{r_k^G + r_k^A(x_k) + \theta_k} \left( \tau - \sum_{k=1}^{K} [r_k^G + r_k^A(x_k) + \theta_k] \right) \right] \right]$$

$$+ (1 - p(x)) \left[ \sum_{k=1}^{K} \left[ \frac{r_k^G + r_k^A(x_k)}{r_k^G + r_k^A(x_k)} \left( \tau - \sum_{k=1}^{K} [r_k^G + r_k^A(x_k)] \right) \right] \right] \quad (1)$$

The superscript $G$, general service terms, are important because they are the primary determinants of who are the large and small providers within the bundled payment episode of care. For example, a hospital’s general services could include surgery preparation, initial surgery, and recovery in the hospital. The general services for a post-discharge provider may only be several follow-up appointments. Thus, the general service charges from the post-discharge provider are likely to be significantly smaller than the hospital’s general service charges. This difference in provider size leads to different levels of motivation to perform readmission reduction effort, $x_i$, which we discuss further in Section 4.

### 3.2. Models of Competition and Payer-desired outcome

In this section we present two models to determine the readmission reduction effort, $x_i$, performed by each provider. The first is where multiple providers compete for the fixed bundled payment amount, $\tau$. The second model considers a cost-effective optimization preferred by payers such as Medicare, which matches the goals of an objective central planner who is concerned with minimizing the overall cost of treatment for an episode of care.

**Multi-provider Bundled Model:** For the sake of exposition, we consider two providers: the hospital, denoted by subscript $h$, and the post-discharge provider, denoted by subscript $pd$. The hospital performs inpatient services, including readmissions. The post-discharge provider includes services performed by skilled nursing facilities, primary care physicians, walk-in clinics, and home health doctors and nurses. We consider a scenario where each provider has the same information, is making decisions at the same time, and the only uncertainty is if a readmission will occur with a particular patient or not. This modeling choice is supported by the recommendation of the Toward Accountable Care Consortium, which suggests, “there needs to be frequent meetings with the payer and data transparency so providers can discover early avoidable leakage.” (Consortium 2014). We believe this
setting is sufficient for our goal of understanding provider interaction under different payment schemes. Each provider will exert the amount of effort, \([x_h, x_{pd}]\), that maximizes their contribution margin, where their contribution margin model is given by Equation 1.

Cost-Effective Payer Model. A payer, such as Medicare, wants to incentivize cost-effective care delivery across the entire treatment cycle. We model this by considering an integrated system (hospital and post-discharge) with a single controller optimizing cost over the entire episode of care. We use this model to represent the desired outcome of bundled payments – i.e. the stated goal of CMS in designing pay-for-performance schemes – to compare with the outcomes of the Multi-provider Bundled Model above. Explicitly, we write the Cost-Effective Payer Model as:

\[
\min_{\mathbf{x}} \left[ E\left[ \pi^{CE}(\mathbf{x}) \right] \right] = \min_{\mathbf{x}} \sum_{k=1}^{K} [c^G_k + c^A_k(x_k) + p(\bar{x})\gamma_k] = \sum_{k=1}^{K} c^G_k + \min_{\mathbf{x}} \sum_{k=1}^{K} [c^A_k(x_k) + p(\bar{x})\gamma_k] \tag{2}
\]

Let \(\bar{x}^{CE} = \sum_k x^{CE}_k\) be the total amount of readmission reduction effort employed in this solution, where \(x^{CE}_k\) is the optimal effort employed by each provider \(k\) in Equation 2.

Model Parameterization. In our numerical analysis, we leverage medical literature and conversations with physicians, hospitals, and other healthcare researchers to determine reasonable ranges for the bundled payment price, \(\tau\), the general procedure charged revenue and costs, \(r^G_i, c^G_i\), and the charged revenue and costs associated with each unit of readmission reduction effort, \(r^A_i(x_i), c^A_i(x_i)\). To capture the readmission probability as a function of provider effort, we draw upon several OM papers that capture readmission reduction as a function of the volume of follow-up appointments. Since we are performing policy level analysis and not developing a data-driven decision support system, we use the follow-up system as a reasonable proxy for the shape and magnitude of readmission reduction, where unit of effort, \(x_i\), corresponds to one follow-up office visit. Using data similar to Helm et al. (2016) we generate low, medium, and high readmission risk profile patients and all patients in aggregate to create the curves in Figure 2. These curves map the amount of effort to the projected readmission probability based on clinical and administrative hospital data. It can be seen that these curves are convex and decreasing, supporting our \(p(\bar{x})\) assumption above.

4. Impact of Bundled Payments on Readmissions

In this section, we analyze whether bundled payment and HRRP pay-for-performance reimbursement initiatives are achieving the goals set out by the primary designer/user of
such plans, CMS. In Section 4.1, we find that bundled payment plans do in fact motivate providers to perform no less than the cost-effective payer-preferred effort level to reduce readmissions. However, we also find that bundled payments can over motivate the post-discharge provider, which results in additional readmission reduction effort that is not cost-effective. In Section 4.2, we demonstrate that the post-discharge provider is often more motivated to perform readmission reduction effort than the hospital. We discuss how bundled payment plans impact vertical integration in Section 4.3 and patient throughput in Section 4.5. Finally, in Section 4.4, we illustrate how combining HRRP penalties with bundled payments can motivate the hospital to perform excessive readmission reduction effort, which could lead to a mis-allocation of healthcare resources.

4.1. Bundled Payments Motivate Cost-Effective Readmission Reduction Effort

One major goal of bundled payment plans is to motivate healthcare providers to mitigate readmissions in a cost-effective manner. In this section, we analyze whether this goal is being achieved by comparing the best-response actions under the multi-provider bundled model, Equation 1, with those from the cost-effective payer model, Equation 2. Theorem 1 shows that providers in a bundled payment plan will do at least as much effort as in the optimal cost-effective setting.

Let $x^B_i(x_j)$ be the best response effort for provider $i$ when the other provider performs $x_j$ units of effort. Theorem 1 shows when either provider will ensure total effort is greater than or equal to the CMS-desired cost-effective total effort level, $\bar{x}^{CE}$, in part (i) and when each provider’s best response will bring the total effort level at least up to the cost-effective total effort level, $\bar{x}^{CE}$, in parts (ii) and (iii). For the theorem, we make the reasonable assumption that the post-discharge provider receives no additional revenue or cost as a result of a readmission, since a readmission is an inpatient intervention. Theorem 1 also
requires minor technical assumptions, in addition to the primary Conditions 3, 4, and 5. These technical assumptions are true with realistic parameter values, as they simply require that \( \tau \) is not excessively large. For typical parameter values, they require that \( \tau \) be less than 20x total general charges in part (ii) and less than 10x total general charges in part (i) and (iii). These technical assumptions assume \( \theta \leq r_h^G + r_{pd}^G \). However, if \( \theta > r_h^G + r_{pd}^G \), they are still typically true, but not by as wide of margin.

**Theorem 1.** Let \( \gamma_{pd} = \theta_{pd} = 0 \), and assume all providers have sufficient capacity to provide the cost-effective level of readmission reduction effort.

(i): Given a total initial effort level below the cost-effective effort level, \( x_h + x_{pd} < \bar{x}^{CE} \), and a technical assumption of \( \tau \leq \frac{2c_h^A[r_h^G + r_{pd}^G]}{\bar{x}^{CE}[p(\bar{x}) - p(\bar{x}^{CE})]} \), if

\[
c^A_{pd}(\bar{x}^{CE}) + c_h^A(\bar{x}^{CE}) < \gamma_h \left[ p(0) - p(\bar{x}^{CE}) \right]
\]

then \( x^B_{pd}(x_h) + x^B_{pd}(x_{pd}) \geq \bar{x}^{CE} \).

(ii): Given a hospital effort level below the cost-effective effort level, \( x_h < \bar{x}^{CE} \), and a technical assumption of \( \tau \leq \frac{4c_h^A[r_h^G + r_{pd}^G]^2}{3r_h^G r_{pd}^G [p(\bar{x}) - p(x^{CE})]} \), if

\[
\tau \left[ \frac{r_{pd}^G}{r_{pd}^G + r_h^G} - \frac{r_{pd}^G}{r_{pd}^G + r_h^G + \theta_h} \right] > \frac{c^A_{pd}(\bar{x}^{CE})}{p(0) - p(x^{CE})},
\]

then \( x^B_{pd}(x_h) \geq \bar{x}^{CE} - x_h \).

(iii): Given a post-discharge effort level below the cost-effective effort level, \( x_{pd} < \bar{x}^{CE} \), and a technical assumption of \( \tau \leq \frac{2c_h^A[r_h^G + r_{pd}^G]}{\bar{x}^{CE}[p(\bar{x}) - p(x^{CE})]} \), if

\[
\gamma_h > \tau \left[ \frac{r_h^G + \theta_h}{r_{pd}^G + r_h^G + \theta_h} - \frac{r_h^G}{r_{pd}^G + r_h^G} \right] + \frac{c_h^A(\bar{x}^{CE})}{p(0) - p(x^{CE})},
\]

then \( x^B_{pd}(x_{pd}) \geq \bar{x}^{CE} - x_{pd} \).

When Condition 3 is true in Theorem 1(i), then at least one of Condition 4 or 5 is true and at any \( \tau \) level at least one of the providers will be motivated to ensure the cost-effective effort level is performed. To determine when each individual provider is motivated to perform at least the cost-effective effort level, the bundled price, \( \tau \), must be included in the condition. Thus, when Condition 4 is true in Theorem 1(ii) (Condition 5 is true in Theorem 1(iii)), the post-discharge provider (hospital) is motivated to perform enough
effort so the total effort is at or above the cost-effective effort level. Analyzing Condition 3 more closely, the left-hand-side is the cost of effort to reduce readmissions, $c_i^A(x^{CE})$, and the right-hand-side is the cost reduction in expected readmissions, $\gamma_h [p(0) - p(\bar{x}^{CE})]$. Therefore, if it is possible to reduce readmissions in a cost-effective manner, the condition will be true and the joint effort will provide at least the cost-effective amount of effort.

Another key observation from Theorem 1 is that readmission reduction behavior is driven by the impact of a readmission on the percentage of the bundled payment allocated to each provider. This can be seen in the left-hand-side (right-hand-side) of Condition 4 (Condition 5), which captures the change in revenue for the post-discharge (hospital) provider when a readmission occurs versus when it does not occur. For the post-discharge provider, a readmission causes them to lose some percent share of the bundled payment. Hence, the larger the percent share they lose (larger left-hand-side of Condition 4) the more likely they are to engage in cost-effective readmission reduction. The percent share lost is a function of the amount a hospital charges for readmissions, $\theta_h$, as well as the “size” of the post-discharge provider relative to the hospital in terms of the fraction of the general charges for the episode of care that belong to the post-discharge provider, $\frac{r_{pd}^G}{r_{pd}^G + r_h}$. The larger the initial fraction, the larger the percent lost if there is a readmission. Hence, the larger the post-discharge provider (as a portion of the total bundle charges) the more likely they are to be motivated to perform readmission reduction effort up to (or beyond as we will see) CMS’ goal of cost-effective reduction effort.

For the hospital, the impact of a readmission is very different. When a readmission occurs it increases the hospital’s reimbursable charges and likewise their revenue percentage. Therefore, a readmission may still benefit the hospital (as in FFS) as it subsequently receives a higher percentage of the fixed bundled payment, illustrated in the first term on the right-hand-side of Condition 5. However, as long as the cost of a readmission for the hospital, $\gamma_h$, is greater than the expected loss in revenue by reducing readmissions plus the cost effectiveness of their readmission reduction effort, the hospital will ensure cost-effective effort levels are maintained.

Interestingly, when the bundled payment, $\tau$, is small, the hospital will be motivated to reduce readmissions to avoid a penalty, whereas the post-discharge organization will do nothing to avoid getting a larger share of the penalty. However, when $\tau$ becomes large then the expected loss in revenue by reducing readmissions increases and the hospital may still
benefit from a readmission in direct opposition with the purpose of pay-for-performance reimbursement. Meanwhile, the post-discharge provider may do excessive effort to capture a larger percent of the bonus by preventing readmission charges from the hospital (See Fig. 3). This means the bundled payment price also plays an important role in the effort to reduce readmissions.

To illustrate this phenomenon for both providers, we rearrange the conditions in Theorem 1. The post-discharge provider will be motivated to perform at least the cost-effective effort level as long as \( \tau \) is set according to Condition 6. As long as \( \tau \) is less than the right-hand-side of Condition 7, the hospital will be motivated to perform at least the cost-effective effort level.

\[
\tau > \frac{(r^G_{pd} + r^G_h) (r^G_{pd} + r^G_h + \theta_h)}{r^G_{pd} \theta_h} \left[ \frac{c^A_{pd}(\bar{x}^{CE})}{p(0) - p(\bar{x}^{CE})} \right]
\]

(6)

\[
\tau < \frac{(r^G_{pd} + r^G_h) (r^G_{pd} + r^G_h + \theta_h)}{r^G_{pd} \theta_h} \left[ \gamma_h - \frac{c^A_{pd}(\bar{x}^{CE})}{p(0) - p(\bar{x}^{CE})} \right]
\]

(7)

These conditions illustrate that the hospital is more motivated by avoiding a penalty, while the post-discharge provider is more motivated by increasing their share of any bonus. From Condition 6, we see that if the bundled payment price is too small the post-discharge provider will not be interested in performing effort to reduce readmissions, because performing effort would increase their share of the penalty. However, if the bundled payment price is high, the post-discharge provider will perform extra effort to gain a larger share of the bonus. Further, as \( r^G_{pd}, r^G_h, c^A_{pd} \) or \( \theta_h \) increases, the \( \tau \) required to motivate the post-discharge provider to exert cost-effective effort increases. Thus, if the revenues received for caring for a patient are high, a larger bundled payment price is needed to motivate the post-discharge provider to reduce readmissions. Conversely, as the slope of the readmission probability curve gets steeper (more effective readmission reduction interventions), the \( \tau \) required to motivate cost-effective post-discharge provider effort decreases. This is because the post-discharge provider is better able to prevent the hospital from taking extra revenue from the bundle, increasing the post-discharge provider’s percent of the bonus or ability to avoid a penalty.

In Condition 7, we see that the hospital is motivated by lower bundled payment prices; i.e. the bundled payment price, \( \tau \) should not be too large or the hospital will
not be motivated to reduce readmissions; gaining greater benefit from simply allowing readmissions to occur. Further as $r_{pd}^G$, $r_h^G$, $\gamma_h$, $\theta_h$, or the slope of the readmission probability curve increase, the maximum $\tau$ that still motivates the hospital to achieve cost-effective effort increases. Thus, when readmission reduction efforts are more effective, the hospital is more likely to engage in readmission reduction to avoid a penalty.

Next, we present a numerical experiment where we parameterize our model with reasonable values from publicly available healthcare data (https://www.cms.gov), healthcare research literature (Costantino et al. 2013, Helm et al. 2016), and interviews with hospital and home health professionals to illustrate these concepts. Figure 3 provides one example of our experiments, plotting the best response effort levels of the hospital and the post-discharge provider with the cost-effective effort level and the joint response effort level as $\tau$ is varied from $2,000$ to $20,000$. The joint response and cost-effective effort curves result from Theorem 1. However, the individual best response curves are when the other provider performs zero readmission reduction effort and are used to illustrate the components of the joint response curve. Thus, these curves do not apply to Theorem 1 results.

The general charges are set at $9,000$ for the hospital and $1,000$ for the post-discharge provider and the charges for a readmission are $20,000$. We assume the cost of one unit of readmission reduction effort, $x_i$, is $220$ with $10\%$ margins. Figure 3 illustrates that the hospital is motivated to perform the cost-effective effort level for the entire range of $\tau$. In fact, the hospital is motivated to achieve cost-effective effort levels until $\tau$ is greater than $100,000$ in this scenario, which is unlikely to be seen in practice. This value is so large because the hospital typically wants to perform cost-effective effort to ensure the minimum penalty or the maximum bonus. This is because the hospital wants to maximize (minimize) the bundled payment bonus (penalty), since increasing their percentage of the bundled payment typically has less effect on their total profit than increasing (decreasing) the size of the bonus (penalty) by avoiding a readmission. Figure 3 also illustrates a scenario where Condition 3 is true and for any value of $\tau$ either the hospital or the post-discharge provider will be motivated to perform at least the cost-effective effort level.

In practice, payers should choose a $\tau$ that motivates both providers to achieve cost-effective effort levels. From our numerical analysis in Figure 3, we see that this value of $\tau$ falls between the $r_{total}^G$ and $r_{total}^G + \theta_h p(0)$ lines. The $r_{total}^G$ line illustrates the total value of all general services performed for every client, where $r_{total}^G = r_{pd}^G + r_h^G$. The $r_{total}^G + \theta_h p(0)$
Figure 3 Impact of increasing bundled payment price on the best response of each provider. The best response assumes the other provider performs zero effort.

line illustrates the total value of general services plus the expected readmission charges if no readmission avoidance effort is performed.

In Figure 3, when the bundled payment price, $\tau$, is low the post-discharge provider is not motivated to perform readmission reduction effort because performing more effort will increase their share of the bundled payment penalty. As $\tau$ increases a bonus will eventually result, at which point the post-discharge provider becomes motivated to perform readmission reduction effort since it will increase their share of the bonus. This illustrates that the post-discharge effort is non-decreasing in $\tau$. Thus, as $\tau$ continues to increase, the post-discharge provider actually ends up performing excessive/wasteful effort, beyond what is cost-effective, to capture a larger share of the bundle payment bonus. This results in the post-discharge provider behaving more like a provider in the FFS context, which payers are trying to avoid. We discuss this post-discharge provider over-motivation further in the next section.

4.2. More Readmission Reduction Effort from the Post-Discharge Provider

In this section, we study the differing effort of the post-discharge provider versus the hospital in a bundled payment plan. Specifically, we show that the post-discharge provider will perform more readmission reduction effort than the hospital when $\tau$ is large and there is a large difference in the charged revenue size of each provider.

We formalize this result by focusing on the amount of readmission reduction effort each provider is motivated to perform. We find the best response effort of both providers when their marginal revenue crosses their marginal cost line. To facilitate exposition, we define $Z(x) = r^G_{pd} + r^A_{pd}(x_{pd}) + r^G_h + r^A_h(x_h)$ as all the general and additional effort revenue
charges made against that bundled payment. With $Z(x)$ established, we are able to show in Theorem 2 that the post-discharge provider’s best response readmission reduction effort, $x_{pd}^B(y)$, will be greater than or equal to the hospital’s best response readmission reduction effort, $x_{h}^B(y)$, when Condition 8 is true and the other provider performs $y$ units of (not necessarily best-response effort) effort. We cannot categorically prove that the derivative of provider revenue is always decreasing or unimodal. However, we are confident in this assumption because it holds throughout our extensive numerical analysis.

**Theorem 2.** Let $\gamma_{pd} = 0$, $r^A = r_{pd}^A = r_h^A$, and $c_{pd}^A \leq c_h^A$. Assume an exponential readmission probability curve, such that $p(\bar{x}) = \lambda_0 e^{-\lambda \bar{x}}$, and the derivative of provider revenue is either decreasing or unimodal. If for all such $x$:

$$
\tau r^A p(\bar{x}) \left( \theta_h + \frac{r^G_h - r_{pd}^G}{Z(x) + \theta} + \tau r^A \left[ 1 - p(\bar{x}) \right] \frac{r^G_h - r_{pd}^G}{Z(x)} \right) > \lambda p(\bar{x}) \left( \gamma - \theta \right) \frac{\tau \left[ 2 r_{pd}^G + r_{pd}^G(x_{pd} + y) \right]}{Z(x)(Z(x) + \theta)},
$$

then $x_{pd}^B(y) \geq x_{h}^B(y)$.

While the Condition 8 appears complex, it is true in many real-world scenarios. The first Condition 8 ratio term is the probability a readmission occurs times the difference in hospital and post-discharge provider non-effort charges when a readmission occurs, divided by total squared charges when a readmission occurs. The second ratio term is the probability a readmission does not occur times the difference in hospital and post-discharge provider general charges when a readmission is avoided, divided by total squared charges when a readmission is avoided. On the right-hand-side of Condition 8, there is the readmission probability times $\lambda$ times “roughly” the marginal contribution margin from a readmission occurring. We say “roughly” because the ratio multiplied with $\theta$ is often close to one. This condition can be interpreted as follows. The post-discharge provider is more likely to perform more effort when: 1) there is larger value of $\tau$ (as seen before), 2) there is larger difference in charged revenue size between the hospital and the post-discharge provider, 3) there is smaller slope on the probability of a readmission, and 4) there is a larger margin on an unplanned readmission for the hospital.

To better understand Theorem 2 consider the following reasons why the post-discharge provider is more likely to perform more effort. For the post-discharge provider, increasing readmission reduction effort (in the event of a bundled payment bonus) always has a double-positive influence on their revenue. First, increasing post-discharge provider effort increases...
their percentage of charged revenue, which increases their revenue from the bundle. Second, increasing post-discharge provider effort will reduce the hospital’s readmission revenue charges, which increases post-discharge percentage of charged revenue, which increases their revenue from the bundle.

Conversely, the hospital experiences contrasting effects from reducing readmissions. Increasing effort results in 1) a revenue benefit by increasing their percentage of revenue charges, but 2) a revenue reduction by decreasing their expected readmission charges, which decreases their percentage of revenue charges. As a possible solution for these differences, we next examine the effect of bundled payments on a vertically integrated system, such as an Accountable Care Organization.

4.3. Bundled Payments with Vertical Integration

Healthcare mergers are increasing in the United States because of new reimbursement plans and other initiatives in the Affordable Care Act (Ginsburg 2016, Fulton 2017). One reason this occurs is so that a single healthcare system has responsibility for all activities surrounding a patient’s health and wellness. This can be beneficial because when all providers are controlled by one vertically integrated healthcare system, our research shows that the optimal readmission reduction effort of all providers, $\bar{x}^{VI}$, aligns with the optimal cost-effective effort level, $\bar{x}^{CE}$, as defined in Equation 2.

**Theorem 3.** Let $\bar{x}^{VI} = x^{VI}_{pd} + x^{VI}_{h}$ be the optimal readmission effort of the post-discharge provider and hospital when both are owned by the same vertically integrated healthcare system. Then $\bar{x}^{VI} = \bar{x}^{CE}$.

This alignment occurs because maximizing the contribution margin of the vertically integrated healthcare system is equivalent to minimizing the total costs to achieve the optimal cost-effective effort level in Equation 2. Thus, with increased vertical integration, we could see increasing effectiveness of bundled payment plans in achieving their goals of cost-effective readmission reduction. However, in the next section we discuss how combining bundled payments with the Hospital Readmission Reduction Program (HRRP) can misalign the best response and cost-effective effort.
4.4. Bundled Payments Plans with HRRP Penalties

In addition to bundled payments, HRRP was another pay-for-performance model introduced in 2013. HRRP can penalize hospitals up to 3% of their total Medicare reimbursement when they exceed readmission thresholds in specific categories. In 2018, 80% of hospitals subject to HRRP experienced some reduction in Medicare reimbursement with fines estimated to total $564 million (NEJMGroup 2018). While HRRP only enforces a penalty, bundled payment plans have a penalty when readmissions occur, but also offer the opportunity for a bonus if cost-effective methods are used to reduce readmissions. When bundled payments and HRRP are both employed, the contribution margin for the hospital becomes Equation 9, which adds a HRRP penalty term, \( HRRP(\bar{x}) \), to Equation 1. The contribution margin for the post-discharge provider remains unchanged from Equation 1, since HRRP only impacts the hospital. We define \( x_{B+HRRP}^i \) as the best response of provider \( i \) when bundled payment plans and HRRP are used together.

\[
E[\pi_{i}^{HRRP}(x)] = (r_i^G - c_i^G) + \left( r_i^A(x_i) - c_i^A(x_i) \right) + (p(\bar{x})\theta_i - p(\bar{x})\gamma_i) \\
+ p(\bar{x}) \left[ \frac{r_i^G + r_i^A(x_i) + \theta_i}{\sum_k \left[ r_k^G + r_k^A(x_k) + \theta_k \right]} \left( \tau - \sum_k \left[ r_k^G + r_k^A(x_k) + \theta_k \right] \right) \right] \\
+ (1 - p(\bar{x})) \left[ \frac{r_i^G + r_i^A(x_i)}{\sum_k \left[ r_k^G + r_k^A(x_k) \right]} \left( \tau - \sum_k \left[ r_k^G + r_k^A(x_k) \right] \right) \right] - HRRP(\bar{x}) \quad (9)
\]

**Theorem 4.** Let \( x_{pd} \) be the post-discharge provider’s effort and \( x_{Bh}^i(x_{pd}) \) be the hospital’s best response effort given \( x_{pd} \) with bundled payments but without HRRP. Similarly, let \( x_{B+HRRP}^i(x_{pd}) \) be the hospital’s best response effort given \( x_{pd} \) with bundled payments and HRRP. Then \( x_{B+HRRP}^i(x_{pd}) \geq x_{Bh}^i(x_{pd}) \).

Theorem 4 says that adding HRRP penalties to a bundled payment plan could result in readmission reduction effort levels greater than the bundled payment plan alone. However, Theorem 1 and the associated numerical analysis illustrate that bundled payment plans already incentivize providers to perform readmission reduction effort that is at least at cost-effective effort levels. Therefore, adding HRRP risks over-motivating the hospital to perform excessive and inefficient levels of readmission reduction, similar to what we saw with the post-discharge provider under a large bundle payment price. In Theorem 3, we found that vertical integration could properly align readmission reduction effort under
bundled payments. However, Theorem 5 shows that adding HRRP may mis-align effort, even for a vertically integrated organization.

**Theorem 5.** Let \( \bar{x}^{VI,HRRP} \) be the optimal readmission reduction effort of the post-discharge provider and hospital when both are owned by the same vertically integrated healthcare system subject to HRRP. If the cost-effective effort is defined as in Equation 2 and there is a positive HRRP penalty with zero readmission reduction effort, then \( \bar{x}^{VI,HRRP} > \bar{x}^{CE} \).

Since the optimal vertically integrated effort with HRRP, \( \bar{x}^{VI,HRRP} \), is larger than the cost-effective effort level, \( \bar{x}^{CE} \), combining HRRP penalties with bundled payment plans re-introduces the mis-alignment between provider’s effort and cost-effective effort that vertical integration had eliminated. While reducing readmissions is a positive overall outcome, the resources wasted on excessive readmission reduction could be better allocated elsewhere.

To illustrate how HRRP penalties impact a bundled payment plan, Figure 4 depicts the readmission reduction effort with and without HRRP penalties under the same parameterization as Figure 3 except this time the charges for a readmission, \( \theta_h \), are $10,000. We assume that the HRRP penalty is convex and decreasing in readmission reduction effort similar to curves found in Figure 2. The left “Joint Response” bar represents the total effort of the hospital and post-discharge providers if they were separate organizations using the multi-provider bundled model. The middle “Vertically Integrated Effort” bar represents the total effort if the hospital and post-discharge providers were owned by one vertically integrated healthcare system. Finally, the right “Cost Effective Effort” bar represents the amount of effort a payer, such as Medicare, would like performed to minimize total costs and ensure cost-effective readmission reduction effort is performed (Equation 2).

The joint response effort is three (four) units above the cost-effective effort level without (with) HRRP which reinforces previous insights that state providers would achieve the cost-effective effort level, but could be over-motivated to perform more effort to improve their contribution margin. Comparing the joint response effort with and without HRRP also provides an example of when the effort increases by adding HRRP to bundled payment plans, as discussed in Theorem 4. Without HRRP, note that vertical integration aligns the total effort with the cost-effective CMS-desired effort level. However, adding HRRP causes the vertically integrated organization to perform excessive effort similar to the joint response effort, as discussed in Theorem 5.
Figure 4 Readmission avoidance effort levels for multiple providers, vertically integrated entities, and the cost-effective payer models, with and without HRRP penalties

Although we have claimed that over-motivation to perform excessive readmission reduction is inefficient from a cost perspective, it may have ancillary benefits for the hospital system in terms of reducing overall congestion and patient waiting times. In the next section, we show that these ancillary benefits are not sufficient to consider over-motivation a desirable outcome from a public policy perspective.

4.5. Bundled Payments Impact on Patient Throughput

From a cost standpoint, the phenomenon of over-motivation to perform excessive and inefficient readmission reduction effort is considered inappropriate. However, it may have additional benefits for the hospital system in terms of reduced congestion and patient wait times. In this section, we show that this additional readmission reduction benefit is not significant enough to consider over-motivation in a positive light. The figures in this section use the same parameterization as Figure 3 with additional values for the external patient arrival rate and service rates for the hospital and post-discharge providers. Figure 5 shows the impact of bundled payment price on readmission probability and hospital waiting times using a fluid model to capture the queueing network dynamics between hospitals and post-discharge care (Whitt 2006). Figure 5(a) shows that a higher bundle price eventually motivates the post-discharge provider to perform more effort, which reduces readmissions but that there are, as expected, diminishing marginal returns to this effort.

Figure 5(b) shows that increasing $\tau$ results in a decrease of the steady state waiting time of an at-capacity hospital. However, this decrease is unlikely to have a substantial impact on hospital operations: even large increases in $\tau$ provide minimal reduction in waiting time. This highlights the fact that over-motivation to excessive readmission reduction effort may carry few additional benefits relative to the cost of these efforts.
5. Aligning Incentives through Risk Profiled Bundled Payments

In previous sections, we illustrate various conditions when bundled payment plans motivate healthcare providers to exert readmission reduction effort that differs from cost-effective effort levels. To promote cost-effective effort, we formulate an alternative bundled payment plan that uses risk-adjusted bundled payment prices, \( \tau_{Lo} \) and \( \tau_{Hi} \), as opposed to current bundled payment plans with a single bundled payment price, \( \tau_{Ag} \). This means determining a risk metric and the dividing line between risk profiles needs to be part of the bundled payment negotiations between healthcare providers and payers. However, current Medicare readmission penalties, such as HRRP, utilize patient risk adjustment, so any patient above a certain risk score on this metric could be defined as high risk and given a high bundled payment price, \( \tau_{Hi} \). Various healthcare research studies, such as Helm et al. (2016) and Bardhan et al. (2014), also present readmission risk metrics that could be used to segment patients. Additionally, determining the value of one price to cover all patients could be more difficult than setting multiple prices for patients at different risk levels because providers will feel they are being better compensated for high risk patients, that are more likely to be readmitted.

Throughout Section 5, we refer to values under a single aggregate bundled payment price with an \( Ag \) subscript to clarify them from values with risk-adjusted bundled payment prices. Therefore, connecting this section to other sections in the paper, \( \tau_{Ag} = \tau \), \( \bar{x}^{CE}_{Ag} = \bar{x}^{CE} \), etc. In a similar manner, we will include the subscripts \( Lo \) and \( Hi \) on parameters to indicate low and high risk. In addition, we assume \( r_i^G \) and \( c_i^A \) are the same across low and high risk patients for provider \( i \).
We need to understand the conditions when a risk-profiled bundled payment plan is an improvement over the current aggregate bundled payment plan. To accomplish this, we compare the best response effort under risk-profiled and aggregate bundled payment prices to the cost-effective effort, $\bar{x}^\text{CE}_r$, where $r \in \{Hi, Lo\}$ is the risk level of the patient. In the remainder of this section, we only consider the post-discharge provider’s effort, since (from Theorem 2 and Figure 3) they are more likely to deviate from the cost-effective effort level due to changes in the negotiated bundled payment price, $\tau_t$, where $t \in \{Ag, Hi, Lo\}$ is the type of bundled payment price for aggregate, high-risk, and low-risk bundled payments.

Thus, we need to determine the best response effort of the post-discharge provider given the various negotiated bundled payment prices, $\tau_t$. Since providers know the risk-level of the patient, they will base their actions off the appropriate readmission probability curve, $p(\bar{x}_r)$, for patient risk level $r$. Therefore, for a low (or high) risk patient, the only difference in the revenue terms across various $\tau$ values is the effort difference motivated by the different $\tau$ values. The best response effort occurs when the derivative of the post-discharge revenue (from Equation 1) equals the cost of one unit of additional post-discharge effort, $c^A_{pd}$, as shown in Equation 10. We define $Z = Z(x) = r^G_h + r^A_h(x_h) + r^G_{pd} + r^A_{pd}(x_{pd})$ and suppress the $(x)$ for notational compactness.

$$\tau_t \left[ p(x_r) \lambda^r \left( \frac{r^G_{pd} + r^A_{pd}x_{pd}}{Z(\bar{x} + \theta)} \right) + p(x_r) \left( \frac{\theta r^A_{pd} [Z^2 - 2r^G_h Z - r^G_h \theta]}{Z^2(Z + \theta)^2} \right) + \frac{r^G_{pd} x_{pd}}{Z^2} \right] = c^A_{pd} \quad (10)$$

For simplicity of future discussion, we let the Equation 10 terms in brackets be $q_{pd,r}(x_{pd,r})$, as shown in Equation 11.

$$q_{pd,r}(x_{pd,r}) = \left[ p(x_r) \lambda^r \left( \frac{r^G_{pd} + r^A_{pd}x_{Lo}}{Z(\bar{x} + \theta)} \right) + p(x_r) \left( \frac{\theta r^A_{pd} [Z^2 - 2r^G_h Z - r^G_h \theta]}{Z^2(Z + \theta)^2} \right) + \frac{r^G_{pd} x_{pd}}{Z^2} \right] \quad (11)$$

Given the best response effort for bundled payment price $\tau_t$, $x^\tau_{pd,r} = (q_{pd,r}(c^A_{pd}/\tau_t))^{-1}$ and the cost-effective effort for risk level $r$, $x^\text{CE}_{pd,r}$, risk-profiling is a better option when the effort distance between these values is lower for the risk profiled plan. This occurs when $\alpha |x^\tau_{pd,Lo} - x^\text{CE}_{pd,Lo}| + (1 - \alpha) |x^\tau_{pd,Hi} - x^\text{CE}_{pd,Hi}| < \alpha |x^\tau_{pd,Lo} - x^\text{CE}_{pd,Lo}| + (1 - \alpha) |x^\tau_{pd,Hi} - x^\text{CE}_{pd,Hi}|$, where $\alpha$ is the percentage of low risk patients. Absolute value signs are required so that the positive distance is represented, no matter if excessive (positive) or insufficient (negative) effort is performed. This condition can be expanded, resulting in Condition 12. Under this
condition, using two risk-profiled bundled payment prices, \( \tau_{Hi} \) and \( \tau_{Lo} \), results in effort closer to cost-effective effort than using one aggregate bundled payment price, \( \tau_{Ag} \), when we assume \( q_{pd,r}(x_{pd,r}) \) is invertible.

\[
\alpha \left| q_{pd,Lo} \left( \frac{c_A}{\tau_{Lo}} \right) - x_{pd,Lo}^{CE} \right| + (1 - \alpha) \left| q_{pd,Hi} \left( \frac{c_A}{\tau_{Hi}} \right) - x_{pd,Hi}^{CE} \right| < \alpha \left| q_{pd,Lo} \left( \frac{c_A}{\tau_{Ag}} \right) - x_{pd,Lo}^{CE} \right| + (1 - \alpha) \left| q_{pd,Hi} \left( \frac{c_A}{\tau_{Ag}} \right) - x_{pd,Hi}^{CE} \right|
\]

(12)

Even with an exponential readmission probability, \( q_{pd,r}(x_{pd,r}) \) is still too complicated to obtain a closed-form inverse function. However, we know from extensive numerical analysis that both decreasing linear and exponential functional forms closely approximate \( q_{pd,r}(x_{pd,r}) \). First, we analyze the case where \( q_{pd,r}(x_{pd,r}) \) is linear in Theorem 6. Numerical analysis shows this to be a very good approximation for most parameter values, especially for low risk patients. We define \( \tau_{CE}^{Lo} \) as the bundled payment price that motivates the cost effective effort to be performed. We describe conditions based on cases of where negotiated bundled payment prices, \( \tau_i \), are set in relation to \( \tau_{CE}^{Lo} \). In both Theorem 6 and 7, there is a fourth relational case, \( \tau_{CE}^{Lo} < \tau_{Lo} < \tau_{Ag} < \tau_{Hi} < \tau_{CE}^{Hi} \), where by definition, risk-profiling is always closer to cost-effective effort.

**THEOREM 6.** Let \( \alpha \) be the percentage of low risk patients. Assume \( p(x_r) = \lambda^r e^{-\lambda^r x_r} \) and the post-discharge provider performs all additional effort. Assume \( \tau_{CE}^{Lo} < \tau_{Ag} < \tau_{CE}^{Hi} \) and \( q_{pd,r}(x_{pd,r}) \) is linear, such that \( q_{pd,Lo}(x_{pd,Lo}) = -m_{Lo}x_{pd,Lo} + b_{Lo} \) for low risk patients and \( q_{pd,Hi}(x_{pd,Hi}) = -m_{Hi}x_{pd,Hi} + b_{Hi} \) for high risk patients.

(i) When \( \tau_{Lo} < \tau_{CE}^{Lo} \) and \( \tau_{CE}^{Hi} < \tau_{Hi} \) (Case 1), if

\[
\alpha \left( \frac{1}{\tau_{Ag}m_{Lo}} - \frac{1}{\tau_{Lo}m_{Hi}} \right) - (1 - \alpha) \left( \frac{1}{\tau_{Hi}m_{Hi}} \right) < \frac{2}{c_A} \left[ (1 - \alpha) \left( x_{pd,Hi}^{CE} - \frac{b_{Hi}}{m_{Hi}} \right) + \alpha \left( \frac{b_{Lo}}{m_{Lo}} - x_{pd,Lo}^{CE} \right) \right],
\]

(13)

then using two risk-profiled bundled payment prices, \( \tau_{Hi} \) and \( \tau_{Lo} \), results in effort closer to cost-effective effort than using one aggregate bundled payment price, \( \tau_{Ag} \).

(ii) When \( \tau_{CE}^{Lo} < \tau_{Lo} \) and \( \tau_{CE}^{Hi} < \tau_{Hi} \) (Case 2), if

\[
\alpha \left( \frac{1}{\tau_{Ag}m_{Lo}} - \frac{1}{\tau_{Lo}m_{Hi}} \right) - (1 - \alpha) \left( \frac{1}{\tau_{Hi}m_{Hi}} \right) < \frac{2(1 - \alpha)}{c_A} \left( x_{pd,Hi}^{CE} - \frac{b_{Hi}}{m_{Hi}} \right),
\]

(14)

then using two risk-profiled bundled payment prices, \( \tau_{Hi} \) and \( \tau_{Lo} \), results in effort closer to cost-effective effort than using one aggregate bundled payment price, \( \tau_{Ag} \).
(iii) When $\tau_L < \tau_{CE,Lo}^{\text{Hi}}$ and $\tau_H < \tau_{CE,Hi}^{\text{Hi}}$ (Case 3), if
\[
\frac{\alpha}{\tau_{Ag,mLm}} + \frac{\alpha}{\tau_{Lo,mLm}} - \frac{(1 - \alpha)}{\tau_{Ag,mHi}} + \frac{(1 - \alpha)}{\tau_{Hi,mHi}} < \frac{2\alpha}{c_{pd}^{\text{Hi}}} \left[ \frac{b_{Lo}}{m_{Lo}} - x_{pd,Lo}^{CE} \right],
\] then using two risk-profiled bundled payment prices, $\tau_H$ and $\tau_L$, results in effort closer to
cost-effective effort than using one aggregate bundled payment price, $\tau_{Ag}$.

Second, we assume $q_{pd,r}(x_{pd,r})$ is exponential in Theorem 7, which is a good approximation
in parameter scenarios where post-discharge provider additional effort and general
charges are relatively large.

**Theorem 7.** Let $\alpha$ be the percentage of low risk patients. Assume $p(\bar{x}_r) = \lambda e^{-\lambda x_r}$ and
the post-discharge provider performs all additional effort. Assume $\tau_{CE,Lo}^{\text{Hi}} < \tau_{Ag} < \tau_{CE,Hi}^{\text{Hi}}$ and
$q_{pd,r}(x_{pd,r})$ is exponential, such that $q_{pd,Lo}(x_{pd,Lo}) = \beta_0^{\text{Lo}} e^{-\beta_0^{\text{Lo}} x_{pd,Lo}}$ for low risk patients and
$q_{pd,Hi}(x_{pd,Hi}) = \beta_0^{\text{Hi}} e^{-\beta_0^{\text{Hi}} x_{pd,Hi}}$ for high risk patients.

(i) When $\tau_L < \tau_{CE,Lo}^{\text{Hi}}$ and $\tau_{CE,Lo}^{\text{Hi}} < \tau_H$ (Case 1), if
\[
(\tau_H - \beta_0^{\text{Hi}})^{\alpha} - (\tau_{Lo} - \beta_0^{\text{Lo}})^{\alpha} + (\tau_{Ag} - \beta_0^{\text{Lo}})^{\alpha} < (c_{pd}^{\text{Hi}})^{2(1 \alpha^{\text{Hi}}) + e^{2(1 \beta_0^{\text{Lo}} -(1 \alpha^{\text{CE,Lo}}))}}
\] then using two risk-profiled bundled payment prices, $\tau_H$ and $\tau_L$, results in effort closer to
cost-effective effort than using one aggregate bundled payment price, $\tau_{Ag}$.

(ii) When $\tau_{CE,Lo}^{\text{Hi}} < \tau_L$ and $\tau_{CE,Hi}^{\text{Hi}} < \tau_H$ (Case 2), if
\[
(\tau_H - \beta_0^{\text{Hi}})^{\alpha} + (\tau_{Lo} - \beta_0^{\text{Lo}})^{\alpha} + (\tau_{Ag} - \beta_0^{\text{Lo}})^{\alpha} < (c_{pd}^{\text{Hi}})^{2(1 \alpha^{\text{Hi}}) + e^{2(1 \beta_0^{\text{Lo}} -(1 \alpha^{\text{CE,Hi}}))}}
\] then using two risk-profiled bundled payment prices, $\tau_H$ and $\tau_L$, results in effort closer to
cost-effective effort than using one aggregate bundled payment price, $\tau_{Ag}$.

(iii) When $\tau_L < \tau_{CE,Lo}^{\text{Hi}}$ and $\tau_H < \tau_{CE,Hi}^{\text{Hi}}$ (Case 3), if
\[
- (\tau_H - \beta_0^{\text{Hi}})^{\alpha} - (\tau_{Lo} - \beta_0^{\text{Lo}})^{\alpha} + (\tau_{Ag} - \beta_0^{\text{Lo}})^{\alpha} - (\tau_{Ag} - \beta_0^{\text{Lo}})^{\alpha} < (c_{pd}^{\text{Hi}})^{2(1 \alpha^{\text{Hi}}) + e^{2(1 \beta_0^{\text{Lo}} -(1 \alpha^{\text{CE,Lo}}))}}
\] then using two risk-profiled bundled payment prices, $\tau_H$ and $\tau_L$, results in effort closer to
cost-effective effort than using one aggregate bundled payment price, $\tau_{Ag}$.

In both Theorems 6 and 7, risk-profiling is a better option in scenarios when the difference
between $x_{pd,Hi}^{CE}$ and $x_{pd,Lo}^{CE}$ is larger. This makes intuitive sense, because if there is a large
difference in cost-effective effort between high risk and low risk patients, then there is
more benefit for risk-profiling, which allows risk-levels to be treated differently. Thus risk-profiled bundled payments are better for health episodes where low risk patients require minimal effort and high risk patients require significant effort. Similarly, risk profiling is better if there is a substantial portion of high risk patients. Thus when there is only a small percentage of high risk patients, risk profiling will not be as impactful because there are not enough high risk patients to make a significant difference in readmission rates.

The bundled payment prices (τ values) obviously have a large influence on the impact of risk-profiling also. However, they are case dependent based on where the negotiated τ values are in relation to each other and the cost-effective values. Obviously, risk-profiling is the best when negotiated τ values are the same as the cost-effective values. However, it can be difficult to know the true cost-effective value and this is where Theorems 6 and 7 provide some interesting insights.

For example, risk-profiled bundled payments are more likely to incentivize cost-effective readmission reduction efforts if the payout for high risk patients is more generous (slightly above the cost-effective payment price). This additional cost to the payer can be offset by setting a lower price for low risk patients, as in Case 1. Hence, if the payer wants to be more aggressive in their negotiations, it is best to focus on reducing the payment for low risk patients, τ_{Lo}, which results in Case 1 where risk-profiling is typically better. Case 1 is likely to be more agreeable to healthcare providers because a low τ_{Lo} value means little to no effort will be performed on low risk patients, but they will also rarely be readmitted and providers can focus on being more cost-effective with these patients. However, since τ_{Hi} is slightly above τ_{Hi}\text{CE} in Case 1, there is more money for readmission reduction effort and possible readmissions on high risk patients who are more likely to be readmitted. Therefore, providers can focus on more readmission prevention with these patients. These results extend to multiple risk levels, however the conditions become more complicated and there are many more cases to consider.

In summary, using two risk-adjusted bundled payment prices generally avoids insufficient effort on high risk patients and allows them to receive the additional readmission reduction effort they need. At the same time, risk-profiling reduces excessive effort on low risk patients, which means healthcare resources can be reallocated to more cost-effective initiatives.
6. Conclusions

As healthcare systems develop policies and reimbursement plans to reduce readmissions, they need to be understood how these actions will actually motivate various healthcare providers. We develop a contribution margin model to determine when various healthcare providers will be motivated to perform readmission reduction effort and how this effort compares to the cost-effective readmission reduction effort level.

This paper provides several contributions to healthcare literature. First, we show how bundled payment plans incentivize different providers in different ways to reduce readmissions. The good news is that one of the providers will be motivated to perform at least the cost-effective readmission reduction effort level. However, we also find that the post-discharge provider can be over-motivated to perform excessive effort because more effort results in a larger percentage of the bundled payment. This means the post-discharge provider can improve their bottom line at the expense of the hospital. In addition, this excessive effort is not cost-effective and is mis-allocating healthcare resources that could be better utilized elsewhere.

Second, this research should impact how bundled payments are designed and administered by illustrating that well-designed bundled payment reimbursement plans could replace HRRP penalties, while better motivating readmission reduction. This is especially true if the post-discharge provider has a lower cost of reducing readmissions than the hospital, since HRRP only impacts the hospital. Third, we develop a range on the bundled payment price that incentivizes a socially desirable cost-effective readmission reduction effort. Ideally, the bundled payment price would be set so that all healthcare providers in the bundle are motivated to perform readmission reduction effort, but not too high so that the post-discharge provider is motivated to perform excessive effort and possibly waste resources. Fourth, we provide insights to policy makers and payers by proposing a new risk-profiled approach to bundled payments that can better align a provider’s readmission reduction effort with the cost-effective effort preferred by payers.

In future research, it would be interesting to determine alternative methods to allocate the bonus/penalty from bundled payment reimbursement plans. These allocations could be based more on effort than revenue or they could include maximums or minimums to better moderate the advantage to a single provider. It would also be useful to connect our models to financial risk models to determine the risk exposure of various providers in...
bundled payment plans. This would help providers understand how much of their revenue is at risk and even create the need for new insurance to cover these risks.

References


Ginsburg PB (2016) Health care market consolidations: Impacts on costs, quality and access. California Legislature, Senate Committee on Health Informational Hearing; March, 16.


Appendix A: Detailed Proofs for Online Supplement

Proof for Theorem 1

We first prove part (iii), then part (ii), and finally combining these conditions to prove part (i). Let the initial effort be $x = [x_{pd}, x_h]$, where $\bar{x} = x_{pd} + x_h < x_{CE} = x_{pd}^* + x_h^{CE}$. Define $x_{CE} = [x_{pd}^*, x_h^{CE}]$, where $x_i^{CE}$ is the optimal (cost-effective) effort employed by provider $i$ in Equation 2. If Equation 19 is true for a given $x$, a provider could improve their contribution margin by performing more effort and returning the system to a cost-effective effort level.

$$E[\pi_i(x^{CE})] - E[\pi_i(x)] > 0$$ (19)

Substitute Equation 1 into Condition 19, and using the notation $RP_i(x)$ and $RP_i^h(x)$ to represent the revenue percentage of provider $i$ with and without a readmission, we get:

$$RP_i^h(x) \equiv \frac{r_i^G + r_i^A(x_i) + \theta_i}{\sum_k [r_i^G + r_i^A(x_k) + \theta_k]}$$

$$RP_i(x) \equiv \frac{r_i^G + r_i^A(x_i)}{\sum_k [r_i^G + r_i^A(x_k)]}$$ (20)

$$\tau \left(p(\bar{x}^{CE}) [RP_i^h(x^{CE}) - RP_i^h(x)] + (1 - p(\bar{x}^{CE})) [RP_i(x^{CE}) - RP_i(x)]\right)$$

$$+ [p(\bar{x}) - p(\bar{x}^{CE})] \left[\tau [RP_i(x) - RP_i^h(x)] + \gamma_i \right] - c_i^A(x^{CE} - x_h) + c_i^A(x_i) > 0$$ (21)

Part (iii) The first two terms of Condition 21 are always positive for the hospital because the revenue percentage for a provider that exerts more effort to achieve cost-effective effort, which increases the numerator and denominator equally, is always larger than the revenue percentage when that provider exerts less effort below the cost-effective level. Therefore, if the third, fourth, and fifth terms of Condition 21 are positive, then this is a sufficient condition for Condition 19 to be true. Rearranging these three terms we get:

$$\gamma_h > \tau [RP_i^h(x) - RP_i(x)] + \frac{c_h^A(\bar{x}^{CE} - x_{pd}) - c_h^A(x_h)}{p(\bar{x}) - p(\bar{x}^{CE})}$$ (22)

Note that $p(\bar{x}) - p(\bar{x}^{CE}) > 0$ and, for the hospital, $[RP_i(x) - RP_i^h(x)]$ will be negative because the second term is simply the first term with $\theta$ added to both the numerator and denominator.

Next, through the following propositions, we show that if Condition 22 is true at $x = 0$, then it is true for all $x$. Define $f_h(x_{pd}, x_h) = \gamma_h + \tau [RP_h(x_{pd}, x_h) - RP_h^h(x_{pd}, x_h)]$, $g_h(x_{pd}, x_h) = \frac{c_h^A(\bar{x}^{CE} - x_{pd}) - c_h^A(x_h)}{p(x) - p(\bar{x}^{CE})}$, and $h_h(x_{pd}, x_h) = f_h(x_{pd}, x_h) - g_h(x_{pd}, x_h)$. Thus, we can write Condition 22 as $h_h(x_{pd}, x_h) > 0$. In the following propositions, we use the notation $(-x_{pd}, x_h)$ to specify the case of $x_{pd}$ being decreased by some $\epsilon$ and $x_h$ increased by the same $\epsilon$.

**Proposition 1.** Assume $c_i^A(x_i)$ is linear. Then $g_h(x_{pd}, x_h) = g_h(x_{pd} + x_h, 0)$.

$$g_h(x_{pd}, x_h) = \frac{c_h^A(\bar{x}^{CE} - x_{pd}) - c_h^A(x_h)}{p(x_{pd} + x_h) - p(\bar{x}^{CE})} = \frac{c_h^A(\bar{x}^{CE} - x_{pd} - x_h) - c_h^A(0)}{p(x_{pd} + x_h) - p(\bar{x}^{CE})} = g_h(x_{pd} + x_h, 0)$$ (23)

□
Proposition 2. \( f_h(x_{pd}, x_h) = \gamma_h + \tau [R_{PD}(x_{pd}, x_h) - R_{PD}^h(x_{pd}, x_h)] \) is increasing in \((-x_{pd}, x_h)\).

Let \( Z(x_{pd}, x_h) = r_h^G + r_h^A(x_h) + r_{pd} \), where \( r_{pd} = r_{pd}^G + r_{pd}^A(x_{pd}) \)

\[
R_{PD}(x_{pd}, x_h) - R_{PD}^h(x_{pd}, x_h) = \frac{r_h^G + r_h^A(x_h)}{r_h^G + r_h^A(x_h) + r_{pd}} - \frac{r_h^G + r_h^A(x_h) + \theta}{r_h^G + r_h^A(x_h) + r_{pd} + \theta}
\]

\[
= \frac{(r_h^G + r_h^A(x_h))(Z(x_{pd}, x_h) + \theta) - (r_h^G + r_h^A(x_h) + \theta)Z(x_{pd}, x_h)}{Z(x_{pd}, x_h)(Z(x_{pd}, x_h) + \theta)} = \frac{-\theta(r_h^G + r_h^A(x_h))}{Z(x_{pd}, x_h)(Z(x_{pd}, x_h) + \theta)}
\]

Assuming \( r_h^A \geq r_h^A \), then \( Z(x_{pd}, x_h - \epsilon, x_h + \epsilon) \geq Z(x_{pd}, x_h) \), if we take the derivative of \( f \) on the diagonal in the direction \((-\epsilon, \epsilon)\), after some algebra this reduces to:

\[
\lim_{\epsilon \to 0} \frac{\tau r_h^A \theta}{Z(x_{pd}, x_h - \epsilon, x_h + \epsilon)Z(x_{pd}, x_h - \epsilon, x_h + \epsilon) + \theta} + \frac{\tau \theta(r_h^A - r_h^A)[r_h^G + r_h^A(x_{pd})][2Z(x_{pd}, x_h) + \epsilon(r_h^A - r_h^A) + \theta]}{Z(x_{pd}, x_h - \epsilon, x_h + \epsilon)Z(x_{pd}, x_h - \epsilon, x_h + \epsilon) + \theta} Z(x_{pd}, x_h)Z(x_{pd}, x_h) > 0
\]

\[\square\]

Proposition 3. \( h_h(y, 0) \) is increasing in \( y \).

First, \( h_h(y, 0) = f_h(y, 0) - g_h(y, 0) \), before effort is done by the hospital to achieve \( x^{CE} \). Thus,

\[
f_h(y, 0) = \gamma_h + \tau \left[ \frac{r_h^G}{r_h^G + r_h^A(y) + r_{pd}(y) + \theta} \right]
\]

\[
g(y, 0) = \frac{c_h^A(x^{CE} - y)}{p(\bar{x}) - p(x^{CE})}
\]

The partial derivative of \( f_h(y, 0) \) with respect to \( y \) as shown below is negative for small \( y \) values, but will become positive as \( y \) grows.

\[
\frac{\partial f_h(y, 0)}{\partial y} = \tau r_h^A \left[ \frac{r_h^G + \theta}{r_h^G + r_h^A(y) + \theta} - \frac{r_h^G}{r_h^G + r_h^A(y) + r_{pd}(y) + \theta} \right]
\]

Recall we assume that \( p(\bar{x}) \) is convex and decreasing in \( \bar{x} \). Also, define \( p^\Delta = p(\bar{x}) - p(x^{CE}) \). Therefore, the partial derivative of \(-g_h(y, 0)\) with respect to \( y \) as shown below creates two positive terms because \( p^\Delta > 0 \)

\[
- \frac{\partial g_h(y, 0)}{\partial y} = - \left[ -c_h^A \frac{p^\Delta - p^\Delta c_h^A(x^{CE} - y)}{(p^\Delta)^2} \right] = c_h^A \frac{p^\Delta - p^\Delta c_h^A(x^{CE} - y)}{(p^\Delta)^2} > 0
\]

The highest magnitude negative term in Equation 28 occurs when \( y = 0 \), so we use this scenario in Condition 30. Note that the magnitude of this term is less than the magnitude of the positive terms in

\[
\frac{c_h^A}{p^\Delta} + \left( -p^\Delta \right) c_h^A \frac{\bar{x}^{CE}}{(p^\Delta)^2} + \tau r_h^A \left[ \frac{r_h^G}{r_h^G + r_h^A(y) + r_{pd}(y) + \theta} - \frac{r_h^G + \theta}{r_h^G + r_h^A(y) + r_{pd}(y) + \theta} \right]
\]

Let \( \theta \leq r_h^G + r_{pd} \) and perform some algebra, so Condition 30 can be reduced to

\[
\tau \leq \frac{2c_h^A |r_h^G + r_{pd}|^2}{r_{pd}^2 p^\Delta r_h^G} \]

\[\]
Thus, as long as the technical assumption in Condition 31 is true (which it is for any realistic parameter values), the positive terms will always dominate, making the derivative positive, and \( h_h(y,0) \) increasing in \( y \). □

If \( h_h(x_{pd}, x_h) > 0 \) then the profit from returning to cost-effective effort is higher than the profit at the point \((x_{pd}, x_h)\). We know that 
\[
h_h(x_{pd}, x_h) \geq h_h(x_h + x_{pd}, 0) \geq h_h(0,0)
\]

(32)

The first inequality holds from Propositions 1 and 2. The second inequality holds from Proposition 3. Thus we only need to show that \( h_h(0,0) > 0 \) for Theorem 1 to hold, which is precisely Condition 5, which proves part (iii) of Theorem 1. Next we prove part (ii) of Theorem 1 for the post-discharge provider. □

Part (ii) Part (ii) uses the same approach as part (iii), as such we suppress much of the details. For post-discharge, \( \gamma_{pd} = 0 \). In Condition 21 for the post-discharge provider, the first two \( \tau \) terms are always positive because the \( x^{CE} \) term with more effort in both the numerator and denominator is larger than the \( x \) term being subtracted from it. The \( RP_{pd}(x) - RP_{pd}^g(x) \) subtraction pair in the third term is also positive, because the only difference between these values is \( \theta_h \), which is only in the denominator, since \( \theta_{pd} = 0 \). Ignoring the obviously positive terms, as we did for the hospital, we get the following sufficient condition for post-discharge provider to meet cost-effective efforts.

\[
\tau \left[ RP_{pd}(x) - RP_{pd}^g(x) \right] > \frac{c^A_{pd}(\bar{x}^{CE} - x_h) - c^A_{pd}(x_{pd})}{p(\bar{x}) - p(\bar{x}^{CE})}
\]

(33)

Next, we show that if Condition 33 is true at \( x = 0 \), then it is true for all \( x \). Using similar functions for \( f, g, \) and \( h \) and the same technique, we can show the same properties of \( f, g, \) and \( h \) exist, with the algebra left to the reader for brevity.

Using this approach we conclude that 
\[
h_{pd}(x_{pd}, x_h) > 0
\]

\[
h_{pd}(x_{pd}, x_h) \geq h_{pd}(0, x_h + x_{pd}) \geq h_h(0,0)
\]

(34)

Hence the profit from returning to cost-effective effort is higher than the profit at the point \((x_{pd}, x_h)\). Thus, we only need to show that \( h_{pd}(0,0) > 0 \) for the theorem to hold, which is precisely Condition 4, which proves part (ii) of Theorem 1. □

Part (i): This follows directly from examining the overlap of Conditions 6 and 7. Specifically, we have

\[
\left[ \frac{c^A_{pd}(\bar{x}^{CE})}{p(0) - p(\bar{x}^{CE})} \right] < \left[ \frac{c^A_{h} (\bar{x}^{CE})}{p(0) - p(\bar{x}^{CE})} \right]
\]

(35)

\[
\implies c^A_{pd}(\bar{x}^{CE}) + c^A_{pd}(x^{CE}) < \gamma_h \left[ p(0) - p(\bar{x}^{CE}) \right]
\]

(36)

□

Proof for Theorem 2:

Summary of Proof. Recall that \( Z(x) = r_{pd}^G + r_{pd}^A(x_{pd}) + r_{h}^G + r_{h}^A(x_h) \). Let 
\( Rev_i = \tau (RP_i + p(\bar{x})(RP_i^g - RP_i)) - p(\bar{x})\gamma_i \). Given that \( Rev_i \) is either decreasing or unimodal (increasing then decreasing), we know that the best response will occur as the \( Rev_i \) term approaches the cost, \( c^A_i \) from above. In the unimodal case, this is guaranteed because if \( Rev_i \) approaches \( c^A_i \) below this must not be optimal, since doing zero effort because would result in higher profit because prior to reaching \( c^A_i \), the
marginal revenue would be smaller than the marginal cost. Thus, the second time \( Rev'_i \) hits \( c_i^A \), it will be from above, which implies that either this point or zero will be optimal. Given that the best response will have \( Rev'_i \) approaching from above, we see that if \( Rev'_h < Rev'_pd \) then \( Rev'_h \) will hit \( c_i^A \) at a lower effort level than \( Rev'_pd \) since the \( Rev'_pd \) curve will be above the \( Rev'_h \) curve. Thus the best effort from the post-discharge provider must be greater at the best effort point. Specifically, if \( Rev'_pd - Rev'_h > 0 \) for all \( x \) then it is true for the best response effort level that \( x_{pd}^R \geq x_{h}^R \). However, the result continues to hold when \( c_{pd}^A \leq c_{h}^A \), as is typically the case in practice. For notational simplicity, we will write the functions \( Rev(x) \), \( Z(x) \), \( RP_i^\theta(x) \) and \( RP_i(x) \) as \( Rev'_i \), \( Z \), \( RP_i^\theta \) and \( RP_i \).

**Post-discharge.** Taking the derivative for the post-discharge provider:

\[
Rev'_pd = \tau \left( p'(\bar{x})(RP_{pd}^\theta - RP_{pd}) + p(\bar{x})(RP_{pd}^\theta - RP_{pd})' + RP_{pd}' \right).
\]

\[
RP_{pd}(x_{pd}, 0) = \frac{r_{G_{pd}}^G + r_{A_{pd}}^A(x_{pd})}{Z} \quad RP_{pd}'(x_{pd}, 0) = \frac{r_{G_{pd}}^G r_{A_{pd}}^A}{Z^2} \tag{38}
\]

\[
[RP_{pd}^\theta(x_{pd}, 0) - RP_{pd}(x_{pd}, 0)]' = \frac{\theta [r_{pd}^G + r_{A_{pd}}^A(x_{pd})] r_{A_{pd}}^A (2Z + \theta) - \theta r_{pd}^A (Z^2 + \theta Z)}{Z(Z + \theta)^2}
\]

\[
= \frac{\theta r_{pd}^A (2Z + \theta)(Z - r_{pd}^A) - Z^2 - Z\theta}{Z(Z + \theta)^2} = \frac{\theta r_{pd}^A (Z^2 - 2Z^2 - Z\theta)}{Z(Z + \theta)^2} = \frac{\theta r_{pd}^A}{Z^2} \left( 1 - \frac{2r_{pd}^G}{Z} - \frac{\theta r_{pd}^A}{Z^2} \right) \tag{39}
\]

Substituting Equations 38 - 39 into Equation 37 and rearranging terms yields:

\[
Rev'_pd = \tau \left[ p(\bar{x}) \left( \tau \left( Z - r_{pd}^G \theta \right) \right) + p(\bar{x}) \left( \frac{\theta r_{pd}^A (Z^2 - 2Z^2 - Z\theta)}{Z^2} \right) \right] \tag{40}
\]

**Hospital.** Taking the derivative for the hospital provider:

\[
Rev'_h = \tau \left( RP_h^\theta(x_{pd}^R) + p'(\bar{x}) [RP_h^\theta(x_{pd}^R) - RP_h(x_{pd}^R)] + p(\bar{x}) [RP_h^\theta(x_{pd}^R) - RP_h(x_{pd}^R)]' \right) - \gamma_h p'(\bar{x}). \tag{41}
\]

\[
RP_h(0, x_{h}) = \frac{r_{G_{h}}^G + r_{A_{h}}^A(x_{h})}{Z} \quad RP'_h(0, x_{h}) = \frac{r_{G_{h}}^G r_{A_{h}}^A}{Z^2} > 0 \tag{42}
\]

\[
[RP_h^\theta(0, x_{h}) - RP_h(0, x_{h})]' = \frac{-\theta r_{pd}^G r_{A_{h}}^A}{Z(Z + \theta)} \left( \frac{1}{Z} + \frac{1}{Z + \theta} \right) = \frac{-\theta r_{pd}^G r_{A_{h}}^A (2Z + \theta)}{Z(Z + \theta)^2} < 0 \tag{43}
\]

Substitute Equations 42 - 43 into Equation 41, resulting in Equation 44.

\[
Rev'_h = \tau \frac{r_{pd}^G r_{A_{h}}^A}{Z^2} - \lambda p(\bar{x}) \left( \frac{\theta r_{pd}^G}{Z(Z + \theta)} - \gamma_h \right) - \tau p(\bar{x}) \frac{\theta r_{pd}^G r_{A_{h}}^A (2Z + \theta)}{Z(Z + \theta)^2} \tag{44}
\]

**Derivation of Condition 8:**

Given Equations 40 and 44 above, \( Rev'_pd - Rev'_h > 0 \), results in
\[ \text{Rev}'_{pd} - \text{Rev}'_h = \frac{\tau}{Z^2} \left( r_G^{G,A} - r_G^{G,A} r_h^A \right) + \tau \theta p(\bar{x}) \left[ \frac{2n_G^{G,A} + r_A^{G,A}(x_{pd})}{Z(\bar{Z} + \theta)} \right] - \gamma \lambda p(\bar{x}) \]

\[ + \frac{\tau \theta p(\bar{x})}{Z + \theta} \left[ \frac{r_A^{G,A} - 2n_G^{G,A} r_h^G}{Z(\bar{Z} + \theta)} - \frac{\theta r_A^{G,A}}{Z(\bar{Z} + \theta)} + \frac{r_G^{G,A}}{Z(Z + \theta)} \right] > 0 \quad (45) \]

After algebraic manipulation, Condition 45 results in Condition 46

\[ p(\bar{x}) \frac{r_A^{G,A}}{(Z + \theta)^2} + p(\bar{x}) \frac{r_A^{G,A} - 2n_G^{G,A} r_G^{G,A}}{Z} > \lambda \left( \frac{Z}{\tau} - \theta \frac{2n_G^{G,A} + r_A^{G,A}(x_{pd})}{Z(\bar{Z} + \theta)} \right) \]

With the assumption \( r_A^{G,A} = r_A^{G,A} \), we can pull out \( r_A^{G,A} \) to better isolate the difference in revenue size of providers. While assuming \( r_A^{G,A} \) values are equal is somewhat restrictive, the values should be similar in practice and it helps with intuitive understanding. We also move \( \tau \) to the left-hand-side. Therefore, Condition 46 becomes Condition 47.

\[ \tau r_A^{G,A} \left( \frac{p(\bar{x}) \frac{\theta h + r_A^{G,A} - 2n_G^{G,A} r_G^{G,A}}{(Z + \theta)^2} + [1 - p(\bar{x})] \frac{r_A^{G,A} - 2n_G^{G,A} r_G^{G,A}}{Z^2} \right) > \lambda p(\bar{x}) \left( \frac{Z}{\tau} - \theta \frac{2n_G^{G,A} + r_A^{G,A}(x_{pd})}{Z(\bar{Z} + \theta)} \right) \]

**Technical Note with \( y \) effort:** If the other provider performs \( y \) units of effort (instead of zero units), then Equation 47 only has one small change. The \( +r_A^{G,A}(x_{pd}) \) term becomes \( +r_A^{G,A}(x_{pd} + y) \). \( \square \)

**Proof for Theorem 3:**

The vertically integrated entity looks to maximize the fixed payment \( \tau \) minus the cost of all providers performing care \( c_i(\bar{x}) \), as shown in Equation 48. In this case, the \( K \) providers are various provider groups with different cost structures within one vertically integrated organization.

\[ \max_x \left[ E[\pi^{+J}(\bar{x})] \right] = \max_x \left[ \tau - \left( \sum_{k=1}^{K} c_k^G + c_k^A(x_{pd}) + p(\bar{x}) \gamma_k \right) \right] \quad (48) \]

Because \( \tau \) is a constant, maximizing Equation 48 is equivalent to minimizing Equation 2. \( \square \)

**Proof for Theorem 4:**

By definition \( x_B^H(x_{pd}) \) is the best response of the hospital without any HRRP penalty, given post-discharge provider effort of \( x_{pd} \). Therefore, the contribution margin of the hospital without HRRP is maximized at \( \pi_h(\bar{x}_{pd}, x_B^H(x_{pd})) \).

From Equation 9, we also know that \( \pi_h^{HRRP}(\bar{x}) = \pi_h(\bar{x}) - HRRP(\bar{x}) \), which is simply subtracting the HRRP penalty from the contribution margin of the hospital without HRRP.

Since \( HRRP(\bar{x}) \geq 0 \) and decreasing in \( \bar{x} \), if \( x_{pd}^{H+HRRP} < x_{pd}^H(\bar{x}) \), then \( \pi_h(\bar{x}_{pd}, x_{pd}^{H+HRRP}(\bar{x}_{pd})) \leq \pi_h(\bar{x}_{pd}, x_B^H(x_{pd})) \) because \( x_B^H \) is the optimal solution to \( \pi_h \). Further \( HRRP(\bar{x}_{pd} + x_B^H(x_{pd})) > HRRP(\bar{x}_{pd} + x_B^H(x_{pd})) \), because \( HRRP(\bar{x}) \) is decreasing in \( \bar{x} \) and we assume \( x_{pd}^{H+HRRP} < x_B^H(\bar{x}_{pd}) \).

Therefore, \( \pi_h(\bar{x}_{pd}, x_{pd}^{H+HRRP}(\bar{x}_{pd})) - HRRP(\bar{x}_{pd} + x_{pd}^{H+HRRP}(\bar{x}_{pd})) < \pi_h(\bar{x}_{pd} + x_B^H(\bar{x}_{pd})) - HRRP(\bar{x}_{pd} + x_B^H(\bar{x}_{pd})) \). This contradicts the fact that \( x_{pd}^{H+HRRP}(\bar{x}_{pd}) \) is an optimal solution to \( \pi_{h}^{HRRP}(\bar{x}) \). Thus, \( x_{pd}^{H+HRRP}(\bar{x}_{pd}) \geq x_B^H(\bar{x}_{pd}) \). \( \square \)
Proof for Theorem 5:

Recall that we assume effort costs are linear in readmission reduction effort, \( x \), and the readmission probability, \( p(\bar{x}) \), is convex in total effort, \( \bar{x} \). \( HRRP(\bar{x}) \) reacts in a similar manner to the readmission probability curve and is therefore also convex in total effort. Let \( c^{A,Tot}(\bar{x}) \) be the cost of the total effort, allowing us to rewrite Equation 2 into Equation 49 below.

\[
\min_{x} \left[ E[\pi^{CE}(x)] \right] = \sum_{k=1}^{K} c^G_k + \min_{x} \left[ c^{A,Tot}(\bar{x}) + \sum_{k=1}^{K} p(\bar{x}) \gamma_k \right]
\] (49)

From the first order conditions of Equation 49 for the cost-effective effort level, the optimal effort \( \bar{x}^{CE} \) is found where

\[
c^{A,Tot}(\bar{x}^{CE}) = -p'(\bar{x}^{CE}) \gamma_k
\] (50)

However, including the HRRP penalties with the vertically integrated contribution margin from Equation 48 results in Equation 51.

\[
\max_{x} \left[ E[\pi^{VI,HRRP}(x)] \right] = \max_{x} \left[ \tau - \left( \sum_{k=1}^{K} [c^G_k + c^A_k(x_k) + p(\bar{x}) \gamma_k + HRRP(\bar{x})] \right) \right]
\] (51)

Thus, an optimal effort level of \( \bar{x}^{VI,HRRP} \) is found where

\[
c^{A,Tot}(\bar{x}^{VI,HRRP}) + HRRP'(\bar{x}^{VI,HRRP}) = -p'(\bar{x}^{VI,HRRP}) \gamma_k
\] (52)

Since \( HRRP'(\bar{x}) < 0 \),

\[
c^{A,Tot}(\bar{x}^{CE}) + HRRP'(\bar{x}^{CE}) < -p'(\bar{x}^{CE}) \gamma_k
\] (53)

The left-hand-side terms are non-decreasing in readmission effort \( \bar{x} \) and the right-hand-side term is decreasing in readmission effort \( \bar{x} \). Thus to make Condition 53 an equality again at the optimal effort with HRRP penalties it must be that \( \bar{x}^{VI,HRRP} > \bar{x}^{CE} \). \( \square \)

Proof for Theorem 6:

Substitute \( q_{pd,Lo}(x_{pd,Lo}) = -m_{Lo}x_{pd,Lo} + b_{Lo} \) for low risk patients and \( q_{pd,Hi}(x_{pd,Hi}) = -m_{Hi}x_{pd,Hi} + b_{Hi} \) for high risk patients into Condition 12 and adjust the signs on the amount of excessive or insufficient effort to match the conditions in Cases (i), (ii), and (iii). Finally, perform algebra to rearrange terms, resulting in the case-specific conditions shown. \( \square \)

Proof for Theorem 7:

Substitute \( q_{pd,Lo}(x_{pd,Lo}) = \beta_{Lo}^{pd} e^{-\beta_{Lo}^{pd} x_{pd,Lo}} \) for low risk patients and \( q_{pd,Hi}(x_{pd,Hi}) = \beta_{Hi}^{pd} e^{-\beta_{Hi}^{pd} x_{pd,Hi}} \) for high risk patients into Condition 12 and adjust the signs on the amount of excessive or insufficient effort to match the conditions in Case (i), (ii), and (iii). Finally, perform algebra to rearrange terms, resulting in the case-specific conditions shown. \( \square \)