Healthcare Reimbursement Policy Impact on Multiple-Provider Readmission Reduction Programs

(Authors’ names blinded for peer review)

In many new alternative healthcare reimbursement plans, readmissions are treated as non-value added quality failures. We examine if and how these alternative reimbursement plans (such as bundled payments and value-based billing) motivate healthcare providers to reduce readmissions. We develop analytical models of provider behavior under different bundled payment models, where healthcare providers are paid a fixed payment for an entire episode of care. We find that, while bundled payments incentivize providers to perform effort that is at least cost-effective, competition in a gain-sharing bundled payment plan can cause excessive effort that more closely resembles fee-for-service payment models. Further, we show that the smaller provider in a bundle is highly sensitive to price, and based on their cost structure they may regularly perform either insufficient or excessive effort, which occurs in 81% of the scenarios in our numerical study, even for the same bundled payment price. This implies that pricing alone cannot close this gap. Instead, a structural redesign of bundled payment models is required. We discuss two redesigned structures to address the issues with gain-sharing bundled payment models. First, we show that a model with a single-controlling provider will always align with the payer-desired cost-effective readmission reduction effort, though combining any bundled payment model with value-based billing initiatives again incentivizes excessive effort. Second, we show that risk-adjusted bundled payment models can also significantly improve cost-effective care delivery. In a large-scale numerical study, we show that these new models can result in substantial cost savings per contract.

Key words: Alternative healthcare reimbursement policy, hospital readmissions, bundled payment plans, readmission reduction effort, value-based billing

1. Introduction

Policy makers and payers are developing alternative reimbursement models that transfer the financial risk of readmissions to healthcare providers, which has increased readmissions concern among these providers (Cheney 2016, Guduguntla et al. 2018, McCarthy and Pandey 2018, Executive-Team 2019). 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 $26 billion in 2014 (Rau 2014). Researchers and policy makers believe that approximately 75% of Medicare readmissions are potentially preventable
Bundled payment plans are an increasingly prevalent alternative reimbursement model currently being tested and developed by the Center for Medicare and Medicaid Services (CMS). Under bundled payment plans, healthcare providers are paid a fixed payment ($\tau$ in this paper) 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 CMS. 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, $\tau$, as a bonus for achieving more cost efficient quality care (CMS 2015).

Bundled payment plans introduce interesting dynamics among healthcare providers, especially as originally formulated in a gain-sharing arrangement between CMS and multiple healthcare providers in Bundled Payment-Classic Models 2 and 4, which are discussed in more detail in Section 1.1. The healthcare industry typically thinks of gain-sharing as payments from hospitals to physicians. However, in line with the general definition, we expand the definition of gain-sharing in this paper, to be when provider participants in a bundled payment plan determine how much effort they exert to optimize their individual contribution margin as they 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 exceed the bundled payment price, however, providers may prefer a lower percentage of charged revenue since the “penalty” is split similarly among providers. We find that this tension can cause misalignment between performed readmission reduction effort and the cost-effective effort level. Interestingly, recent bundled payment plans (see Bundled Payment-Advanced and Bundled Payment-CJR in Section 1.1) have moved away from gain-sharing and focus more on a single controlling healthcare provider, which our analysis and results show will better motivate cost-effective readmission reduction effort.

Other alternative reimbursement models include value-based billing (VBB) programs. These programs are based on annual performance rather than individual episodes of care and can incur penalties, such as the Hospital Readmission Reduction Program (HRRP), as well as penalties and incentives, such as value-based purchasing (VBP) programs. One goal of most alternative
reimbursement plans is to motivate healthcare providers to engage in cost-effective readmission reduction effort (Cheney 2016, Guduguntla et al. 2018). For example, from discussions with home health executives, we found that hospitals and physician groups operating under bundled payment plans are pushing patient care from skilled nursing facilities to home health providers because of lower costs and low readmission rates (Executive-Team 2019).

1.1. Bundled Payment Models and Value-Based Billing (VBB) Programs
Alternative reimbursement plans are being implemented across the United States and around the world by both public and private payers (CMS 2015, Porter et al. 2016, Fan et al. 2018, Guo et al. 2019, Struijs and Baan 2011). In this section, we provide a summary of key CMS bundled payment models and VBB programs that motivated our research. There are similar models from private payers and other countries also (Porter et al. 2016, Spinks et al. 2018). The Bundled Payment-Classic models are from the initial CMS trial of bundled payment plans from roughly 2013-2019, while Bundled Payment-Advanced and Bundled Payment-CJR are relatively new CMS bundled payment initiatives (CMS 2015). Bundled payment models fall into two main categories: multiple provider gain-sharing (initial models) and now transitioning to single provider with no gain-sharing with CMS. Gain-sharing involves splitting the bonus/penalty among multiple providers whereas in non gain-sharing models the single provider accepts sole financial responsibility for the bonus/penalty. Our research analyzes all the models described below, expect Model 1, which does not include readmissions. We show that these bundled payment structures are key to aligning (or not) incentives for cost-effective readmission reduction. Table 1 at the end of this section summarizes the models below.

Bundled Payment-Classic Model 1 (Acute Care Hospital Stay Only): The episode of care is the inpatient stay in an acute care hospital. CMS pays the hospital a discounted amount based on the Inpatient Prospective Payment System payment rates. This model does not include readmissions.

Bundled Payment-Classic Model 2 (Acute & Post-Acute Care Episode): In this gain-sharing model, CMS makes fee-for-service (FFS) payments to multiple providers for services that include the inpatient stay, post-acute care, and all related services (including readmissions) during the episode of care, which can be 30, 60, or 90 days after hospital discharge. Total charges are later reconciled against a bundled payment price, resulting in either a bonus or penalty that is allocated proportionally to each provider’s total charges.

Bundled Payment-Classic Model 3 (Post-Acute Care Only): This model is similar in structure to BP-Classic 2, except that the episode of care begins after discharge from an inpatient hospital stay. Services include skilled nursing, inpatient rehabilitation, long-term care, home health,
or physician services. This model also includes readmissions and potential *gain-sharing* between multiple providers.

**Bundled Payment-Classic Model 4 (Prospective Acute Care Hospital Stay Only):** CMS makes a single, prospectively determined bundled payment that encompasses all services furnished by the hospital, physicians, and other providers and includes related readmissions within 30 days of discharge from the hospital. Here, the hospital controls the payments to physicians and other providers based on “no-pay” FFS claims submitted to CMS, which mimics *gain-sharing* between multiple providers.

**Bundled Payment-Advanced:** Similarly to BP-Classic 2 and 3, CMS makes a retrospective bonus or penalty payment after comparing total charges to the bundled payment price. However, the contract is with a *single provider*, either an acute care hospital or large physician group, that is solely responsible for the inpatient stay plus the 90 days after discharge from the hospital. The single provider may, at their discretion, either enter a gain-sharing agreement with other providers or assume full fiscal responsibility for the bundle.

**Bundled Payment-Comprehensive Care for Joint Replacement (BP-CJR):** Like BP-Advanced, a single provider (a hospital in this case) is financially accountable for a joint replacement episode of care, which includes the inpatient stay plus the 90 days post-discharge. The hospital can assume full fiscal responsibility for the bundle or may, at their discretion, enter into a gain-sharing agreement with other providers. Different from the other models, CMS allows simple *risk stratification* for hip fractures, setting different bundled payment prices dependent on the presence of a fracture.

**Value-Based Billing (VBB) Programs:** In contrast to bundled payment programs, VBB programs typically apply a bonus or penalty based on overall annual performance. HRRP penalizes hospitals up to 3% of their total Medicare reimbursement for excessive readmissions (Zhang et al. 2016). Value-based purchasing (VBP) programs for hospitals, skilled nursing facilities, and (proposed) home health providers, withhold 2% of Medicare payments to fund value-based incentive payments based on performance metrics that include readmissions.

<table>
<thead>
<tr>
<th>Program</th>
<th>Payment Approach</th>
<th>Episode Coverage</th>
<th>Providers in Bundle</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP 1</td>
<td>Single provider</td>
<td>Inpatient</td>
<td>Hospital</td>
</tr>
<tr>
<td>BP 2</td>
<td>Gain-sharing (Retro.)</td>
<td>Inpatient, post-discharge 30, 60, 90 days</td>
<td>Hospital, post-discharge</td>
</tr>
<tr>
<td>BP 3</td>
<td>Gain-sharing</td>
<td>Post-discharge 30, 60, 90 days</td>
<td>Hospital, post-discharge</td>
</tr>
<tr>
<td>BP 4</td>
<td>Gain-sharing (Prosp.)</td>
<td>Inpatient, post-discharge 30, 60, 90 days</td>
<td>Hospital, post-discharge</td>
</tr>
<tr>
<td>BP-Adv</td>
<td>Single provider</td>
<td>Inpatient, post-discharge 90 days</td>
<td>Hospital or physician group</td>
</tr>
<tr>
<td>BP-CJR</td>
<td>Single, risk-adjusted</td>
<td>Inpatient, post-discharge 90 days</td>
<td>Hospital</td>
</tr>
<tr>
<td>HRRP</td>
<td>Penalty</td>
<td>30 day readmissions</td>
<td>Hospital</td>
</tr>
<tr>
<td>VBP</td>
<td>Bonus or Penalty</td>
<td>Quality (including readmissions)</td>
<td>Hospital, post-discharge</td>
</tr>
</tbody>
</table>

Table 1 Summary of CMS bundled payment and VBB programs (Note: Single provider payment approaches may include separate gain-sharing agreements with other providers at their discretion.)
In this paper, we analyze the impact of these different bundled payment models on readmission reduction effort, discussing relevant literature in Section 2. In Section 3, we develop a general contribution margin model for providers operating within a bundled payment plan. We begin by highlighting some of the drawbacks of gain-sharing models such as BP-Classic Models 2 and 4 in Section 4.1. We discuss how some of the newer payment models, such as BP-Advanced and BP-CJR, mitigate the drawbacks of the older models, in Section 4.2. However, in Section 4.3 we show that even these newer models can lead to mis-aligned incentives when combined with VBB programs. In Section 5 we discuss the benefits and argue for broader use of risk-stratified models such as BP-CJR. We present insights from an extensive numerical study in Section 6 and conclude in Section 7.

1.2. Research Questions and Contributions
In this paper, we analyze if (and under what conditions), alternative reimbursement plans incentivize a cost-effective approach to readmission reduction. We focus our analysis on hospitals (including physician groups) and post-discharge providers to determine how they interact and how their interactions impact the motivation to reduce readmissions. We specifically study:

- Will bundled payment plans motivate cost-effective care delivery with respect to readmissions?
- Which types of providers respond best to different types of bundled payment models?
- What is the effect of combining VBB programs with bundled payment plans?
- What bundled payment structure can best motivate cost-effective readmission reduction?

To the best of our knowledge, this paper is one of the first in the operations management literature to study the impact of bundled payment and VBB programs on the joint efforts and behaviors of multiple providers to reduce readmissions. Our model simultaneously considers (1) providers working together to reduce readmissions, and (2) competing with one another for a fixed payout. In this article, we develop and analyze an analytical model of these joint efforts based on provider contribution margin functions. These functions are ill-behaved due to the split of the bundled payment bonus/penalty that is based both on the fraction of revenue charged by each individual provider as well as the total joint provider efforts. As such, traditional optimization and analysis methods cannot be applied.

We initially find that bundled payment plans are capable of motivating providers to perform readmission reduction effort that, at a minimum, achieves cost-effective levels. However, we find that competition for the bundled payment often causes providers to return to fee-for-service (FFS) type behavior that bundled payments were designed to avoid. Further, we show that this excessive effort is primarily motivated by increasing their contribution margin and provides little benefit to patient quality of care or hospital operations. We prove that the smaller post-discharge provider is highly sensitive to the bundled payment price and is therefore far more likely to behave in this
FFS manner (Proposition 1). This price sensitivity, which also depends on provider cost structures, makes it very difficult to align incentives in a gain-sharing bundle with pricing alone. Interestingly, we show that providers are more profitable when they perform cost-effective effort (Proposition 2) leading to a payer-provider win-win, though the current competition for the bundled payout often causes providers to deviate from this preferred solution. As such, structural changes to bundled payment models may be required to both achieve payer efficiency goals and improve provider profitability.

Figure 1 provides a conceptual visualization of insufficient and excessive effort compared with cost-effective effort as a function of the bundled payment price, $\tau$. Here, the post-discharge provider (dashed line) performs insufficient readmission reduction effort when the bundled payment price is low. As the price increases, the post-discharge provider quickly moves toward excessive effort (Theorem 1 and Proposition 2). Conversely, the hospital is robust to pricing and performs near cost-effective effort for a wide range of bundled payment prices (Theorem 1).

![Diagram](image)

**Figure 1** Impact of increasing bundled payment price, $\tau$, on the readmission reduction effort of healthcare providers under a bundled payment plan.

We propose two structural solutions to mitigate excessive effort and FFS behavior. First, eliminating gain-sharing in favor of a contract with a single controlling healthcare provider aligns the provider’s effort with the payer-desired cost-effective effort (Proposition 3 and Corollary 1). This could be one reason for the development of the BP-Advanced and BP-CJR models. While this new type of model has promise, we also find that integrating bundled payments with VBB programs can be counterproductive, resulting in excessive effort even in the newer single provider models (Corollary 3), though this misalignment is even more severe in the older gain-sharing models (Proposition 4). Second, we show that risk-adjusted bundled payments (e.g. BP-CJR currently being piloted) can mitigate some of the incentive to perform excessive effort (Theorem 3). Both of these new structures can save the payer millions of dollars per contract in efficiency gains (see Section 6 for this large numerical study) and should be considered more broadly as new plans are developed. In summary, our research suggests moving away from gain-sharing toward single controlling provider models and more broadly implementing risk-adjusted bundled payments.
2. Literature Review

Currently, most readmission reduction research focuses on one particular readmission avoidance tactic in a specific healthcare system. These readmission avoidance tactics cover pre-discharge and post-discharge initiatives. Pre-discharge initiatives involve nurse staffing ratios (McHugh and Ma 2013), day of week influences, risk prediction models (Shi et al. 2020, Kansagara et al. 2011, van Walraven et al. 2010), and discharge timing (Shi et al. 2020). Post-discharge research includes scheduling and impact of phone calls, tele-medicine and follow-up appointments (Helm et al. 2016, Liu et al. 2018, Cardozo and Steinberg 2010, Harrison et al. 2011), dedicated post-discharge management teams, nursing facilities and home health services (Tao et al. 2012), and greater levels of patient education by their primary care provider (Costantino et al. 2013, Wallmann et al. 2013, Watson et al. 2011, 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. However, our contribution is on studying how various reimbursement policy initiatives will motivate providers to perform additional effort to achieve cost-effective readmission reduction.

Medicare and other payers have experimented with alternative reimbursement policies for many years to motivate improved healthcare quality at lower costs. Lee and Zenios (2012) investigate Medicare’s End-Stage Renal Disease program for dialysis providers and find that the scheme proposed by Medicare would not provide the desired incentives. Ata et al. (2013) use 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) study optimal contracts for scheduling with a principal-agent framework for outpatient medical services. Their results show 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) use a strategic queuing framework to study imaging test orders and find that over-testing can occur due to distorted price signals, even without FFS payment systems. Aswani et al. (2019) study contracts in the Medicare Shared Savings Program and find 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 study the impact of bundled payments on readmissions.

Recently, several more alternative 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) show 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) demonstrate that adding a financial bonus to HRRP could further incentivize hospitals to reduce readmissions. Zhang et al. (2016) use a game theory approach to determine how hospitals will react to HRRP penalties, assuming all hospitals maximize their contribution margins. They determine that many hospitals will simply accept the penalty as a cost of doing business and not make any extra effort to reduce readmissions. In contrast, 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 VBB analysis, the focus of our paper is bundled payment reimbursement plans.

Recent research on bundled payment plans focuses 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). Liu et al. (2020) develops a Partially Observable Markov Decision Process model of pre- and post-discharge readmission reduction programs controlled by a single provider and shows that the length of the episode of care is the key driving factor behind adoption of such plans. Medical research on bundled payment plan implementation suggests that bundled payment plans in orthopedics results in cost savings and quality improvement (Rana and Bozic 2015, Iorio et al. 2017, McAsey et al. 2019), but results for other health concerns are mixed (Spinks et al. 2018, Joynt Maddox et al. 2018). In the operations management literature, Gupta and Mehrotra (2015) focus on 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 consider bundled payment agreements by looking at patient selection and the provider’s utility and financial risk. They find that performance of bundled payments is sensitive to the bundled payment price and the provider’s risk aversion. Vlachy et al. (2020) use game theory models to compare FFS and bundled payment plans with respect to quality and cost while varying the integration between hospitals and physician groups, which are the two players in their model. However, none of these papers consider the impact of bundled payment plans on readmissions, which is a contribution and focus of our paper.

Guo et al. (2019) compare bundled payment and FFS reimbursement plans on how they impact social welfare, readmissions, and patient waiting time. They find 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 between patient, provider, and payer. Andritsos and Tang (2018) study readmission reduction with respect to different payment plans, including FFS, 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. In contrast to these papers, we study the interaction between providers (e.g. hospital and post-discharge) both working together to prevent readmissions and also competing for a fixed bundled payment rather than provider-patient or provider-payer relationships.

3. Bundled Payment Models for Readmission Reduction

We consider a system in which hospitals and post-discharge providers exert additional readmission reduction effort to reduce readmissions beyond standard 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 the ICU or Progressive Care Unit), improving the nurse-to-patient ratios, improving patient education, using monitoring technologies, or implementing more targeted scheduling of discharge follow-up visits by a doctor. Additional efforts of post-discharge providers could include increasing the frequency of patient visits, providing additional home care, using tele-medicine monitoring, or making frequent phone calls. For examples of additional initiatives to reduce readmissions, see Project RED (Jack et al. 2013) and related reports from Boutwell et al. (2016).

Consider a bundle with $I$ providers. Provider $i \in \{1, \ldots, I\}$ receives a revenue $r_i^A(x_i)$ for performing $x_i$ units of readmission reduction effort and incurs $c_i^A(x_i)$ in cost. Let $p(\bar{x})$ be a scalar function that maps the total joint readmission reduction effort of all providers, $\bar{x} = \sum_{i=1}^{I} x_i$, to a probability of readmission. Define $r_i^G$ and $c_i^G$ as the revenue and cost for standard care not related to readmission reduction. Let $\theta_i$ and $\gamma_i$ be the revenue and cost for services performed during unscheduled hospital readmissions, with $\theta_i = \gamma_i = 0$ for non-hospital providers since readmissions by definition involve only hospital inpatient care. Denote the joint vector of readmission reduction effort by all providers in the bundle as $\mathbf{x} = (x_1, x_2, \ldots, x_I)$, where non-boldface $x$ represent the effort of an individual provider. We introduce the following notation for exposition.

- **General Charges**: Let $G_i = r_i^G - c_i^G$ be the contribution margin of provider $i$ for standard care services.

- **Readmission Prevention Charges**: Let $\rho_i(x_i) = r_i^A(x_i) - c_i^A(x_i)$ be the contribution margin of provider $i$ for readmission prevention services.

- **Total Charges against Bundle**: Let $Z(\mathbf{x}) = \sum_{i=1}^{I} (r_i^G + r_i^A(x_i))$ be the total charges against the bundle without a readmission.
bundled payments. The Cost-Effective (CE) model is given by:

\[
\text{minimize } \sum_{i=1}^{I} \left( p(x_i) \right) - \theta_i R_i^0(x) (\tau - Z(x))
\]

This model represents the desired outcome of healthcare market dynamics.

Modeling Assumptions:

1. Fraction of Bundle: Let \( R_i^q(x) = \frac{r_i^q + r_i^A(x_i) + \theta_i}{\bar{Z}(x) + \sum_{i=1}^{I} \theta_i} \) and \( R_i^{-\theta}(x) = \frac{r_i^G + r_i^A(x_i)}{\bar{Z}(x)} \) be the fraction of the total charges attributed to provider \( i \) with and without a readmission respectively.

2. Marginal Revenue: Let the marginal revenue of provider \( i \) be defined as

\[
q_i(x) = \frac{\partial}{\partial x_i} \left[ p(x_i) + p(x_i) \right] \theta_i + R_i^0(x) \tau - Z(x) + \sum_{i=1}^{I} \theta_i \]
4. \( p(\bar{x}) \) is convex and decreasing since readmission reduction efforts intuitively have decreasing marginal returns. This is consistent with various models of readmission reduction in the literature for both pre- and post-discharge interventions (Helm et al. 2016, Shi et al. 2020). We provide our own example in Section 6.1.

5. Bundled payment price is not excessively large; \( \tau \leq \frac{2c_A^i[r_h^G+r_{pd}^G]}{r_{pd}^i[p(0)-p(\bar{x}^{CE})]} \). This condition holds for all realistic parameter values we consider.

6. Hospital charges for a readmission are not more than the general procedure charges for that episode of care; \( \theta_h \leq r_h^G + r_{pd}^G \). This assumption is generally reasonable, see Urish et al. (2018) and SID (2005-2009). Even if this assumption does not hold, the insights and results in Theorem 1, where it is used, are still valid but require more complex conditions.

To conclude this section, we present two results regarding the optimization problem given in Equation (2). First, Lemma 1 characterizes the optimal solution to maximizing (2) in \( x_i \).

**Lemma 1.** Suppose \( p(\bar{x}) = \lambda_1 e^{-\lambda_2 \bar{x}} \). If

\[
\lambda_2 > O\left(\frac{1}{\sqrt{\gamma Z(0)}}\right)
\]

Equation (2) has a unique optimal solution that is either

(i) Equal to zero, or

(ii) The right-most point (of which there are at most two) that satisfies the first order condition

\[
q_i(x) = \frac{d}{dx_i}[c_A^i(x_i) + p(\bar{x})\gamma_i] = 0.
\]

The proof for Lemma 1 can be found in Appendix EC.2, with all proofs for this paper. We use big O notation in the condition of Lemma 1 since the resulting right-hand-side is so small such that any coefficients or constants would not change the magnitude significantly. For example, consider the practical worst case for the condition, which is a lower bound on the parameters found in the literature (\( \theta = r_h^G = 5000, r_{pd}^G = 500, \gamma = 4000 \), see Section 6). In this scenario, the condition would become \( \lambda_2 > 2 \cdot 10^{-8} \), which means that \( p(x) \) would essentially be a horizontal line and any readmission reduction effort would have almost no effect. All of the \( \lambda_2 \) values from our data for low, medium, and high risk patients fall between 0.3 and 0.4. For the condition not to be satisfied at 0.3, all charges would need to be less than $1.

The lemma still holds under more general structures of \( p \) but with more complicated, but nonrestrictive conditions on the first and second derivatives of \( p \) that should always hold under all but the most degenerate parameter sets. Finally, note that our analytical results rely on the structure of the contribution margin and cost-functions around the first order conditions and are agnostic to the existence of a Nash equilibrium. In the case of discrete effort levels, we can find a counter example where a Nash equilibrium does not exist.
4. Impact of Bundled Payments on Readmissions

In this section, we first show (in Section 4.1) that bundled payment plans do in fact motivate providers to perform no less than the cost-effective effort level to reduce readmissions. However, we also find that bundled payment plans can over motivate the post-discharge provider (who is more sensitive to the bundled payment price than the hospital), which results in excessive readmission reduction effort that is not cost-effective. In Section 4.2 we discuss how bundled payment plan results differ in a single controlling healthcare provider setting and how the cost-effective effort maximizes the total contribution margin for all providers. Finally, in Section 4.3 we discuss how adding VBB programs to bundled payment plans increases effort misalignment for all bundled payment models.

4.1. Gain-sharing Bundled Payments

One major goal of bundled payment plans is to reduce costs while maintaining the overall quality of care and treatment. To accomplish this healthcare providers need to reduce readmissions in a cost-effective manner (Cheney 2016, Guduguntla et al. 2018, Executive-Team 2019). In this section, we refer to a provider as being more motivated if the parameter space in which they will perform readmission reduction effort is larger. For example, provider 1 would be said to be more motivated than provider 2 if provider 1’s optimal response would perform readmission reduction effort for bundled payment prices $\tau \in [\$10,000, \infty)$ and provider 2 would only perform reduction effort for prices $\tau \in [\$30,000, \infty)$. Here, we compare the best-response actions under gain-sharing bundled payment models, such as BP-Classic Models 2 and 4, or BP-Advanced if the hospital decides to form gain-sharing agreements with other providers, with optimal actions from the cost-effective model in (3). Theorem 1 partially answers the question regarding the impact of new payment models on healthcare costs by showing that gain-sharing bundled payment models are indeed motivating providers to reduce readmissions when it is cost-effective to do so. While the conditions in the theorem are sufficient but not necessary, our numerical study indicates that the hospital condition is relatively tight; e.g., the hospital condition holds in 92% of all the scenarios in our numerical analysis.

**Theorem 1.** Let $x^B_i(x_j)$ be the best response effort for provider $i$ when the other provider performs $x_j$ units of effort.

(i): if

$$c_{pd}^A(x^{CE}) + c_h^A(x^{CE}) < \gamma_h \left[p(0) - p(x^{CE})\right].$$

then $x_{pd}^B(x_h) + x_h^B(x_{pd}) \geq x^{CE}$.

(ii): Given a hospital effort level below the cost-effective effort level, $x_h < x^{CE}$, if
\[
\tau \left[ \frac{r_{pd}^G}{r_{pd}^G + r_h^G} - \frac{r_{pd}^G}{r_{pd}^G + r_h^G + \theta_h} \right] > \frac{c_{pd}^A(\bar{x}_{CE})}{p(0) - p(\bar{x}_{CE})}, \tag{5}
\]

then \(x_{Bpd}(x_h) \geq \bar{x}_{CE} - x_h\).

(iii): Given a post-discharge effort level below the cost-effective effort level, \(x_{pd} < \bar{x}_{CE}\), if
\[
\gamma_h > \tau \left[ \frac{r_{pd}^G + \theta_h}{r_{pd}^G + r_h^G + \theta_h} - \frac{r_h^G}{r_{pd}^G + r_h^G} \right] + \frac{c_{pd}^A(\bar{x}_{CE})}{p(0) - p(\bar{x}_{CE})}, \tag{6}
\]

then \(x_{Bh}(x_{pd}) \geq \bar{x}_{CE} - x_{pd}\).

Theorem 1 provides several insights into which factors are motivating each provider. Understanding provider motivation can provide guidance to payers for how to more effectively deploy existing bundled payment models and design new ones. First, hospitals are motivated to reduce readmissions by avoiding the bundled payment penalty whereas the post-discharge provider is motivated by the bonus. This motivation helps explain provider behavior in gain-sharing models. Hospital effort is most often close to cost-effective for reasonable prices since readmissions tend to drive the providers toward a penalty, whereas the post-discharge provider most often performs extra effort when any bonus is possible, which is common if readmissions are sufficiently reduced.

Second, post-discharge providers that are more involved in patient care have greater motivation to reduce readmissions; left-hand-side of (5) is increasing in \(r_{pd}^G\). Third, as shown in the following corollaries, single controlling provider post-discharge and hospital bundled payment models provide the greatest motivation to perform (at a minimum) the cost-effective readmission reduction effort.

For the post-discharge provider, a readmission causes them to lose some of their share of the bundled payment (left-hand-side of (5)) since it increases the hospital’s fraction of total chargers. The percent share lost is a function of the “size” of the post-discharge provider relative to the hospital, which we define as the fraction of the general charges for the episode of care from the post-discharge provider, \(\frac{r_{pd}^G}{r_{pd}^G + r_h^G}\). The larger the portion of the total bundle charges, the more motivation the post-discharge provider will have to perform readmission reduction effort up to (or beyond) CMS’ desired cost-effective level. Extending this insight, the following corollary shows that the post-discharge provider is most motivated by models in which they are solely responsible financially for patient costs and outcomes.

**Corollary 1.** The post-discharge provider is more motivated to perform readmission reduction effort in a bundle that only encompasses post-discharge care and readmissions (i.e., \(r_h^G = 0\)), than in gain-sharing with a hospital.
In contrast to the post-discharge provider, a readmission increases the hospital’s reimbursable charges and likewise their revenue percentage. From the first term on the right-hand-side of (6), it can be seen that a readmission may still benefit the hospital (as in FFS) when the bundled payment price is too high, which reduces the incentives to reduce readmissions. Corollary 2 shows that these misaligned incentives can be mitigated by single controlling provider bundled payment models such as BP-CJR and BP-Advanced.

**Corollary 2.** If a bundle only includes the hospital stay and readmission reduction efforts both pre- and post-discharge \( r^G_{pd} = 0 \), then \( \bar{x}_h \geq \bar{x}^{CE} \).

Rearranging the conditions in Theorem 1 as (7) and (8) highlights motivation for the post-discharge provider and hospital respectively, with respect to the bundled payment payout, \( \tau \). From these equations, it can be deduced that the hospital is motivated by avoiding a penalty, while the post-discharge provider is motivated by increasing their share of any bonus. This can be seen by the directions of the inequalities: the hospital in (8) is motivated by smaller bundled payment prices where readmissions result in a penalty, whereas the post-discharge provider in (7) is motivated by larger prices. We elaborate on the sensitivity of providers to the bundled payment price in Proposition 1 and Proposition 2.

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

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

Proposition 1 shows that the post-discharge provider is more sensitive to the bundled payment price than the hospital. Our numerical study shows that this results in the post-discharge provider performing too little effort at low bundled payment prices and then quickly switching to excessive effort at high bundled payment prices, while the hospital remains close to cost-effective effort levels for a large range of bundled payment prices. This suggests that pricing in gain-sharing models with both hospital and post-discharge providers can be challenging and perhaps gain-sharing should be avoided. Note, the set \( \xi \) implies that if the post-discharge provider performs zero effort it will be in the set. Further, as the hospital performs more effort, the post-discharge provider is more sensitive.

**Proposition 1.** Let \( r^A_{pd} = r^A_{h} \) and \( \xi = \{ x : r^G_{h} > r^G_{pd} + r^A_{x_{pd} - x_{h}} \} \). For \( x \in \xi \),

\[
\frac{\partial E[\pi_{pd}(x+1)]}{\partial \tau} - \frac{\partial E[\pi_{pd}(x)]}{\partial \tau} > \frac{\partial E[\pi_{h}(x+1)]}{\partial \tau} - \frac{\partial E[\pi_{h}(x)]}{\partial \tau}.
\]

One explanation is that increasing readmission reduction effort has a compounded impact on post-discharge revenue – reducing hospital share by reducing readmissions and increasing their
share through increased charges – whereas the hospital increases their share by doing more effort but this effort reduces their share due to fewer readmission charges.

Proposition 2 shows that the price sensitivity in Proposition 1 eventually results in the post-discharge provider performing more effort than the hospital, which, in our numerical study typically leads to excessive effort since the hospital most often performs effort near cost-effective effort levels.

**Proposition 2.** Let \( r^A = r^A_{pd} = r^A_h \). Let \( y \) be an amount of effort performed by a single provider.

(i) \( \exists \tilde{x} \) s.t. if \( x^B_h(y) \geq \tilde{x} \), then \( x^B_{pd}(y) \geq x^B_h(y) \)

(ii) \( \tilde{x} \) is decreasing in \( \tau \)

(iii) \( \tilde{x} \) is decreasing in \( \tau \)

(iv) \( \exists \tilde{\tau} \) s.t. \( \forall \tau \geq \tilde{\tau} \), \( x^B_{pd}(y) \geq x^B_h(y) \)

Proposition 2 implies that if the hospital’s best response is to perform a sufficient level of effort (beyond \( \tilde{x} \)), then the post-discharge provider would be incentivized to perform at least that much effort. Further, the larger the bundled payment price and/or the larger the difference in charged revenue per episode between the providers, the more often the post-discharge provider will perform more effort than the hospital. This highlights the fact that, the larger the financial disparity between providers in the bundle, the more likely the small provider is to perform excessive effort.

### 4.2. Bundled Payments with a Single Controlling Provider

The new BP-Advanced and BP-CJR models only form a contract with a single participant hospital or large physician group. This results in a single controlling healthcare provider who is solely responsible for a bundled payment episode of care. While it is possible for the BP-Advanced and BP-CJR participant to create their own gain-sharing agreements with other providers, we show that this would be sub-optimal for the contracting provider. In addition, vertically integrated healthcare systems, which are becoming increasingly common in the United States (Ginsburg 2016, Fulton 2017), already behave as a single controlling provider. In this section, we show that this single controlling provider approach not only benefits the payer, it also benefits the provider. The contribution margin for this model is given by

\[
\pi^S = \max_x \left[ \tau - \left( \sum_{i=1}^{l} \left[ c^G_i + c^A_i(x_i) \right] + p(\bar{x})g_h \right) \right].
\]

The following proposition shows that BP-CJR and BP-Advanced achieve the cost-effective outcome as long as the participant hospitals do not independently enter into a gain-sharing contract with other providers.

**Proposition 3.** Let \( \bar{x}^S = x^S_{pd} + x^S_h \) be the optimal readmission effort of the post-discharge provider and hospital when both are directed by a single controlling healthcare provider.

Then \( \bar{x}^S = \bar{x}^{CE} \).
Thus, with increased vertical integration and the new BP-Advanced and BP-CJR models that CMS is testing, we could see increasing effectiveness of bundled payment plans achieving cost-effective readmission reduction without instances of excessive or insufficient effort. Next, we show that this cost-effective approach actually benefits participant contribution margin. This creates a win-win scenario for payers and providers. Let $\pi^{BJR} = \pi_{pd}(x_{pd}^B(x_h^B), x_h(x_{pd}^B)) + \pi_h(x_{pd}^B(x_h^B), x_h^B(x_{pd}^B))$ be the total contribution margin of the hospital and post-discharge provider when both are performing their joint best response effort levels and $\pi^{CE} = \tau - \min_x(\sum_{i=1}^{I}[c_i^G + c_i^A(x_i)] + p(\bar{x})\gamma_h)$ be the contribution margin when performing the cost-effective effort as in (3).

**Theorem 2.** $\pi^{CE} \geq \pi^{BJR}$.

Theorem 2 illustrates that regardless of the organizational structure of the providers or the type of bundled payment reimbursement model, a cost-effective approach results in increased overall contribution margin compared to gain-sharing. Thus, single controller models induce cost-effective care delivery and improve provider contribution margins. In the next section, we discuss how combining bundled payment plans with various VBB programs can once again mis-align the best response and cost-effective effort levels even with a single controlling provide scheme.

4.3. Bundled Payment Plans with Value-Based Billing (VBB) Programs

In addition to bundled payments, there are VBB programs such as the hospital readmission reduction program (HRRP) or value-based purchasing programs for hospitals, skilled nursing facilities, and home health agencies, which we summarize in Appendix EC.1. However, we show that combining these VBB programs with bundled payment plans may not have the intended impact.

Let $\mathcal{V}_i(\bar{x})$ be the function for the VBB bonus/penalty for provider $i$. We make the reasonable assumption that $\mathcal{V}_i(\bar{x})$ is convex decreasing for penalties and concave increasing for a bonus. The contribution margin for each provider can be determined by adding $\mathcal{V}_i(\bar{x})$ to (2). We define $x_i^{B+V}(\cdot)$ as the best response of provider $i$ when bundled payment plans and VBB programs are used together. Recall that $x_i^B(\cdot)$ is the best response function without VBB. The following proposition shows that combining VBB programs with bundled payments may result in excessive effort.

**Proposition 4.** Given effort $x_j$ of provider $j$, $x_i^{B+V}(x_j) \geq x_i^B(x_j)$.

Recall that Theorem 1 shows that bundled payment plans already incentivize providers to perform readmission reduction effort that is at (or more than) cost-effective effort levels. Hence, Proposition 4 shows that combining a VBB program with a gain-sharing bundled payment plan could result in effort greater than the bundled payment plan alone. In Proposition 3, we found that using a single controlling healthcare provider could properly align readmission reduction effort under bundled payments. However, Corollary 3 shows that adding a VBB program can mis-align
this effort, even for the case of a single controlling healthcare provider. Let $\bar{x}^{S,V}$ be the optimal readmission reduction effort of the post-discharge provider and hospital when both are directed by a single controlling healthcare provider subject to $V_i(\bar{x})$. Recall that $\bar{x}^{CE}$ is the cost-effective effort from (3).

**COROLLARY 3.** $\bar{x}^{S,V} > \bar{x}^{CE}$.

5. **Aligning Incentives through Risk Profiled Bundled Payments**

In previous sections, we illustrate various conditions when gain-sharing bundled payment plans motivate healthcare providers to exert excessive or insufficient readmission reduction effort. To promote cost-effective effort, we consider an alternative bundled payment plan with risk-adjusted prices, $\tau_{Lo}$ and $\tau_{Hi}$, as opposed to current bundled payment plans with a single bundled payment price, $\tau_{Ag}$. This has been suggested in literature (Porter et al. 2016, Iorio et al. 2017), but prior CMS models were not risk-adjusted in price, though the new CMS BP-CJR sets risk-adjusted prices for the presence of a hip fracture when performing major joint replacement (CMS 2017, Cairns et al. 2018). The results of this section suggest that more broadly employing risk-adjusted bundled payment models could have significant benefits for better aligning incentives to reduce readmissions, which has also been suggested in Liu et al. (2020).

In the rest of this section, we compare the best response effort under risk-profiled and aggregate bundled payment prices to the cost-effective effort, $\bar{x}^{CE}$, where $r \in \{Hi, Lo\}$ is the risk level of the patient, which we assume the providers and payer are able to determine using risk models. A risk $r$ patient has readmission risk reduction curve, $p_r(\bar{x}_r) = \lambda r e^{-\lambda_r \bar{x}_r}$. Let $\tau_t$, where $t \in \{Ag, Hi, Lo\}$ be the bundled payment price for aggregate, high-risk, and low-risk patients. Let $\alpha$ and $1 - \alpha$ represent the fraction of low and high risk patients that a provider serves respectively. We define the provider contribution margin functions as $\pi^{rsk}_r$ for risk adjusted and $\pi^{Ag}_r$ for non-risk adjusted bundled payments, analogous to (2). The bundled payment contribution margin is

$$\pi^{rsk}_i(x_{Lo}, x_{Hi}) = \alpha \pi^{rsk,Lo}_i(x_{Lo}) + (1 - \alpha) \pi^{rsk,Hi}_i(x_{Hi}) \tag{11}$$

$$\pi^{rsk}_i(x) = G_i + p_i(x_i,r) + p_r(x_r)(\theta_i - \gamma_i + R_i^0(x_r)(\tau_r - [Z(x_r) + \sum_{i=1}^{L} \theta_i]))$$

$$+ (1 - p_r(x_r))R_i^{-\theta}(x_r)(\tau_r - Z(x_r)), \quad r \in \{Lo, Hi\}. \tag{12}$$

The contribution margin function for $\pi^{Ag}_i$ is defined similarly by replacing $\tau_{Lo}$ and $\tau_{Hi}$ with $\tau_{Ag}$. In the remainder of this section, we focus on the post-discharge provider’s effort, since (from Proposition 2 and the numerical study in Section 6), the post-discharge provider is more sensitive to the negotiated bundled payment price and more likely to deviate from cost-effective effort.

Next, we compare the difference in effort between the gain-sharing bundled payment model and the cost effective effort level with and without risk adjustment. Let $q_{i,r}(x_r)$ be the risk adjusted
marginal revenue of provider $i$, analogous to (1). Note that $q_{pd,r}(x_r)$ is invertible around the optimal point, since the function is strictly decreasing within that region. Given the best response effort for bundled payment price $\tau_t$, $x_{pd,r}^* = \left( q_{pd,r}(c_{pd}^A/\tau_t) \right)^{-1}$ and the cost-effective effort for risk level $r$, $x_{pd,r}^{CE}$, risk-profiling is a better option when the effort distance between these values is lower for the risk profiled plan. Let $f(\tau_{t1}, \tau_{t2}) = \alpha |x_{pd,Lo} - x_{pd,Lo}^{CE}| + (1 - \alpha) |x_{pd,Hi} - x_{pd,Hi}^{CE}|$ represent the difference between the gain-sharing and cost-effective effort levels when the bundled prices for low and high risk patients are $\tau_{t1}$ and $\tau_{t2}$ respectively. Risk-adjusted bundled payments are closer to cost-effective when $f(\tau_{Lo}, \tau_{Hi}) < f(\tau_{Ag}, \tau_{Ag})$.

We cannot solve the closed form inverse of $q_{pd,r}(x_r)$. However, it can be seen from the form of $q_{pd,r}(x_r)$ that the function should have a shape similar to exponential, as the exponential term dominates on the decreasing portion of the curve. To confirm this intuition, we performed extensive numerical experiments that demonstrate that an exponential function is a good fit for $q_{pd,r}(x_r)$ for a broad range of parameters. Thus, we approximate $q_{pd,r} \approx \beta_r e^{-\beta x_{pd,r}}$. We define $\tau_{x_r}^{CE}$ as the bundled payment price that motivates the cost-effective effort to be performed for risk class $r$.

In practice, it may be difficult to determine the exact price that induces cost-effective effort, and/or there may be other mitigating factors that cause the negotiated price to deviate from the cost-effective price. Theorem 3 shows that pricing structure has a significant impact on how beneficial risk-adjustment can be. Further, we find that the fraction of low risk patients, $\alpha$, plays a key role in addition to three components of readmission reduction efforts: efficiency, effectiveness, and cost.

**Definition 1. Efficiency** of readmission reduction efforts is defined as, $\varepsilon_0 = \gamma/c_{pd}^A$. Efficiency increases in $\gamma$ since the marginal reduction in readmission costs is larger when $\gamma$ is larger and hence each unit of effort is more efficient. Similarly, efficiency increases as the cost of readmission reduction effort decreases relative to $\gamma$, which makes these efforts more efficient.

**Definition 2. Effectiveness** of readmission reduction efforts is defined as, $\varepsilon_1 = \lambda_1 \cdot \lambda_2$. As effectiveness increases this implies that the initial risk is higher (more room for improvement) and/or the slope of improvement is steeper (more impact of readmission reduction efforts). Next, we present a theorem that relates the impact of risk-adjusted bundled payments to the key factors described above.

**Theorem 3 (Risk-profiling effectiveness.)** If $\tau_{x_r}^{CE} < \tau_{Lo} < \tau_{Hi} < \tau_{x_r}^{CE}$, and $q_{pd,r}(x_r)$ is approximated by an exponential function, $\beta_r e^{-\beta x_{pd,r}}$, then

(i) If $\tau_{x_r}^{CE} < \tau_{Lo}$, $\tau_{Hi} < \tau_{x_r}^{CE}$ then $f(\tau_{Lo}, \tau_{Hi}) < f(\tau_{Ag}, \tau_{Ag})$

(ii) If $\tau_{x_r}^{CE} < \tau_{Lo}$, $\tau_{Hi} > \tau_{x_r}^{CE}$, then there exists a threshold, $\tilde{\alpha}_1$, s.t. $f(\tau_{Lo}, \tau_{Hi}) < f(\tau_{Ag}, \tau_{Ag}) \forall \alpha > \tilde{\alpha}_1$
(iii) $\tilde{\alpha}_1$ is decreasing in $\epsilon_0, \epsilon_{1i}^H$, and $c_{pd}^A$ and is increasing in $\epsilon_{1i}^L$

(iv) If $\tau_{Lo} < \tau_{CE Lo}, \tau_{CE Hi} < \tau_{Hi}$ then there exists a threshold, $f(\tau_{Lo}, \tau_{Hi}) < f(\tau_{Ag}, \tau_{Ag}) \forall \alpha > \tilde{\alpha}_2$

(v) $\tilde{\alpha}_2$ is decreasing in $\epsilon_0$ and $c_{pd}^A$

(vi) If $\tau_{Lo} < \tau_{CE Lo}, \tau_{Hi} < \tau_{CE Hi}$ and $(c_{pd}^A)^2 < \max\{\tau_{Ag}^2 \beta_{Lo}^2, 2 \cdot \tau_{Lo}^2 \beta_{Lo}^2\}$ then there exists a threshold, $\tilde{\epsilon}_0$, s.t. $f(\tau_{Lo}, \tau_{Hi}) < f(\tau_{Ag}, \tau_{Ag}) \forall \epsilon_0 < \tilde{\epsilon}_0$

Theorem 3 highlights the key features that contribute to the success of risk-adjusted bundled payments: bundled payment price, fraction of low risk patients in the bundle, and efficiency, effectiveness, and cost of readmission reduction efforts.

**Bundle Price.** When the risk-adjusted prices are sandwiched between the cost-effective prices this always favors risk-adjustment. This occurs because the aggregate price is between the low and high prices, and thus a hospital with aggregate pricing will perform too much effort for low risk and too little for high risk patients. This implies that for risk-adjustment to be most effective, hospitals should err on the side of being generous in their payments for low risk patients and more stingy in their payments for high risk patients.

**Efficiency and Effectiveness.** If the bundled price for high risk patients is set too high, risk-adjusted payments are still more effective if there are sufficiently many low risk patients. The effects of this pricing deviation are further mitigated if readmission reduction efforts are cheaper, more efficient, or more effective (for high risk patients if the low risk price is also high). The larger payment for high risk patients may induce excessive effort. However, the more efficient the effort, the less excessive effort will need to be performed to increase the post-discharge share of the bundled payment payout. To see this, note that reducing readmissions takes share away from the hospital by reducing their total charges. Thus, with more efficient (effective) efforts, less effort is required to capture a larger share of the payout. Thus, accurately choosing the best bundled payment price is less important in systems that have sufficiently many low risk patients where readmission reduction efforts are efficient, effective, and/or cheap.

In contrast, aggressive pricing schemes where the prices are set low (case (vi)) only result in risk-adjusted payments outperforming an aggregate payment when readmission reduction efforts are sufficiently *inefficient*. In this scenario, a risk-adjusted payment better aligns with cost-effective effort for high risk patients but not for low risk patients. Inefficient readmission reduction efforts drives cost-effective effort level toward zero, narrowing the gap between cost-effective and risk-adjustment payment. This scenario still may have overall benefits since it is generally better to focus more effort on high risk patients.

We evaluated the conditions that lead to the conclusions above in the proof of Theorem 3 for all the scenarios in our numerical study (See Section 6) for cases (ii), (iv) and (vi), considering
nine combinations of parameters: three values of $\tau_A$ and three values of $\alpha$. Ignoring the ordering of $\tau_r$ with respect to $\tau_{CE}$, case (ii) holds in all scenarios, case (iv) holds in 95% of scenarios and case (vi) holds in none of the scenarios. This gives a sense of whether the chosen values of the key parameters satisfy the thresholds in Theorem 3. Hence, it may be prudent to avoid under pricing low risk patients while overpricing high risk patients.

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. In general, there is concern that high risk patients could be left out of readmission reduction policies due to the high cost of keeping these patients out of the hospital (Liu et al. 2020), and tailoring payments to incentivize hospitals to better care for these patients is not only appropriate but also aligned with cost-effective payer outcomes. At the same time, risk-profiling reduces gain-sharing excessive effort on low risk patients, which means healthcare resources can be reallocated to more cost-effective initiatives. This analysis and result could partially explain the CMS policy change to allow some level of risk-stratification in the current BP-CJR models. As bundled payment models become more prevalent, CMS and other payers should consider expanding risk stratification to all bundled payment models to ensure providers are motivated to perform the appropriate effort on all patients. These results extend to multiple risk levels, however the conditions become more complicated and there are many more cases to consider.

6. Numerical Analysis
In our numerical analysis, we use a wide range of parameters to explain how bundled payment models perform in a variety of scenarios. We design a test suite of 10,800 simulation scenarios and analyze them, in Section 6.2 to show how gain-sharing between hospitals and post-discharge providers makes it difficult to achieve cost-effective readmission reduction effort levels. In Section 6.3, we demonstrate the financial impact of combining VBB with bundled payment plans. In Section 6.4, we demonstrate the benefits of risk-adjusted bundled payments models and how often they outperform non-risk-adjusted models. Finally, in Section 6.5, we illustrate how different bundled payment models will impact patient throughput and readmission levels in a healthcare system.

6.1. Numerical Parameterization
We leverage publicly available healthcare data (https://www.cms.gov), healthcare research literature (Costantino et al. 2013, Helm et al. 2016), and conversations with physicians, hospitals, and other healthcare researchers to determine reasonable ranges for the bundled payment price ($\tau_r$), the general procedure charged revenue and costs ($r_G^{i,c}$), and the charged revenue and costs associated with each unit of readmission reduction effort ($r_A^A(x_i), c_A^A(x_i)$).

Since we are performing policy level analysis, we use readmission reduction as a function of the volume of follow-up appointments as a reasonable proxy for the shape and magnitude of readmission
reduction as a function of effort, $x_i$. Using data similar to Helm et al. (2016) we generate the curves in Figure 2 for low and high risk patients and all patients in aggregate. These curves map the amount of effort to the projected readmission probability based on clinical and administrative hospital data.

![Figure 2](image)

**Figure 2** Impact that units of readmission reduction effort, $x_i$, (in this case the number of optimal follow-up appointments) can have on readmission probability, $p(\bar{x})$.

In the test suite, we vary general procedure charged revenue for the hospital ($r_{Gh}$) from $5,000 to $30,000 and for the post-discharge provider ($r_{Gpd}$) from $500 to $7,000. The cost of an additional unit of readmission reduction effort for both the hospital and the post-discharge provider is varied from $100 to $500. Finally, we vary the charged revenue of an unscheduled hospital readmission ($\theta_h$) from $5,000 to $30,000. For all of these charged revenue values for both providers, the associated cost to the provider is set at several percentages (80%, 90%, 95%) of the respective charged revenue value to simulate varying contribution margins. These values are all varied in relation to each other. We solve each scenario for $\tau \in [5,000, 54,000]$ in increments of $1,000. The range of values in these scenarios are representative of many different kinds of procedures, including the episode payment range for joint replacement (Cairns et al. 2018) and the mean payment amount for coronary artery bypass grafting (Guduguntla et al. 2018).

### 6.2. Motivating Cost-Effective Readmission Reduction Effort with Gain-Sharing

Figure 3 provides an illustrative example of one scenario, 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 in the x-axis. The joint response and cost-effective effort curves result from Theorem 1 and Equation 3, respectively. The individual best response curves are when the other provider performs zero readmission reduction effort and are used to illustrate the motivation of the individual providers. In Figure 3, general procedure charges are $9,000 for the hospital and $1,000 for the post-discharge provider and the charges for an unscheduled readmission are $20,000. The cost of one unit of readmission reduction effort, $x_i$, is $220 with 10% contribution margins. We believe that reasonable payment models would likely have a price somewhere between the
total general charges \( (r^G_{total} = r^G_h + r^G_{pd}) \) and total general charges plus the charge for a readmission \( (r^G_{total} + \theta_h) \). In the remainder of this section we study the impact of gain-sharing bundled payments on provider behavior for the hospital, the post-discharge provider, and the joint effort between both providers.

Hospital Robustness. In Figure 3, the hospital is motivated to perform the cost-effective effort level for the entire range of bundled payment prices. This robustness to price is further supported by the result that the hospital best response effort is within one unit of the cost-effective effort in 99.9% of our experimental scenarios when the bundled payment price is set to the expected total general charges for an episode \( (\tau = r^G_{total}) \). Additionally, the hospital best response effort is within one unit of the cost-effective effort in 84.29% of scenarios at the upper bound price of \( \tau = r^G_{total} + \theta_h \). Table 2 shows more generally the percent of our experimental scenarios in which the hospital deviates and the associated additional cost to the payer.

<table>
<thead>
<tr>
<th>Units of Effort above ( \bar{x}^{CE} )</th>
<th>( \tau = r^G_{total} )</th>
<th>( \tau = r^G_{total} + \theta_h )</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of Scenarios</td>
<td>Extra Annual Cost ($M)</td>
<td>% of Scenarios</td>
</tr>
<tr>
<td>0</td>
<td>79.9%</td>
<td>$0</td>
</tr>
<tr>
<td>1</td>
<td>20.0%</td>
<td>$0.2M</td>
</tr>
<tr>
<td>2</td>
<td>0.1%</td>
<td>$0.7M</td>
</tr>
<tr>
<td>3</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>5</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>6-10</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>11-15</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>16+</td>
<td>0.0%</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 2 Hospital best response effort above the cost-effective effort and the extra annual cost (millions of $) this could cause in a hospital with 10,000 annual patient arrivals at \( \tau = r^G_{total} \) and \( \tau = r^G_{total} + \theta_h \).
When $\tau = r^G_{total}$ only 11 of the 10,800 scenarios are more than one effort unit above cost-effective effort. Even setting the price as high as $r^G_{total} + \theta_h$, which is likely higher than in practice, the hospital still performs well in most scenarios. This illustrates our claim that the larger hospital provider is more likely to behave in a payer-desired manner. Further, based on CMS’ (and likely other payers) history of payment tightening to incentivize more efficient care delivery, in addition to the fact that bundled payments were designed to avoid paying for readmissions, we feel that the bundled payment price is likely to be much closer to $r^G_{total}$ than $r^G_{total} + \theta_h$. Thus, we focus most of our analysis on $\tau = r^G_{total}$, using $\tau = r^G_{total} + \theta_h$ as an (likely extreme) upper bound.

**Post-Discharge Deviance.** In contrast to the hospital, Figure 3 demonstrates that the post-discharge provider is highly sensitive to the bundled payment price as an illustration of Proposition 1. This sensitivity makes it difficult to avoid excessive effort by pricing alone, as small changes can significantly impact the post-discharge provider behavior. Table 3 repeats Table 2 for the post-discharge provider. In comparison with the hospital, the post-discharge provider effort is within one unit of cost-effective effort in 53% of the scenarios when $\tau = r^G_{total}$ and in nearly 57% of the scenarios they would actually perform insufficient effort. The avoidable readmissions caused by insufficient effort can cost the payer millions of dollars annually per contract. In the other direction, the post-discharge provider exerts more than one unit of excessive effort in 16% of the scenarios, again resulting in millions of dollars of extra cost. Even more concerning is the ramifications if the bundled price is set too high. At $\tau = r^G_{total} + \theta_h$, the post-discharge provider best response effort is more than 10 units above cost-effective effort (excessive effort) in 73% of the scenarios. This would result in more than $30 million of extra cost to payers annually if all patients in a medium size healthcare system (10,000 arrivals per year) experienced this level of excessive effort.

The difficulty in choosing an effective bundled payment price is further illustrated in Table 4, which maps the ratio of price to general charges, $\tau/r^G_{total}$, to the percentage of scenarios that a post-discharge provider would achieve several milestone effort points. This presentation method normalizes the bundled payment price to illustrate it as a percent of the general charges to display how behaviors change as a function of how generous or stingy the bundled payment price is relative to the general cost of caring for the patient. In Table 4, if the bundled payment price is 1.1 times the total general charges, then in 32% of the scenarios the post-discharge provider performs insufficient effort and in 28% of the scenarios the post-discharge provider performs effort in excess of 50% above cost-effective. In the final column of the table we choose the threshold of eight since any effort beyond this has a miniscule impact on readmissions, which implies excessive effort solely to gain a larger share of the bundled payment. This analysis suggests that the solution is more complex than simply identifying an effective bundled payment price, and requires structural change to
Table 3 Post-discharge provider best response effort above the cost-effective effort and the extra annual cost (in millions of $) this could cause in a hospital with 10,000 annual patient arrivals at $\tau = r_{\text{G total}}$ and $\tau = r_{\text{G total}}^{\text{CE}} + \theta_h$.

<table>
<thead>
<tr>
<th>Units of Effort above $\bar{x}_{\text{CE}}$</th>
<th>% of Scenarios</th>
<th>Extra Annual Cost ($\text{M}$)</th>
<th>% of Scenarios</th>
<th>Extra Annual Cost ($\text{M}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4 or lower</td>
<td>3.0%</td>
<td>$12.8$M</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>-3</td>
<td>8.4%</td>
<td>$6.1$M</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>-2</td>
<td>19.4%</td>
<td>$2.3$M</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>-1</td>
<td>25.6%</td>
<td>$0.5$M</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>0</td>
<td>18.6%</td>
<td>$0.0$M</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>1</td>
<td>9.0%</td>
<td>$0.4$M</td>
<td>0.03%</td>
<td>$0.9$M</td>
</tr>
<tr>
<td>2</td>
<td>5.9%</td>
<td>$1.4$M</td>
<td>0.6%</td>
<td>$2.6$M</td>
</tr>
<tr>
<td>3</td>
<td>3.6%</td>
<td>$2.7$M</td>
<td>1.7%</td>
<td>$4.4$M</td>
</tr>
<tr>
<td>4</td>
<td>1.9%</td>
<td>$4.2$M</td>
<td>2.3%</td>
<td>$6.6$M</td>
</tr>
<tr>
<td>5</td>
<td>1.6%</td>
<td>$5.3$M</td>
<td>3.2%</td>
<td>$9.8$M</td>
</tr>
<tr>
<td>6-10</td>
<td>2.4%</td>
<td>$7.1$M</td>
<td>19.2%</td>
<td>$18.5$M</td>
</tr>
<tr>
<td>11-15</td>
<td>0.3%</td>
<td>$11.4$M</td>
<td>17.5%</td>
<td>$31.8$M</td>
</tr>
<tr>
<td>16-20</td>
<td>0.2%</td>
<td>$12.9$M</td>
<td>51.0%</td>
<td>$35.1$M</td>
</tr>
<tr>
<td>21+</td>
<td>0.0%</td>
<td>NA</td>
<td>4.7%</td>
<td>$45.1$M</td>
</tr>
</tbody>
</table>

Table 4 Cumulative percentage of scenarios when post-discharge provider effort reaches specified milestone effort levels as the bundled price ($\tau$) to total general charges ($r_{\text{G total}}$) ratio increases.

<table>
<thead>
<tr>
<th>$\tau/r_{\text{G total}}$</th>
<th>PD$&lt;CE$</th>
<th>PD$&gt;CE$</th>
<th>PD$&gt;H$</th>
<th>PD$=1.5*CE$</th>
<th>PD$=2*CE$</th>
<th>PD$&gt;8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>96.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>0.9</td>
<td>87.0%</td>
<td>2.7%</td>
<td>3.3%</td>
<td>6.0%</td>
<td>5.0%</td>
<td>0.2%</td>
</tr>
<tr>
<td>1.0</td>
<td>64.6%</td>
<td>19.9%</td>
<td>20.0%</td>
<td>9.2%</td>
<td>5.7%</td>
<td>3.2%</td>
</tr>
<tr>
<td>1.1</td>
<td>31.8%</td>
<td>49.9%</td>
<td>47.3%</td>
<td>28.1%</td>
<td>19.9%</td>
<td>12.3%</td>
</tr>
<tr>
<td>1.2</td>
<td>16.3%</td>
<td>72.3%</td>
<td>68.5%</td>
<td>50.9%</td>
<td>39.4%</td>
<td>28.1%</td>
</tr>
<tr>
<td>1.3</td>
<td>8.6%</td>
<td>83.1%</td>
<td>78.7%</td>
<td>67.6%</td>
<td>57.2%</td>
<td>44.3%</td>
</tr>
<tr>
<td>1.4</td>
<td>4.3%</td>
<td>89.4%</td>
<td>85.0%</td>
<td>76.9%</td>
<td>68.9%</td>
<td>58.1%</td>
</tr>
<tr>
<td>1.5</td>
<td>2.0%</td>
<td>93.6%</td>
<td>89.2%</td>
<td>83.3%</td>
<td>76.9%</td>
<td>68.2%</td>
</tr>
<tr>
<td>2.0</td>
<td>0.0%</td>
<td>99.8%</td>
<td>97.4%</td>
<td>97.1%</td>
<td>93.3%</td>
<td>89.9%</td>
</tr>
<tr>
<td>2.5</td>
<td>0.0%</td>
<td>100.0%</td>
<td>98.6%</td>
<td>99.8%</td>
<td>98.6%</td>
<td>97.1%</td>
</tr>
<tr>
<td>3.0</td>
<td>0.0%</td>
<td>100.0%</td>
<td>98.8%</td>
<td>100.0%</td>
<td>99.9%</td>
<td>99.6%</td>
</tr>
</tbody>
</table>

Joint Effort. Table 5 shows the joint response of both providers in a gain-sharing model. One positive is that the joint response is typically near the cost-effective level when $\tau = r_{\text{G total}}^{\text{CE}}$ (within
one unit of cost-effective in 84.01% of all scenarios) since at this price point the hospital will most often perform near-cost-effective effort and the post-discharge provider is not incentivized to perform effort. However, even at \( \tau = r^G_{total} \), the joint response drifts more than one unit of effort above the cost-effective level in 16% of scenarios with significant implications for cost to the payer.

In a medium (large) healthcare system with 10,000 (30,000) patient arrivals per year and contracts where \( \tau = r^G_{total} \), if the effort is more than one unit of effort above the cost-effective level, then the payer would incur unnecessary costs in excess of $1.4 million ($4.2 million) per contract. If the price \( \tau = r^G_{total} + \theta_h \), the performance is much more costly. This excess effort in these scenarios is again driven by the post-discharge (smaller) provider, which further suggests eliminating gain-sharing in favor of a single-provider contract.

<table>
<thead>
<tr>
<th>Units of Effort above ( \bar{X}^{CE} )</th>
<th>( \tau = r^G_{total} )</th>
<th>( \tau = r^G_{total} + \theta_h )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( % ) of Scenarios</td>
<td>% of Extra Annual Cost ($M)</td>
<td>% of Scenarios</td>
</tr>
<tr>
<td>0</td>
<td>56.8%</td>
<td>$0</td>
</tr>
<tr>
<td>1</td>
<td>27.2%</td>
<td>$0.3M</td>
</tr>
<tr>
<td>2</td>
<td>6.0%</td>
<td>$1.4M</td>
</tr>
<tr>
<td>3</td>
<td>3.6%</td>
<td>$2.7M</td>
</tr>
<tr>
<td>4</td>
<td>1.9%</td>
<td>$4.2M</td>
</tr>
<tr>
<td>5</td>
<td>1.6%</td>
<td>$5.3M</td>
</tr>
<tr>
<td>6-10</td>
<td>2.4%</td>
<td>$7.1M</td>
</tr>
<tr>
<td>11-15</td>
<td>0.3%</td>
<td>$11.4M</td>
</tr>
<tr>
<td>16-20</td>
<td>0.2%</td>
<td>$12.9M</td>
</tr>
<tr>
<td>21-25</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>26-30</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>31-35</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>36-40</td>
<td>0.0%</td>
<td>NA</td>
</tr>
<tr>
<td>41+</td>
<td>0.0%</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 5 Joint best response effort above the cost-effective effort and the extra annual cost (in millions of $) this could cause in a healthcare system with 10,000 annual patient arrivals at \( \tau = r^G_{total} \) and \( \tau = r^G_{total} + \theta_h \).

![Figure 4](attachment:image.png)

**Figure 4** Comparison of extra per episode costs between the joint response, hospital only response, and the post-discharge provider only response at \( \tau = r^G_{total} \) and \( \tau = r^G_{total} + \theta_h \).

In the new BP-Advanced and BP-CJR models being tested by CMS, only a hospital or large physician group is allowed to enter into a contract with CMS. The results in this section may be
one explanation for this shift. This is illustrated in Figure 4, which compares the percentage of scenarios that would be cost-effective or have extra costs at two bundled payment price points. Even at this range of bundled payment price values, the hospital is motivated to achieve cost-effective effort a majority of the time and rarely incurs extra costs exceeding $500 per episode of care. However, the post-discharge provider and the joint response (driven by post-discharge provider effort) will experience extra costs in a high percentage of scenarios. These results provide insights into effective implementation of the new, single-controlling provider models. In these new models, hospitals should avoid entering into their own gain-sharing arrangements with other providers and should instead simply contract out these auxiliary services, providing guidelines for how these services should be performed. This approach will achieve the win-win of minimizing cost to society as well as increased contribution margins for the hospital (see Theorem 2).

6.3. VBB Impact on Bundled Payment Plans

From Proposition 3, a single controlling healthcare provider without gain-sharing will perform cost-effective effort. This controlling healthcare provider could be a vertically integrated healthcare system or it could simply be a large hospital that contracts with various post-discharge providers directly (without gain-sharing) under BP-Advanced or BP-CJR models. Our numerical experiments demonstrate the magnitude increase in contribution margin (mean 19.0%, median 11.3%) by doing cost-effective effort from Theorem 2, as opposed to the joint-response total effort.

However, when VBB programs are combined with bundled payments (as in Section 4.3), even the single provider will perform excess effort in 93% of scenarios. In 23% of the scenarios the best response effort would be two or more units above the cost-effective effort, which incurs millions in excess cost. VBB programs only increase the best response efforts of a single controlling provider by at most three units and do not grow unchecked as in gain-sharing models.

Figure 5 depicts the readmission reduction effort with and without VBB under the same parameterization as Figure 3 when \( \tau = $12,000 \), where charges for a readmission, \( \theta_h \), are $8,000 and $11,000. This figure highlights the fact that a payer should consider not combining VBB with bundled payments. However, if there are other benefits of doing so, it is best to implement VBB within a non-gain-sharing bundled payment model, as Figure 6 shows that single-provider models are far more robust to the misaligned incentives.

For Figure 6, we rerun our numerical experiments adding a VBB function to the contribution margin equations. VBB programs shift a large percentage of scenarios toward higher levels of excessive effort. The additional units of effort are more noticeable when \( \tau = r^{G}_{total} \), as adding a VBB program is enough to push the majority of scenarios away from the cost-effective effort and be 1-4 units of effort above. This is illustrated by the gray (box marker) Joint Response with VBB line being further to the right than the red (circle marker) Joint Response (No VBB) line.
Figure 5  Readmission reduction effort levels for gain-sharing, single controlling healthcare provider, and cost-effective models, with and without VBB penalties.

Figure 6  VBB program addition comparison of the cumulative percentage of scenarios that experience varying units of effort above the cost-effective effort at $\tau = r^G_{\text{total}}$ and $\tau = r^G_{\text{total}} + \theta_h$.

Of particular interest is that, even with VBB, the single controlling provider model with VBB performs nearly as well or better than the gain-sharing model without the incentive mis-alignment caused by VBB. This single provider model is also robust to the bundled payment price, $\tau$, where as the gain-sharing models perform poorly at higher price points (compare Figure 6(b) with 6(a)). This highlights another benefit of the single controlling provider model.

6.4. Risk Profiling Impact on Gain-Sharing Bundled Payment Plans
In this section, we investigate the magnitude of improvement that can be achieved using risk-adjusted bundled payments such as BP-CJR and argue that these improvements are significant enough to consider expanding the use of risk-adjustment in other bundled payment models. We perform these experiments under the gain-sharing model to demonstrate that risk-adjustment can mitigate the excessive effort observed in these models. As in our previous experiments, we vary
a baseline bundled price, $\tau_{\text{base}}$ from $5,000 to $54,000 in increments of $1,000 for each of 10,800 parameter scenarios. The general procedure and effort charged revenue and cost terms are the same as in Section 6.1. $\tau_{Hi}$ and $\tau_{Lo}$ are set by adding (subtracting) the change in expected readmission cost at the cost-effective effort level due to the high (low) readmission probabilities. Low, baseline, and high risk patients had an initial risk of readmission of 5%, 19%, and 73% respectively, in accordance with the literature (see Helm et al. (2016), Hu et al. (2014)). Let $\Delta p_r = p_r(x_r^{CE}) - p_b(x_b^{XE})$, where $r \in Hi, Lo$ and $p_b(x_b^{XE})$ is the readmission risk for the baseline risk curve $p_b$ used in all other experiments. Using this notation, $\tau_r = \tau_{\text{base}} + \Delta p_r \gamma$. We set $\tau_Ag = (1 - \alpha)\tau_{Hi} + \alpha\tau_{Lo}$ based on the fraction of high and low risk patients in the bundle. With this formulation, $\tau_{Lo} \leq \tau_{\text{base}} \leq \tau_{Hi}$.

We vary the proportion of low risk patients $\alpha \in \{0.5, 0.65, 0.85\}$. We use $\alpha = 0.5$ as an extreme lower bound for comparison purposes, as it would be extremely unlikely for 50% of patients of any condition to be high risk. In reality, $\alpha = 0.85$ is the most realistic parameter based on the literature. For example, by applying a personalized risk prediction model to a large data set containing the majority of admissions/readmissions for four different states in the United States, Helm et al. (2016) found that only about 20% of patients were determined to be high risk even though they studied some of the highest readmission risk diagnoses (bladder, kidney, and prostate cancer). As such, we focus primarily on $\alpha = 0.85$, with the other values as comparison to demonstrate robustness to the fraction of low risk patients. Note that we did not optimize any of the bundled payment prices to provide a fair comparison.

(a) Realistic: $\tau_r = \Delta p_r \gamma$

(b) Wide range: $\tau_r = (p_r(0) - p_b(0))\theta$

Figure 7 Percentage of scenarios where the risk-profiled bundled payment effort is closer to cost-effective effort than the aggregate effort levels for varying low risk bundled payment prices divided by the total general charges, $r_{\text{total}}^G$.

Figure 7 plots the percentage of scenarios in which risk-adjusting better aligns incentives as a function of the low risk bundled payment price, $\tau_{Lo}$, normalized by the general chargers, $r_{\text{total}}^G$, as in Table 4. 7a uses risk-adjusted prices based on the formula presented above. Fig 7b plots...
the same results for poorly chosen bundled payment prices, where the spread between high and low risk prices is significantly (unreasonably) widened. In 7a, risk-profiling effort is closer to cost-effective effort in more than 99% of scenarios at any price to total charges ratio. In 7b, we illustrate what can happen when risk-adjusted bundled prices are chosen poorly, which demonstrates the performance of risk-profiling when bundled payment prices are haphazardly set, such that the negotiated $\tau_{Hi}$ value is obviously motivating excessive readmission reduction effort on high risk patients. Even in this extreme case, risk-profiling motivates effort closer to the cost-effective effort in more than 70% of the scenarios across all parameters settings. The drop when the bundled price is set near to the total charges (i.e. close to one on the x-axis) occurs because in this range $\tau_{Hi}$ is so large that it motivates excessive effort from the risk-adjusted providers, whereas $\tau_{Ag}$ is low enough that the providers incur negative contribution margin and thus will perform cost-effective effort to limit their losses. While this highlights the importance of properly choosing risk-adjusted pricing schemes, it requires significantly deviant pricing to observe this kind of behavior.

Figure 8 provides a more detailed analysis of the magnitude of improvement and associated cost savings of risk-adjusted bundled payments as a function of the normalized bundled payment price. Figure 8a shows the difference in the distance from cost-effective effort between risk-adjusted and aggregate bundled payments. Figure 8b show the difference in annual cost for a medium sized hospital between risk-adjusted and aggregate bundled payments. The cost savings are greatest (≈$600K-$1M) when the price is slightly above total general charges, which is likely to be the most reasonable price point. When all scenarios with all bundled payment prices are considered the cost savings per episode is $19 ($34) with 85% (65%) low risk patients, which becomes an annual savings of $193,000 ($344,000) for a hospital with 10,000 patient arrivals. However, if a more appropriate range is selected, such as when $\tau_{Lo}$ is within 10% of $r_{tot}^G$, then these cost savings increase to $36
($70) with 85% (65%) low risk patients, which becomes an annual savings of $360,000 ($700,000) for a hospital with 10,000 patient arrivals. Figure 8 also illustrates that the risk-profiled effort improvement and cost benefit are lower with a larger low risk percentage of patients because low risk patients typically only require little readmission reduction effort and thus only a small effort improvement and cost benefit can actually occur for low risk patients. This analysis suggests further investigation of expanding risk-adjusted bundled payments after the initial pilot of BP-CJR.

6.5. 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 healthcare system in terms of reduced congestion and patient wait times. In this section, we show that this additional readmission reduction benefit is not large enough to consider effort over-motivation in a positive light.

The figures in this section use the same illustrative parameter values as Figure 3 with additional values for the external patient arrival rate and service rates for the hospital and post-discharge providers. Figure 9 shows the impact of bundled payment price on readmission probability and hospital waiting times using a fluid model to capture the queuing network dynamics between hospitals and post-discharge providers (Whitt 2006). Note the joint response initially follows the hospital response and then switches to the post-discharge response curve when it dips below the hospital’s best response in both figures. Figure 9(a) shows that a higher bundled 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 9](image_url)  
**Figure 9** Impact of increasing bundled payment price and subsequent increase in readmission reduction effort on (a) readmission probability, (b) hospital waiting time (weeks) in an at-capacity hospital. The best response assumes the other provider performs zero effort.

Figure 9(b) shows that the additional effort performed as a result of increasing the bundled payment price, $\tau$, 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 operational benefits relative to the cost of these efforts, which means that over-motivated effort may be primarily based on increasing contribution margins and not on quality of care.

7. Conclusions
As policy makers develop reimbursement plans to maintain quality levels and reduce costs, they need to understand how these policies will actually motivate readmission reduction, which is a major cost of healthcare. To help policy makers understand this phenomenon, we develop a contribution margin gain-sharing model and add elements of VBB programs and risk-profiling to understand the behavior and incentives of various healthcare providers with respect to efforts to reduce readmissions. We compare the effort under these different payment plans with the cost-effective readmission reduction effort level desired by a payer.

This paper derives five main insights that contribute to the healthcare literature. (1) We show how gain-sharing bundled payment plans incentivize different providers in different ways to reduce readmissions. A positive result is that typically at least one of the providers will be motivated to perform no less than the cost-effective readmission reduction effort level. However, (2) we also find that the post-discharge provider (primarily) can be over-motivated to perform excessive effort to garner a larger percentage of the bundled payment, improving their contribution margin at the expense of the hospital and the payer. We show that this excessive effort is primarily driven by increasing their contribution margin, and has little benefit on quality of care. Along these lines, we show that setting bundled payment prices to reduce readmissions can be extremely difficult in a gain-sharing environment because the smaller post-discharge provider can be highly sensitive to changes in the bundled payment price, leading to both insufficient and excessive effort, even at the same bundled payment price, depending on the cost structure of the providers in the bundled.

The hospital (as the larger provider), on the other hand, is relatively insensitive to changes in the bundled payment price and is generally better aligned with payer-desired cost-effective readmission reduction efforts. This analysis suggests that pricing alone may not be sufficient to motivate cost-effective effort and that a structural change in the design of bundled payment models may be needed. (3) Bundled payment models with a single controlling provider, such as the new BP-Advanced and BP-CJR CMS models, can mitigate these mis-aligned incentives, resulting in payer-desired cost-effective readmission reduction effort. However, (4) combining VBB programs with bundled payments will once again mis-align incentives and encourage excessive effort and overcharging similar to the FFS schemes that bundled payments were designed to avoid. This is true even for a single controlling provider bundle, though the effect is far more severe in a gain-sharing
bundle. Instead, well-designed bundled payment reimbursement plans could replace readmission-targeting VBB programs, such as HRRP, while better motivating readmission reduction.

Finally, (5) we find that risk-adjusted bundled payment pricing, such as BP-CJR, can better align a provider’s readmission reduction effort even in gain-sharing contexts, with significant cost savings over traditional bundled payment models. In summary, the design of the bundled payment structure, not just pricing, is critical to driving proper behavior of providers. Payers designing new bundled payment models should consider (1) moving away from gain-sharing contracts with multiple providers, (2) avoid combining VBB with bundled payments, and (3) expand their implementation of risk-adjusted bundled payment pricing.

References


Executive-Team (2019) Personal interviews with home health provider executive leadership team.


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


E-companion for Healthcare Reimbursement Policy Impact on Multiple-Provider Readmission Reduction Programs

EC.1. Value-Based Billing (VBB) Program Descriptions

There are several VBB programs with similar characteristics and impact from a provider perspective, even if their mechanics and focus vary. We discuss the primary CMS VBB programs below.

**Hospital Readmission Reduction Program (HRRP):** HRRP was a VBB program 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). We assume this HRRP penalty term is convex decreasing in total readmission reduction effort similarly to $p(\bar{x})$.

**Hospital Value-Based Purchasing Program:** In effect since 2013, VBP withholds 2% of participating hospitals’ Medicare payments and uses this withholding to fund value-based incentive payments to hospitals based on their performance metrics. Hospitals are judged on outcome measures such as: mortality and complications (readmissions), healthcare-associated infections, patient safety, patient experience, efficiency and cost reduction. It is possible for a hospital to earn back an incentive payment that is less than, equal to, or more than the 2% withholding for that fiscal year based on the range and distribution of all eligible hospitals’ total performance scores (CMS 2020c). While HRRP only enforces a penalty, hospital value-based purchasing allows the option for a bonus or penalty based on a total performance score that includes readmissions. In the 2020 fiscal year, 56% of hospitals received a bonus while the remaining hospitals received a penalty (CMS 2020b).

**Skilled Nursing Facility Value-Based Purchasing Program:** This program rewards skilled nursing facilities (SNFs) with incentive payments based on the quality of care they provide to Medicare beneficiaries, as measured by a hospital readmissions measure. As required by statute, CMS withholds 2% of SNFs’ FFS payments to fund the program. CMS redistributes 60% of the total withholding amount to SNFs as incentive payments (CMS 2020d). This program allows the option for a penalty or bonus based on readmission scores.

**Home Health Value-Based Purchasing Program:** This program was proposed in 2015 to provide Medicare-certified home health agencies (HHAs) incentives to give higher quality and more efficient care. HHAs in the program will compete on value, where payment is tied to quality performance, in which all-condition risk-adjusted potentially preventable hospital readmission rates are one of the quality measures. CMS will adjust payments to HHAs for services based on the quality of care, not just quantity of the services they provide in a given performance period. It was
to be implemented in several states on a trial period, but final withholding percentages were not determined (CMS 2020a). This program also allows for a penalty or bonus based on readmissions.

EC.2. Paper Proofs

**Lemma 1.** Suppose \( p(\bar{x}) = \lambda_1 e^{-\lambda_2 \bar{x}} \). If

\[
\lambda_2 > o\left(\frac{1}{\sqrt{\gamma_2(Z(0))}}\right)
\]

Equation (2) has a unique optimal solution that is either

(i) Equal to zero, or

(ii) The right-most point (of which there are at most two) that satisfies the first order condition

\[
q_i(x) = \frac{d}{dx_i}[c_i^A(x_i) + p(\bar{x})\gamma_i] = 0.
\]

**Proof for Lemma 1.** For notational simplicity, we write the function \( Z(x) \) as \( Z \). We assume the target provider (we first analyze the post-discharge provider and next the hospital) performs \( x \), which we take as the variable of interest, and the non-target provider performs \( y \), which we take as a constant. Unless explicitly stated, let \( \bar{x} = x + y \) be the total effort performed by both providers.

**Post-discharge provider.** We begin with the first order condition of Equation 2 for the post-discharge provider, \( q_{pd}(x) = c_{pd}^A \). First we define the components of \( q_{pd} \). Recall that

\[
q_{pd}(x) = \tau \left( p'(\bar{x})(R^\theta_{pd}(x) - R^{-\theta}_{pd}(x)) + p(\bar{x}) \frac{\partial}{\partial x_{pd}} [R^\theta_{pd}(x) - R^{-\theta}_{pd}(x)] + \frac{\partial}{\partial x_{pd}} R^{-\theta}_{pd}(x) \right).
\]

(\text{EC.1})

\[
R^{-\theta}_{pd}(x) = \frac{r_{pd}^G + r_{pd}^A(x_{pd})}{Z}
\]

\[
\frac{\partial}{\partial x_{pd}} R^{-\theta}_{pd}(x) = \frac{(r_{pd}^G + r_{pd}^A(x_{pd}))r_{pd}^A}{Z^2}
\]

(\text{EC.2})

\[
R^\theta(x) - R^{-\theta}(x) = \frac{r_{pd}^G + r_{pd}^A(x_{pd})}{Z + \theta_h} - \frac{r_{pd}^G + r_{pd}^A(x_{pd})}{Z}
\]

\[
= \frac{(r_{pd}^G + r_{pd}^A(x_{pd}))(Z + \theta_h)}{Z(Z + \theta_h)} - \frac{(r_{pd}^G + r_{pd}^A(x_{pd}))(Z)}{Z(Z + \theta_h)}
\]

\[
= \frac{(r_{pd}^G + r_{pd}^A(x_{pd}))((Z - Z - \theta_h))}{Z(Z + \theta_h)} = -\theta_h (r_{pd}^G + r_{pd}^A(x_{pd}))
\]

\[
= \frac{\partial}{\partial x_{pd}} (R^\theta_{pd}(x) - R^{-\theta}_{pd}(x)) = \theta_h [r_{pd}^G + r_{pd}^A(x_{pd})] r_{pd}^A (2Z + \theta_h) - \theta_h r_{pd}^A (Z^2 + \theta_h Z)
\]

\[
= \theta_h r_{pd}^A [(r_{pd}^G + r_{pd}^A(x_{pd}))(2Z + \theta_h) - Z^2 - Z\theta_h]
\]

(\text{EC.3})

(\text{EC.4})

(\text{EC.5})

(\text{EC.6})

(\text{EC.7})

Substituting Equations EC.2 - EC.7 into Equation EC.1 and rearranging terms yields:
\[ q_{pd}(x) = \tau \left[ -p'(\bar{x}) \left( \frac{(r^G_{pd} + r^A_{pd}(x_{pd}))\theta_h}{Z(Z + \theta_h)} \right) + \frac{(r^G_{h} + r^A_{h}(x_h))r^A_{pd}}{Z^2} + p(\bar{x}) \left( \frac{\theta_h r^A_{pd}((r^G_{pd} + r^A_{pd}(x_{pd}))(2Z + \theta_h) - Z^2 - Z\theta_h)}{[Z(Z + \theta_h)]^2} \right) \right] \] (EC.8)

Further,
\[ q'_{pd}(x)/\tau = -p''(\bar{x}) \left( \frac{(r^G_{pd} + r^A_{pd}(x_{pd}))\theta_h}{Z(Z + \theta_h)} \right) + 2p'(\bar{x}) \left( \frac{\theta_h r^A_{pd}[(r^G_{pd} + r^A_{pd}(x_{pd}))(2Z + \theta_h) - Z^2 - Z\theta_h]}{[Z(Z + \theta_h)]^2} \right) \]
\[ + p(\bar{x}) \left( \frac{\theta_h (r^A_{pd})^2[2(r^G_{pd} + r^A_{pd}(x))]Z(Z + \theta_h)^2 - [(r^G_{pd} + r^A_{pd}(x))(2Z + \theta_h) - Z^2 - Z\theta_h][4Z^3 + 6Z^2\theta_h + 2Z\theta^2_h]}{[Z(Z + \theta_h)]^4} \right) + \frac{-2(r^G_{h} + r^A_{h}(x_h))(r^A_{pd})^2}{Z^3} \] (EC.9)

Let \( f^- \) represent the 1st and 4th terms in (EC.9) that are always negative and increasing. While \( f^+ \) represent the 2nd and 3rd terms, which could be positive decreasing or negative depending on parameters. Also note that \( \lim_{x \to -\infty} f^-(x) = 0 \) and \( \lim_{x \to -\infty} f^+(x) = 0 \) If \( f^+ \) approaches zero faster than \( f^- \), this implies that, \( q' \) will cross the origin at most once, since eventually the negative terms, \( f^- \), will dominate \( f^+ \) as \( x \) grows. To show this, for notational simplicity we can bound \( f^+ \) by \( C_1 e^{-\lambda x} \) and \( f^- \) by \( C_2 e^{-\lambda x} + \frac{c_3}{Z(x)^3} \). Then
\[ \lim_{x \to -\infty} f^+(x) \leq \frac{C_2}{C_1} + \frac{c_3}{Z(x)^3} e^{-\lambda x} = -\infty \]
where the limit can easily be shown using L’Hospital’s rule. Thus, by our argument, \( q' \) crosses the origin at most once and becomes negative as \( x \) grows. This shows that \( q \) is either increasing and then decreasing or strictly decreasing.

Having established the unimodality of \( q_{pd} \) in \( x_{pd} \) we next show that the optimal point will either be the right most point that satisfies the FOC or zero. Without loss of generality, assume the hospital performs \( y \) effort. If there are no points that satisfy the FOC, then it must be the case that \( q_{pd}(y, x) < c_{pd} \ \forall x \) or \( \frac{q_{pd}(y, x)}{c_{pd}} > c_{pd} \ \forall x \). The second case is impossible because it implies that the post-discharge provider could have unbounded contribution margin, which is impossible since the contribution margin is bounded by the bundle price, \( \tau \). In the first case, choose any point \( \bar{x} \). Then
\[ \pi_{pd}(y, \bar{x}) = \pi_{pd}(y, 0) + \int_0^{\bar{x}} q_{pd}(y, s) - c_{pd}^A ds < \pi_{pd}(y, 0) \] (EC.10)
since \( q_{pd}(y, s) - c_{pd}^A < 0 \ \forall s \). Thus \( \pi_{pd}(y, 0) > \pi_{pd}(y, \bar{x}) \ \forall \bar{x} \), which implies that \( x_{pd}^F(y) = 0 \).

In the case where there is only one point satisfying the FOC, \( x_{h}^{F_1} \), by the unimodality of \( q_{pd} \), either (i) \( x_{h}^{F_1} \) is the peak of the curve, i.e. \( q_{pd}(y, x) < c_{pd} \ \forall x \neq x_{h}^{F_1} \), or (ii) \( q_{pd} > c_{pd} \ \forall x < x_{h}^{F_1} \). In case (i)
\[ \pi_{pd}(y, x_{pd}^{F_1}) = \pi_{pd}(y, 0) + \int_0^{x_{pd}^{F_1}} q_{pd}(y, s) - c_{pd}^A ds < \pi_{pd}(y, 0) \] (EC.11)
since \( q_{pd}(y, s) - c_{pd}^{A} ds < 0 \) \( \forall s \neq x_{pd}^{F_1} \). Thus, \( \pi_{pd}(y, 0) > \pi_{pd}(y, x_{pd}^{F_1}) \), which implies that \( x_{pd}^{F_2}(y) = 0 \). For case (ii) we have that

\[
\pi_{pd}(y, x_{pd}^{F_1}) = \pi_{pd}(y, 0) + \int_{0}^{x_{pd}^{F_1}} q_{pd}(y, s) - c_{pd}^{A} ds > \pi_{pd}(y, 0)
\]  

(EC.12)

since \( q_{pd}(y, s) - c_{pd}^{A} ds > 0 \) \( \forall s < x_{pd}^{F_1} \). Thus, \( \pi_{pd}(y, 0) < \pi_{pd}(y, x_{pd}^{F_1}) \). Further, by unimodality, \( q_{pd}(y, x) < c_{pd} \) \( \forall x > x_{pd} \) which implies that \( x_{pd}^{B}(y) = x_{pd}^{F_1} \) using the same integration arguments as above. Further, since \( q_{pd}(y, x) > c_{pd}^{A} \) \( \forall x \in [0, x_{pd}^{B}(y)) \) and \( q_{pd}(y, x) < c_{pd}^{A} \) \( \forall x > x_{pd}^{B}(y) \), we have that if \( q_{pd}(y, x) > c_{pd} \) then \( x_{pd}^{F_2}(y) > x \).

In the case where there are two points that satisfy the first order condition, we can eliminate one of the points. Specifically, let \( x_{pd}^{F_1} < x_{pd}^{F_2} \) where \( q_{pd}(y, x_{pd}^{F_1}) = c_{pd}^{A} \) and \( q_{pd}(y, x_{pd}^{F_2}) = c_{pd}^{A} \). For \( x < x_{pd}^{F_1} \) we have that \( q_{pd}(y, x) < c_{pd}^{A} \) by unimodality (first increasing then decreasing) of \( q_{pd} \). Thus either

\[
\pi_{pd}(y, x_{pd}^{F_1}) = \int_{0}^{x_{pd}^{F_1}} q_{pd}(y, s) - c_{pd}^{A} ds
\]  

(EC.13)

\[
< \int_{0}^{x_{pd}^{F_1}} q_{pd}(y, s) - c_{pd}^{A} ds + \int_{x_{pd}^{F_1}}^{x_{pd}^{F_2}} q_{pd}(y, s) - c_{pd}^{A} ds
\]  

(EC.14)

\[
= \int_{0}^{x_{pd}^{F_1}} q_{pd}(y, s) - c_{pd}^{A} - p'(s + y)\gamma_{pd} ds = \pi_{pd}(x_{pd}^{F_2}, y)
\]  

(EC.15)

Since \( \int_{0}^{x_{pd}^{F_1}} q_{pd}(y, s) - c_{pd}^{A} ds < 0 < \int_{x_{pd}^{F_1}}^{x_{pd}^{F_2}} q_{pd}(y, s) - c_{pd}^{A} ds \). Thus \( x_{pd}^{F_1} \) cannot be optimal and the optimal solution is either \( x_{pd}^{F_2} \) or zero as argued previously. If \( x_{pd}^{B}(y) = x_{pd}^{F_2} \) we showed by unimodality that \( q_{pd}(x) > c_{pd} \) \( \forall x \in (x_{pd}^{F_1}, x_{pd}^{F_2}) \) and \( q_{pd}(x) < c_{pd} \) \( \forall x \not\in \{x_{pd}^{F_1}, x_{pd}^{F_2}\} \). Thus \( \forall x : q_{pd}(x) > c_{pd} \), we have that \( x_{pd}^{F_2} > x \).

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

\[
q_{h}(x) = \tau \left( \frac{\partial}{\partial x_{h}} R_{h}^{\theta}(x) + p'(\bar{x}) \left[ R_{h}^{G}(x) - R_{h}^{-\theta}(x) \right] + p(\bar{x}) \frac{\partial}{\partial x_{h}} \left[ R_{h}^{G}(x) - R_{h}^{-\theta}(x) \right] \right)
\]  

(EC.16)

\[
R_{h}^{\theta} = \frac{r_{h}^{G} + r_{h}^{A}(x_{h})}{Z}
\]  

(\( Z > 0 \))

\[
\frac{\partial}{\partial x_{h}} R_{h}^{\theta}(x) = \frac{(r_{pd}^{G} + r_{pd}^{A}(x_{pd}))r_{h}^{A}}{Z^{2}} > 0
\]  

(EC.17)

\[
R_{h}^{G}(x) - R_{h}^{-\theta}(x) = \frac{r_{h}^{G} + r_{h}^{A}(x_{h}) + \theta_{h}}{Z + \theta_{h}} - \frac{r_{h}^{G} + r_{h}^{A}(x_{h})}{Z}
\]  

\[
= \frac{\theta_{h}(r_{pd}^{G} + r_{pd}^{A}(x_{pd}))}{Z(Z + \theta_{h})}
\]  

\[
\frac{\partial}{\partial x_{h}} [R_{h}^{G}(x) - R_{h}^{-\theta}(x)] = -\theta_{h}(r_{pd}^{G} + r_{pd}^{A}(x_{pd}))(2Z + \theta_{h}) < 0
\]  

(EC.18)
Substitute Equations EC.17 - EC.18 into Equation EC.16, resulting in Equation EC.19.

\[ q_h(x) = \tau(r^G_{pd} + r^A_{pd}(x_{pd})) \left( \frac{r^A_h}{Z^2} + p'(\bar{x}) \frac{\theta_h}{Z(Z + \theta_h)} - p(\bar{x}) \frac{\theta_h r^A_h(2Z + \theta_h)}{[Z(Z + \theta_h)]^2} \right) \]  

(EC.19)

Note the first term is positive decreasing and the second two are negative increasing. Using the same technique as for the post-discharge provider we take \( q' \).

\[ \frac{q'_h(x)}{\tau(r^G_{pd} + r^A_{pd}(x_{pd}))} = \frac{-2(r^A_h)^2}{Z^3} + p''(\bar{x}) \frac{\theta_h}{Z(Z + \theta_h)} - 2p'(\bar{x}) \frac{\theta_h r^A_h(2Z + \theta_h)}{[Z(Z + \theta_h)]^2} - p(\bar{x}) \frac{2\theta_h(r^A_h)^2[Z(Z + \theta_h)]^2 - \theta_h(r^A_h)^2(2Z + \theta_h)[4Z^3 + 6Z^2\theta_h + 2Z\theta_h^2]}{[Z(Z + \theta_h)]^4} \]  

(EC.20)

The first term of \( q' \) is clearly negative and the last three terms are positive. Using the same limiting argument as with the post-discharge provider, we can show that (EC.20) is either strictly decreasing or unimodal (first increasing then decreasing) since the positive terms go to zero faster than the negative terms. For the hospital, however, we must also consider the marginal cost, which is a function of hospital effort as well. The marginal cost is strictly increasing in hospital effort \( x \).

First, consider the case where there exists a point, \( w^+ \), s.t. \( q(w^+, y) = p'(w^+ + y)\gamma + c \in \{ z : q'(w, y) > 0 \} \), i.e. the marginal cost and marginal revenue curves intersect while marginal revenue is increasing. To show that there are at most two such points, we compare the rates of change, \( q' \) and \( \frac{\partial}{\partial z} p'(x, y) + c_h = p''(x, y) \). First, we rewrite \( q' \) as

\[ q'_h(x, y) = \tau(r^G_{pd} + r^A_{pd}(y)) \left( \frac{-2(r^A_h)^2}{Z^3} + p''(\bar{x}) \frac{\theta_h}{Z(Z + \theta_h)} - 2p'(\bar{x}) \frac{\theta_h r^A_h(2Z + \theta_h)}{[Z(Z + \theta_h)]^2} - p(\bar{x}) \frac{2\theta_h(r^A_h)^2[Z(Z + \theta_h)]^2 - \theta_h(r^A_h)^2(2Z + \theta_h)[4Z^3 + 6Z^2\theta_h + 2Z\theta_h^2]}{[Z(Z + \theta_h)]^4} \right) \]

\[ = \frac{p''(\bar{x})}{\tau(r^G_{pd} + r^A_{pd}(y))} \left( \frac{\theta_h}{Z(Z + \theta_h)} + 2\frac{\theta_h r^A_h(2Z + \theta_h)}{\lambda_2[Z(Z + \theta_h)]^2} \frac{2\theta_h(r^A_h)^2[Z(Z + \theta_h)]^2 - \theta_h(r^A_h)^2(2Z + \theta_h)[4Z^3 + 6Z^2\theta_h + 2Z\theta_h^2]}{(\lambda_2)^2[Z(Z + \theta_h)]^4} \right) - \tau(r^G_{pd} + r^A_{pd}(y)) \frac{2(r^A_h)^2}{Z^3} \]  

(EC.21)

To understand the behavior of the marginal cost and marginal revenue terms, we compare their slopes: (EC.21) and \( p''(x + y) \). We begin by comparing the coefficients of \( p''(x + y) \). For the marginal cost, the coefficient is a constant, \( \gamma \). Define the coefficient of \( p''(x + y) \) in (EC.21) as \( m(x) \). The following supporting lemma provides properties of \( m(x) \) that we will use to show the result that the marginal revenue and cost equations intersect at most twice.

**Lemma EC.1.** \( m(x) \) is positive and decreasing in \( x \) and \( \lim_{x \to \infty} = 0 \).
Proof. First, note that the first two terms of $m(x)$ are positive. For the third term, we remove the coefficient that is common to both the positive and negative terms in the numerator, $\theta_h(r_h^d)^2$. The positive term then becomes $2Z^4 + 4Z^3\theta_h + 2Z^2\theta_h^2$, whereas the negative term is greater than $8Z^4 + 12Z^3\theta_h + 4Z^2\theta_h^2$, which shows that the numerator is negative and hence the fraction is positive.

To show that the function is decreasing toward zero. It is sufficient to show that the numerator is dominated by the denominator. This is clear for the first term since the variable, $x$ does not appear in the numerator. For the second term we get

$$\frac{2\theta_h r_h^{d^2}(2Z + \theta_h)}{\lambda_2 Z^2(Z + \theta_h)^2} = \frac{2\theta_h r_h^{d^2}}{\lambda_2 Z^2(Z + \theta_h)^2} + 2 \frac{\theta_h r_h^{d^2}}{\lambda_2 Z^2(Z + \theta_h)},$$

which is also clearly decreasing in $x$. For the third term we have

$$\frac{\theta_h (r_h^d)^2(2Z + \theta_h)(4Z^3 + 6Z^2\theta_h + 2Z\theta_h^2)}{[Z(Z + \theta_h)^4]} = \frac{8}{(Z + \theta_h)^4} + 16\frac{\theta_h}{Z^2(Z + \theta_h)^4} + 10\frac{\theta_h^2}{Z^3(Z + \theta_h)^4} + \frac{2\theta_h^3}{Z^4(Z + \theta_h)^4},$$

which again is clearly decreasing and approaches zero as $x \to \infty$. □

To complete the proof, we first define the set of values where $q$ is increasing as follows. Let $w_o = \arg\max\{w : q'(w, y) > 0\}$. Now consider the following cases.

Case 1: $q'(x, y) < p''(x + y)\gamma \forall x$.

Case 1(a): $q(0, y) \geq p'(y)\gamma + c_h$. If $p'(w_0 + y)\gamma + c_h > q(w_0, y)$ then $\exists w^+ \text{ s.t. } p'(w^+ + y)\gamma + c_h = q(w^+, y)$ by the intermediate value theorem. Further $w^+$ is unique since for $x > w^+$, we have that

$$q(x, y) = q(w^+ + y) + \int_{w^+}^{x} q'(s, y)ds = p'(w^+ + y)\gamma + c_h + \int_{w^+}^{x} q'(s, y)ds < p'(w^+ + y)\gamma + c_h + \int_{w^+}^{x} p'(s, y)ds = p'(x + y)\gamma + c_h$$

Thus the FOC has a unique solution and this solution is optimal by the arguments made previously for the post-discharge provider.

If $p'(w_0 + y)\gamma + c_h \leq q(w_0, y)$ then $\exists w^- \text{ s.t. } p'(w^- + y)\gamma + c_h = q(w^-, y)$ and this point is unique by the fact that $q$ is decreasing on $[w_0, \infty)$ and $p'(x + y)\gamma + c_h$ is monotone increasing. Further, $\lim_{x \to \infty} = 0$ so, by the intermediate value theorem, the two curves must intersect. Thus the FOC has a unique solution and this solution is optimal by the arguments made previously for the post-discharge provider.

Case 1(b): $q(0, y) < p'(y)\gamma + c_h$. Then the optimal solution is zero, since the marginal revenue curve $q$ will always lie below the marginal cost curve using the same arguments from Case 1(a).

Case 2: $q'(0, y) \geq p''(y)\gamma$. First, we leverage Lemma EC.1 to show that $p''(x)\gamma$ goes to zero faster than $q'(x, y)$:

$$\lim_{x \to \infty} \frac{\tau(r_{pd}^{G} + r_{pd}^{A}(y))}{p'(x + y)\gamma} \left(\frac{p''(x + y)m(x) - \frac{2(r_h^{d^2})^2}{Z(x, y)^3}}{p'(x + y)\gamma}\right) = \lim_{x \to \infty} \frac{p''(x + y)m(x)}{p'(x + y)\gamma} - \frac{2(r_h^{d^2})^2}{Z(x, y)^3}$$
then $x$ satisfy the FOC. First, we know that $x$ exists since $\lim_{x \to \infty} q(x,y) = 0 < p'(x + y)\gamma + c_h$, there are two points $x_0 < x_1$ that satisfy the FOC. If $q(w_0, y) \geq p'(y + w_0)\gamma + c_h$, there are two points $x_0 < x_1$ that satisfy the FOC. First, we know that $x_0$ exists by the intermediate value theorem and that $x_1$ exists since $\lim_{x \to \infty} q(x,y) = 0 < p'(x + y)\gamma + c_h$. $q'(x,y) \geq p''(x + y)$ for $x \in [0, x_0]$ and $q'(x_1, y) \leq p''(x_1 + y)$, otherwise $q(x_1, y) > p'(x_1 + y)\gamma + c_h$. Since we showed that $p''(x + y)\gamma$ goes to zero faster than $q'(x,y)$, then $q'(x,y) < p''(x + y)\gamma \forall x > x_1$, which establishes the uniqueness of $x_1$ using the same arguments as for the post-discharge provider.

If $q(w_0, y) < p'(y + w_0)\gamma + c_h$, then the optimal solution is zero since $q$ is decreasing after $w_0$.

**Case 2(b):** $q(0, y) \geq p'(y)\gamma + c_h$ then there will be a unique solution, $x_1$ to the FOC. If $x_1 \in [0, w_0]$ then $q'(x_1, y) \leq p''(x_1 + y)$, otherwise $q(x_1, y) > p'(x_1 + y)\gamma + c_h$. Since we showed that $p''(x + y)\gamma$ goes to zero faster than $q'(x,y)$, then $q'(x,y) < p''(x + y)\gamma \forall x > x_1$, which establishes the uniqueness of $x_1$ using the same arguments as for the post-discharge provider. If $x_1 > w_0$, then it is unique because on this interval $q'(x,y)$ is decreasing while $p'(x + y)\gamma + c_h$ is increasing.

This completes the proof. □

**Theorem 1.** Let $x_i^B(x_j)$ be the best response effort for provider $i$ when the other provider performs $x_j$ units of effort

(i): if

$$c_h^A(x^C) + c_h^A(x^C) < \gamma_h \left[p(0) - p(x^C)\right].$$

then $x^B_{pd}(x_h) + x^B_{pd}(x_{pd}) \geq x^C$.

(ii): Given a hospital effort level below the cost-effective effort level, $x_h < x^C$, if

$$\frac{r^G_{pd}}{r^G_{pd} + r^G_{h}} - \frac{r^G_{pd}}{r^G_{pd} + r^G_{h} + \theta_h} > \frac{c_h^A(x^C)}{p(0) - p(x^C)},$$

then $x^B_{pd}(x_h) \geq x^C - x_h$.

(iii): Given a post-discharge effort level below the cost-effective effort level, $x_{pd} < x^C$, if

$$\gamma_h \geq \frac{r^G_{h} + \theta_h}{r^G_{pd} + r^G_{h} + \theta_h} - \frac{r^G_{h}}{r^G_{pd} + r^G_{h}} + \frac{c_h^A(x^C)}{p(0) - p(x^C)},$$

then $x^B_{h}(x_{pd}) \geq x^C - x_{pd}$.

**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 $x \equiv [x_{pd}, x_h]$, where $x = x_{pd} + x_h < x^C = x^C_{pd} + x^C_h$. Define $x^C \equiv [x^C_{pd}, x^C_h]$, then

$$m(x) = \lim_{x \to \infty} \frac{2(r^A_{pd}r^A_{h})^2}{\tilde{z}(x,y)^3} = -\infty$$

$$\text{Case 2(a):} \quad q(0, y) < p'(y)\gamma + c_h.$$ If $q(w_0, y) \geq p'(y + w_0)\gamma + c_h$, there are two points $x_0 < x_1$ that satisfy the FOC. First, we know that $x_0$ exists by the intermediate value theorem and that $x_1$ exists since $\lim_{x \to \infty} q(x,y) = 0 < p'(x + y)\gamma + c_h$. $q'(x,y) \geq p''(x + y)$ for $x \in [0, x_0]$ and $q'(x_1, y) \leq p''(x_1 + y)$, otherwise $q(x_1, y) > p'(x_1 + y)\gamma + c_h$. Since we showed that $p''(x + y)\gamma$ goes to zero faster than $q'(x,y)$, then $q'(x,y) < p''(x + y)\gamma \forall x > x_1$, which establishes the uniqueness of $x_1$ using the same arguments as for the post-discharge provider.
where $x_{i}^{CE}$ is the optimal (cost-effective) effort employed by provider $i$ in Equation 3. If Equation EC.25 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.

$$E[\pi_{i}(x^{CE})] - E[\pi_{i}(x)] > 0 \quad \text{(EC.25)}$$

Substitute Equation 2 into Condition EC.25.

$$\tau \left( p(\bar{x}) \left[ R_{h}^{0}(x^{CE}) - R_{h}^{\theta}(x) \right] + (1 - p(\bar{x})) \left[ R_{i}^{-\theta}(x^{CE}) - R_{i}^{\theta}(x) \right] \right)$$

$$+ \left[ p(\bar{x}) - p(\bar{x}^{CE}) \right] \left( \tau \left[ R_{i}^{-\theta}(x) - R_{h}^{0}(x) \right] + \gamma_{h} \right) - c_{i}^{A}(\bar{x}^{CE} - x_{j}) + c_{i}^{A}(x_{i}) > 0 \quad \text{(EC.26)}$$

**Part (iii)** The first two terms of Condition EC.26 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 EC.26 are positive, then this is a sufficient condition for Condition EC.25 to be true. Rearranging these three terms we get:

$$\gamma_{h} > \tau \left[ R_{h}^{0}(x) - R_{h}^{\theta}(x) \right] + \frac{c_{i}^{A}(\bar{x}^{CE} - x_{pd}) - c_{i}^{A}(x_{h})}{p(\bar{x}) - p(\bar{x}^{CE})}$$

$$\text{(EC.27)}$$

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

Next, through the following propositions, we show that if Condition EC.27 is true at $x = 0$, then it is true for all $x$. Define $f_{h}(x_{pd}, x_{h}) = \gamma_{h} + \tau \left[ R_{h}^{0}(x_{pd}, x_{h}) - R_{h}^{\theta}(x_{pd}, x_{h}) \right]$, $g_{h}(x_{pd}, x_{h}) = \frac{c_{i}^{A}(\bar{x}^{CE} - x_{pd}) - c_{i}^{A}(x_{h})}{p(\bar{x}) - p(\bar{x}^{CE})}$, and $h_{h}(x_{pd}, x_{h}) = f_{h}(x_{pd}, x_{h}) - g_{h}(x_{pd}, x_{h})$. Thus, we can write Condition EC.27 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 EC.1.** Assume $c_{i}^{A}(x_{i})$ is linear. Then $g_{h}(x_{pd}, x_{h}) = g_{h}(x_{pd} + x_{h}, 0)$.

$$g_{h}(x_{pd}, x_{h}) = \frac{c_{i}^{A}(\bar{x}^{CE} - x_{pd}) - c_{i}^{A}(x_{h})}{p(x_{pd} + x_{h}) - p(\bar{x}^{CE})} = \frac{c_{i}^{A}(\bar{x}^{CE} - x_{pd} - x_{h}) - c_{i}^{A}(0)}{p(x_{pd} + x_{h}) - p(\bar{x}^{CE})} = g_{h}(x_{pd} + x_{h}, 0) \quad \text{(EC.28)}$$

**PROPOSITION EC.2.** $f_{h}(x_{pd}, x_{h}) = \gamma_{h} + \tau \left[ R_{h}^{0}(x_{pd}, x_{h}) - R_{h}^{\theta}(x_{pd}, x_{h}) \right]$ is increasing in $(-x_{pd}, x_{h})$.

Let $Z(x_{pd}, x_{h}) = r_{h}^{G} + r_{h}^{A}(x_{h}) + r_{pd}$, where $r_{pd} = r_{pd}^{G} + r_{pd}^{A}(x_{pd})$

$$R_{h}^{\theta}(x_{pd}, x_{h}) - R_{h}^{0}(x_{pd}, x_{h}) = \frac{r_{h}^{G}(x_{h})}{r_{h}^{G}(x_{h}) + r_{pd}} - \frac{r_{h}^{G} + r_{h}^{A}(x_{h}) + \theta_{h}}{r_{h}^{G} + r_{h}^{A}(x_{h}) + r_{pd} + \theta_{h}}$$
\[
\left( r_h^G + r_h^A(x_h) \right) \left( Z(x_{pd}, x_h) + \theta_h \right) - \left( -r_h^G + r_h^A(x_h) + \theta_h \right) Z(x_{pd}, x_h) \right) = -\theta_h \left( r_{pd}^G + r_{pd}^A(x_{pd}) \right) \left( Z(x_{pd}, x_h) + \theta_h \right) \]

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

\[
\lim_{\epsilon \to 0} \frac{\tau_r^A \theta_h}{\tau\theta_h} + \tau \theta_h \left( r_h^G - r_{pd}^A \right) \left( r_{pd}^G + r_{pd}^A \right) \left( 2Z(x_{pd}, x_h) + \epsilon (r_h^A - r_{pd}^A) + \theta_h \right) = Z(x_{pd} - \epsilon, x_h + \epsilon) \left[ Z(x_{pd} - \epsilon, x_h + \epsilon) + \theta_h \right] Z(x_{pd}, x_h) \geq 0 \quad (EC.30)
\]

\[\square\]

**Proposition EC.3.** \( h_y(y, 0) \) is increasing in \( y \).

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

\[
f_h(y, 0) = \gamma_h + \tau \left[ \frac{r_h^G}{r_h^G + r_{pd}^G + r_{pd}^A(y)} - \frac{r_h^G + \theta_h}{r_h^G + r_{pd}^G + r_{pd}^A(y) + \theta_h} \right] \quad (EC.31)
\]

\[
g(y, 0) = c_h^A(\bar{x}^{CE} - y) / p(\bar{x}) - p(\bar{x}^{CE}) \quad (EC.32)
\]

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.

\[
\frac{\partial f_h(y, 0)}{\partial y} = \tau r_p^A \left[ \frac{r_h^G + \theta_h}{r_h^G + r_{pd}^G + r_{pd}^A(y) + \theta_h} - \frac{r_h^G}{r_h^G + r_{pd}^G + r_{pd}^A(y)} \right] \quad (EC.33)
\]

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

\[
-\frac{\partial g_h(y, 0)}{\partial y} = \left[ -\frac{c_h^A p^A - p^A c_h^A(\bar{x}^{CE} - y)}{(p^A)^2} \right] = \frac{c_h^A}{p^A} - \frac{p^A c_h^A(\bar{x}^{CE} - y)}{(p^A)^2} > 0 \quad (EC.34)
\]

The highest magnitude negative term in Equation EC.33 occurs when \( y = 0 \), so we use this scenario in Condition EC.35. Note that the magnitude of this term is less than the magnitude of the positive terms in Equation EC.34 as long as

\[
\frac{c_h^A}{p^A} + \frac{\left( -p^A \right)c_h^A \bar{x}^{CE}}{(p^A)^2} > \tau r_p^A \left[ \frac{r_h^G}{r_h^G + r_{pd}^G + r_{pd}^A(y)} - \frac{r_h^G + \theta_h}{r_h^G + r_{pd}^G + r_{pd}^A(y) + \theta_h} \right] \quad (EC.35)
\]

Let \( \theta_h \leq r_h^G + r_{pd}^G \) and perform some algebra, so Condition EC.35 can be reduced to

\[
\tau \leq \frac{2r_h^A \left( r_h^G + r_{pd}^G \right)}{r_{pd}^A p^A r_h^G} \leq \frac{2r_h^A \left( r_h^G + r_{pd}^G \right)^2}{r_{pd}^A p^A r_h^G} \quad (EC.36)
\]
Thus, as long as the technical assumption in Condition EC.36 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 contribution margin from returning to cost-effective effort is higher than the contribution margin at the point \((x_{pd}, x_h)\). We know that
\[
h_h(x_{pd}, x_h) \geq h_h(x_h + x_{pd}, 0) \geq h_h(0, 0)
\] (EC.37)

The first inequality holds from Propositions EC.1 and EC.2. The second inequality holds from Proposition EC.3. Thus we only need to show that \( h_h(0, 0) > 0 \) for Theorem 1 to hold, which is precisely Condition 6, 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. In Condition EC.26 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 \( R_{pd}^{-\theta}(x) - R_{pd}^{\theta}(x) \) 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. 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.

\[
\tau \left[ R_{pd}^{-\theta}(x) - R_{pd}^{\theta}(x) \right] > \frac{c_A^A(\bar{x}^{CE} - x_h) - c_A^A(x_{pd})}{p(\bar{x}) - p(\bar{x}^{CE})}
\] (EC.38)

Next, we show that if Condition EC.38 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 \( h_{pd}(x_{pd}, x_h) > 0 \)
\[
h_{pd}(x_{pd}, x_h) \geq h_{pd}(0, x_h + x_{pd}) \geq h_h(0, 0)
\] (EC.39)

Hence the contribution margin from returning to cost-effective effort is higher than the contribution margin 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 5, which proves part (ii) of Theorem 1. □

**Part (i):** This follows directly from examining the overlap of Conditions 7 and 8.

\[
\left[ \frac{c_A^A(\bar{x}^{CE})}{p(0) - p(\bar{x}^{CE})} \right] < \left[ \frac{c_h^A(\bar{x}^{CE})}{p(0) - p(\bar{x}^{CE})} \right]
\] (EC.40)

\[\iff c_A^A(\bar{x}^{CE}) + c_h^A(\bar{x}^{CE}) < \gamma_h \left[ p(0) - p(\bar{x}^{CE}) \right]\] (EC.41)

This results in the condition above which matches Condition 4. □
COROLLARY 1. The post-discharge provider is more motivated to perform readmission reduction effort in a bundle that only encompasses post-discharge care and readmissions (i.e., \( r^G_h = 0 \)), than in gain-sharing with a hospital.

Proof for Corollary 1: Let \( r^G_h = 0 \). Then (5) becomes

\[
\tau (p(0) - \bar{p}(x_{CE})) \left[ \frac{\theta_h}{r^A_{pd} + \theta_h} \right] = c^A_{pd}(\bar{x}_{CE}).
\]

Further,

\[
\tau (p(0) - \bar{p}(x_{CE})) \left[ \frac{r^G_{pd}}{r^G_{pd} + r^G_h} - \frac{r^G_{pd}}{r^G_{pd} + r^G_h + \theta_h} \right] = \tau (p(0) - \bar{p}(x_{CE})) \left[ \frac{r^G_{pd} \theta_h}{(r^G_{pd} + r^G_h)(r^G_{pd} + r^G_h + \theta_h)} \right]
\]

\[
\leq \tau (p(0) - \bar{p}(x_{CE})) \left[ \frac{\theta_h}{r^G_{pd} + r^G_h + \theta_h} \right] \leq \tau (p(0) - \bar{p}(x_{CE})) \left[ \frac{\theta_h}{r^G_{pd} + \theta_h} \right]
\]

Combining (EC.43) with Theorem 1 shows that, in any scenario in which the post-discharge provider would perform at least cost-effective effort in a gain-sharing bundle, they would also perform cost-effective effort without including the hospital in the gain-sharing. □

COROLLARY 2. If a bundle only includes the hospital stay and readmission reduction efforts both pre- and post-discharge (\( r^G_{pd} = 0 \)), then \( x^B_h \geq \bar{x}_{CE} \).

Proof for Corollary 2:

Corollary 2 follows by setting \( r^G_{pd} = 0 \) in (6), which results in the condition that \( \gamma_h(p(0) - \bar{p}(x_{CE})) > c^A_{pd}(x_{pd}) + c^A_h(\bar{x}_{CE} - x^A_{pd}) \). This condition has the cost savings of fewer readmissions on the left-hand-side and the cost of effort on the right-hand-side. From the properties of the minimization of (3), this condition must be true at \( \bar{x}_{CE} \), because the cost of readmission effort should be less than the cost of readmission savings. □

PROPOSITION 1. Let \( r^A = r^A_{pd} = r^A_h \) and \( \xi = \{ x : r^G_h > r^G_{pd} + r^A(x_{pd} - x_h) \} \). For \( x \in \xi \),

\[
\frac{\partial E[\pi_{pd}(x+1)]}{\partial \tau} - \frac{\partial E[\pi_{pd}(x)]}{\partial \tau} > \frac{\partial E[\pi_h(x+1)]}{\partial \tau} - \frac{\partial E[\pi_h(x)]}{\partial \tau}.
\]

Proof for Proposition 1:

Taking the derivative of Equation 2 for the hospital and post-discharge provider with respect to \( \tau \) yields the following:

\[
\frac{\partial E[\pi_h(x)]}{\partial \tau} = \frac{p(x) [r^G_h + r^A_h(x_h) + \theta_h]}{r^G_{pd} + r^G_{pd}(x_{pd}) + r^G_h + r^A_h(x_h) + \theta_h} + \frac{(1 - p(x)) [r^G_h + r^A_h(x_h)]}{r^G_{pd} + r^A_{pd}(x_{pd}) + r^G_h + r^A_h(x_h)}
\]

(EC.45)
\[
\frac{\partial E[\pi_{pd}(x)]}{\partial \tau} = \frac{p(\bar{x}) \left[ r_{pd}^G + r_{pd}^A(x_{pd}) \right]}{r_{pd}^G + r_{pd}^A(x_{pd}) + r_h^G + r_h^A(x_h) + \theta_h} + \frac{(1 - p(\bar{x})) \left[ r_{pd}^G + r_{pd}^A(x_{pd}) \right]}{r_{pd}^G + r_{pd}^A(x_{pd}) + r_h^G + r_h^A(x_h)} \quad \text{(EC.46)}
\]

Next evaluate the difference of each equation at (x+1) and x and to allow more simplification we assume \( r^A = r_{pd}^A = r_h^A \).

\[
\frac{\partial E[\pi_h(x+1)]}{\partial \tau} - \frac{\partial E[\pi_h(x)]}{\partial \tau} = \frac{p(\bar{x} + 1) \left[ r_h^G + r_h^A(x_h + 1) + \theta_h \right]}{r_{pd}^G + r_{pd}^A(x_{pd} + 1)} - \frac{p(\bar{x}) \left[ r_h^G + r_h^A(x_h) + \theta_h \right]}{r_{pd}^G + r_{pd}^A(x_{pd} + 1)}
\]

\[
+ \frac{(1 - p(\bar{x} + 1)) \left[ r_h^G + r_h^A(x_h + 1) \right]}{r_{pd}^G + r_{pd}^A(x_{pd} + 1)} - \frac{(1 - p(\bar{x})) \left[ r_h^G + r_h^A(x_h) \right]}{r_{pd}^G + r_{pd}^A(x_{pd})} \quad \text{(EC.47)}
\]

\[
\frac{\partial E[\pi_{pd}(x+1)]}{\partial \tau} - \frac{\partial E[\pi_{pd}(x)]}{\partial \tau} = \frac{p(\bar{x} + 1) \left[ r_{pd}^G + r_{pd}^A(x_{pd} + 1) \right]}{r_{pd}^G + r_{pd}^A(x_{pd} + 1)} - \frac{p(\bar{x}) \left[ r_{pd}^G + r_{pd}^A(x_{pd}) \right]}{r_{pd}^G + r_{pd}^A(x_{pd})}
\]

\[
+ \frac{(1 - p(\bar{x} + 1)) \left[ r_{pd}^G + r_{pd}^A(x_{pd} + 1) \right]}{r_{pd}^G + r_{pd}^A(x_{pd} + 1)} - \frac{(1 - p(\bar{x})) \left[ r_{pd}^G + r_{pd}^A(x_{pd}) \right]}{r_{pd}^G + r_{pd}^A(x_{pd})} \quad \text{(EC.48)}
\]

Next, take Equation EC.48 for post-discharge minus Equation EC.47 for the hospital

\[
\left[ \frac{\partial E[\pi_{pd}(x+1)]}{\partial \tau} - \frac{\partial E[\pi_{pd}(x)]}{\partial \tau} \right] - \left[ \frac{\partial E[\pi_h(x+1)]}{\partial \tau} - \frac{\partial E[\pi_h(x)]}{\partial \tau} \right] =
\]

\[
\left[ r_{pd}^G - r_h^G + r_h^A(x_{pd} - x_h) - \theta_h \right] \left( \frac{p(\bar{x} + 1)}{r_{pd}^G + r_{pd}^A(x_{pd} + 1)} - \frac{p(\bar{x})}{r_{pd}^G + r_{pd}^A(x_{pd})} \right)
\]

\[
+ \left[ r_{pd}^G - r_h^G + r_h^A(x_{pd} - x_h) \right] \left( \frac{(1 - p(\bar{x} + 1))}{r_{pd}^G + r_{pd}^A(x_{pd} + 1)} - \frac{(1 - p(\bar{x}))}{r_{pd}^G + r_{pd}^A(x_{pd})} \right) \quad \text{(EC.49)}
\]

Then, to prove that Equation EC.49 is positive and the post-charge provider slope is greater than the hospital slope of contribution margin to bundled payment price, \( \tau \), we show that Equation EC.49 is positive for all values of \( \theta_h \). To do so, we first show Equation EC.49 is positive at \( \theta_h = 0 \) and second, that it never becomes negative for \( \theta_h > 0 \).

Consider Equation EC.49 when \( \theta_h = 0 \), which results in:

\[
\left[ r_{pd}^G - r_h^G + r_h^A(x_{pd} - x_h) \right] \left( \frac{1}{r_{pd}^G + r_{pd}^A(x_{pd} + 1)} - \frac{1}{r_{pd}^G + r_{pd}^A(x_{pd})} \right) \quad \text{(EC.50)}
\]

The large parenthesis term is always negative and the term in brackets is always negative, given our assumption that \( r_h^G > r_{pd}^G + r_h^A(x_{pd} - x_h) \). Thus, Equation EC.49 is positive at \( \theta_h = 0 \).
Next, we solve for the roots of (EC.49) as a function of $\theta_h$. For exposition, we make the following notational substitutions. Let $B = r_{pd}^G + r_h^G + r_A(x_h + x_{pd})$ and $C = r_{pd}^G - r_h^G + r_A(x_{pd} - x_h)$, which is negative for $x \in \xi$. Note that $B + C \geq 0$. To solve for the roots, we must solve

$$
(C - \theta_h) \left( \frac{p(\bar{x} + 1)}{B + \theta_h + r_A} - \frac{p(\bar{x})}{B + \theta_h} \right) + C \left( \frac{(1 - p(\bar{x} + 1))}{B + r_A} - \frac{(1 - p(\bar{x}))}{B} \right) = 0 \quad (EC.51)
$$

Multiplying through by the two denominators of the first term yields

$$
(C - \theta_h) (p(\bar{x} + 1)(B + \theta_h) - p(\bar{x})(B + \theta_h + r_A)) + C(B + \theta_h)(B + \theta_h + r_A) \left( \frac{(1 - p(\bar{x} + 1))}{B + r_A} - \frac{(1 - p(\bar{x}))}{B} \right) = 0 \quad (EC.52)
$$

It is clear that (EC.52) is a quadratic equation that can be solved using the quadratic formula

$$
\frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (EC.53)
$$

We first show that $-b \leq 0$ and next that $b \geq \sqrt{b^2 - 4ac}$. This suffices to show that all the roots of the quadratic equation are not positive. This in turn shows that the equation (EC.49) never becomes negative for positive values of $\theta_h$, which directly implies the result of the theorem.

After some algebra,

$$
b = \frac{(p(\bar{x}) - p(\bar{x} + 1)) \cdot B^3 + (r_A^2(2 \cdot p(\bar{x}) - p(\bar{x} + 1)) + C \cdot (p(\bar{x}) - p(\bar{x} + 1))) \cdot B^2}{(B^2 + B)}
$$

$$
\quad + \frac{(p(\bar{x}) (r_A^2 + 2 \cdot C(p(\bar{x}) - 1) r_A^2) \cdot B + C \cdot (p(\bar{x}) - 1) r_A^2)^2}{(B^2 + B)} \geq 0 \quad (EC.54)
$$

To see why (EC.54) is true, note the following global assumption $(r_{pd}^G > r_{pd}^A > 1$ and $r_h^G > r_h^A > 1)$, which simply insures that general charges are more than two units of effort and that charged revenue for one unit of effort is greater than $\$1$.

$$
0 \geq (r_A^2(2 \cdot p(\bar{x}) - p(\bar{x} + 1)) + C \cdot (p(\bar{x}) - p(\bar{x} + 1))) \cdot B^2
$$

$$
\geq CB^2 \cdot (p(\bar{x}) - p(\bar{x} + 1))
$$

$$
\geq -(p(\bar{x}) - p(\bar{x} + 1)) \cdot B^3.
$$

The last inequality holds because $B + C \geq 0$ so $-B \leq C$. This implies that

$$
(p(\bar{x}) - p(\bar{x} + 1)) \cdot B^3 + (r_A^2(2 \cdot p(\bar{x}) - p(\bar{x} + 1)) + C \cdot (p(\bar{x}) - p(\bar{x} + 1))) \cdot B^2 \geq 0
$$

Further, since $C \leq 0$ for all $x_0 \in \xi$, the second term (on the second line) of (EC.54) is clearly positive. Thus, $-b \leq 0$. Next we show that $a \geq 0$.

$$
a = \frac{(p(\bar{x}) - p(\bar{x} + 1))B^2 + (p(\bar{x}) - p(\bar{x} + 1))(C + r_A)(B + C r_A)(p(\bar{x}) - 1)}{B(B + r_A)}
$$

$$
\geq \frac{(p(\bar{x}) - p(\bar{x} + 1))B^2 + (p(\bar{x}) - p(\bar{x} + 1))CB}{B(B + r_A)} \geq 0. \quad (EC.55)
Further, we know that $c = -Cr^A \geq 0$. Hence $4ac \geq 0$ which implies that $\sqrt{b^2 - 4ac} \leq \sqrt{b^2} = b$, which finally implies that $-b + \sqrt{b^2 - 4ac} \leq 0$. Thus either the roots are negative or imaginary. Combining this with the fact that for $\theta_h = 0$
\[
\left[ \frac{\partial E[\pi_{pd}(x+1)]}{\partial \tau} - \frac{\partial E[\pi_{pd}(x)]}{\partial \tau} \right] - \left[ \frac{\partial E[\pi_h(x+1)]}{\partial \tau} - \frac{\partial E[\pi_h(x)]}{\partial \tau} \right] \geq 0,
\]
we have that all $\theta_h \geq 0$, s.t. $x \in \xi$, the marginal increase in contribution margin from an extra unit of effort as a function of $\tau$ is greater for the post-discharge organization than for the hospital, since the function will not cross zero as long as $\theta_h \geq 0$ and $x \in \xi$. \hfill \Box

**Proposition 2.** Let $r^A = r^A_{pd} = r^A_h$. Let $y$ be an amount of effort performed by a single provider. 
(i) $\exists \bar{x}$ s.t. if $x_h^B(y) \geq \bar{x}$, then $x_{pd}^B(y) \geq x_h^B(y)$
(ii) $\bar{x}$ is decreasing in $\tau$
(iii) $\bar{x}$ is decreasing in $r^C_h - r^C_{pd}$
(iv) $\exists \tilde{\tau}$ s.t. $\forall \tau \geq \tilde{\tau}, x_{pd}^B(y) \geq x_h^B(y)$

**Proof for Proposition 2:**

To give a sketch of the proof argument we fix one provider’s effort at $y$ and analyze the best response of the other provider. For illustration purposes, without loss of generality fix the post-discharge provider’s effort at $y$. From Lemma 1, we know the best response of the hospital will occur either at 0 or where $q_h(x, y) = c_h^A + p'(\bar{x})\gamma_h$.

Given that the best response will have $q_h$ approaching from above, if $q_h(x_h^B(y), y) < q_{pd}(x_h^B(y), y)$ then $q_{pd}(x_h^B(y), y) > c^A_{pd} + p'(\bar{x})\gamma_{pd}, x_{pd}^B(x) > y$.

$q_h$ will cross $c^A_{pd} + p'(\bar{x})\gamma_{pd}$ at a lower effort level than $q_{pd}$ since the $q_{pd}$ curve will be above the $q_h$ curve. Thus, the best effort from the post-discharge provider must be greater at the best effort point. For notational simplicity, we will write the functions $q_i(x)$ and $Z(x)$ as $q_i$ and $Z$, respectively.

We know that $q_h(x^*_h) \geq c^A_h - p'(x)\gamma_h$. Therefore if $q_{pd}(x^*_h) \geq q_h(x^*_h) - p'(x)\gamma_h$ then $q_{pd}(x^*_h) \geq c^A_{pd} + p'(x)\gamma_{pd} - p'(x)\gamma_h = c^A_h$. Since $q_{pd} \geq c^A_{pd}$, we know from Lemma 1 there are at most two values of $x$ where $q_{pd} = c^A_{pd}$. If $q_{pd}$ is always less than $c^A_{pd}$ then the optimal effort, $x^*_pd = 0$, we exclude this case. If there is one point, then $q_{pd}(x) \geq c^A_{pd} \forall x \in [0, x^*_pd]$ and $q_{pd}(x) < c^A_{pd} \forall x > x^*_pd$. If there are two points, $x_0$ and $x_1$, then $q_{pd}(x) \geq c^A_{pd} \forall x \in [x_0, x_1]$ and $q_{pd}(x) < c^A_{pd} \forall x < x_0$ and $x > x_1$.

Further, we know $x^*_pd = x_1$ since approaching the first order condition from below would not result in this solution being optimal. Thus, in all cases where $x^*_pd > 0$, we see that $x^*_pd$ always lies at the rightmost boundary of a single region $\{x : q(x) \geq c\}$. Thus, from the hospital and post-discharge provider scenario originally described, we know that $x^*_pd \geq x^*_h$. 

The last inequality holds because $B \geq -C$ hence $B^2 + CB \geq 0$. It directly follows that $-b/2a \leq 0$. To complete the proof, note that $c = -Cr^A \geq 0$. Hence $4ac \geq 0$ which implies that $\sqrt{b^2 - 4ac} \leq \sqrt{b^2} = b$, which finally implies that $-b + \sqrt{b^2 - 4ac} \leq 0$. Thus either the roots are negative or imaginary.
**Difference** The $q_{pd}$ and $q_h$ terms below come from the Lemma 1 proof for the value of $q_i(x)$ in Equation EC.19 and Equation EC.9, respectively. Each is calculated when the opposing provider performs zero or non-zero effort and we assume $r^A = r^A_{pd} = r^A_{h}$.

\[
q_{pd} - q_h = \tau p'(x) \left( -2\theta_h(x_r^G + r^A_{pd}(x_{pd})) \right) + \tau \frac{r^A[r^G_r - r^G_{pd} + r^A_h(x_h) - r^A_{pd}(x_{pd})]}{Z^2} + \tau p(x)\theta_h r^A \left( \frac{2(r^G_{pd} + r^A_{pd}(x_{pd}))(2Z + \theta_h) - Z^2 - Z\theta_h}{Z(Z + \theta_h)^2} \right)
\]

(EC.56)

Taking $q_{pd} - q_h \geq c^A_{pd} - c^A_h - p'(x)\gamma_h$ results in the expression below.

\[
\tau \frac{r^A[r^G_h - r^G_{pd} + r^A_h(x_h) - r^A_{pd}(x_{pd})]}{Z^2} + \tau p(x)\theta_h r^A \left( \frac{2(r^G_{pd} + r^A_{pd}(x_{pd}))(2Z + \theta_h) - Z^2 - Z\theta_h}{Z(Z + \theta_h)^2} \right) \geq -p'(x) \left( \gamma_h - \frac{2\theta_h (r^G_{pd} + r^A_{pd}(x_{pd}))}{Z(Z + \theta_h)} \right) + (c^A_{pd} - c^A_h)
\]

(EC.57)

which could be rearranged into the expression below.

\[
\tau r^A \left( p(x)\theta_h^2 [r^G_{pd} + r^A_{pd}(x_{pd}) - r^G_h - r^A_h(x_h)] + \theta_h [3r^G_{pd} + 3r^A_{pd}(x_{pd}) - r^G_h - r^A_h(x_h)](r^G_{pd} + r^A_{pd}(x_{pd}) + r^G_h + r^A_h(x_h)) \right.
\]

\[
\left. + \frac{[r^G_h + r^A_h(x_h) - r^G_{pd} - r^A_{pd}(x_{pd})]}{Z^2} \right) \geq -p'(x) \left( \gamma_h - \frac{2\theta_h (r^G_{pd} + r^A_{pd}(x_{pd}))}{Z(Z + \theta_h)} \right) + (c^A_{pd} - c^A_h)
\]

(EC.58)

For notational simplicity let $\eta(x)$ be the complex first term in the large parentheses multiplied by $p(x)$ and replace the second $p(x)'$ term on the right-hand-side with one. Thus we have

\[
\tau r^A \left( p(x)\eta(x) + \frac{[r^G_h + r^A_h(x_h) - r^G_{pd} - r^A_{pd}(x_{pd})]}{Z^2} \right) \geq -p'(x)(\gamma + 1) + (c^A_{pd} - c^A_h)
\]

(EC.59)

Finally, note that both sides of (EC.59) are decreasing in $x$. However, if we use an exponential curve for $p(x)$ the right-hand-side is decreasing faster than the left-hand-side. This can easily be seen by the fact that the right-hand-side is bounded below by $\frac{1}{(Z + \theta_h)^2}$ and $\max_{x \to \infty} \frac{1}{(Z + \theta_h)^2} / -p'(x) = \infty$. Thus, $\exists \hat{x} : \forall x \geq \hat{x}$ the left-hand-side of (EC.59) is greater than or equal to $-p'(x)(\gamma + 1)$, and hence if $x^*_h \geq \hat{x}$ then $x^*_{pd} \geq x^*_h$.

Since $\tau$ only appears on the left-hand-side, there is also a $\hat{\tau} : \forall \tau \geq \hat{\tau}$ the condition is true. Finally, $\hat{x}$ is decreasing in $\tau \square$

**PROPOSITION 3.** Let $\bar{x}^S = x^S_{pd} + x^S_h$ be the optimal readmission effort of the post-discharge provider and hospital when both are directed by a single controlling healthcare provider.

Then $\bar{x}^S = \bar{x}_{CE}$.
Proof for Proposition 3:

The single controlling healthcare provider looks to maximize the fixed payment $\tau$ minus the cost of all providers performing care $c_i(x)$, as shown in Equation 10. In this case, the $I$ providers are various provider groups with different cost structures under the direction of a single controlling healthcare provider.

Because $\tau$ is a constant, maximizing contribution margin in Equation 10 is equivalent to minimizing cost in Equation 3. □

Theorem 2. $\pi^{CE} \geq \pi^{BJR}$.

Proof for Theorem 2:

From Proposition 3, we know that $\bar{x}_S = \bar{x}^{CE}$. Since the revenue, $\tau$ is fixed and the effort decision variables for all cost terms are the same, this implies that $\pi^S = \pi^{CE}$.

If $\pi^{BJR} > \pi^S$, where $\pi^S = \pi_{pd}(x^S_{pd},x^S_h) + \pi_h(x^S_{pd},x^S_h)$, then there exists some combination of best response effort terms available to providers in a joint response not available to those same providers in a centrally controlled health system. However, a centrally controlled health system can always direct their providers to perform the joint best response effort level, which contradicts our initial inequality, proving $\pi^S \geq \pi^{BJR}$.

Thus, $\pi^{CE} = \pi^S \geq \pi^{BJR}$. □

Proposition 4. Given effort $x_j$ of provider $j$, $x_i^{B+B}(x_j) \geq x_i^B(x_j)$.

Proof for Proposition 4:

By definition $x_i^B(x_j)$ is the best response of a provider without any VBB term, given the other provider effort is $x_j$. Therefore, the contribution margin of a provider without VBB is maximized at $\pi_i(x_j, x_i^B(x_j))$. From Section 4.3, we also know that $\pi_{VBB,i}(x) = \pi_i(x) - \mathcal{V}_i(\bar{x})$.

If $x_i^{B+B}(x_j) < x_i^B(x_j)$, then $\pi_i(x_j, x_i^{B+B}(x_j)) \leq \pi_i(x_j, x_i^B(x_j))$ because $x^B$ is the optimal solution to $\pi_i$. Further, since $\mathcal{V}_i(\bar{x}) \geq 0$ and decreasing in $\bar{x}$, $\mathcal{V}_i(x_j + x_i^{B+B}(x_j)) > \mathcal{V}_i(x_j + x_i^B(x_j))$, because we assume $x_i^{B+B}(x_j) < x_i^B(x_j)$.

Therefore, $\pi_i(x_j, x_i^{B+B}(x_j)) - \mathcal{V}_i(\bar{x} + x_i^{B+B}(x_j)) < \pi_i(x_j, x_i^B(x_j)) - \mathcal{V}_i(\bar{x} + x_i^B(x_j))$. This contradicts the fact that $x_i^{B+B}(x_j)$ is an optimal solution to $\pi_{VBB,i}(x)$. Thus, $x_i^{B+B}(x_j) \geq x_i^B(x_j)$. □

Corollary 3. $\bar{x}^{S,V} > \bar{x}^{CE}$.

Proof for Corollary 3:

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, $\bar{x}$. $\mathcal{V}(\bar{x})$ reacts in a similar manner to the
readmission probability curve and is therefore also convex in total effort. Since both providers are controlled/owned by one entity it does not matter which has a VBB arrangement. Let $c^{A,Tot}(\bar{x})$ be the cost of the total effort, allowing us to rewrite Equation 3 into Equation EC.60 below.

$$\min_x \left[ E[\omega(x)] \right] = \sum_{i=1}^{I} c_i^G + \min_x [c^{A,Tot}(\bar{x}) + p(\bar{x})\gamma_h] \quad \text{(EC.60)}$$

From the first order conditions of Equation EC.60 for the cost-effective effort level, the optimal effort $\bar{x}^{CE}$ is found where

$$c^{A,Tot}(\bar{x}^{CE}) = -p'(\bar{x}^{CE})\gamma_h \quad \text{(EC.61)}$$

However, including the VBB value(s) with the single controlling healthcare provider contribution margin results in Equation EC.62.

$$\max_x \left[ \tau - \left( \sum_{i=1}^{I} \left[ c_i^G + c_i^A(x_i) + p(\bar{x})\gamma_h + V(\bar{x}) \right] \right) \right] \quad \text{(EC.62)}$$

Thus, an optimal effort level of $\bar{x}^{S,V}$ is found where

$$c^{A,Tot}(\bar{x}^{S,V}) + V'(\bar{x}^{S,V}) = -p'(\bar{x}^{S,V})\gamma_h \quad \text{(EC.63)}$$

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

$$c^{A,Tot}(\bar{x}^{CE}) + V'(\bar{x}^{CE}) < -p'(\bar{x}^{CE})\gamma_h \quad \text{(EC.64)}$$

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 EC.64 an equality again at the optimal effort with a VBB program it must be that $\bar{x}^{S,V} > \bar{x}^{CE}$.

**Risk-profiling effectiveness.** If $\tau_{z_{Lo}} < \tau_{Ag} < \tau_{z_{Hi}}$, and $q_{pd,r}(x_r)$ is approximated by an exponential function, $\beta_{rl}^0 e^{-\beta^r x_{pd,r}}$, then

(i) If $\tau_{z_{Lo}} < \tau_{Lo}, \tau_{Hi} < \tau_{z_{Hi}}$ then $f(\tau_{Lo}, \tau_{Hi}) < f(\tau_{Ag}, \tau_{Ag})$

(ii) If $\tau_{z_{Lo}} < \tau_{Lo}, \tau_{Hi} > \tau_{z_{Hi}}$, then there exists a threshold, $\tilde{\alpha}_1$, s.t. $f(\tau_{Lo}, \tau_{Hi}) < f(\tau_{Ag}, \tau_{Ag}) \forall \alpha > \tilde{\alpha}_1$

(iii) $\tilde{\alpha}_1$ is decreasing in $\epsilon_0$ and $\epsilon_{Hi}^1$, and $c_{pd}^A$ and is increasing in $\epsilon_{Lo}^1$

(iv) If $\tau_{Lo} < \tau_{z_{Lo}} < \tau_{Hi}$ then there exists a threshold, $f(\tau_{Lo}, \tau_{Hi}) < f(\tau_{Ag}, \tau_{Ag}) \forall \alpha > \tilde{\alpha}_2$

(v) $\tilde{\alpha}_2$ is decreasing in $\epsilon_0$ and $c_{pd}^A$

(vi) If $\tau_{Lo} < \tau_{z_{Lo}}, \tau_{Hi} < \tau_{z_{Hi}}$ and $(c_{pd}^A)^2 < \max\{\tau_{Ag}^2, 2 \cdot \tau_{Lo}^2\} \text{ then there exists a threshold,}$

$\tilde{\epsilon}_0$, s.t. $f(\tau_{Lo}, \tau_{Hi}) < f(\tau_{Ag}, \tau_{Ag}) \forall \epsilon_0 < \tilde{\epsilon}_0$
Proof for Theorem 3: (i) For fixed effort by the hospital, post-discharge effort is increasing in \( \tau \). Thus, since \( \tau_{xCE} < \tau_{Lo} < \tau_{Ag} \), we have that \( 0 < x_{pd,Lo} - \bar{x}_{Lo} < x_{pd,Ag} - \bar{x}_{CE,Lo} \). The same argument applies for \( \tau_{Hi} \).

Next, we consider the cases (ii) - (vi). We begin with the first order condition for the post-discharge provider which determines their optimal effort level: \( q_{pd,r}(x_h, x_{pd}) = c^A_{pd} \). Inverting the function with respect to \( x_{pd} \) (since we leave \( x_h \) as a constant), we get that \( x_{pd}^B(x_h) = q_{pd}^{-1}(c^A_{pd}, x_h) \). For the remainder of the proof, we slightly abuse notation by dropping the dependence on \( x_h \) for compactness. Plugging in the best response for the post-discharge provider, we obtain the following condition for \( f(\tau_{Lo}, \tau_{Hi}) < f(\tau_{Ag}, \tau_{Ag}) \):

\[
\alpha \left| q_{pd,Lo} \left( \frac{c^A_{pd}}{\tau_{Lo}} \right)^{-1} - x_{CE,Lo} \right| + (1 - \alpha) \left| q_{pd,Hi} \left( \frac{c^A_{pd}}{\tau_{Hi}} \right)^{-1} - x_{CE,Hi} \right| < \alpha \left| q_{pd,Lo} \left( \frac{c^A_{pd}}{\tau_{Ag}} \right)^{-1} - x_{CE,Lo} \right| + (1 - \alpha) \left| q_{pd,Hi} \left( \frac{c^A_{pd}}{\tau_{Ag}} \right)^{-1} - x_{CE,Hi} \right| \quad \text{(EC.65)}
\]

Further, we can solve for \( x_{pd}^{CE} \) using the first order condition \( p'(x) = c^A_{pd} \) given that post-discharge provider’s cost of effort is no greater than the hospital’s costs (as per our assumption). Solving for \( x_{pd} \), we get

\[
x_{pd,CE} = -\frac{\ln \left( \frac{c_{pd}}{\lambda_2} \right)}{\lambda_2}.
\]  

(Substituting \( q_{pd,r}(x_{pd,r}) = \beta_0^e \beta^r x_{pd,r} \) for patients of risk level \( r \) and substituting \( x_{pd,CE}^{CE} \) into Condition EC.65 risk-adjusted bundled payments are closer to cost effective under the following conditions.

(ii) - (iii) If \( \tau_{Lo} < \tau_{xCE}^{CE} < \tau_{Hi} \), and

\[
(\tau_{Hi}^0)^{(1 - \alpha)} \beta^{Lo} - (\tau_{Lo}^0)^{(1 - \alpha)} \beta^{Hi} + \left( \tau_{Ag}^0 \beta_0^Hi \right)^{(1 - \alpha)} \beta^{Lo} - \left( \tau_{Ag}^0 \beta_0^Hi \right)^{(1 - \alpha)} \beta^{Hi} < (c^A_{pd}^2 (1 - \alpha) \beta^{Lo} - \alpha \beta^{Hi}) + \varepsilon_0 (1 - \alpha) \beta^{Lo} \beta^{Hi} (1/\lambda_2 - 1/\lambda_2^0) \left( \frac{\varepsilon_1^{Hi} (1 - \alpha) \beta^{Lo} \beta^{Hi}}{\lambda_2^0} \right),
\]  

Then \( f(\tau_{Lo}, \tau_{Hi}) < f(\tau_{Ag}, \tau_{Ag}) \).

Denote the left-hand-side of (EC.67) by \( h(\alpha) \) and the right-hand-side by \( g(\alpha, \varepsilon_0, \varepsilon_1^{Hi}, \varepsilon_1^{Lo}) \). It is clear that \( h(\alpha) \) is decreasing in \( \alpha \) and setting \( \alpha = 1 \) results in \( - (\tau_{Hi}^0)^{(1 - \alpha)} \beta^{Lo} - (\tau_{Ag}^0 \beta_0^Hi \beta^{Hi} (1/\lambda_2 - 1/\lambda_2^0) \left( \frac{\varepsilon_1^{Hi} (1 - \alpha) \beta^{Lo} \beta^{Hi}}{\lambda_2^0} \right) < 0. \) Since the left-hand-side is a continuous function of \( \alpha \), then by the intermediate value theorem there must exist some \( \hat{\alpha} \) such that the left-hand-side of (EC.67) equals zero. Finally, since the left-hand-side is strictly decreasing in \( \alpha \), then for all \( \alpha > \hat{\alpha} \), \( h(\alpha) < 0. \) Finally, \( g(\alpha, \varepsilon_0, \varepsilon_1^{Hi}, \varepsilon_1^{Lo}) \) is increasing
in $\varepsilon_0$, $\varepsilon_1^{Hi}$, and $c_{pd}^A$ and decreasing in $\varepsilon_1^{Lo}$. Thus, $\tilde{\alpha}$ is decreasing in $\varepsilon_0$, $\varepsilon_1^{Hi}$, and $c_{pd}^A$ and increasing in $\varepsilon_1^{Lo}$ since the right-hand-side is larger.

(iv) – (v) If $\tau_{Lo}^{CE} < \tau_{Lo}$, $\tau_{Hi}^{CE} < \tau_{Hi}$, and

\[
(\tau_{Hi}^{0} \beta_0^{Hi})(1-\alpha)\beta^{Lo} + (\tau_{Lo}^{0} \beta_0^{Lo})\alpha\beta^{Hi} + (\tau_{Ag}^{0} \beta_0^{Hi})(1-\alpha)\beta^{Lo} - (\tau_{Ag}^{0} \beta_0^{Lo})\alpha\beta^{Hi}
\]

\[
< \left( c_{pd}^A \right)^{2(1-\alpha)}\beta^{Lo} + \left( \varepsilon_0 \varepsilon_1^{Hi} \right)^{2(1-\alpha)}\beta^{Lo} \frac{\beta^{Hi}}{\lambda^{Hi}}. \tag{EC.68}
\]

Then $f(\tau_{Lo}, \tau_{Hi}) < f(\tau_{Ag}, \tau_{Ag})$.

Denote the left-hand-side of (EC.68) by $h(\alpha)$ and the right-hand-side by $g(\alpha, \varepsilon_0, \varepsilon_1^{Hi})$. It is clear that $h(\alpha)$ is decreasing in $\alpha$ since $\tau_{Lo} < \tau_{Ag}$. Further, setting $\alpha = 1$ results in $\tau_{Lo} - \tau_{Ag} < 0$. Using the same arguments as previously we have that there exists an $\tilde{\alpha}_2$. Finally, $g(\alpha, \varepsilon_0, \varepsilon_1^{Hi})$ is increasing in $\varepsilon_0$, $\varepsilon_1^{Hi}$, and $c_{pd}^A$. Thus, $\tilde{\alpha}$ is decreasing in $\varepsilon_0$, $\varepsilon_1^{Hi}$, and $c_{pd}^A$ since the right-hand-side is larger.

(vi) If $\tau_{Lo} < \tau_{Lo}^{CE}$, $\tau_{Hi} < \tau_{Hi}^{CE}$ and

\[
\left( c_{pd}^A \right)^{2(1-\alpha)}\beta^{Hi} \frac{\beta^{Lo}}{\lambda^{Hi}} + \left( \varepsilon_0 \varepsilon_1^{Hi} \right)^{2(1-\alpha)}\beta^{Lo} \frac{\beta^{Hi}}{\lambda^{Hi}}
\]

\[
< (\tau_{Hi}^{0} \beta_0^{Hi})(1-\alpha)\beta^{Lo} + (\tau_{Lo}^{0} \beta_0^{Lo})\alpha\beta^{Hi} - (\tau_{Ag}^{0} \beta_0^{Hi})(1-\alpha)\beta^{Lo} + (\tau_{Ag}^{0} \beta_0^{Lo})\alpha\beta^{Hi}. \tag{EC.69}
\]

Then $f(\tau_{Lo}, \tau_{Hi}) < f(\tau_{Ag}, \tau_{Ag})$.

It is clear that the left-hand-side of (EC.69) is increasing in $\varepsilon_0$ whereas the right-hand-side is independent of $\varepsilon_0$. As $\varepsilon_0$ goes to zero, we are left with $(c_{pd}^A)^{2(1-\alpha)}\beta^{Hi}$. If this is smaller than the right-hand-side, then there exists an $\tilde{\varepsilon}_0$ s.t. the condition holds. A sufficient condition is for $(c_{pd}^A)^{2(1-\alpha)}\beta^{Hi} < (\tau_{Ag}^{0} \beta_0^{Hi})(1-\alpha)\beta^{Lo} + (\tau_{Lo}^{0} \beta_0^{Lo})\alpha\beta^{Hi}$. For this to hold it is sufficient that $(c_{pd}^A)^2 < \max \{ \tau_{Ag}^{0} \beta_0^{Lo}, 2 \cdot \tau_{Lo}^{0} \beta_0^{Lo} \}$.

□