Promoting Collaborative Care: Relative Performance-based Payment Models for Hospitals and Post-acute Care Providers

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

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

Tolga Tezcan
Rice University, Houston, TX, US
tolga.tezcan@rice.edu

Bundled payment models are widely used in healthcare reimbursement but they were primarily designed for conditions managed by a single provider in a centralized manner. Therefore, these models do not account for the complexities of conditions requiring post-acute care (PAC) after hospitalization, resulting in weak incentives for care coordination between hospitals and PAC providers, particularly in decentralized settings. Motivated by the Comprehensive Care for Joint Replacement (CJR) payment model introduced by CMS—which holds hospitals accountable for the cost and quality of the entire episode, including PAC—we examine the effectiveness of existing payment models in decentralized care. Our analysis shows that bundled payment models and the CJR framework, especially without gainsharing agreements, do not fully align provider incentives. To address these shortcomings, we propose a payment model that explicitly links hospital and PAC provider reimbursements. Using a game-theoretical framework, we show that this model results in socially optimal provider actions. Moreover, we establish that similar efficiency gains can be achieved within the CJR framework if hospitals and PAC providers adopt structured gainsharing agreements. Numerical simulations show that implementing this approach could reduce readmission costs by 12%, generating over $150 million in annual savings for Medicare joint replacement procedures alone.

Key words: Healthcare, regulation, information asymmetry, coordination, payment models

1. Introduction

Diagnosis-Related Group (DRG) based payment models have been widely implemented across numerous countries, including the United States, Canada, Australia, New Zealand, Germany, and Sweden, with the goal of promoting efficient and high-quality healthcare delivery (OECD 2019). These systems operate by assigning predetermined payment amounts for specific medical conditions or treatments, based on the average cost of treating patients within the respective DRG. The payment amount serves as a cost benchmark to incentivize cost efficiency improvement—healthcare providers who manage to operate more efficiently than the average are able to generate positive

margins from their services. The use of relative cost performance benchmarks follows the classic principle of yardstick regulation and has been shown to elicit socially optimal actions (in balancing expenditures and benefits) of cost reduction when the regulator is less informed than agents of their cost structures (Shleifer 1985). The implementation of DRG-based payment systems, particularly through the inpatient prospective payment system (PPS) initiated by the Centers for Medicare & Medicaid Services (CMS) in 1983, has resulted in reduced hospital spending growth for Medicare (Davis and Rhodes 1988) and has led to the introduction of PPSs in other settings such as post-acute care (PAC).[1]

However, despite the success in incentivizing cost reduction, there is a legitimate concern that PPSs potentially encourage providers to prioritize cost savings over the quality of care. To address this concern, regulators have introduced outcome-based payment systems, commonly known as pay-for-performance payment models. These models go beyond simply reimbursing providers for services rendered or tasks completed, by linking a portion of DRG-based payments to health outcomes and treatment quality.[2] In these models, the magnitude of reward and penalty payments for each provider is determined based on their relative performance compared to other providers, similar to the approach used in DRG-based payments. Extensive empirical research has been conducted to examine the effects of outcome-based payment models (see Blumenthal et al. (2015) among others). Furthermore, there is a growing body of literature exploring the design and effectiveness of these payment models; see, for example, Arifoğlu et al. (2021), Savva et al. (2019), Chen and Savva (2018), Zhang et al. (2016).

Care coordination: Prospective and outcome-based payment models, which we refer to as single-entity payment models, have proven effective in cases where treatment decisions are made in a *centralized* manner within a single entity (e.g., a hospital or a PAC provider) for a specific episode of care. However, certain medical conditions, such as joint replacement, involve multiple independent providers, leading to *decentralized* treatment decisions.

Following joint replacement surgery, a significant number of Medicare beneficiaries are discharged from hospitals or other acute-care settings to various PAC settings, including skilled nursing facilities, inpatient rehabilitation facilities, and home health agencies (Li et al. 2020, Schwarzkopf et al. 2016). These PAC settings vary in the intensity and complexity of the medical, skilled nursing, and rehabilitative services they provide (Department of Health and Human Services 2017) and the cost of PAC can constitute a substantial portion of the overall care expenses (Barnett et al. 2019).

[1] In PAC settings, patients are classified into resource utilization groups (RUG) not DRGs. However, the PPSs for acute and post-acute care use the same payment structure otherwise. Therefore, we use the term DRG-based models for both settings.

[2] Notable examples of these models include the Hospital Value-Based Purchasing (VBP) Program and the Skilled Nursing Facility Value-Based Purchasing (SNF VBP) programs, as outlined in CMS (2023).

While other conditions, such as stroke, traumatic injuries, and pneumonia, may also require PAC, our primary focus in this paper is on joint replacement due to the recent emphasis placed on these procedures by the CMS (Department of Health and Human Services 2017).

Effective coordination between hospitals and PAC providers is important for achieving favorable outcomes in joint replacement procedures. PAC for joint replacement typically involves a range of services designed to support the patient's recovery, such as physical therapy and occupational therapy (MedPAC 2022). Acute-care providers (usually working in a hospital setting) play a critical role in ensuring successful PAC outcomes as well through coordination and effective transition of care (Arana et al. 2017, Department of Health and Human Services 2021). Hospitals and PAC providers can implement several strategies to improve coordination and transitions, including the use of connected electronic health records and other technology tools, establishing partnerships, scheduling post-discharge visits from hospital physicians, and implementing joint education and training programs (Adler-Milstein et al. 2021, Britton et al. 2017, Cipriano et al. 2018).

In the past, CMS utilized separate PPSs to reimburse hospitals and PAC providers for their individual contributions to joint replacement treatment. However, these payment methods, even with outcome-based adjustments, failed to provide adequate incentives for investments in care coordination (Department of Health and Human Services 2021). This was because additional investments made by one party to enhance the efficiency of the other party were not reimbursed under the traditional DRG payment structure. Furthermore, these payment methods could lead to unintended consequences, such as hospitals discharging patients to PAC settings with unnecessary intensive care in an attempt to reduce readmissions (Zhu et al. 2018) because hospitals are penalized for excess 30-day readmissions under the Hospital Readmissions Reduction Program (HRRP). Consequently, a more comprehensive approach to payment models was needed, one that considers the collaborative nature of care delivery across multiple entities (in a decentralized manner) in the context of joint replacement procedures.

CJR Program: To enhance coordination between hospitals and PAC providers, CMS implemented the Comprehensive Care for Joint Replacement (CJR) model in 2016 across specific geographic areas in the United States. The CJR model aimed to address the challenges of care coordination by holding participating hospitals financially accountable for the entire episode of care, encompassing hospitalization and all PAC services for 90 days post-discharge. CMS establishes a target price for hip and knee replacements based on the average regional treatment cost, with quality of care taken into account through a composite-quality score adjustment of up to 3%. Participating hospitals receive a reconciliation payment if the total episode of care costs during the performance year are below the target price, while repayments may be required if the total episode

of care costs exceed the target price (CMS 2018, Department of Health and Human Services 2021). However, PAC providers continue to be paid using an outcome-based PPS.

The CJR program draws on concepts from bundled-payment models but incorporates unique features. While (conventional) bundled payment models involve a single payment to cover the entire episode of care (a feature CJR program adopted), the CJR program utilizes separate payments to hospitals and PAC providers, recognizing them as distinct entities. This separation introduces challenges in assigning responsibility for costs and quality outcomes across the full episode of care. Unlike single-entity models where one organization oversees all aspects of treatment, hospitals in the CJR program collaborate with multiple PAC providers, complicating the attribution of costs and care quality. This fragmentation can lead to misaligned incentives, where one provider benefits from the efforts of another without contributing proportionally—a phenomenon known as free-riding in multi-provider settings (Salanie 1997, Chapter 5.3.8).

To further encourage care coordination, the CJR program permits hospitals to establish gainsharing agreements with PAC providers, allowing them to share financial rewards and penalties. However, designing effective gainsharing agreements is challenging, particularly due to informational asymmetries between hospitals and PAC providers (Gupta et al. 2021, Ghamat et al. 2021), and CMS has provided limited guidance on this issue. Additionally, hospitals cannot mandate which PAC provider a patient chooses, further complicating coordination efforts (McGarry and Grabowski 2017). Furthermore, despite the intent to encourage collaboration, these agreements have seen minimal adoption in practice. Interviews conducted with CJR hospitals indicate that gainsharing agreements were virtually nonexistent during the program's implementation (Hopewell et al. 2024, Ghamat et al. 2021). As a result, empirical evaluations have found that the CJR program has achieved only modest and statistically insignificant cost savings, primarily through reduced PAC utilization rather than efficiency improvements (CMS 2021a, Barnett et al. 2019, Finkelstein et al. 2018).

Analysis of existing payment models: Given the limitations of the CJR program in effectively incentivizing care coordination we analyze various payment models to understand their impact on provider behavior and patient outcomes. To assess the long-term impact of different payment models, we develop a stylized model where the cost efficiency of care provided by a PAC provider can be improved by additional investment from the hospital that provided acute care to the patient, and the readmissions can be reduced through collective investments.

Using this framework, we first examine the bundled payment model that precedes the CJR program and demonstrate that, while bundled payment improves cost containment compared to PPS, it fails to incentivize hospitals and PAC providers to coordinate care effectively. Specifically, these

models do not provide hospitals with incentives to invest in reducing PAC costs, and provide inadequate incentives for providers to collectively reduce readmissions. As a result, bundled payment results in suboptimal coordination and higher overall costs.

Next, we introduce a CJR-type payment model that captures the key financial incentives of the CJR program, focusing on settings where gainsharing agreements are absent. This payment model extends the bundled payment model by making hospitals responsible for the total episode of care costs, including PAC services and readmissions. We show that, while this model incentivizes hospitals to invest in lowering PAC costs, it does not induce first-best levels of investment in reducing readmissions. The misalignment arises because hospitals bear full financial responsibility for readmissions, while PAC providers remain accountable only for their own costs. This imbalance in financial accountability limits the incentives for PAC providers to collaborate in reducing hospital readmissions and costs, thereby reinforcing fragmentation in post-acute care.

Improving payment models: Recognizing these limitations, we propose a new payment model that explicitly links hospital and PAC provider payments to each other's performance. In contrast to the CJR-type payment model, which adjusts hospital payments based on total episode costs while leaving PAC provider payments unchanged, our payment model introduces an outcome-based adjustment that ensures both hospitals and PAC providers share financial responsibility for readmission costs. This structure eliminates the incentive misalignment and encourages collaboration between hospitals and PAC providers to improve patient outcomes.

We prove that our payment model elicits first-best provider actions, i.e., both hospitals and PAC providers invest at socially optimal levels in reducing costs of care and readmissions. By aligning financial incentives across entities, our model eliminates the incentive misalignment under existing bundled payment and CJR-type models. Furthermore, we demonstrate that these first-best outcomes can also be achieved within the existing CJR framework if hospitals and PAC providers enter into a specific form of gainsharing agreement. Additionally, we demonstrate the flexibility of our payment model in incorporating various practical considerations and discuss the information requirements necessary to achieve care coordination in both the CJR and our proposed models, emphasizing the critical role of accurate PAC cost estimates.

Through numerical experiments calibrated to Medicare data, we quantify the cost-saving potential of our proposed payment model. Our results show that, compared to the bundled payment model, our payment model reduces readmission costs by 12%, translating into estimated annual savings exceeding $150 million for Medicare joint replacement procedures alone. Moreover, even partial implementation of our model—where hospitals and PAC providers share 50% of readmission costs—produces substantial improvements, nearly matching first-best levels of care coordination.

As policymakers consider expanding bundled payment models beyond joint replacement to other multi-entity care settings, our framework provides a strong foundation for designing more effective payment models that can drive broader improvements in care quality and cost efficiency.

2. Literature review

In this section we review the relevant literature on healthcare payment models and describe the contribution of our research to different streams in this literature. We also explain our contribution to the literature on moral hazard in teams.

Research on payment models for a single-entity setting: Our research contributes to the literature on designing healthcare payment models (see for example So and Tang (2000), Jiang et al. (2012, 2020, 2021), Ata et al. (2013), Adida et al. (2016), Bastani et al. (2016), Aswani et al. (2019), Bavafa et al. (2021), Adida and Bravo (2023), Hwang et al. (2023), Savva et al. (2023), Goodman and Dai (2024)), particularly those focused on relative performance-based payment models currently employed by CMS. Savva et al. (2019) show that augmenting cost-based payment models by a waiting time-based adjustment can incentivize waiting time reduction. Arifoğlu et al. (2021) identify inefficiencies of the HRRP and propose a bundled payment model to elicit the socially optimal levels of cost and readmission reductions.

Our proposed payment model similarly draws on the relative (to other providers) performance to determine each provider's payment. However, while existing literature focuses on centralized care settings where a single provider manages the entire care episode, our paper examines decentralized care settings where different providers independently manage segments of a patient's care, such as acute and post-acute care. This necessitates a new modeling approach and novel payment models to explore interactions and achieve coordination between these providers, and assess the effects of prevailing payment models on cost and quality of *collaborative care*. We use our model to show that single-entity payment models, as discussed in previous studies, are inadequate for promoting care coordination and achieving optimal outcomes. Additionally, we examine a CJR-type payment model and unveil its limitations in incentivizing care coordination.

Research on payment models for care coordination: Our research makes a significant contribution to the literature on payment models for care coordination across different settings. This body of work primarily focuses on comparing the potential cost and quality of care improvements resulting from a transition from traditional fee-for-service payment models to episode-based payment models, such as bundled payments.

Adida and Bravo (2019) study reimbursement contracts between a managing organization offering basic care, and an external provider providing advanced care, which is reimbursed by the former. Bravo et al. (2023) consider patient referral by an accountable care organization to a preferred

external provider under the Medicare Shared Saving Program, and design cost- and risk-sharing contracts to improve coordination in referral markets. Rajagopalan and Tong (2022) consider a general practitioner's patient referral to a specialist with limited capacity. They find that making a bundled payment shared by different providers leads to higher referral rates and lower service time by the specialist, and propose to use a variable payment per unit of service time to achieve first-best outcomes. Vlachy et al. (2023) explore the impact of bundled payments on care intensity co-production by a hospital and a physician. Gupta and Mehrotra (2015) analyzes a regulator's optimal selection of healthcare bundles with key parameters chosen by proposers.

A key distinction between this literature and our study lies in our focus on relative performance-based payment models as currently employed by CMS. In contrast, the existing literature primarily explores payment models with exogenously determined reimbursement amounts and performance targets, assuming the regulator possesses the required information to establish these measures.

Additionally, there is a body of literature that examines the design of gainsharing contracts in existing payment models. Gupta et al. (2021) investigate the design of gainsharing contracts between hospitals and physicians when the overall payment for care is based on a bundled payment model. Similarly, Ghamat et al. (2021) explore the design of gainsharing contracts between hospitals and PAC providers under CJR. In contrast, our study focuses on the design of payment models for scenarios where care is delivered by multiple providers, rather than solely addressing the allocation of a single bundled payment among different providers. In addition, similar to other papers reviewed above, Gupta et al. (2021) and Ghamat et al. (2021) do not utilize relative performance-based payment models.

One relevant study in this literature is Zorc et al. (2017). Their research also addresses the design of payment models to promote care coordination, with a specific focus on the care pathway between general practitioners (GPs) and specialists, aiming to reduce delays in accessing specialist treatment. As in our work, they explore the design of relative performance-based payment models and demonstrate how model parameters can be determined using the performance of other providers, such as through yardstick regulation, under various scenarios.

However, a key distinction between their study and ours lies in the focus of provider pairs. While they concentrate on the coordination between GPs and specialists, our study considers a broader network of hospital and PAC providers. As a result, their model is more applicable to care coordination between GPs and specialists, as their underlying assumptions differ from ours. Specifically, in our model, we incorporate the notion that effective PAC treatment requires additional investment from hospitals directly, whereas their model assumes that the quality of care at each stage is solely determined by the provider operating at that specific stage.

Research on the impact of CJR: A body of literature empirically demonstrates the impact of the CJR program on provider actions (Ellimoottil et al. 2016, Finkelstein et al. 2018, Barnett et al. 2019, Haas et al. 2019, Einav et al. 2022, Chen and Delana 2025). These studies show the early effects of CJR on patient discharge destinations and highlight some of the incentive issues inherent in the CJR program. This literature further motivates our research on designing more effective payment models for the decentralized multi-entity settings.

Research on moral hazard in teams: Our paper contributes to the broader theory of moral hazard in teams. Holmstrom (1982) demonstrates that, in team production from agents' unobservable inputs, free-riding arises as agents input less than the first-best level. He further designs group incentives that elicit socially optimal actions by basing agent payments on team production. Mookherjee (1984) provides necessary and sufficient conditions for the optimality of independent contracts and rank order tournaments. Although our proposed payment model also provides group incentives, a crucial distinction exists: in this literature, the regulator knows cost structures but not agent inputs, so it can calculate the first-best inputs and contract based on the resulting team production. However, in our setting the regulator does not know cost structures and cannot calculate the first-best actions. We overcome this information asymmetry with yardstick competition as in Shleifer (1985) while extending it from single-entity settings to a multi-entity setting for collaborative care. We show that the bundled payment and CJR-type payment models fail to elicit first-best actions, and develop an innovative payment model that achieves the first best by adjusting provider payments using other providers' performance for the entire episode of care.

3. Model and first-best outcome

We examine an episode of care for a specific condition, such as joint replacement, which consists of two stages. The first stage involves acute care delivered in a hospital, followed by PAC provided by another entity like a skilled nursing facility (SNF). To account for the impact of care quality decisions, we assume that unsuccessful care may lead to patient readmission to the hospital, requiring the patient to restart the care process there (approximately 23% of Medicare patients discharged to SNFs experience readmission to the hospital within 30 days (Britton et al. 2017)).

We consider three types of decision-makers—the regulator, hospitals, and PAC providers. The regulator sets the payment model that dictates how providers are reimbursed for providing acute and post-acute care, with the objective of maximizing total welfare. Informed by the terms of the reimbursement scheme, hospitals and PAC providers choose their investments (or expenditures) in reducing the cost of care and readmission probability to maximize their expected profits. We establish the Nash equilibrium of this game to assess the long-term impact of different reimbursement schemes on the cost and quality performance of collaborative care.

In the rest of this section, we first describe the care model and the objective function of each decision-maker. Then, we establish the first-best outcome, i.e., provider actions that maximize total welfare. Finally, we discuss the information asymmetry between the regulator and providers, how it prevents the direct enforcement of the first-best outcome, and has led healthcare regulators to adopt relative performance-based payment models.

Care model: Each patient receives an episode of care that includes both acute and post-acute stages with (expected) associated costs C^h and C^s incurred by the hospital and the PAC provider, respectively. A patient is readmitted with probability R, which measures the quality of care. Hospitals and PAC providers make investments to reduce the cost and improve the quality of care as described next.

We define the cost of acute care per patient as $C^h : [0, \Gamma] \to \mathbb{R}^+$, a function of the action a^h taken by the hospital. The action a^h represents the total expenditures by the hospital to reduce the acute care cost C^h per patient, with an upper bound of $\Gamma > 0$. For example, hospitals may invest in developing health information technologies, improving care delivery processes, and other initiatives aimed at enhancing cost-efficiency.

After receiving acute care, a patient is discharged from the hospital to receive post-acute care by a PAC provider.[3] The per-patient *cost of PAC*, denoted by $C^s : [0, \Gamma] \times [0, \Gamma] \to \mathbb{R}^+$, depends on the hospital's action b^h and the PAC provider's action b^s. Similar to acute care, b^h and b^s are considered in monetary terms with upper bound Γ, without loss of generality. For example, hospitals and PAC providers may invest in dedicating more staff time, or hiring additional staff to coordinate care and improving coordination through shared electronic medical record systems.

After receiving acute and post-acute care for the initial (i.e., index) admission, a patient is rehospitalized with probability $R : [0, \Gamma] \times [0, \Gamma] \to [0, 1]$. This probability depends on the per-patient investments of the hospital (e^h) and the PAC provider (e^s) towards reducing readmissions. If readmitted, the patient receives acute and post-acute care from the same providers as in the initial admission, with fixed expected costs denoted by ξ^h and $\xi^s > 0$, respectively.[4]

Our model can incorporate endogenous costs of care for readmitted patients, as discussed in Appendix F. In addition, for expositional simplicity, we present a simple model where the actions of the providers are captured by one-dimensional variables, i.e., e^h, e^s, b^h, and b^s. We present a more detailed model in Appendix J and prove that our results are valid in that more general setting as well.

[3] Equivalently, we capture a setting where a fixed (i.e., exogenous) fraction of patients require PAC, and our main model focuses exclusively on these patients. We extend the model in Appendix E to consider different types of PAC providers and endogenous discharge decisions.

[4] Since ξ^h is the expected cost of acute care among all readmitted patients, our model accounts for cases where some patients are readmitted directly to the PAC setting.

Providers (hospitals and PAC providers): We consider a provider network consisting of N hospitals indexed by $i = 1, \ldots, N$, and M PAC providers indexed by $j = 1, \ldots, M$, with the fraction of patients discharged from hospital i to PAC provider j denoted by $p_{ij} \geqslant 0$. For notational simplicity, we denote $\mathcal{K} \equiv \{1, \ldots, K\}$ and $\mathcal{K}_i \equiv \mathcal{K} \setminus i$ for any given integer K throughout the paper.

For each hospital $i \in \mathcal{N}$, we use $p_i \equiv \sum_{j \in \mathcal{M}} p_{ij}$ to denote the fraction of patients receiving acute care from the hospital and assume that $p_i > 0$ for all $i \in \mathcal{N}$, i.e., each hospital provides care to some patients. Similarly, for each PAC provider $j \in \mathcal{M}$, we use $\tilde{p}_j \equiv \sum_{i \in \mathcal{N}} p_{ij}$ to denote the fraction of patients receiving PAC from the provider and assume that $\tilde{p}_j > 0$ for all $j \in \mathcal{M}$, i.e., each PAC provider treats some patients. Without loss of generality, we normalize the total patient population to one, i.e., $\sum_{i \in \mathcal{N}} p_i = \sum_{j \in \mathcal{M}} \tilde{p}_j = 1$.

Next, we derive the objectives of all hospitals $i \in \mathcal{N}$ and PAC providers $j \in \mathcal{M}$. We append subscripts to each decision-making parameter, as introduced earlier, to denote association with specific providers. Specifically, a_i^h represents the investment by hospital i to reduce acute care costs. The investments by hospital i to reduce PAC costs and readmissions for provider j are denoted by b_{ij}^h and e_{ij}^h, respectively, while b_{ij}^s and e_{ij}^s denote the corresponding investments by PAC provider j. For notational simplicity, we denote the decisions of hospital i by $\mathbf{h}_i = (a_i^h, b_{ij}^h, e_{ij}^h, j \in \mathcal{M})$ and the decisions of PAC provider j by $\mathbf{s}_j = (b_{ij}^s, e_{ij}^s, i \in \mathcal{N})$.

The objective of hospital i, denoted by Π_i^h, can be expressed as follows:

$$\Pi_i^h(\mathbf{h}_i) = T_i^h - \mathcal{C}_i^h(\mathbf{h}_i), \tag{1}$$

where T_i^h represents the reimbursement amount received by the hospital (determined by the regulator) and $\mathcal{C}_i^h(\mathbf{h}_i)$ is the total cost of hospital i given by

$$\mathcal{C}_i^h(\mathbf{h}_i) = p_i \left[C^h(a_i^h) + a_i^h\right] + \sum_{j \in \mathcal{M}} p_{ij} \left[R(e_{ij}^h, e_{ij}^s)\xi^h + b_{ij}^h + e_{ij}^h\right]. \tag{2}$$

We assume that each hospital invests the same level in reducing the costs of acute care for all patients, regardless of the PAC destination (i.e., a_i^h is constant for all $i \in \mathcal{N}$ and does not depend on j), as hospitals are unlikely to vary their treatment decisions based on the anticipated PAC provider. However, the investment in reducing the expected costs of PAC care and readmissions (i.e., b_{ij}^h and e_{ij}^h) varies depending on the specific PAC provider. This difference in investment levels is essential because managing costs effectively in these areas involves collaborative efforts between entities, such as utilizing shared healthcare records or implementing coordinated care protocols in the PAC setting.

The objective of PAC provider j, denoted by Π_j^s, can be expressed as follows:

$$\Pi_j^s(\mathbf{s}_j) = T_j^s - \mathcal{C}_j^s(\mathbf{s}_j), \tag{3}$$

where T_j^s denotes the payment received by the PAC provider from the regulator, and $\mathcal{C}_j^s(\mathbf{s}_j)$ is the total cost of the PAC provider j given by

$$\mathcal{C}_j^s(\mathbf{s}_j) = \sum_{i \in \mathcal{N}} p_{ij} \left[C^s(b_{ij}^h, b_{ij}^s) + R(e_{ij}^h, e_{ij}^s)\xi^s + b_{ij}^s + e_{ij}^s \right] \tag{4}$$

In this section, and throughout the rest of the paper, costs C^h and C^s should be understood as expected values, since the cost of treating each patient can be random. Payment amounts to each provider can be based on the realization of these costs. In general, we assume that providers are maximizing the expected value of their objective.

Regulator: The regulator aims to maximize the total welfare W, which is equal to the patient surplus minus the total cost, i.e.,

$$W = \upsilon - \sum_{i \in \mathcal{N}} \mathcal{C}_i^h(\mathbf{h}_i) - \sum_{j \in \mathcal{M}} \mathcal{C}_j^s(\mathbf{s}_j). \tag{5}$$

Here, υ represents the patient surplus from receiving treatment, and the remaining terms constitute the expected total cost of providing care as described earlier; see (2) and (4). For simplicity, we assume that the patient receives a fixed benefit υ from treatment, independent of readmissions. However, this assumption can be relaxed to incorporate patient disutility from readmissions, as discussed in a similar manner in Section 5.3 of Arifoğlu et al. (2021). As υ is fixed, the regulator's problem of maximizing total welfare is equivalent to minimizing the total expected cost of a care episode. Additionally, we have proven that our main result remains valid when the regulator assigns different relative weights to patients' and providers' utilities.

Next, we outline our assumptions regarding socially optimal actions. Specifically, we assume that the regulator's objective function has a unique optimizer, referred to as "first-best" from hereon. Additionally, we assume that the cost functions C^h and C^s, along with the readmission probability function R, are twice differentiable, and that the first-best actions are characterized by the following first-order conditions (FOCs) obtained from the regulator's objective function (5):

$$\frac{dC^h(a_h^*)}{da_h} + 1 = 0, \tag{6}$$

$$\frac{\partial C^s(b_h^*, b_s^*)}{\partial b^h} + 1 = 0, \tag{7}$$

$$\frac{\partial C^s(b_h^*, b_s^*)}{\partial b^s} + 1 = 0, \tag{8}$$

$$\frac{\partial R(e_h^*, e_s^*)}{\partial e^h}(\xi^h + \xi^s) + 1 = 0, \tag{9}$$

$$\frac{\partial R(e_h^*, e_s^*)}{\partial e^s}(\xi^h + \xi^s) + 1 = 0, \tag{10}$$

In Appendix A, we provide sufficient conditions on the cost of care and readmission probability functions for these assumptions to hold. Under these assumptions, the socially optimal actions for

all hospitals, denoted by (a_h^*, b_h^*, e_h^*), are identical to each other, and similarly, the first-best actions for all PAC providers, denoted by (b_s^*, e_s^*), are also identical.

On a technical note, when $p_{ij} = 0$, i.e., no patients are discharged from hospital i to PAC provider j, the associated cost of care is zero, regardless of the values of $b_{ij}^h, b_{ij}^s, e_{ij}^h, e_{ij}^s$. In that case, we assume that the first-best actions are $a_h^*, b_h^*, b_s^*, e_h^*, e_s^*$, and hospital i and PAC provider j choose first-best actions in equilibrium under any reimbursement scheme. This treatment eases exposition and is without loss of generality because any provider actions lead to zero cost of care from PAC provider j because $p_{ij} = 0$, see (4).

Information asymmetry: In an ideal scenario, if the regulator has perfect knowledge of the expected cost of care and readmission probability functions (C^a, C^s, and R), it can determine the first-best actions and enforce providers to implement them through strict penalties for any deviation. However, obtaining such precise information on the cost and readmission probability functions is prohibitively difficult in practice due to the intricate and evolving nature of healthcare technologies and processes.

This information asymmetry has compelled healthcare regulators, such as CMS in the United States and National Health Service (NHS) in the United Kingdom, to adopt yardstick competition-type payment schemes such as the CJR, HRRP, and VBP programs. These schemes reimburse providers based on their relative performance compared to other providers, as measured by documented costs and other observable characteristics of hospitals and patients (CMS 2023, Cots et al. 2011).

Therefore, unlike the extant literature on payment models for care coordination—which typically assumes that the regulator has full information—we consider a setting in which the regulator does not observe C^a, C^s, or R. Instead, we focus on relative performance-based payment models, which are commonly used to address such information asymmetries. We assume the regulator has access to ex-post outcomes, including realized costs, investments, and readmission probabilities. This allows for the design of relative performance metrics, as implemented in the CJR program (see Section 6.2) and in single-entity settings (Shleifer 1985, Arifoğlu et al. 2021, Savva et al. 2023). Notably, this assumption requires that the regulator observes providers' ex-post investments in cost and readmission reduction, but not the detailed effort components underlying those investments. For simplicity, we abstract away from modeling these underlying efforts, and show in Appendix J that our results remain valid even when providers make multidimensional, unobservable efforts.

4. Payment models and care coordination

Healthcare payment models have been widely implemented to contain costs and reduce readmissions, but most are designed for single-entity settings with centralized decision-making. To examine

the challenges in the CJR program, we first analyze existing bundled payment models in multi-entity settings using the framework introduced earlier. In Section 4.1, we show that these models fail to achieve first-best outcomes. In Section 4.2, we introduce a CJR-type payment model that captures key incentive mechanisms in the CJR program and identify its limitations in incentivizing care coordination. In Section 4.3, we discuss challenges in designing coordinating payment models, which establishes the foundation for our subsequent analysis in Section 5.

4.1. Bundled payment model in multi-entity settings

Before implementing the CJR program, CMS used separate DRG-based PPSs to reimburse hospitals and PAC providers for their respective contributions to joint replacement treatment. (Similar payment systems are commonly used by both public and private insurance companies worldwide.)

Under PPS, providers receive a fixed reimbursement that covers treatment costs, including those for readmitted patients. These reimbursement rates are determined based on the average cost across all similar providers, creating an incentive for cost reduction (CMS 2024).

In our modeling framework, PPS payment model can be defined as follows. Let

$$\bar{C}_i^h = \frac{\sum_{k \in \mathcal{N}_i} p_k C^h(a_k^h)}{1 - p_i}$$

denote the average acute care cost,

$$\bar{a}_i^h = \frac{\sum_{k \in \mathcal{N}_i} p_k a_k^h}{1 - p_i}$$

denote the average investments to reduce acute care costs,

$$\bar{b}_i^h = \frac{\sum_{k \in \mathcal{N}_i} \sum_{j \in \mathcal{M}} p_{kj} b_{kj}^h}{1 - p_i},$$

denote the average investments to reduce post-acute care costs and

$$\bar{e}_i^h = \frac{\sum_{k \in \mathcal{N}_i} \sum_{j \in \mathcal{M}} p_{kj} e_{kj}^h}{1 - p_i},$$

denote the average investments to reduce readmission probability, among all other hospitals excluding hospital i. The payment amount to hospital i is then given by

$$T_i^h = p_i[\bar{C}_i^h + \bar{a}_i^h + \bar{b}_i^h + \bar{e}_i^h] + \sum_{j \in \mathcal{M}} p_{ij} R(e_{ij}^h, e_{ij}^s)\xi^h. \tag{11}$$

Thus, hospital i receives payment $(\bar{C}_i^h + \bar{a}_i^h + \bar{b}_i^h + \bar{e}_i^h)$ for an initial admission equal to the average cost and investments, and receives payment ξ^h for a readmission. The payment $[\bar{C}_i^h + \bar{a}_i^h + \bar{b}_i^h + \bar{e}_i^h]$

can be thought of as the estimated cost of care and investment to provide effective care. The last component (i.e, $\sum_{j\in\mathcal{M}} p_{ij}R(e^h_{ij}, e^s_{ij})\xi^h$) covers the cost of treating readmitted patients.

Similarly, the payment amount to PAC provider j is

$$T^s_j = \tilde{p}_j[\bar{C}^s_j + \bar{b}^s_j + \bar{e}^s_j] + \sum_{i\in\mathcal{N}} p_{ij}R(e^h_{ij}, e^s_{ij})\xi^s, \tag{12}$$

where

$$\bar{C}^s_j = \frac{\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} p_{ik}C^s(b^h_{ik}, b^s_{ik})}{1-\tilde{p}_j}, \bar{b}^s_j = \frac{\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} p_{ik}b^s_{ik}}{1-\tilde{p}_j}, \text{ and } \bar{e}^s_j = \frac{\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} p_{ik}e^s_{ik}}{1-\tilde{p}_j},$$

respectively, represent the average PAC cost, investments to reduce PAC cost and readmission probability, among all other PAC providers.

Even in a single-entity setting, PPS fails to provide the right incentives to reduce readmissions, as providers continue to receive payments for each readmitted patient (Arifoğlu et al. 2021). Naturally, the same holds in equilibrium for the multi-entity setting. Indeed, we show (under the assumptions in Appendix A) that PPS leads to a unique equilibrium in which hospitals and PAC providers do not invest in reducing readmissions, i.e., $e^h_{ij} = e^s_{ij} = 0$, for each hospital $i \in \mathcal{N}$ and PAC provider $j \in \mathcal{M}$. Moreover, in the multi-entity setting, PPS fails to incentivize sufficient investment in care coordination: hospitals do not invest in reducing the PAC cost, and PAC providers invest in reducing the PAC cost at suboptimal levels, consistent with the findings in Arifoğlu et al. (2021) for the single-entity setting.

These limitations highlight the need for alternative payment models that actively encourage providers to reduce readmissions. One such approach is the bundled payment model, which, in a single entity setting, refers to a payment arrangement in which a provider—such as a hospital or outpatient surgery center—receives a fixed payment covering all services delivered within that setting during a defined episode of care. In this structure, each provider is reimbursed for only the portion of the episode they deliver. For example, the hospital receives a bundled payment for the inpatient stay, while post-acute care services are paid separately if not included in the bundle. However, in some designs, bundled payments can also cover both the initial admission and any related readmissions that occur within the episode window.

This structure intuitively aligns incentives by rewarding providers for keeping patients healthy: they receive a fixed payment regardless of whether a readmission occurs, so avoiding complications and readmissions improves their margins. Recognizing these potential benefits, CMS implemented bundled payment models through programs such as HRRP and SNF VBP Program.

Next we introduce the bundled payment model in our multi-entity setting, following (Arifoğlu et al. 2021). Define

$$\bar{R}_i^h = \frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}} p_{kj}R(e_{kj}^h, e_{kj}^s)}{1-p_i} \text{ and } \bar{R}_j^s = \frac{\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} p_{ik}R(e_{ik}^h, e_{ik}^s)}{1-\tilde{p}_j} \tag{13}$$

as the average readmission benchmark for hospital i and PAC provider j, respectively. The payment amounts for hospital i and PAC provider j can be modeled, respectively, as follows:

$$T_i^h = p_i\Big[\bar{C}_i^h + \bar{a}_i^h + \bar{b}_i^h + \bar{e}_i^h + \bar{R}_i^h\xi^h\Big], \tag{14}$$

$$T_j^s = \tilde{p}_j\Big[\bar{C}_j^s + \bar{b}_j^s + \bar{e}_j^s + \bar{R}_j^s\xi^s\Big]. \tag{15}$$

In words, each provider receives a payment for the portion of the episode of care delivered within their setting, equal to the average cost and investments required to cover both the initial admission and any potential readmission in that same setting.

We highlight the fact that the main difference between the bundled payments and PPS is that a fixed payment for each patient (i.e., $p_i\bar{R}_i^h\xi^h$ in (14)) replaces the payment for caring for readmitted patients (i.e, $\sum_{j\in\mathcal{M}} p_{ij}R(e_{ij}^h, e_{ij}^s)\xi^h$ in (11).) Because this payment under bundled payment is fixed regardless of whether the patient is readmitted or not, it incentivizes hospitals to reduce readmissions. Similarly, bundled payments pay a fixed amount (i.e., $\tilde{p}_j\bar{R}_j^s\xi^s$ in (15)) to PAC providers.

The bundled payment model elicits first-best cost and quality outcomes in single-entity settings (Arifoğlu et al. 2021). Notably, the payments (14) and (15) are fixed and do not influence providers' incentives. These payments are designed to ensure that providers break even in equilibrium while minimizing overall payment levels. However, in multi-entity settings, the bundled payment model alone fails to elicit first-best actions because it does not incentivize enhanced care coordination between entities. We next formally establish this limitation.

Proposition 1. *If the regulator uses the bundled payment model defined in* (14) *to reimburse hospitals and* (15) *to reimburse PAC providers, then the unique Nash equilibrium is for each hospital* $i\in\mathcal{N}$ *to pick* $a_i^h = a_h^*$*, and for each hospital* $i\in\mathcal{N}$ *and PAC provider* $j\in\mathcal{M}$ *such that* $p_{ij}>0$*, to pick* $b_{ij}^h = 0, e_{ij}^h = \breve{e}_h$ *and* $b_{ij}^s = g(0), e_{ij}^s = \breve{e}_s$*, respectively, for some* $\breve{e}_h$ *and* $\breve{e}_s$ *defined in the proof in Appendix B and* g *is defined in* (A-5)*. In addition, costs of care and investments associated with readmissions are higher than the first-best level:*

$$R(\breve{e}_h, \breve{e}_s)(\xi^h+\xi^s) + \breve{e}_h + \breve{e}_s > R(e_h^*, e_s^*)(\xi^h+\xi^s) + e_h^* + e_s^*. \tag{16}$$

The main implication of Proposition 1 is that, similar to PPS, hospitals do not invest in reducing PAC costs, i.e., $b_{ij}^h = 0$. This occurs because the bundled payment model does not provide incentives

for hospitals to lower the costs incurred by PAC providers they collaborate with. Moreover, hospitals and PAC providers invest inadequately in reducing readmission probabilities, resulting in higher total costs of care and investments associated with readmissions than the first-best level, as shown in (16). This inefficiency arises because the bundled payment model holds hospitals and PAC providers financially responsible only for readmission costs within their own settings rather than for the total cost of the entire episode of care, even though their actions impact overall costs.

Thus, the bundled payment model for single-entity settings does not incentivize hospitals to coordinate care in reducing PAC costs or readmission probabilities. This lack of coordination results in suboptimal investments in cost and readmission reductions, ultimately driving up the overall cost of care. Recognizing these shortcomings, the CJR program builds on the bundled payment model by holding participating hospitals financially accountable for the cost of both acute and post-acute care, including readmissions. We evaluate the effectiveness of the CJR program below.

4.2. CJR-type payment model

The CJR program extends the bundled payment model by introducing incentives to improve coordination between hospitals and PAC providers. It does so by making hospitals accountable for the entire episode of care, linking their reimbursement to the total cost of treatment. Hospitals receive financial rewards if total spending for a patient remains below a predetermined threshold and face penalties if it exceeds that amount. Additionally, CMS permits hospitals to share up to 50% of gains or losses from payment reconciliation with PAC providers through gainsharing agreements. However, this approach has not been effective in practice. Recent research, based on extensive interviews, found that "none of the hospitals interviewed throughout the evaluation said that they had established such agreements" (Hopewell et al. 2024) (see also Ghamat et al. (2021)).

To analyze the incentive mechanisms embedded in the CJR program, we consider a simplified CJR-type payment model that captures its key features. In this model, the regulator reimburses providers under the bundled payment model, as defined in (14)-(15), while adjusting hospital payments based on the total cost of collaborative care—including PAC costs for initial admissions and both acute and post-acute care costs for readmissions. However, PAC providers continue to be reimbursed under bundled payment. We assume that gainsharing agreements between hospitals and PAC providers are absent, both to isolate and assess their impact and because they have not been widely adopted in practice.

We consider a CJR-type payment model where the payment amount for hospital i is given by:

$$\begin{aligned} T_i^h = \; & p_i\Big[\bar{C}_i^h + \bar{R}_i^h\xi^h + \bar{a}_i^h + \bar{b}_i^h + \bar{e}_i^h\Big] \\ & + \sum_{j\in\mathcal{M}} p_{ij}\Big[\bar{C}_i^{sh} - C_s(b_{ij}^h, b_{ij}^s) + \big(\bar{R}_i^h - R(e_{ij}^h, e_{ij}^s)\big)\,(\xi^h + \xi^s)\Big], \end{aligned} \tag{17}$$

where

$$\bar{C}_i^{sh} = \frac{\sum_{k \in \mathcal{N}_i} \sum_{j \in \mathcal{M}} p_{kj} C^s(b_{kj}^h, b_{kj}^s)}{1 - p_i}, \tag{18}$$

is the average cost of PAC for all patients discharged from all hospitals, excluding hospital i. The main difference between this CJR-type payment model and the bundled payment model (see (14)) is the term in the second line in (17). This term captures the fact that hospitals are held responsible for the whole episode of care in the CJR program. The payment amount for PAC provider j is the same as in bundled payment (15), since they continue to be paid under bundled payment model.

Remark 1. Our CJR-type payment model is designed to capture the core financial incentives of the actual CJR program while simplifying certain complexities for analytical clarity. The first key difference lies in the scope of performance evaluation. The actual CJR program assesses hospitals based on various episode quality and cost metrics, incorporating factors such as complications, patient outcomes, and readmissions. In contrast, our simplified model focuses solely on readmissions to specifically examine the joint actions of hospitals and PAC providers in reducing them. By doing so, we isolate coordination incentives without the added complexity of broader quality measures.

Also, we assume that both hospitals and PAC providers receive a fixed payment per patient, including readmissions. In contrast, CMS employs the HRRP for hospitals and various VBP plans for PAC providers. These programs introduce additional complexities beyond the bundled payment models considered in our analysis. Nevertheless, our model captures similar incentive structures, as these programs also encourage readmission reduction but only based on each provider's own costs, without accounting for the financial impact on the other entity involved in patient care. (For further details, see Arifoğlu et al. (2021) and CMS (2023).) Despite these differences, our model preserves the fundamental incentive mechanisms of the CJR program, ensuring that our findings provide meaningful insights into how bundled payments influence hospital and PAC provider behavior.

We next establish the equilibrium outcome under the CJR-type payment model.

Proposition 2. *If the regulator uses* (17) *to reimburse hospitals and* (15) *to reimburse PAC providers, then the unique Nash equilibrium is for each hospital* $i \in \mathcal{N}$ *to pick* $a_i^h = a_h^*$*, and for each hospital* $i \in \mathcal{N}$ *and PAC provider* $j \in \mathcal{M}$ *such that* $p_{ij} > 0$*, to pick* $b_{ij}^h = b_h^*, e_{ij}^h = \tilde{e}_h$ *and* $b_{ij}^s = b_s^*, e_{ij}^s = \tilde{e}_s$*, respectively, for some* $\tilde{e}_h$ *and* $\tilde{e}_s$ *defined in the proof in Appendix C. In addition, costs of care and investments associated with readmissions are higher than the first-best level:*

$$R(\tilde{e}_h, \tilde{e}_s)(\xi^h + \xi^s) + \tilde{e}_h + \tilde{e}_s > R(e_h^*, e_s^*)(\xi^h + \xi^s) + e_h^* + e_s^*. \tag{19}$$

Proposition 2 shows that while the CJR-type payment model achieves some of its intended goals, it falls short in others. On the one hand, hospitals and PAC providers invest at first-best levels to

reduce PAC costs; specifically $b^h_{ij} = b^*_h$ and $b^s_{ij} = b^*_s$, respectively. This result occurs because each hospital is fully responsible for PAC cost performance due to the adjustment term $\sum_{j\in\mathcal{M}} p_{ij}\left[\bar{C}^{sh}_i - C_s(b^h_{ij}, b^s_{ij})\right]$ in (17).

However, hospitals and PAC providers do not invest at first-best levels in reducing readmissions, resulting in higher total costs of care and investments, as shown in (19). This misalignment occurs due to two key factors: (i) Hospitals are disproportionately accountable for readmission costs–they receive payment adjustments based on both acute and post-acute care costs of readmissions yet already bear the acute care costs under fixed per-episode payments; (ii) PAC providers are not accountable for the acute care cost of readmissions, limiting their incentives to invest in reducing readmissions.

4.3. Discussion

In summary, while the CJR-type payment model successfully incentivizes hospitals to reduce PAC costs, it fails to induce first-best investment from both hospitals and PAC providers in reducing readmissions. Although gainsharing agreements have the potential to improve coordination in theory, their limited adoption in practice raises concerns about the program's overall effectiveness. Addressing these challenges is the focus of this paper.

However, several key issues must be considered. First, the CJR program extends the bundled payment model from a single-entity to a multi-entity setting by introducing additional payment adjustments for hospitals. Whether these adjustments successfully create incentives for care coordination remains a priori unclear. Thus, our first goal is to identify payment models that can achieve first-best outcomes in a multi-entity setting—a challenge that, to the best of our knowledge, has not been formally addressed in the literature.

Second, the CJR program's reliance on gainsharing agreements may reflect additional objectives or constraints that are not explicitly captured in our model. A deeper understanding of the structure of coordinating incentive mechanisms can provide valuable insights into how gainsharing agreements should be designed from a regulatory perspective. However, structuring these agreements is inherently complex due to several factors: (i) hospitals and PAC providers operate independently and may have conflicting objectives (e.g., extended patient follow-ups may lower PAC costs but increase hospital costs); (ii) hospitals exercise discretion over PAC referrals, and this discretion may impact the effectiveness of gainsharing agreements; (iii) given the diversity of PAC providers (e.g., SNFs, home health agencies), patients may choose different post-acute care options, making it unclear how gainsharing agreements should account for these complex referral networks.

Given these complexities, the next section introduces a payment model structurally similar to the CJR-type model and demonstrates its ability to induce first-best provider actions under various

modeling considerations. Within this framework, we also examine the role of gainsharing agreements in shaping coordination incentives and establish how they can be structured to achieve first-best outcomes in multi-entity care settings.

5. Coordinating reimbursement schemes

To address the limitations of the CJR-type payment model, we propose an alternative payment structure that explicitly incentivizes care coordination between hospitals and PAC providers in Section 5.1. In Section 5.2 we show how gainsharing agreements can be structured to achieve first-best actions, building on our results in Section 5.1. We present the results of a numerical analysis to quantify the cost-saving potential of the proposed payment models in Section 5.3.

5.1. A new payment model

Our proposed payment model consists of two key components: (i) a payment to cover the costs of care and investments to improve care; and (ii) an outcome-based adjustment to promote coordination, both of which are calculated based on the performances of other providers. The outcome-based adjustment ensures that hospitals and PAC providers share financial responsibility for care coordination by linking reimbursements to each other's performance.

The payment amount for hospital i is given by:

$$T_i^h = \underbrace{p_i\left[\bar{C}_i^h + \bar{a}_i^h + \bar{b}_i^h + \bar{e}_i^h + \bar{R}_i^h\xi^h\right]}_{\text{Cost of care}} + \underbrace{\sum_{j\in\mathcal{M}} p_{ij}\left[\bar{C}_i^{sh} - C^s(b_{ij}^h, b_{ij}^s) + \left(\bar{R}_i^h - R(e_{ij}^h, e_{ij}^s)\right)\xi^s\right]}_{\text{Outcome-based adjustment for care coordination}}. \quad (20)$$

The payment for PAC provider j is given by:

$$T_j^s = \underbrace{\tilde{p}_j\left[\bar{C}_j^s + \bar{b}_j^s + \bar{e}_j^s + \bar{R}_j^s\xi^s\right]}_{\text{Cost of care}} + \underbrace{\sum_{i\in\mathcal{N}} p_{ij}\left[\bar{R}_j^s - R(e_{ij}^h, e_{ij}^s)\right]\xi^h}_{\text{Outcome-based adjustment for care coordination}}. \quad (21)$$

The "Cost of care" component follows the bundled payment structure in (14)-(15), reimbursing providers for the costs of care delivery and investments in cost and readmission reductions for both initial admissions and potential readmissions. The "Outcome-based adjustment for care coordination" component differentiates this payment model from the bundled payment model by explicitly aligning incentives between hospitals and PAC providers. In (20), hospitals are rewarded or penalized based on the average PAC cost for their discharged patients and the average PAC cost for their readmitted patients, relative to other hospitals. Similarly, in (21), PAC providers are rewarded or penalized based on the average acute-care cost for readmitted patients, relative to other PAC providers. The introduction of these payment adjustments ensures that both hospitals and PAC providers internalize the financial consequences of their coordination investments, thereby incentivizing collaborative strategies to enhance patient care and reduce overall costs as we show next (the proof is presented in Appendix D).

Theorem 1. *If the regulator uses* (20) *to reimburse hospitals and* (21) *to reimburse PAC providers, then the unique Nash equilibrium is for each each hospital* $i \in \mathcal{N}$ *and PAC provider* $j \in \mathcal{M}$ *to pick first-best actions* $a_i^h = a_h^*, b_{ij}^h = b_h^*, e_{ij}^h = e_h^*$, *and* $b_{ij}^s = b_s^*, e_{ij}^s = e_s^*$, *respectively. In addition, all providers break even in this equilibrium.*

This result demonstrates that the proposed payment model successfully elicits first-best investments from both hospitals and PAC providers in reducing care costs and readmission probabilities. As discussed following Proposition 2, the CJR-type payment model fails to induce first-best investments in reducing readmissions, as hospitals are held disproportionately accountable, while PAC providers bear no responsibility for the acute care costs of treating readmitted patients. The proposed model eliminates this misalignment by excluding hospitals' own readmission costs from payment adjustments and linking PAC providers' payments to the readmission costs incurred by their partner hospitals.

In addition, Theorem 1 offers key insights into the design of relative performance-based payment models for multi-entity healthcare systems. While bundled payments alone may incentivize efficient care in single-entity settings (see Section 4.1 for a detailed discussion), additional adjustments are necessary in multi-entity settings, where patients receive acute and post-acute care from different providers. Our proposed model bridges this gap through an outcome-based adjustment that ties each provider's payment to the performance of their collaborating providers, thereby fostering coordinated and efficient care delivery.

In terms of implementation, our results above show that care coordination in multi-entity settings can be achieved without gainsharing agreements—our proposed payment model elicits the first-best outcome by separately adjusting payments to hospitals and PAC providers. Next, we derive further practical insights by showing that our proposed payment model (i) can be implemented within the existing CJR program in Section 5.2, and (ii) generate significant cost savings for CMS in Section 5.3 through numerical experiments.

5.2. Implementation using the CJR program

Proposition 2 established that the CJR-type payment model does not lead to first-best actions in the absence of gaingainsharing agreements. This section examines how gaingainsharing agreements can be structured to achieve first-best actions, building on the results in Theorem 1.

The primary difference between our payment model and the CJR-type payment model without gainsharing agreements lies in the allocation of readmission costs. In the CJR-type payment model, PAC providers are not financially responsible for hospital readmissions (see (15)). In contrast, our payment model links their payments to the acute care costs associated with readmitted patients

through the "Outcome-based adjustment for care coordination" in (21). To address this discrepancy, gainsharing agreements should be designed to make PAC providers accountable for these costs.

Currently, CMS permits gainsharing agreements derived from hospitals' reconciliation payments based on costs of the entire episode of care, without specifying which cost components should be included. However, as discussed above, to achieve the first-best outcome, gainsharing agreements should focus on the acute-care cost of readmitted patients, and exclude the PAC cost for initial or readmitted patients. This structure provides incentives for PAC providers to invest in reducing readmissions while maintaining hospitals' incentives to reduce the PAC cost.

Building on this insight, we now consider a modified CJR-type payment model in which hospitals share a portion of the payment adjustments for the acute-care cost of readmitted patients with PAC providers. We consider the full range of the sharing portion as denoted by $\theta \in [0,1]$. In practice, offering flexibility in the sharing portion allows healthcare providers to design gainsharing agreements that reflect their specific patient populations and care settings. Moreover, given the limited adoption of gainsharing agreements, our analysis across the full range of θ will demonstrate the potential of (partial and full) gainsharing agreements in promote collaborative care.

To formalize this framework, assume that hospitals bear a fraction $(1-\theta)$ of the payment adjustment associated with acute-care readmission costs, while PAC providers share the remaining θ. Under this modified CJR-type payment model, the hospital payment for hospital i is given by:

$$\begin{aligned} T_i^h = & p_i \Big[\bar{C}_i^h + \bar{a}_i^h + \bar{b}_i^h + \bar{e}_i^h + \bar{R}_i^h \xi^h \Big] \\ & + \sum_{j \in \mathcal{M}} p_{ij} \Big[\bar{C}_i^{sh} - C_s(b_{ij}^h, b_{ij}^s) + \big(\bar{R}_i^h - R(e_{ij}^h, e_{ij}^s) \big) \left((1-\theta)\xi^h + \xi^s \right) \Big]. \end{aligned} \tag{22}$$

The corresponding payment for PAC provider j is given by:

$$T_j^s = \tilde{p}_j \Big[\bar{C}_j^s + \bar{b}_j^s + \bar{e}_j^s + \bar{R}_j^s \xi^s \Big] + \sum_{i \in \mathcal{N}} p_{ij} \Big[\bar{R}_j^s - R(e_{ij}^h, e_{ij}^s) \Big] \theta \xi^h. \tag{23}$$

With flexible gainsharing agreements, the modified CJR-type payment model encompasses several payment models as special cases. When $\theta = 0$, it reduces to the original CJR-type payment model (15) and (17), in which hospitals do not share any payment adjustments with PAC providers. When $\theta = 1$, PAC providers become fully accountable for hospital readmission costs, making this model identical to our proposed payment model in (20)-(21). This result highlights that full alignment of provider incentives can be achieved through appropriately structured gainsharing agreements.

We now analyze the equilibrium outcomes under this modified CJR-type payment model for all values of $\theta \in [0,1]$, allowing us to assess the impact of different sharing levels on provider incentives. Section 5.3 provides numerical results to illustrate the effects of θ on cost and readmission outcomes.

Proposition 3. *If the regulator uses* (22) *to reimburse hospitals and* (23) *to reimburse PAC providers, then the unique Nash equilibrium is for each hospital* $i \in \mathcal{N}$ *to pick* $a_i^h = a_h^*$*, and for each hospital* $i \in \mathcal{N}$ *and PAC provider* $j \in \mathcal{M}$ *such that* $p_{ij} > 0$*, to pick* $b_{ij}^h = b_h^*, e_{ij}^h = \tilde{e}_h(\theta)$ *and* $b_{ij}^s = b_s^*, e_{ij}^s = \tilde{e}_s(\theta)$*, respectively, for some* $\tilde{e}_h$ *and* $\tilde{e}_s$ *defined in the proof in Appendix C. We have* $\tilde{e}_h(1) = e_h^*$*,* $\tilde{e}_s(1) = e_s^*$*. If* $\partial^2 R(e^h, e^s)/\partial e^h \partial e^s \geqslant 0$*,* $\tilde{e}_h(\theta)$ *decreases in* θ *and* $\tilde{e}_s(\theta)$ *increases in* θ*.*

Proposition 3 characterizes providers' readmission-reduction investments across the full range of sharing portion θ. As explained above, when hospitals fully share payment adjustments with PAC providers (i.e., $\theta = 1$), the modified CJR-type payment is the same as our proposed payment model, thus providers invest at the first-best levels. To further study partial sharing, we consider $\partial^2 R(e^h, e^s)/\partial e^h \partial e^s \geqslant 0$, i.e., hospital investment is less effective at reducing readmissions at higher PAC provider investments. Under this condition, hospitals invest more while PAC providers invest less than the first-best levels, i.e., $\tilde{e}_h(\theta) \geqslant e_h^*$ and $\tilde{e}_s(\theta) \leqslant e_s^*$. This is expected as hospitals partially absorb the payment adjustment associated with acute-care readmission costs that should be borne by PAC providers. Furthermore, with greater sharing portion θ, providers invest closer to the first-best levels—hospitals invest less and PAC providers invest more in reducing readmissions. As we demonstrate in the next section, this leads to significant cost savings and reductions in readmissions.

5.3. Numerical Experiments

To complement our theoretical findings, we conduct a numerical analysis to quantify the cost-saving potential of our proposed payment model. Using real-world data, we calibrate our model and evaluate its impact on PAC costs and readmission rates, comparing it to the alternative payment models discussed earlier. This analysis provides empirical support for the theoretical results, demonstrating how the proposed approach enhances financial and operational efficiency in multi-entity healthcare settings.

In this section, we first outline the model calibration process, detailing the parameter selection and data sources used in the analysis. We then present the results from a series of numerical experiments, highlighting the comparative performance of different payment models in terms of cost reduction, incentive alignment, and readmission outcomes.

Model calibration: We estimate model parameters using Medicare claims data three years prior to CJR (CMS 2021b), a period during which providers were reimbursed under PPS. The estimation of model parameters relies on the baseline (i.e., three years prior to CJR) risk-adjusted average costs among hospitals whose participation in the CJR model was mandatory from Exhibit D-1 of CMS (2021b). We assume that all providers adopt equilibrium actions consistent with PPS.

First, we use the "Anchor payments" in this exhibit to estimate the expected total cost for acute care of a new admission, i.e., $C^h(a_h^*) + a_h^* = \$12,190$, the "Readmission payments" to estimate the expected acute care cost for the potential readmission, i.e., $R(0,0)\xi^h = \$1,225$, and the "SNF payments" to estimate the PAC cost in an entire episode, i.e., $C^s(0,g(0)) + R(0,0)\xi^s = \$6,142$. Second, we estimate the readmission probability in PPS using the 90-day readmission rate for SNF discharges prior to CJR implementation from Welsh et al. (2017), i.e., $R(0,0) = 0.121$. Third, since the PAC cost per episode of care is much higher for patients who were readmitted versus those who were not (Phillips et al. 2019), we assume that the PAC cost of a readmitted patient is 60% of the patient's total PAC cost, i.e., $\xi^s/(C^s(0,g(0)) + \xi^s) = 40\%$ (our results are robust to other reasonable PAC cost allocations). Fourth, we use the following functions:

$$C^h(a^h) = \tau^h \exp\left(1 - \frac{a^h}{\tau^h}\right), \tag{24}$$

$$C^s(b^h, b^s) = \tau^s \left(\exp\left(1 - \frac{b^h}{\tau^s}\right) + \exp\left(1 - \frac{b^s}{\tau^s}\right)\right), \tag{25}$$

$$R(e^h, e^s) = \frac{\tau}{\xi^h + \xi^s}\left(\exp\left(1 - \frac{e^h}{\tau}\right) + \exp\left(1 - \frac{e^s}{\tau}\right)\right), \tag{26}$$

for constants $\tau^h, \tau^s, \tau > 0$, where $\exp(\cdot)$ is the exponential function. These functions satisfy all assumptions and yield reasonable equilibrium outcomes under different payment models calibrated to data (see below).[5] We use FOCs to determine, in simple closed forms, the first-best actions:

$$a_h^* = \tau^h, b_h^* = b_s^* = \tau^s, e_h^* = e_s^* = \tau, \tag{27}$$

and the following equilibrium actions under different payment models:

$$g(0) = \tau^s, \breve{e}_h = \tau\left(1 + \ln\left(\frac{\xi^h}{\xi^h + \xi^s}\right)\right), \breve{e}_s = \tau\left(1 + \ln\left(\frac{\xi^s}{\xi^h + \xi^s}\right)\right), \tag{28}$$

$$\tilde{e}_h = \tau\left(1 + \ln\left(\frac{(2-\theta)\xi^h + \xi^s}{\xi^h + \xi^s}\right)\right), \tilde{e}_s = \tau\left(1 + \ln\left(\frac{\theta\xi^h + \xi^s}{\xi^h + \xi^s}\right)\right). \tag{29}$$

Fifth, we plug equilibrium actions into (24)-(26) and use estimates in the first three steps to obtain the following calibrated parameter values:

$$\tau^h = 6,095,\ \tau^s = 1,102,\ \tau = 399,\ \xi^h = 10,124,\ \xi^s = 7,798. \tag{30}$$

Results: Using the calibrated parameter values and the equilibrium actions for different payment models summarized in Table 1, we evaluate the differences in three aspects; (i) the expected total PAC costs and investments for an initial admission, given by $C^s(b^h, b^s) + b^h + b^s$, (ii) the readmission probability, given by $R(e^h, e^s)$, and (iii) the expected total costs of care and investments for a readmission, given by $R(e^h, e^s)(\xi^h + \xi^s) + e^h + e^s$ and referred to as "readmission costs" below.

[5] Since there is no consensus in the literature regarding the functional forms for the cost and quality of collaborative acute and post-acute care, we have explored several options. We chose the exponential form because it yields more realistic equilibrium outcomes compared to other functions used in the literature, such as the inverse functions utilized in Arifoğlu et al. (2021) and the Cobb-Douglas function employed in Andritsos and Tang (2018).

Table 1 Equilibrium outcomes under different payment models

Payment model	a^h	b^h	b^s	e^h	e^s
Prospective payment system	a_h^*	0	$g(0)$	0	0
Bundled payment model	a_h^*	0	$g(0)$	$\breve{e}_h$	$\breve{e}_s$
CJR-type payment model	a_h^*	b_h^*	b_s^*	$\tilde{e}_h(> e_h^*)$	$\tilde{e}_s(< e_s^*)$
Our proposed payment model	a_h^*	b_h^*	b_s^*	e_h^*	e_s^*

Note: For each parameter a superscript $*$ denotes first-best outcomes.

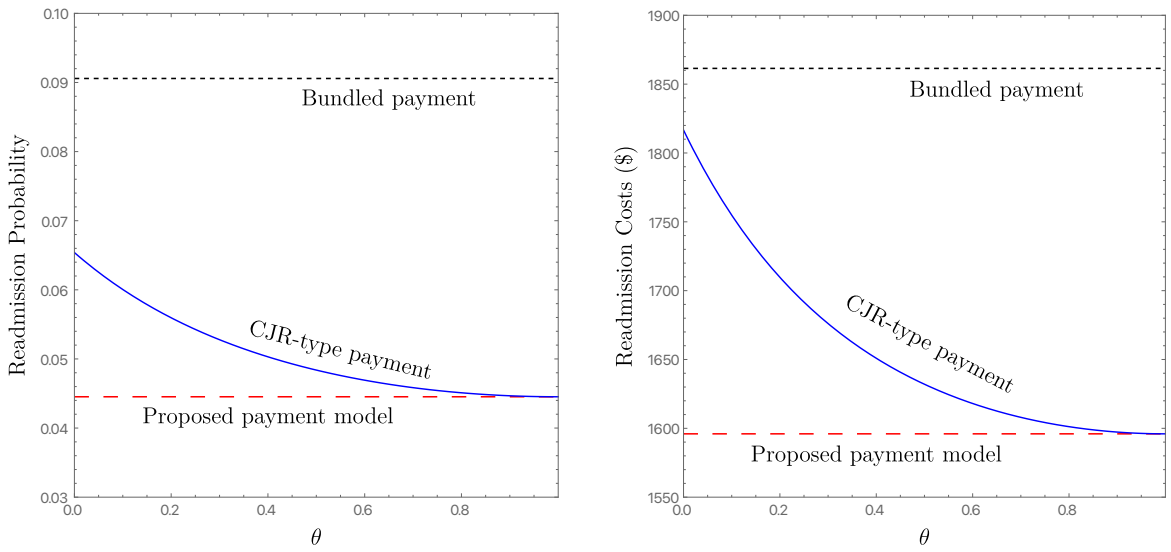
The numerical results show that single-entity payment models lead to 18% higher total PAC costs, amounting to $5,200 per patient compared to $4,408 under both the proposed payment model and the CJR-type payment model. Furthermore, as shown in Figure 1, the proposed payment model induces the socially optimal readmission probability of 4.5%, whereas the bundled payment model results in a readmission probability of 9.1%, and the CJR-type payment model at $\theta = 0$ leads to a readmission probability of 6.6%. The lower readmission probability under the proposed payment model translates into reductions in readmission costs of 14% and 12% compared to the bundled payment and the CJR-type payment model, respectively. These reductions correspond to per-patient cost savings of $265 and $221, respectively.

The aggregate financial impact of these savings is substantial. Extrapolating the per-patient savings to the entire Medicare population, using patient volumes from 2019, yields estimated annual savings of approximately $197 million under the proposed payment model, compared to the bundled payment model, and $164 million compared to the CJR-type payment model. Projections indicate that these savings could reach $325 million and $271 million annually by 2030, assuming similar baseline costs for regions currently not subject to the CJR payment model (Shichman et al. 2023).

Surprisingly, even partial cost-sharing under the CJR-type payment model significantly improves outcomes. When hospitals and PAC providers share just 50% of readmission costs ($\theta = 0.5$), the readmission probability drops below 4.8%, approaching the socially optimal level, while readmission costs fall within 2% of first-best levels. These results highlight the potential of well-structured gainsharing agreements to promote collaborative care—even without full cost-sharing.

Beyond cost reduction, the proposed payment models have additional benefits that contribute to overall healthcare system efficiency. By substantially reducing readmission probabilities, it lowers patient health risks, minimizes the disruption and inconvenience associated with readmission episodes, and optimizes the utilization of healthcare resources. The improved alignment of financial incentives between hospitals and PAC providers fosters a more coordinated approach to patient care, leading to enhanced outcomes and a more efficient allocation of healthcare expenditures.

Figure 1 **Equilibrium readmission probabilities and costs under different payment models ($\tau^h = 6,095$, $\tau^s = 1,102$, $\tau = 399$, $\xi^h = 10,124$, $\xi^s = 7,798$; cost and readmission probability functions are given by (24)-(26))**



(a) Readmission Probability

(b) Readmission Costs

6. Practical considerations

While our analysis has focused on incentive mechanisms for payment models to promote collaborative care, implementing these payment models requires additional considerations like data availability and cost estimation. This section focuses on the practical aspects by examining how the proposed payment model can be effectively implemented and flexibly adjusted to accommodate various practical considerations.

We begin by demonstrating our proposed payment model's adaptability in using total cost per patient as a cost benchmark for each provider in Section 6.1. This modification facilitates the use of readily available data for cost benchmarking, simplifying implementation. In Section 6.2, we address parameter estimation, highlighting the feasibility of leveraging existing data from the CJR program. Next, we discuss the model's flexibility, particularly its ability to incorporate aspects of the CJR framework in Section 6.3. Finally, we show in Section 6.4 that our payment model continues to yield first-best outcomes even when some of our modeling assumptions are relaxed.

6.1. Total costs

It is possible to simplify the implementation of the proposed payment model by reducing the number of cost components regulators need to monitor.

To compute PAC provider payment T_j^s in (21), regulators only need to observe the total PAC cost for each patient, expressed as $C_{ij}^s + b_{ij}^s + e_{ij}^s$, along with readmission probabilities and costs. However, hospital payment T_i^h in (20) is determined solely by PAC costs $C^s(b_{ij}^h, b_{ij}^s)$ and does not

account for costs b^s_{ij} and e^s_{ij}. When it is difficult to observe individual cost components separately, we propose adjusting the hospital's payment structure to be based on the total PAC cost.

We define the total cost of PAC provider j's treatment for patients discharged from hospital i as $\mathscr{C}^{sh}_{ij} = C^s(b^h_{ij}, b^s_{ij}) + b^s_{ij} + e^s_{ij}$, and let $\bar{\mathscr{C}}^{sh}_i$ denote the average total cost for hospital i as given by:

$$\mathscr{C}^{sh}_{ij} = C^s(b^h_{ij}, b^s_{ij}) + b^s_{ij} + e^s_{ij}, \tag{31}$$

and let

$$\bar{\mathscr{C}}^{sh}_i = \frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}} p_{kj}\mathscr{C}^{sh}_{kj}}{1-p_i}. \tag{32}$$

The payment amount for hospital i is then given by

$$\mathscr{T}^h_i = p_i\Big[\bar{C}^h_i + \bar{a}^h_i + \bar{b}^h_i + \bar{e}^h_i + \bar{R}^h_i\xi^h\Big] + \sum_{j\in\mathcal{M}} p_{ij}\Big[\bar{\mathscr{C}}^{sh}_i - \mathscr{C}^{sh}_{ij}\Big] + \sum_{j\in\mathcal{M}} p_{ij}\Big[\bar{R}^h_i - R(e^h_{ij}, e^s_{ij})\Big]\xi^s. \tag{33}$$

Note that the first component above is the average total cost of all hospitals except i. We can show that Theorem 1 remains valid when T^s_j defined in (21) is used along with $\mathscr{T}^h_i$. The implementation of the proposed model within the CJR program (see Section 5.2) can be similarly modified.

6.2. Informational requirements

In this section, we explain how the proposed payment model can be implemented while minimizing additional information requirements for CMS, building on the observations from the previous section. The CJR program already collects data to determine payments to providers, and this information can also be used to calculate payment amounts in our model. Additionally, other methods can be employed when more data are available. We explain these methods and provide guidance on implementation where applicable.

Our payment model requires two types of data: patient flow and readmission data and cost estimates.

Patient flow and readmissions data: This includes the number of patients treated by each provider, along with the pair of providers (where applicable) who treated them, and readmission status (corresponding to p_i, $\tilde{p}_j$, p_{ij}, and R in (33) and (21)). The CJR program already relies on patient flow and readmissions data collected by CMS from providers (CMS 2018, Ko et al. 2022). Thus, the proposed payment model does not induce additional information burden regarding patient flows.

Cost Estimates: Our model extends the single-entity DRG-based payment models explored in prior research (Savva et al. 2019, Arifoğlu et al. 2021), which use prospective payments. In these models, payment amounts do not require precise cost estimates as long as they ensure providers can break even (see the discussion in the paragraph preceeding Proposition 1). This is because payments are fixed and independent of provider actions. Similarly, parts of our payment model follow this structure and do not require highly accurate cost estimates. In particular, the cost-of-care payments ((33) and (21)) for hospital costs operate in the same way, allowing them to rely on approximate estimates without affecting the model's overall functionality.

However, the payment term $\mathscr{C}_{ij}^{sh}$ in (33), which reflects the PAC provider's treatment cost, directly influences hospital decisions. As a result, obtaining precise cost estimates for PAC providers is important, while other cost components do not require the same level of accuracy. Next, we present estimation methods used in the CJR program and other applications that align with these requirements.

CMS determines bonus payments for hospitals under the CJR program based on payment data rather than based on their actual cost data. While this method does not perfectly reflect actual provider costs, it is the approach currently used and can be incorporated into our proposed payment model. In addition, providers retain some discretion in cost reduction even with the DRG-based payment system. For example, institutional PAC providers can reduce their CMS payments by shortening patient lengths of stay. Additionally, these providers receive separate payments for certain non-therapy ancillary services, such as medications, respiratory therapy, and specialized equipment (CMS 2024). Because CMS already uses these payments to evaluate cost efficiency in the CJR program, they provide a practical baseline for cost estimation in our proposed payment model. Specifically, payments made to hospitals and PAC providers can be used to approximate:

$$C_i^h + a_i^h + b_i^h + e_i^h, \text{ and } \xi^h$$

for hospitals and

$$C_{ij}^s + b_{ij}^s + e_{ij}^s, \text{ and } \xi^s$$

for PAC providers. These values can then be used in our payment calculations (33) and (21), as well as the implementation within the CJR program (see Section 5.2).

Although payment-based estimates align with CMS practices, they may not fully capture cost efficiencies in provider expenses.[6] Therefore, more refined methods are needed to improve accuracy. CMS already collects claims and cost reports from providers, which offer a more detailed view of

[6] Discrepancies between CMS cost estimates and actual provider costs may influence provider actions under both the CJR program and our payment model. However, a detailed examination of these effects is beyond the scope of this paper and is left for future research.

treatment costs beyond what payment data alone can provide. These reports include financial and operational details that facilitate cost-to-charge ratio adjustments, as demonstrated in Oomer et al. (2017), Taira et al. (2003), Salemi et al. (2013). This approach provides a more precise measure of provider costs than relying solely on payments.

An alternative cost estimation framework is used by the NHS in the UK (Amies-Cull et al. 2023). The NHS calculates costs at the provider level using the Patient Level Information and Costing Systems (PLICS). This system assigns costs, such as staff salaries, equipment usage, medications, and overheads, to individual patient care episodes. By grouping similar episodes under Healthcare Resource Groups (HRGs) and aggregating their costs, the NHS determines the average cost of delivering services for each HRG. These calculations adhere to nationally mandated costing standards, ensuring consistency and comparability across providers. The resulting data informs payment mechanisms, resource allocation, and benchmarking across the healthcare system.

6.3. Flexibility of the proposed payment model

The proposed payment model is adaptable to incorporate some of the features used in CJR.

Regionalization: The CJR program employs a regionalized payment approach, setting target prices for each region to reflect cost variations, as noted by (Department of Health and Human Services 2021). Our model can also adapt to regional cost variations by offering flexibility in calculating payment amounts and benchmark parameters. For instance, a hospital's payment could be determined based on its performance relative to similar hospitals in the same region, combined with the average performance of a selected provider group. This method allows for the setting of flexible payment amounts and performance targets that take regional cost differences into account.

Risk exposure: Despite its benefits, our model increases risk exposure for PAC providers, particularly in the context of excess acute-care costs from readmissions. This risk could be mitigated by implementing strategies such as limiting the initial share of penalties and rewards and increasing them gradually, akin to the approach used in the CJR payment model for hospitals. Such an approach would give PAC providers time to adapt and optimize their collaboration with acute-care providers.

6.4. Extensions

We extend our proposed payment model by incorporating several practical considerations and demonstrate that it continues to elicit first-best outcomes with appropriate modifications.

Endogenous discharge decisions: We consider different types of PAC settings, e.g., SNFs, home care, and inpatient rehabilitation centers, with different care intensity as proxied by the PAC cost and readmission functions. Patients are heterogeneous in their risks for readmissions and hospitals decide the fraction of their patients discharged to each PAC setting, with riskier patients discharged

to more intensive PAC settings, e.g., SNFs than home care. In each PAC setting, discharged patients receive PAC from providers in that PAC setting with exogenous discharge probabilities, as in the main model. See Appendix E for details.

The discharge decisions have significant implications for PAC providers, hospitals, and social welfare. Specifically, these decisions determine the readmission risk of patients discharged to each PAC setting, thereby affecting the collaborative investments of hospitals and PAC providers in reducing readmissions and the PAC costs. In addition, the socially optimal discharge decisions are made by weighing the benefits of reduced readmissions from discharging more patients to more intense PAC settings against the increased PAC costs.

Despite these intricacies, we show that our proposed payment model continues to incentivize hospitals and PAC providers to make socially optimal decisions, with the following modifications. First, each hospital is reimbursed for their collaborative investments in reducing PAC costs and readmissions based on the average fraction of patients discharged to the same PAC setting among other hospitals. The reimbursement no longer depends on the hospital's own discharge fractions which could distort its collaborative investment decisions. Second, hospital payments are adjusted by the PAC providers' investments to reduce the PAC cost and readmissions relative to other PAC providers. This new payment adjustment incentivizes hospitals to consider the implications of their discharge decisions on PAC providers' investments in reducing costs and readmissions.

Endogenous readmission treatment costs: We extend our original model by considering that, the cost of treating a patient who is readmitted depends on the providers' actions. Specifically, we define C^h and C^s as the treatment costs for hospitals and PAC providers, respectively, for both the initial admission and readmission. We update our proposed payment model by replacing the (exogenous) readmission treatment costs ξ^h and ξ^s with the corresponding average treatment costs among all other providers. We prove that the payment model continues to elicit socially optimal decisions. See Appendix F for details.

Uniform investments: In our original model, we assume that hospitals make different PAC provider dependent investments to reduce the PAC costs, and PAC providers make different hospital dependent investments. In this extension, we assume that hospitals invest the same amount for all PAC providers, and that PAC providers invest the same amount for all hospitals. We update the reimbursements for these investments in our proposed payment model and show that it continues to elicit socially optimal decisions. See Appendix G for details.

Fixed investments: Our original model focuses on variable investments which are accounted on a per-patient basis. We extend it by considering lump-sum investments from hospitals and PAC providers. We consider two separate cases in which the hospital/PAC provider lump-sum investment is uniform or specific to PAC provider/hospital. In each case, we update the reimbursements for

provider investments in our proposed payment model and show that it continues to elicit socially optimal decisions. See Appendix H for details.

Non-identical providers: We extend our original model by considering heterogeneity among providers and patients across multiple dimensions, such as geographic and demographic factors. If these factors are observable by the regulator and exogenous to the providers, our proposed payment model can be adapted to accommodate this heterogeneity and continue to elicit socially optimal decisions. Specifically, the regulator calculates relative-performance benchmarks based on each provider's observable characteristics through an estimation procedure (e.g., linear regression). This aligns with the CJR program's risk-adjustment procedure, which uses a linear regression model with variables on patient characteristics and health status to adjust the target price for different patients. See Appendix I for details.

7. Conclusions

Payment models play a critical role in shaping healthcare providers' incentives to deliver high-quality, cost-effective care. While DRG- and performance-based payment models have demonstrated effectiveness in single-entity settings, their ability to promote care coordination is limited in settings involving multiple independent entities, such as hospitals and PAC providers in joint replacement surgeries. To enhance care coordination, CMS implemented the CJR program, which holds participating hospitals financially accountable for the entire episode of care, encompassing hospitalization and PAC services. However, despite this effort, the program has yielded only modest and statistically insignificant cost savings. This raises the question of which payment models are capable of improving care coordination and how they might be effectively integrated within existing healthcare reimbursement structures.

This paper develops an analytical framework to examine incentive mechanisms underpinning multi-entity payment models. Our analysis shows that conventional bundled payment models and the CJR framework, especially without gainsharing agreements, do not fully align provider incentives. To address these limitations, we propose a payment model that explicitly links hospital and PAC provider reimbursements. Using a game-theoretical framework, we show that this model results in socially optimal provider actions. We further show that the proposed payment model can be seamlessly incorporated into the existing CJR framework through structured gainsharing agreements, and that the proposed model can be flexibly adjusted to incorporate various practical considerations.

Limitations and future research: Our model assumes a sufficiently high patient volume to stabilize performance estimates, which may not hold for smaller institutions or providers in less populated regions. Addressing variability in such settings requires further refinement, potentially

through adjustments in risk-sharing mechanisms. Additionally, while our analysis primarily focuses on readmissions as a quality measure, the CJR model incorporates a broader set of indicators, including complication rates and patient-reported outcomes. Future research could explore the integration of these additional metrics into our framework.

Like other bundled payment systems, our model does not address potential over-treatment or provider risk selection behaviors, which could skew care provision towards healthier patients. These issues underline the need for ongoing monitoring and potential model adjustments to align provider incentives more closely with patient health outcomes. Developing payment models that curtail incentives for over-treatment and risk selection remains a crucial area of research to improve collaborative care delivery. Additionally, the CJR payment model was made voluntary in certain regions and the effects of voluntary participation in payment models and their implications for provider selection and patient outcomes present valuable research avenues (see Einav et al. (2022)).

References

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

Adida, E. and F. Bravo (2023). Primary care first initiative: Impact on care delivery and outcomes. *Manufacturing & Service Operations Management 25*(4), 1471–1488.

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

Adler-Milstein, J., K. Raphael, T. A. O'Malley, and D. A. Cross (2021). Information Sharing Practices Between US Hospitals and Skilled Nursing Facilities to Support Care Transitions. *JAMA Network Open 4*(1), e2033980–e2033980.

Amies-Cull, B. et al. (2023). NHS Reference Costs: A History and Cautionary Note. *Health Economics Review 13*(1), 54.

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

Arana, M., L. Harper, H. Qin, and J. Mabrey (2017). Reducing length of stay, direct cost, and readmissions in total joint arthroplasty patients with an outcomes manager-led interprofessional team. *Orthopedic Nursing 36*(4), 279–284.

Arifoğlu, K., H. Ren, and T. Tezcan (2021). Hospital Readmissions Reduction Program does not provide the right incentives: Issues and remedies. *Management Science 67*(4), 2191–2210.

Aswani, A., Z.-J. M. Shen, and A. Siddiq (2019). Data-driven incentive design in the medicare shared savings program. *Operations Research 67*(4), 1002–1026.

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

Barnett, M. L., A. Wilcock, J. M. McWilliams, A. M. Epstein, K. E. Joynt Maddox, E. J. Orav, D. C. Grabowski, and A. Mehrotra (2019). Two-year evaluation of mandatory bundled payments for joint replacement. *New England Journal of Medicine 380*(3), 252–262.

Bastani, H., M. Bayati, M. Braverman, R. Gummadi, and R. Johari (2016). Analysis of Medicare pay-for-performance contracts. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2839143, last accessed on July 14, 2023.

Bavafa, H., S. Savin, and C. Terwiesch (2021). Customizing primary care delivery using e-visits. *Production and Operations Management 30*(11), 4306–4327.

Blumenthal, D., M. K. Abrams, and R. Nuzum (2015). The Affordable Care Act at 5 years. *New England Journal of Medicine 372*(25), 2451–2458.

Bravo, F., R. Levi, G. Perakis, and G. Romero (2023). Care coordination for healthcare referrals under a shared-savings program. *Production and Operations Management 32*(1), 189–206.

Britton, M. C., G. M. Ouellet, K. E. Minges, M. Gawel, B. Hodshon, and S. I. Chaudhry (2017). Care transitions between hospitals and skilled nursing facilities: Perspectives of sending and receiving providers. *The Joint Commission Journal on Quality and Patient Safety 43*(11), 565–572.

Chen, C. and K. Delana (2025). The role of physician integration in alternative payment models: The case of the comprehensive joint replacement program. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4195640, last accessed on April 02, 2025.

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

Cipriano, C. A., N. Brown, and S. D. Holubar (2018). Comprehensive perioperative care for total joint arthroplasty: a narrative review. *Journal of Arthroplasty 33*(8), 2444–2450.

CMS (2018). Overview of CJR quality measures, composite quality score, and pay-for-performance methodology. https://innovation.cms.gov/files/x/cjr-qualsup.pdf, last accessed on July 14, 2023.

CMS (2021a). CMS comprehensive care for joint replacement model: Performance year 4 evaluation report. https://tinyurl.com/2021-cjr-py4-annual-report, last accessed on July 24, 2024.

CMS (2021b). CMS comprehensive care for joint replacement model: Performance year 4 evaluation report – appendices. https://tinyurl.com/2021-cjr-py4-app, last accessed on May 28, 2024.

CMS (2023). Value-based programs. https://www.cms.gov/medicare/quality/value-based-programs, last accessed on July 14, 2023.

CMS (2024). Inpatient prospective payment system (ipps). Accessed: Mar. 1, 2025.

CMS (2024). Medicare Program; Prospective Payment System and Consolidated Billing for Skilled Nursing Facilities; Updates to the Quality Reporting Program and Value-Based Purchasing Program for Federal Fiscal Year 2025. *Federal Register (April 3, 2024) 89*(65), 23424–23495.

Cots, F., P. Chiarello, X. Salvador, X. Castells, and W. Quentin (2011). DRG-Based hospital payment: Intended and unintended consequences. In W. Q. Reinhard Busse, Alexander Geissler and M. Wiley (Eds.), *Diagnosis-Related Groups in Europe: Moving towards Transparency, Efficiency and Quality in Hospitals*, pp. 75–92. Maidenhead: McGraw Hill Open University Press.

Davis, C. and D. J. Rhodes (1988). The impact of DRGs on the cost and quality of health care in the united states. *Health Policy 9*(2), 117–131.

Department of Health and Human Services (2017). Medicare Program; Cancellation of advancing care coordination through episode payment and cardiac rehabilitation incentive payment models. *Federal Register (December 1, 2017) 82*(230), 57066 –57104.

Department of Health and Human Services (2021). Medicare Program: Comprehensive Care for Joint Replacement Model three-year extension and changes to episode definition and pricing. *Federal Register (May 3, 2021) 86*(83), 23496–23576.

Einav, L., A. Finkelstein, Y. Ji, and N. Mahoney (2022). Voluntary regulation: Evidence from medicare payment reform. *The Quarterly Journal of Economics 137*(1), 565–618.

Ellimoottil, C., A. M. Ryan, H. Hou, J. Dupree, B. Hallstrom, and D. C. Miller (2016). Medicare's new bundled payment for joint replacement may penalize hospitals that treat medically complex patients. *Health Affairs 35*(9), 1651–1657.

Finkelstein, A., Y. Ji, N. Mahoney, and J. Skinner (2018). Mandatory Medicare Bundled Payment Program for Lower Extremity Joint Replacement and Discharge to Institutional Postacute Care: Interim Analysis of the First Year of a 5-Year Randomized Trial. *JAMA 320*(9), 892–900.

Ghamat, S., G. S. Zaric, and H. Pun (2021). Care-coordination: Gain-sharing agreements in bundled payment models. *Production and Operations Management 30*(5), 1457–1474.

Goodman, E. and T. Dai (2024). Impact of physician payment scheme on diagnostic effort and testing. To appear in Management Science.

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

Gupta, D., M. Mehrotra, and X. Tang (2021). Gainsharing Contracts for CMS' Episode-Based Payment Models. *Production and Operations Management 30*(5), 1290–1312.

Haas, D. A., X. Zhang, R. S. Kaplan, and Z. Song (2019). Evaluation of economic and clinical outcomes under centers for medicare & medicaid services mandatory bundled payments for joint replacements. *JAMA Internal Medicine 179*(7), 924–931.

Holmstrom, B. (1982). Moral hazard in teams. *The Bell Journal of Economics 13*(2), 324–340.

Hopewell, C., A. Cowell, K. Olzenak, A. Heinzerling, C. Simon, A. Markovitz, L. Alecxih, and A. Ackerman (2024). Drivers of care transformation. https://www.cms.gov/priorities/innovation/data-and-reports/2024/cjr-py6-ar-drivers-transformation, last accessed February 03, 2025.

Hwang, W., J. O. Jónasson, and H. Peura (2023). Favorable risk selection in Medicare Advantage: The effect of allowing non-medical services. Available at SSRN: https://tinyurl.com/ssrnFavRiskSel, last accessed July 8, 2024.

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

Jiang, H., Z. Pang, and S. Savin (2020). Performance incentives and competition in health care markets. *Production and operations management 29*(5), 1145–1164.

Jiang, H., Z. Pang, and S. Savin (2021). How should payers respond to consolidation in healthcare markets? Available at SSRN: https://tinyurl.com/ssrnConsolidation, last accessed July 08, 2024.

Ko, H., B. I. Martin, R. E. Nelson, and C. E. Pelt (2022). Patient Selection in the Comprehensive Care for Joint Replacement Model. *Health Services Research 57*(1), 72–90.

Li, Y., M. Ying, X. Cai, Y. Kim, and C. P. Thirukumaran (2020). Trends in Postacute Care Use and Outcomes After Hip and Knee Replacements in Dual-Eligible Medicare and Medicaid Beneficiaries, 2013-2016. *JAMA Network Open 3*(3), e200368–e200368.

McGarry, B. E. and D. C. Grabowski (2017). Helping patients make more informed postacute care choices. Health Affairs Blog, https://tinyurl.com/HealthABlog1, last accessed on July 14, 2023.

MedPAC (2022). Report to the congress: Medicare payment policy (march 2022). https://tinyurl.com/MedpacMarch22, last accessed on April 21, 2023.

Mookherjee, D. (1984). Optimal incentive schemes with many agents. *The Review of Economic Studies 51*(3), 433–446.

OECD (2019). *Health at a Glance 2019: OECD Indicators*. Paris: OECD Publishing.

Ok, E. A. (2007). *Real analysis with economic applications*, Volume 10. Princeton University Press.

Oomer, N. M., M. J. Ingber, and L. Coots (2017, January). *Using Medicare Cost Reports to Calculate Costs for Post-Acute Care Claims*. Research Triangle Park (NC): RTI Press.

Phillips, J. L., A. J. Rondon, C. Vannello, Y. A. Fillingham, M. S. Austin, and P. M. Courtney (2019). How much does a readmission cost the bundle following primary hip and knee arthroplasty? *The Journal of Arthroplasty 34*(5), 819–823.

Rajagopalan, S. and C. Tong (2022). Payment models to coordinate healthcare providers with partial attribution of outcome costs. *Manufacturing & Service Operations Management 24*(1), 600–616.

Salanie, B. (1997). *The economics of contracts: A primer*. MIT Press.

Salemi, J. L., M. M. Comins, K. Chandler, M. F. Mogos, and H. M. Salihu (2013). A practical approach for calculating reliable cost estimates from observational data: Application to cost analyses in maternal and child health. *Applied Health Economics and Health Policy 11*(4), 343–357.

Savva, N., L. Debo, and R. A. Shumsky (2023). Hospital reimbursement in the presence of cherry picking and upcoding. *Management Science 69*(11), 6777–6799.

Savva, N., T. Tezcan, and Ö. Yıldız (2019). Can yardstick competition reduce waiting times? *Management Science 65*(7), 3196–3215.

Schwarzkopf, R., J. Ho, J. R. Quinn, N. Snir, and D. Mukamel (2016). Factors influencing discharge destination after total knee arthroplasty: A database analysis. *Geriatric Orthopaedic Surgery & Rehabilitation 7*(2), 95–99.

Shichman, I., M. Roof, N. Askew, L. Nherera, J. C. Rozell, T. M. Seyler, and R. Schwarzkopf (2023). Projections and epidemiology of primary hip and knee arthroplasty in medicare patients to 2040-2060. *The Journal of Bone and Joint Surgery 8*(1), e22.00112.

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

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

Sundaram, R. K. (1996). *A first course in optimization theory*. Cambridge University Press.

Taira, D. A., T. B. Seto, R. Siegrist, R. Cosgrove, R. Berezin, and D. J. Cohen (2003, Mar). Comparison of analytic approaches for the economic evaluation of new technologies alongside multicenter clinical trials. *American Heart Journal 145*(3), 452–458.

Vlachy, J., T. Ayer, M. Ayvaci, and S. Raghunathan (2023). The business of healthcare: The role of physician integration in bundled payments. To appear in Manufacturing & Service Operations Management.

Welsh, R. L., J. E. Graham, A. M. Karmarkar, N. E. Leland, J. G. Baillargeon, D. L. Wild, and K. J. Ottenbacher (2017). Effects of postacute settings on readmission rates and reasons for readmission following total knee arthroplasty. *JAMDA 18*(4), 367.e1–367.e10.

Werner, R. M., N. B. Coe, M. Qi, and R. T. Konetzka (2019). Patient outcomes after hospital discharge to home with home health care vs to a skilled nursing facility. *JAMA Internal Medicine 179*(5), 617–623.

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

Zhu, J., V. Patel, J. Shea, M. Neuman, and R. Werner (2018). Hospitals using bundled payment report reducing skilled nursing facility use and improving care integration. *Health Affairs 37*, 1282–1289.

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

Appendix

A. Conditions for unique socially optimal actions determined by FOCs

In this section we prove that the regulator has unique optimal actions that can be determined by FOCs (6)–(10) under certain conditions.

Assumption A-1. *(i) The expected acute care cost is strictly decreasing and strictly convex in hospital investment, i.e.,* $dC^h(a^h)/da^h < 0$ *and* $d^2C^h(a^h)/d(a^h)^2 > 0$*, and the following boundary conditions hold.*

$$\lim_{a^h \downarrow 0} \frac{dC^h(a^h)}{da^h} < -1 < \lim_{a^h \uparrow \Gamma} \frac{dC^h(a^h)}{da^h}.$$

(ii) The expected PAC cost is decreasing and jointly strictly convex in hospital and SNF investments, i.e.,

$$\frac{\partial C^s(b^h,b^s)}{\partial b^i} < 0 \text{ and } \frac{\partial^2 C^s(b^h,b^s)}{\partial (b^i)^2} > 0 \text{ for } i = h,s, \ \frac{\partial^2 C^s(b^h,b^s)}{\partial (b^h)^2}\frac{\partial^2 C^s(b^h,b^s)}{\partial (b^s)^2} > \left(\frac{\partial^2 C^s(b^h,b^s)}{\partial b^h \partial b^s}\right)^2,$$

and the following boundary conditions hold.

$$\lim_{b^h \downarrow 0} \frac{\partial C^s(b^h,b^s)}{\partial b^h} < -1 < \lim_{b^h \uparrow \Gamma} \frac{\partial C^s(b^h,b^s)}{\partial b^h} \text{ for all } b^s \in [0,\Gamma],$$
$$\lim_{b^s \downarrow 0} \frac{\partial C^s(b^h,b^s)}{\partial b^s} < -1 < \lim_{b^s \uparrow \Gamma} \frac{\partial C^s(b^h,b^s)}{\partial b^s} \text{ for all } b^h \in [0,\Gamma].$$

(iii) Readmission probability is strictly decreasing and jointly strictly convex in hospital and SNF investments, i.e.,

$$\frac{\partial R(e^h,e^s)}{\partial e^i} < 0 \text{ and } \frac{\partial^2 R(e^h,e^s)}{\partial (e^i)^2} > 0 \text{ for } i = h,s, \ \frac{\partial^2 R(e^h,e^s)}{\partial (e^h)^2}\frac{\partial^2 R(e^h,e^s)}{\partial (e^s)^2} > \left(\frac{\partial^2 R(e^h,e^s)}{\partial e^h \partial e^s}\right)^2,$$

and the following boundary conditions hold.

$$\lim_{e^h \downarrow 0} \frac{\partial R(e^h,e^s)}{\partial e^h} < -\frac{1}{\xi^h + \xi^s} < \lim_{e^h \uparrow \Gamma} \frac{\partial R(e^h,e^s)}{\partial e^h} \text{ for all } e^s \in [0,\Gamma], \tag{A-1}$$
$$\lim_{e^s \downarrow 0} \frac{\partial R(e^h,e^s)}{\partial e^s} < -\frac{1}{\xi^h + \xi^s} < \lim_{e^s \uparrow \Gamma} \frac{\partial R(e^h,e^s)}{\partial e^s} \text{ for all } e^h \in [0,\Gamma]. \tag{A-2}$$

Under these assumptions, the socially optimal (or first-best) actions, denoted by $(a_h^*, b_h^*, b_s^*, e_h^*, e_s^*)$, are unique and can be characterized using the FOCs.

Lemma A-1 (First-best benchmark). *The regulator's objective in* (5) *has a unique maximizer in which* $a_i^h = a_h^*$ *for each hospital* $i \in \mathcal{N}$*, and for each PAC provider* $j \in \mathcal{M}$ *such that* $p_{ij} > 0$*,* $b_{ij}^h = b_h^*, b_{ij}^s = b_s^*, e_{ij}^h = e_h^*$*, and* $e_{ij}^s = e_s^*$*, where* $a_h^*, b_h^*, b_s^*, e_h^*, e_s^* \in (0,\Gamma)$ *satisfy FOCs* (6)–(10)*.*

Proof of Lemma A-1. Let $\vec{\mathbf{h}} = \{\mathbf{h}_i, i \in \mathcal{N}\}$ and $\vec{\mathbf{s}} = \{\mathbf{s}_j, j \in \mathcal{M}\}$ denote the actions of all hospitals and all PAC providers, respectively. Thus, total welfare W is a function of $\vec{\mathbf{h}}$ and $\vec{\mathbf{s}}$, and by (2), (4), and (5), is given by

$$W = \upsilon - \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} p_{ij}\Big[C^h(a_i^h) + a_i^h + C^s(b_{ij}^h, b_{ij}^s) + b_{ij}^h + b_{ij}^s + R(e_{ij}^h, e_{ij}^s)(\xi^h + \xi^s) + e_{ij}^h + e_{ij}^s\Big]. \quad \text{(A-3)}$$

For notational simplicity, we will drop the arguments when it is clear from the context.

First, we characterize a_h^*, i.e., hospital's first-best investment to reduce the expected acute care cost. Taking the first and second partial derivatives of W in (A-3) with respect to a_i^h, we have

$$\frac{\partial W}{\partial a_i^h} = -p_i\left[\frac{dC^h(a_i^h)}{da_i^h} + 1\right],$$
$$\frac{\partial^2 W}{\partial (a_i^h)^2} = -p_i\frac{d^2C^h(a_i^h)}{d(a_i^h)^2}.$$

By Assumption A-1(i), we have $\partial^2 W/\partial(a_i^h)^2 < 0$, $\lim_{a_i^h\downarrow 0}\partial W/\partial a_i^h > 0$, and $\lim_{a_i^h\uparrow\Gamma}\partial W/\partial a_i^h < 0$. Thus, there exists a unique $a_h^* \in (0,\Gamma)$ that satisfies FOC (6) and

$$a_h^* = \underset{a_i^h\in[0,\Gamma]}{\arg\max}\, W(\vec{\mathbf{h}}, \vec{\mathbf{s}}) \text{ for each } i \in \mathcal{N} \text{ and any fixed } (\vec{\mathbf{h}}, \vec{\mathbf{s}}) \setminus \{a_i^h\} \quad \text{(A-4)}$$

Second, we characterize (b_h^*, b_s^*), i.e., first-best investments made by the hospital and PAC provider to reduce the expected PAC cost. When $p_{ij} = 0$, W is independent of b_{ij}^h and b_{ij}^s. Without loss of generality (WLOG) we assume that first-best actions are taken (see the last paragraph of Section 3 for details), i.e., $b_{ij}^h = b_h^*$ and $b_{ij}^s = b_s^*$, where b_h^* and b_s^* are given by (7)-(8). When $p_{ij} > 0$, we take the first and second partial derivatives of W in (A-3) with respect to b_{ij}^s and obtain

$$\frac{\partial W}{\partial b_{ij}^s} = -p_{ij}\left[\frac{\partial C^s(b_{ij}^h, b_{ij}^s)}{\partial b^s} + 1\right],$$
$$\frac{\partial^2 W}{\partial (b_{ij}^s)^2} = -p_{ij}\frac{\partial^2 C^s(b_{ij}^h, b_{ij}^s)}{\partial (b^s)^2}.$$

For any fixed $b_{ij}^h \in [0,\Gamma]$, we have $\partial^2 W/\partial(b_{ij}^s)^2 < 0$, $\lim_{b_{ij}^s\downarrow 0}\partial W/\partial b_{ij}^s > 0$, and $\lim_{b_{ij}^s\uparrow\Gamma}\partial W/\partial b_{ij}^s < 0$ by Assumption A-1(ii). Hence, there exists a unique $g(b_{ij}^h) \in (0,\Gamma)$ that satisfies

$$\frac{\partial C^s(b_{ij}^h, g(b_{ij}^h))}{\partial b^s} + 1 = 0. \quad \text{(A-5)}$$

Applying the Implicit Function Theorem to (A-5), we obtain

$$\frac{dg(b_{ij}^h)}{db_{ij}^h} = -\frac{\partial^2 C^s(b_{ij}^h, g(b_{ij}^h))/\partial b^h \partial b^s}{\partial^2 C^s(b_{ij}^h, g(b_{ij}^h))/\partial (b^s)^2}. \quad \text{(A-6)}$$

Since W is concave in b_{ij}^s by Assumption A-1(ii), we have

$$W|_{b_{ij}^s = g(b_{ij}^h)} = \sup_{b_{ij}^s\in[0,\Gamma]} W.$$

Next we show that for any given $(\vec{\mathbf{h}}, \vec{\mathbf{s}}) \setminus \{b_{ij}^h, b_{ij}^s\}$, there exists a unique $b_h^* \in (0, \Gamma)$ that satisfies

$$W|_{\{b_{ij}^h = b_h^*, b_{ij}^s = g(b_h^*)\}} = \sup_{b_{ij}^h \in [0,\Gamma]} W|_{b_{ij}^s = g(b_{ij}^h)}.$$

Let $W(b_{ij}^h) = W|_{b_{ij}^s = g(b_{ij}^h)}$ for notational simplicity. Then,

$$\begin{aligned}\frac{dW(b_{ij}^h)}{db_{ij}^h} =& -p_{ij}\left[\frac{\partial C^s(b_{ij}^h, g(b_{ij}^h))}{\partial b^h} + \frac{\partial C^s(b_{ij}^h, g(b_{ij}^h))}{\partial b^s}\frac{dg(b_{ij}^h)}{db_{ij}^h} + 1 + \frac{dg(b_{ij}^h)}{db_{ij}^h}\right] \\ =& -p_{ij}\left[\frac{\partial C^s(b_{ij}^h, g(b_{ij}^h))}{\partial b^h} + 1\right], \end{aligned} \tag{A-7}$$

where the second equality follows from (A-5).

$$\begin{aligned}\frac{d^2W(b_{ij}^h)}{d(b_{ij}^h)^2} =& -p_{ij}\left[\frac{\partial^2 C^s(b_{ij}^h, g(b_{ij}^h))}{\partial b^h \partial b^s}\frac{dg(b_{ij}^h)}{db_{ij}^h} + \frac{\partial^2 C^s(b_{ij}^h, g(b_{ij}^h))}{\partial (b^h)^2}\right] \\ =& p_{ij}\left[\frac{\left(\frac{\partial^2 C^s(b_{ij}^h, g(b_{ij}^h))}{\partial b^h \partial b^s}\right)^2}{\frac{\partial^2 C^s(b_{ij}^h, g(b_{ij}^h))}{\partial (b^s)^2}} - \frac{\partial^2 C^s(b_{ij}^h, g(b_{ij}^h))}{\partial (b^h)^2}\right] < 0, \end{aligned}$$

where the second equality by plugging in $dg(b_{ij}^h)/db_{ij}^h$ from (A-6), and the inequality follows from Assumption A-1(ii). Moreover, we have $\lim_{b_{ij}^h \downarrow 0} dW(b_{ij}^h)/db_{ij}^h > 0$ and $\lim_{b_{ij}^h \uparrow \Gamma} dW(b_{ij}^h)/db_{ij}^h < 0$ by Assumption A-1(ii). Thus, $W(b_{ij}^h)$ has a unique maximizer $b_h^* \in (0, \Gamma)$ satisfying the FOC $dW(b_h^*)/db_{ij}^h = 0$, which by (A-7) reduces to

$$\frac{\partial C^s(b_h^*, g(b_h^*))}{\partial b^h} + 1 = 0. \tag{A-8}$$

It then yields (7) by defining

$$b_s^* = g(b_h^*) \in (0, \Gamma), \tag{A-9}$$

and (8) follows by substituting $b_{ij}^h = b_h^*$ into (A-5).

Third, we characterize (e_h^*, e_s^*), i.e., first-best investments made by the hospital and PAC provider to reduce the readmission probability. When $p_{ij} = 0$, W is independent of e_{ij}^h and e_{ij}^s. WLOG we assume that first-best actions are taken (see the last paragraph of Section 3 for details), i.e., $e_{ij}^h = e_h^*$ and $e_{ij}^s = e_s^*$, where e_h^* and e_s^* are given by (9)-(10). When $p_{ij} > 0$, we take the first and second partial derivatives of W in (A-3) with respect to e_{ij}^s and obtain

$$\begin{aligned}\frac{\partial W}{\partial e_{ij}^s} &= -p_{ij}\left[\frac{\partial R(e_{ij}^h, e_{ij}^s)}{\partial e^s}(\xi^h + \xi^s) + 1\right], \\ \frac{\partial^2 W}{\partial (e_{ij}^s)^2} &= -p_{ij}\frac{\partial^2 R(e_{ij}^h, e_{ij}^s)}{\partial (e^s)^2}(\xi^h + \xi^s). \end{aligned}$$

For any fixed $e^h_{ij} \in [0,\Gamma]$, we have $\partial^2 W/\partial (e^s_{ij})^2 < 0$, $\lim_{e^s_{ij}\downarrow 0} \partial W/\partial e^s_{ij} > 0$, and $\lim_{e_s\uparrow\Gamma} \partial W/\partial e^s_{ij} < 0$ by Assumption A-1(iii). Hence there exists a unique $z(e^h_{ij}) \in (0,\Gamma)$ that satisfies

$$\frac{\partial R(e^h_{ij}, z(e^h_{ij}))}{\partial e^s}(\xi^h + \xi^s) + 1 = 0. \tag{A-10}$$

Applying the Implicit Function Theorem, we obtain

$$\frac{dz(e^h_{ij})}{de^h_{ij}} = -\frac{\partial^2 R(e^h_{ij}, z(e^h_{ij}))/\partial e^h \partial e^s}{\partial^2 R(e^h_{ij}, z(e^h_{ij}))/\partial (e^s)^2}. \tag{A-11}$$

Since W is concave in e^s_{ij} by Assumption A-1(iii), we have

$$W|_{e^s_{ij}=z(e^h_{ij})} = \sup_{e^s_{ij}\in[0,\Gamma]} W.$$

Next we show that for any given $(\vec{\mathbf{h}}, \vec{\mathbf{s}}) \setminus \{e^h_{ij}, e^s_{ij}\}$, there exists a unique $e^*_h \in (0,\Gamma)$ that satisfies

$$W|_{\left\{e^h_{ij}=e^*_h, e^s_{ij}=e^*_s\right\}} = \sup_{e^h_{ij}\in[0,\Gamma]} W|_{e^s_{ij}=z(e^h_{ij})}.$$

Let $W(e^h_{ij}) = W|_{e^s_{ij}=z(e^h_{ij})}$ for notational simplicity. Then,

$$\begin{aligned}
\frac{dW(e^h_{ij})}{de^h_{ij}} &= -p_{ij}\left[\frac{\partial R(e^h_{ij}, z(e^h_{ij}))}{\partial e^h}(\xi^h + \xi^s) + \frac{\partial R(e^h_{ij}, z(e^h_{ij}))}{\partial e^s}\frac{dz(e^h_{ij})}{de^h_{ij}}(\xi^h + \xi^s) + 1 + \frac{dz(e^h_{ij})}{de^h_{ij}}\right] \\
&= -p_{ij}\left[\frac{\partial R(e^h_{ij}, z(e^h_{ij}))}{\partial e^h}(\xi^h + \xi^s) + 1\right],
\end{aligned}$$

where the second equality follows from (A-10).

$$\begin{aligned}
\frac{d^2 W(e^h_{ij})}{d(e^h_{ij})^2} =& -p_{ij}\left[\frac{\partial^2 R(e^h_{ij}, z(e^h_{ij}))}{\partial e^h \partial e^s}\frac{dz(e^h_{ij})}{de^h_{ij}} + \frac{\partial^2 R(e^h_{ij}, z(e^h_{ij}))}{\partial (e^h)^2}\right](\xi^h + \xi^s) \\
=& p_{ij}\left[\frac{\left(\frac{\partial^2 R(e^h_{ij}, z(e^h_{ij}))}{\partial e^h \partial e^s}\right)^2}{\frac{\partial^2 R(e^h_{ij}, z(e^h_{ij}))}{\partial (e^s)^2}} - \frac{\partial^2 R(e^h_{ij}, z(e^h_{ij}))}{\partial (e^h)^2}\right](\xi^h + \xi^s) < 0,
\end{aligned}$$

where the second equality follows by plugging in $\frac{dz(e^h_{ij})}{de^h_{ij}}$ from (A-11), and the inequality follows from Assumption A-1(iii). Moreover, we have $\lim_{e^h_{ij}\downarrow 0} dW(e^h_{ij})/de^h_{ij} > 0$ and $\lim_{e^h_{ij}\uparrow\Gamma} dW(e^h_{ij})/de^h_{ij} < 0$ by Assumption A-1(iii). Thus there exists a unique $e^*_h \in (0,\Gamma)$ that satisfies (9) with $e^*_s = z(e^*_h) \in (0,\Gamma)$; (10) follows by substituting $e^h_{ij} = e^*_h$ into (A-10). □

B. Proofs: Bundled payment model

We continue to adopt Assumption A-1 with the first inequalities in (A-1)-(A-2) strengthened respectively into

$$\lim_{e^h\downarrow 0} \frac{\partial R(e^h, e^s)}{\partial e^h} < -\frac{1}{\xi^h} \text{ for all } e^s \in [0,\Gamma], \tag{A-12}$$

$$\lim_{e^s \downarrow 0} \frac{\partial R(e^h, e^s)}{\partial e^s} < -\frac{1}{\xi^s} \text{ for all } e^h \in [0, \Gamma], \tag{A-13}$$

which ensures that under payment model (14)-(15) hospitals and PAC providers have unique best responses that can be determined using FOCs.

Proof of Proposition 1: By (1), (2), and (14), hospital i's objective is

$$\Pi_i^h(\mathbf{h}_i) = \sum_{j \in \mathcal{M}} p_{ij} \left[(\bar{C}_i^h + \bar{R}_i^h \xi^h) - (C^h(a_i^h) + R(e_{ij}^h, e_{ij}^s)\xi^h) + (\bar{a}_i^h + \bar{b}_i^h + \bar{e}_i^h) - (a_i^h + b_{ij}^h + e_{ij}^h) \right]. \tag{A-14}$$

By (3), (4), and (15), PAC provider j's objective is

$$\Pi_j^s(\mathbf{s}_j) = \sum_{i \in \mathcal{N}} p_{ij} \left[(\bar{C}_j^s + \bar{R}_j^s \xi^s) - (C^s(b_{ij}^h, b_{ij}^s) + R(e_{ij}^h, e_{ij}^s)\xi^s) + (\bar{b}_j^s + \bar{e}_j^s) - (b_{ij}^s + e_{ij}^s) \right]. \tag{A-15}$$

We proceed in three steps. First, we prove that each hospital i picks $a_i^h = a_h^*$ for any fixed $(\mathbf{h}_k, \mathbf{s}_j, k \in \mathcal{N}, j \in \mathcal{M}) \setminus \{a_i\}$. Taking the first and second partial derivatives of Π_i^h in (A-14) with respect to a_i^h,

$$\frac{\partial \Pi_i^h}{\partial a_i^h} = -p_i \left[\frac{d(C^h(a_i^h))}{da_i^h} + 1 \right],$$
$$\frac{\partial^2 \Pi_i^h}{\partial (a_i^h)^2} = -p_i \frac{d^2(C^h(a_i^h))}{d(a_i^h)^2}.$$

By Assumption A-1(i), we have $\partial^2 \Pi_i^h / \partial (a_i^h)^2 < 0$, $\lim_{a_i^h \downarrow 0} \partial \Pi_i^h / \partial a_i^h > 0$, and $\lim_{a_i^h \uparrow \Gamma} \partial \Pi_i^h / \partial a_i^h < 0$. Thus, hospital i has a unique optimal action determined by FOC and by (6), we have $a_i^h = a_h^*$.

Second, we analyze hospitals' and PAC providers' investments to reduce the PAC cost in equilibrium. When $p_{ij} = 0$, Π_i^h and Π_j^s are independent of both b_{ij}^h and b_{ij}^s. WLOG we assume that first-best actions are taken (see the last paragraph of Section 3 for details), i.e., $b_{ij}^h = b_h^*$ and $b_{ij}^s = b_s^*$, where b_h^* and b_s^* are given by (7)-(8). When $p_{ij} > 0$, we have $\partial \Pi_i^h / \partial b_{ij}^h = -p_{ij} < 0$, which yields $b_{ij}^h = 0$ in any equilibrium. Plugging in (A-15) and taking the first and second partial derivatives with respect to b_{ij}^s, respectively, we obtain

$$\frac{\partial \Pi_j^s}{\partial b_{ij}^s} = -p_{ij} \left[\frac{\partial C^s(0, b_{ij}^s)}{\partial b^s} + 1 \right],$$
$$\frac{\partial^2 \Pi_j^s}{\partial (b_{ij}^s)^2} = -p_{ij} \frac{\partial^2 C^s(0, b_{ij}^s)}{\partial (b^s)^2}.$$

By Assumption A-1(ii), we have $\partial^2 \Pi_j^s / \partial (b_{ij}^s)^2 < 0$, $\lim_{b_{ij}^s \downarrow 0} \partial \Pi_j^s / \partial b_{ij}^s > 0$, and $\lim_{b_{ij}^s \uparrow \Gamma} \partial \Pi_j^s / \partial b_{ij}^s < 0$. Thus, a unique equilibrium exists in which PAC provider j's action b_{ij}^s is determined by FOC and we have $b_{ij}^s = g(0)$ by (A-5).

Third, we analyze hospitals' and PAC providers' investments to reduce the readmission probability in equilibrium. When $p_{ij} = 0$, Π_i^h and Π_j^s are independent of both e_{ij}^h and e_{ij}^s. WLOG we assume

that first-best actions are taken (see the last paragraph of Section 3 for details), i.e., $e^h_{ij} = e^*_h$ and $e^s_{ij} = e^*_s$, where e^*_h and e^*_s are given by (9)-(10). When $p_{ij} > 0$, we take partial derivatives of Π^h_i and Π^s_j with respect to e^h_{ij} and e^s_{ij}, respectively, and obtain

$$\frac{\partial \Pi^h_i}{\partial e^h_{ij}} = -p_{ij}\left[\frac{\partial R(e^h_{ij}, e^s_{ij})}{\partial e^h}\xi^h + 1\right], \tag{A-16}$$

$$\frac{\partial \Pi^s_j}{\partial e^s_{ij}} = -p_{ij}\left[\frac{\partial R(e^h_{ij}, e^s_{ij})}{\partial e^s}\xi^s + 1\right]. \tag{A-17}$$

We proceed as follows: (i) We show that hospital i and PAC provider j have unique best responses characterized by $e^h_{ij} = \breve{z}_h(e^s_{ij})$ and $e^s_{ij} = \breve{z}_s(e^h_{ij})$. (ii) We establish the existence of a unique equilibrium $(\breve{e}_h, \breve{e}_s)$ as determined by

$$\frac{\partial R(\breve{e}_h, \breve{e}_s)}{\partial e^h}\xi^h + 1 = 0, \tag{A-18}$$

$$\frac{\partial R(\breve{e}_h, \breve{e}_s)}{\partial e^s}\xi^s + 1 = 0. \tag{A-19}$$

(iii) We prove that $(\breve{e}_h, \breve{e}_s) \neq (e^*_h, e^*_s)$ and the total costs of care and investments associated with readmissions, i.e., $\Phi(e^h, e^s) = R(e^h, e^s)(\xi^h + \xi^s) + e^h + e^s$, is strictly convex. The proof is complete by noting that (e^*_h, e^*_s) is the unique unconstrained maximizer of $\Phi(e^h, e^s)$ by Lemma A-1.

(i) For any fixed $e^s_{ij} \in [0, \Gamma]$, by Assumption A-1(iii), (A-12), and (A-16), we have $\partial^2 \Pi^h_i / \partial (e^h_{ij})^2 < 0$, $\lim_{e^h_{ij} \downarrow 0} \partial \Pi^h_i / \partial e^h_{ij} > 0$, and $\lim_{e^h_{ij} \uparrow \Gamma} \partial \Pi^h_i / \partial e^h_{ij} < 0$. Thus there exists a unique $\breve{z}_h(e^s_{ij}) \in (0, \Gamma)$ satisfying

$$\frac{\partial R(\breve{z}_h(e^s_{ij}), e^s_{ij})}{\partial e^h}\xi^h + 1 = 0. \tag{A-20}$$

For any fixed $e^h_{ij} \in [0, \Gamma]$, by Assumption A-1(iii), (A-13), and (A-17), we have $\partial^2 \Pi^s_j / \partial (e^s_{ij})^2 < 0$, $\lim_{e^s_{ij} \downarrow 0} \partial \Pi^s_j / \partial e^s_{ij} > 0$, and $\lim_{e^s_{ij} \uparrow \Gamma} \partial \Pi^s_j / \partial e^s_{ij} < 0$. Thus there exists a unique $\breve{z}_s(e^h_{ij}) \in (0, \Gamma)$ satisfying

$$\frac{\partial R(e^h_{ij}, \breve{z}_s(e^h_{ij}))}{\partial e^s}\xi^s + 1 = 0. \tag{A-21}$$

(ii) By part (i), any equilibrium $(\breve{e}_h, \breve{e}_s)$ must satisfy $\breve{e}_h = \breve{z}_h(\breve{e}_s)$ and $\breve{e}_s = \breve{z}_s(\breve{e}_h)$; this and (A-20)-(A-21) imply (A-18)-(A-19). Below, we establish the existence and uniqueness of $(\breve{e}_h, \breve{e}_s)$. Plugging $\breve{e}_s = \breve{z}_s(\breve{e}_h)$ into (A-20), we can characterize the hospital's equilibrium investment $\breve{e}_h$ by $\breve{\Psi}(\breve{e}_h) = 0$, where

$$\breve{\Psi}(e^h) = \frac{\partial R(e^h, \breve{z}_s(e^h))}{\partial e^h}\xi^h + 1. \tag{A-22}$$

Taking the partial derivative with respect to e^h, we obtain

$$\frac{d\breve{\Psi}(e^h)}{de^h} = \xi^h\left(\frac{\partial^2 R(e^h, \breve{z}_s(e^h))}{\partial (e^h)^2} + \frac{\partial^2 R(e^h, \breve{z}_s(e^h))}{\partial e^s \partial e^h}\frac{d\breve{z}_s(e^h)}{de^h}\right), \tag{A-23}$$

where

$$\frac{d\breve{z}_s(e^h)}{de^h} = -\frac{\partial^2 R(e^h, \breve{z}_s(e^h))/\partial e^h \partial e^s}{\partial^2 R(e^h, \breve{z}_s(e^h))/\partial (e^s)^2} \tag{A-24}$$

by taking the derivative of (A-21) with respect to e^h_{ij}. Plugging (A-24) into (A-23), we have

$$\frac{d\breve{\Psi}(e^h)}{de^h} = \xi^h \left(\frac{\partial^2 R(e^h, \breve{z}_s(e^h))}{\partial (e^h)^2} - \frac{\left(\partial^2 R(e^h, \breve{z}_s(e^h))/\partial e^h \partial e^s\right)^2}{\partial^2 R(e^h, \breve{z}_s(e^h))/\partial (e^s)^2} \right) > 0, \tag{A-25}$$

where the inequality follows from Assumption A-1(iii). We also have $\lim_{e^h \downarrow 0} \breve{\Psi}(e^h) < 0$ and $\lim_{e^h \uparrow \Gamma} \breve{\Psi}(e^h) > 0$ by Assumption A-1(iii) and (A-12). Therefore, there exists a unique $\breve{e}_h = \{e^h \in (0,\Gamma) | \breve{\Psi}(e^h) = 0\}$. This and $\breve{z}_s(e^h) \in (0,\Gamma)$ imply that there exists a unique $\breve{e}_s = \breve{z}_s(\breve{e}_h) \in (0,\Gamma)$.

(iii) We first prove that $(\breve{e}_h, \breve{e}_s) \neq (e^*_h, e^*_s)$ by contradiction. Suppose not, i.e., $\breve{e}_h = e^*_h$ and $\breve{e}_s = e^*_s$, then by (9),

$$\frac{\partial R(\breve{e}_h, \breve{e}_s)}{\partial e^h}(\xi^h + \xi^s) + 1 = 0,$$

which contradicts (A-18) due to $\xi^s > 0$. Thus we have $(\breve{e}_h, \breve{e}_s) \neq (e^*_h, e^*_s)$.

Finally, we prove that $\Phi(e^h, e^s)$ is a jointly strictly convex function.

$$\frac{\partial^2 \Phi}{\partial (e^h)^2} = \frac{\partial^2 R(e^h, e^s)}{\partial (e^h)^2}(\xi^h + \xi^s),\ \frac{\partial^2 \Phi}{\partial (e^s)^2} = \frac{\partial^2 R(e^h, e^s)}{\partial (e^s)^2}(\xi^h + \xi^s),\ \frac{\partial^2 \Phi}{\partial e^h \partial e^s} = \frac{\partial^2 R(e^h, e^s)}{\partial e^h \partial e^s}(\xi^h + \xi^s).$$

The strict convexity of Φ follows from strict convexity of $R(e^h, e^s)$ by Assumption A-1(iii). □

C. Proofs: Equilibrium under CJR-type payment models

We continue to adopt Assumption A-1 with (A-1)-(A-2) strengthened respectively into

$$\lim_{e^h \downarrow 0} \frac{\partial R(e^h, e^s)}{\partial e^h} < -\frac{1}{\xi^h + \xi^s} \text{ and } \lim_{e^h \uparrow \Gamma} \frac{\partial R(e^h, e^s)}{\partial e^h} > -\frac{1}{2\xi^h + \xi^s} \text{ for all } e^s \in [0, \Gamma], \tag{A-26}$$

$$\lim_{e_s \downarrow 0} \frac{\partial R(e^h, e^s)}{\partial e^s} < -\frac{1}{\xi^s} \text{ and } \lim_{e^s \uparrow \Gamma} \frac{\partial R(e^h, e^s)}{\partial e^s} > -\frac{1}{\xi^h + \xi^s} \text{ for all } e^h \in [0, \Gamma], \tag{A-27}$$

which ensures that PAC providers have unique optimal actions that can be determined using FOCs.

Proof of Proposition 2: Since the CJR-type payment model (15)-(17) is a special case of the modified CJR-type payment model (22)-(23) at $\theta = 0$, the equilibrium actions presented in Proposition 2 are equal to the equilibrium actions presented in Proposition 3 at $\theta = 0$. Moreover, (19) can be proven in the same way as that for (16) and hence is omitted.

Proof of Proposition 3: By (1), (2), and (22), hospital i's objective is

$$\begin{aligned}\Pi^h_i(\mathbf{h}_i) =& p_i \left[\bar{C}^h_i + \bar{a}^h_i - C^h(a^h_i) - a^h_i \right] \\ &+ \sum_{j \in \mathcal{M}} p_{ij} \left[\bar{C}^{sh}_i - C_s(b^h_{ij}, b^s_{ij}) + \left(\bar{R}^h_i - R(e^h_{ij}, e^s_{ij}) \right) \left((2-\theta)\xi^h + \xi^s \right) + \bar{b}^h_i - b^h_{ij} + \bar{e}^h_i - e^h_{ij} \right].\end{aligned} \tag{A-28}$$

By (3), (4), and (23), PAC provider j's objective is

$$\Pi_j^s(\mathbf{s}_j) = \sum_{i\in\mathcal{N}} p_{ij}\left[\bar{C}_j^s - C^s(b_{ij}^h, b_{ij}^s) + (\bar{R}_j^s - R(e_{ij}^h, e_{ij}^s))(\theta\xi^h + \xi^s) + \bar{b}_j^s - b_{ij}^s + \bar{e}_j^s - e_{ij}^s\right]. \tag{A-29}$$

It is straightforward to verify that in any equilibrium, we have $a_i^h = a_h^*$ for each hospital $i \in \mathcal{N}$, and for each PAC provider $j \in \mathcal{M}$ such that $p_{ij} > 0$, we have $b_{ij}^h = b_h^*$ and $b_{ij}^s = b_s^*$; the proof is identical to that in Theorem 1. Next, we analyze hospitals' and PAC providers' investments to reduce the readmission probability in equilibrium. When $p_{ij} = 0$, Π_i^h and Π_j^s are independent of both e_{ij}^h and e_{ij}^s. WLOG we assume that first-best actions are taken (see the last paragraph of Section 3 for details), i.e., $e_{ij}^h = e_h^*$ and $e_{ij}^s = e_s^*$, where e_h^* and e_s^* are given by (9)-(10). When $p_{ij} > 0$, we take partial derivatives of Π_i^h and Π_j^s with respect to e_{ij}^h and e_{ij}^s, respectively, and obtain

$$\frac{\partial \Pi_i^h}{\partial e_{ij}^h} = -p_{ij}\left[\frac{\partial R(e_{ij}^h, e_{ij}^s)}{\partial e^h}((2-\theta)\xi^h + \xi^s) + 1\right], \tag{A-30}$$

$$\frac{\partial \Pi_j^s}{\partial e_{ij}^s} = -p_{ij}\left[\frac{\partial R(e_{ij}^h, e_{ij}^s)}{\partial e^s}(\theta\xi^h + \xi^s) + 1\right]. \tag{A-31}$$

We proceed as follows: (i) We show that hospital i and PAC provider j have unique best responses characterized by $e_{ij}^h = z_h(e_{ij}^s)$ and $e_{ij}^s = z_s(e_{ij}^h)$. (ii) We establish the existence of a unique equilibrium $(\tilde{e}_h, \tilde{e}_s)$ for any given $\theta \in [0,1]$, defined by

$$\frac{\partial R(\tilde{e}_h, \tilde{e}_s)}{\partial e^h}((2-\theta)\xi^h + \xi^s) + 1 = 0, \tag{A-32}$$

$$\frac{\partial R(\tilde{e}_h, \tilde{e}_s)}{\partial e^s}(\theta\xi^h + \xi^s) + 1 = 0. \tag{A-33}$$

Finally, we prove that (iii) $\tilde{e}_h$ and $\tilde{e}_s$ are continuous in $\theta \in [0,1]$, and (iv) $d\tilde{e}_h/d\theta < 0$ and $d\tilde{e}_s/d\theta > 0$ for all $\theta \in [0,1]$ if $\partial^2 R(e^h, e^s)/\partial e^h \partial e^s \geqslant 0$. For notational simplicity, we drop argument θ from $\tilde{e}_h$ and $\tilde{e}_s$.

(i) At any fixed $e_{ij}^s \in [0, \Gamma]$, by (A-30) and Assumption A-1(iii), we have $\partial^2 \Pi_i^h/\partial (e_{ij}^h)^2 < 0$, $\lim_{e^h \uparrow \Gamma} \partial \Pi_i^h/\partial e_{ij}^h < 0$, and $\lim_{e_{ij}^h \downarrow 0} \partial \Pi_i^h/\partial e_{ij}^h > 0$ for any $\theta \in (\underline{\theta}_h, \bar{\theta}_h)$, where

$$\underline{\theta}_h = \sup_{e_{ij}^s \in [0,\Gamma]} \left\{2 + \frac{1}{\xi^h}\left[\xi^s + \frac{1}{\lim_{e_{ij}^h \uparrow \Gamma} \partial R(e_{ij}^h, e_{ij}^s)/\partial e^h}\right]\right\} < 0,$$

$$\bar{\theta}_h = \inf_{e_{ij}^s \in [0,\Gamma]} \left\{2 + \frac{1}{\xi^h}\left[\xi^s + \frac{1}{\lim_{e_{ij}^h \downarrow 0} \partial R(e_{ij}^h, e_{ij}^s)/\partial e^h}\right]\right\} > 1$$

by (A-26). Thus, the hospital has a unique best response $z_h(e_{ij}^s) \in (0, \Gamma)$ and is determined by the FOC of Π_i^h, i.e.,

$$\frac{\partial R(z_h(e_{ij}^s), e_{ij}^s)}{\partial e^h}((2-\theta)\xi^h + \xi^s) + 1 = 0. \tag{A-34}$$

At any fixed $e_{ij}^h \in [0,\Gamma]$, by (A-27), (A-31), and Assumption A-1(iii), we have $\partial^2\Pi_j^s/\partial(e_{ij}^s)^2 < 0$, $\lim_{e^s\uparrow\Gamma}\partial\Pi_j^s/\partial e_{ij}^s < 0$, and $\lim_{e^s\downarrow 0}\partial\Pi_j^s/\partial e_{ij}^s > 0$ for any $\theta \in (\underline{\theta}_s, \bar{\theta}_s)$, where

$$\underline{\theta}_s = \sup_{e_{ij}^h\in[0,\Gamma]}\left\{-\frac{1}{\xi^h}\left[\xi^s + \frac{1}{\lim_{e_{ij}^s\downarrow 0}\partial R(e_{ij}^h, e_{ij}^s)/\partial e^s}\right]\right\} < 0,$$
$$\bar{\theta}_s = \inf_{e_{ij}^h\in[0,\Gamma]}\left\{-\frac{1}{\xi^h}\left[\xi^s + \frac{1}{\lim_{e_{ij}^s\uparrow\Gamma}\partial R(e_{ij}^h, e_{ij}^s)/\partial e^s}\right]\right\} > 1$$

by (A-27). Thus, the PAC provider has a unique best response $z_s(e_{ij}^h) \in (0,\Gamma)$ and is determined by the FOC of Π_j^s, i.e.,

$$\frac{\partial R(e_{ij}^h, z_s(e_{ij}^h))}{\partial e^s}(\theta\xi^h + \xi^s) + 1 = 0. \tag{A-35}$$

(ii) Consider any given $\theta \in \Theta = (\underline{\theta}_h, \bar{\theta}_h) \cap (\underline{\theta}_s, \bar{\theta}_s) \supset [0,1]$ due to $\underline{\theta}_h, \underline{\theta}_s < 0$ and $\bar{\theta}_h, \bar{\theta}_s > 0$. Plugging $e_{ij}^s = z_s(e_{ij}^h)$ into (A-34), we can characterize the hospital's equilibrium readmission-reduction investment $\tilde{e}_h$ by $\Psi(\tilde{e}_h) = 0$, where

$$\Psi(e^h) = \frac{\partial R(e^h, z_s(e^h))}{\partial e^h}((2-\theta)\xi^h + \xi^s) + 1. \tag{A-36}$$

Taking the partial derivative with respect to e^h, we obtain

$$\frac{d\Psi(e^h)}{de^h} = ((2-\theta)\xi^h + \xi^s)\left(\frac{\partial^2 R(e^h, z_s(e^h))}{\partial(e^h)^2} + \frac{\partial^2 R(e^h, z_s(e^h))}{\partial e^s \partial e^h}\frac{dz_s(e^h)}{de^h}\right), \tag{A-37}$$

where

$$\frac{dz_s(e^h)}{de^h} = -\frac{\partial^2 R(e^h, z_s(e^h))/\partial e^h \partial e^s}{\partial^2 R(e^h, z_s(e^h))/\partial(e^s)^2} \tag{A-38}$$

by taking the derivative of (A-35) with respect to e_{ij}^h. Plugging (A-38) into (A-37), we have

$$\frac{d\Psi(e^h)}{de^h} = ((2-\theta)\xi^h + \xi^s)\left(\frac{\partial^2 R(e^h, z_s(e^h))}{\partial(e^h)^2} - \frac{(\partial^2 R(e^h, z_s(e^h))/\partial e^h \partial e^s)^2}{\partial^2 R(e^h, z_s(e^h))/\partial(e^s)^2}\right) > 0, \tag{A-39}$$

where the inequality follows from Assumption A-1(iii). We also have $\lim_{e^h\downarrow 0}\Psi(e^h) < 0$ and $\lim_{e^h\uparrow\Gamma}\Psi(e^h) > 0$ by $\theta \in (\underline{\theta}_h, \bar{\theta}_h)$. Therefore, there exists a unique $\tilde{e}_h = \{e^h \in (0,\Gamma) | \Psi(e_h) = 0\}$. This and $z_s(e_h) \in (0,\Gamma)$ imply that there exists a unique $\tilde{e}_s = z_s(\tilde{e}_h) \in (0,\Gamma)$. We thus have proven that, for any $\theta \in \Theta$, there exists a unique equilibrium in which each hospital chooses $e_h = \tilde{e}_h$ and each PAC provider chooses $e_s = \tilde{e}_s$, where $\tilde{e}_h$ and $\tilde{e}_s$ are defined in (A-32)-(A-33).

(iii) Now we show that $\tilde{e}_h$ and $\tilde{e}_s$ are continuous in $\theta \in \Theta$, which implies continuity of $\tilde{e}_h$ and $\tilde{e}_s$ in $\theta \in [0,1] \subset \Theta$. For ease of exposition, we will make explicit the dependence of z_s and Ψ on θ; see (A-35)-(A-36). Since $R(e^h, e^s)$ is twice differentiable and $\partial^2 R(e^h, e^s)/\partial(e^s)^2 > 0$, by the Implicit Function Theorem, $z_s(e^h, \theta)$ defined as in (A-35) is continuous in $e^h \in (0,\Gamma)$ and $\theta \in \Theta$, so as $\Psi(e^h, \theta)$

defined as in (A-36). Since for any $\theta \in \Theta$, a unique $\tilde{e}_h \in (0, \Gamma)$ exists and satisfies $\Psi(\tilde{e}_h, \theta) = 0$, by the Implicit Function Theorem and noting $\partial\Psi(\tilde{e}_h, \theta)/\partial e^h > 0$ by (A-39), $\tilde{e}_h$ is continuous in $\theta \in \Theta$. This and continuity of $z_s(e^h, \theta)$ imply that $\tilde{e}_s = z_s(\tilde{e}_h, \theta)$ is continuous in $\theta \in \Theta$.

(iv) By continuity of $\tilde{e}_h$ and $\tilde{e}_s$ in $\theta \in [0,1]$, to prove $\tilde{e}^h > e_h^* = \lim_{\theta \uparrow 1} \tilde{e}_h$ and $\tilde{e}^s < e_s^* = \lim_{\theta \uparrow 1} \tilde{e}_s$, it suffices to prove $d\tilde{e}_h/d\theta < 0$ and $d\tilde{e}_s/d\theta > 0$ for all $\theta \in [0,1]$. Taking the derivative of $\Psi(\tilde{e}_h(\theta), \theta) = 0$ with respect to θ, we obtain

$$\frac{d\tilde{e}_h}{d\theta} = -\frac{\partial\Psi(\tilde{e}_h, \theta)/\partial\theta}{\partial\Psi(\tilde{e}_h, \theta)/\partial e^h} = -\frac{((2-\theta)\xi^h + \xi^s)\frac{\partial^2 R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h \partial e^s}\frac{\partial z_s(\tilde{e}_h, \theta)}{\partial \theta} - \xi^h \frac{\partial R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h}}{((2-\theta)\xi^h + \xi^s)\left[\frac{\partial^2 R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial (e^h)^2} + \frac{\partial^2 R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h \partial e^s}\frac{\partial z_s(\tilde{e}_h, \theta)}{\partial e^h}\right]}. \tag{A-40}$$

By (A-38) and Assumption A-1(iii), the denominator on the right-hand side of (A-40) is positive for all $\theta \in [0,1]$, thus

$$\operatorname{sgn}\left(\frac{d\tilde{e}_h}{d\theta}\right) = -\operatorname{sgn}\left(((2-\theta)\xi^h + \xi^s)\frac{\partial^2 R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h \partial e^s}\frac{\partial z_s(\tilde{e}_h, \theta)}{\partial \theta} - \xi^h \frac{\partial R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h}\right). \tag{A-41}$$

Taking the derivative of (A-35) with respect to θ, we have

$$\frac{\partial z_s(e^h, \theta)}{\partial \theta} = -\frac{\xi^h \frac{\partial R(e^h, z_s(e^h, \theta))}{\partial e^s}}{(\theta\xi^h + \xi^s)\frac{\partial^2 R(e^h, z_s(e^h, \theta))}{\partial (e^s)^2}}. \tag{A-42}$$

Plugging it into (A-41), we obtain

$$\begin{aligned}
\operatorname{sgn}\left(\frac{d\tilde{e}_h}{d\theta}\right) &= \operatorname{sgn}\left(\frac{(2-\theta)\xi^h + \xi^s}{\theta\xi^h + \xi^s}\frac{\partial^2 R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h \partial e^s}\frac{\frac{\partial R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^s}}{\frac{\partial^2 R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial (e^s)^2}} + \frac{\partial R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h}\right) \\
&= \operatorname{sgn}\left(\frac{(2-\theta)\xi^h + \xi^s}{\theta\xi^h + \xi^s}\frac{\partial^2 R(\tilde{e}_h, \tilde{e}_s)}{\partial e^h \partial e^s}\frac{\frac{\partial R(\tilde{e}_h, \tilde{e}_s)}{\partial e^s}}{\frac{\partial^2 R(\tilde{e}_h, \tilde{e}_s)}{\partial (e^s)^2}} + \frac{\partial R(\tilde{e}_h, \tilde{e}_s)}{\partial e^h}\right) \\
&= \operatorname{sgn}\left(\frac{(2-\theta)\xi^h + \xi^s}{\theta\xi^h + \xi^s}\frac{\partial^2 R(\tilde{e}_h, \tilde{e}_s)}{\partial e^h \partial e^s}\frac{\partial R(\tilde{e}_h, \tilde{e}_s)}{\partial e^s} + \frac{\partial R(\tilde{e}_h, \tilde{e}_s)}{\partial e^h}\frac{\partial^2 R(\tilde{e}_h, \tilde{e}_s)}{\partial (e^s)^2}\right) < 0,
\end{aligned} \tag{A-43}$$

where the second equality follows from $\tilde{e}_s = z_s(\tilde{e}_h, \theta)$ for any θ, the third equality follows from $\partial^2 R(e^h, e^s)/\partial (e^s)^2 > 0$ by Assumption A-1(iii), and the inequality follows from $\partial^2 R(e^h, e^s)/(\partial e^h \partial e^s) \geqslant 0$, $\partial R(e^h, e^s)/\partial e^h < 0$, $\partial R(e^h, e^s)/\partial e^s < 0$, and $\partial^2 r(e_h, e_s)/\partial e_s^2 > 0$ by Assumption A-1(iii).

Taking the derivative of $\tilde{e}_s$ with respect to θ, we obtain

$$\frac{d\tilde{e}_s}{d\theta} = \frac{dz_s(\tilde{e}_h, \theta)}{d\theta} = \frac{\partial z_s(\tilde{e}_h, \theta)}{\partial e^h}\frac{d\tilde{e}_h}{d\theta} + \frac{\partial z_s(\tilde{e}_h, \theta)}{\partial \theta} = -\frac{\frac{\partial^2 R(\tilde{e}_h, \tilde{z}_s)}{\partial e^h \partial e^s}}{\frac{\partial^2 R(\tilde{e}_h, \tilde{z}_s)}{\partial (e^s)^2}}\frac{d\tilde{e}_h}{d\theta} - \frac{\xi^h \frac{\partial R(\tilde{e}_h, \tilde{z}_s)}{\partial e^s}}{(\theta\xi^h + \xi^s)\frac{\partial^2 R(\tilde{e}_h, \tilde{z}_s)}{\partial (e^s)^2}} > 0, \tag{A-44}$$

where the second equality follows from differentiation by parts, the third follows from (A-38) and (A-42), and the inequality follows from $d\tilde{e}_h/d\theta < 0$ by (A-43), $\partial^2 R(e^h, e^s)/(\partial e^h \partial e^s) \geqslant 0$, $\partial R(e^h, e^s)/\partial e^h < 0$, $\partial R(e^h, e^s)/\partial e^s < 0$, and $\partial^2 R(e^h, e^s)/\partial (e^s)^2 > 0$ by Assumption A-1(iii). $\square$

D. Proof of Theorem 1

The proof is based on the observation that under the proposed payment scheme, the difference between a hospital's objective and the regulator's objective is independent of that hospital's actions, and the difference between a PAC provider's objective and the regulator's objective is independent of that PAC provider's actions. More precisely, given the actions of all other hospitals and PAC providers, by (1)-(4), (20), and (21), hospital i's objective is

$$\begin{aligned}\Pi_i^h(\mathbf{h}_i) =& p_i\left[\bar{C}_i^h + \bar{a}_i^h - C^h(a_i^h) - a_i^h\right] \\ &+ \sum_{j\in\mathcal{M}} p_{ij}\left[\bar{C}_i^{sh} - C^s(b_{ij}^h, b_{ij}^s) + (\bar{R}_i^h - R(e_{ij}^h, e_{ij}^s))(\xi^h + \xi^s) + \bar{b}_i^h - b_{ij}^h + \bar{e}_i^h - e_{ij}^h\right], \end{aligned} \tag{A-45}$$

and PAC provider j's objective is

$$\Pi_j^s(\mathbf{s}_j) = \sum_{i\in\mathcal{N}} p_{ij}\left[\bar{C}_j^s - C^s(b_{ij}^h, b_{ij}^s) + (\bar{R}_j^s - R(e_{ij}^h, e_{ij}^s))(\xi^h + \xi^s) + \bar{b}_j^s - b_{ij}^s + \bar{e}_j^s - e_{ij}^s\right]. \tag{A-46}$$

By (A-3), letting $\vec{\mathbf{h}} = \{\mathbf{h}_i, i \in \mathcal{N}\}$ and $\vec{\mathbf{s}} = \{\mathbf{s}_j, j \in \mathcal{M}\}$ denote the actions of all hospitals and all PAC providers, respectively, we have

$$\begin{aligned}\Delta_h = \Pi_i^h(\mathbf{h}_i) - W(\vec{\mathbf{h}}, \vec{\mathbf{v}}) =& p_i\left[\bar{C}_i^h + \bar{a}_i^h\right] + \sum_{j\in\mathcal{M}} p_{ij}\left[\bar{C}_i^{sh} + \bar{R}_i^h(\xi^h + \xi^s) + \bar{b}_i^h + \bar{e}_i^h + b_{ij}^s + e_{ij}^s\right] - v \\ &+ \sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}} p_{kj}\left[C^s(b_{kj}^h, b_{kj}^s) + R(e_{kj}^h, e_{kj}^s)\xi^s + b_{kj}^s + e_{kj}^s\right] + \sum_{k\in\mathcal{N}_i} \mathcal{C}_k^h(\mathbf{h}_k)\end{aligned}$$

does not depend on $\mathbf{h}_i$, and

$$\begin{aligned}\Delta_s = \Pi_j^s(\mathbf{s}_j) - W(\vec{\mathbf{h}}, \vec{\mathbf{v}}) =& \sum_{i\in\mathcal{N}}\Big\{p_{ij}\left[\bar{C}_j^s + \bar{R}_j^s(\xi^h + \xi^s) + \bar{b}_j^s + \bar{e}_j^s + b_{ij}^h + e_{ij}^h\right] + p_i(C^h(a_i^h) + a_i^h)\Big\} - v \\ &+ \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} p_{ik}\left[C^s(b_{ik}^h, b_{ik}^s) + R(e_{ik}^h, e_{ik}^s)(\xi^h + \xi^s) + b_{ik}^h + b_{ik}^s + e_{ik}^h + e_{ik}^s\right]\end{aligned}$$

does not depend on $\mathbf{s}_j$. It then follows that hospital i's problem

$$\underset{\mathbf{h}_i\in[0,\Gamma]^{2M+1}}{\text{maximize}}\ \Pi_i^h(\mathbf{h}_i)$$

is equivalent to

$$\underset{\mathbf{h}_i\in[0,\Gamma]^{2M+1}}{\text{maximize}}\ W(\mathbf{h}_i|\mathbf{h}_k, \mathbf{s}_j, k\in\mathcal{N}_i, j\in\mathcal{M}) \tag{A-47}$$

because the two objectives differ by Δ_h which does not depend on the hospital's decisions $\mathbf{h}_i$. Similarly, PAC provider j's problem, i.e.,

$$\underset{\mathbf{s}_j\in[0,\Gamma]^{2N}}{\text{maximize}}\ \Pi_j^s(\mathbf{s}_j)$$

is equivalent to

$$\underset{\mathbf{s}_j \in [0,\Gamma]^{2N}}{\text{maximize}}\ W(\mathbf{s}_j | \mathbf{h}_i, \mathbf{s}_k, i \in \mathcal{N}, k \in \mathcal{M}_j). \tag{A-48}$$

because the two objectives differ by Δ_s which does not depend on the PAC provider's decisions $\mathbf{s}_j$. By Lemma A-1, the regulator's problem, i.e.,

$$\underset{\mathbf{h}_i \in [0,\Gamma]^{2M+1}, \mathbf{s}_j \in [0,\Gamma]^{2N}, i \in \mathcal{N}, j \in \mathcal{M}}{\text{maximize}} W(\mathbf{h}_i, \mathbf{s}_j, i \in \mathcal{N}, j \in \mathcal{M}), \tag{A-49}$$

has a unique maximizer given by $\mathbf{h}_i = \mathbf{h}^*$ for each hospital $i \in \mathcal{N}$ and $\mathbf{s}_j = \mathbf{s}^*$ for each PAC provider $j \in \mathcal{M}$, where

$$\mathbf{h}^* = (a^*, \underbrace{b_h^*, \ldots, b_h^*}_{M \text{ times}}, \underbrace{e_h^*, \ldots, e_h^*}_{M \text{ times}}) \text{ and } \mathbf{s}^* = (\underbrace{b_s^*, \ldots, b_s^*}_{N \text{ times}}, \underbrace{e_s^*, \ldots, e_s^*}_{N \text{ times}})$$

are the first-best actions for each hospital and each PAC provider, respectively. It then follows that for each hospital $i \in \mathcal{N}$,

$$\mathbf{h}^* = \underset{\mathbf{h}_i \in [0,\Gamma]^{2M+1}}{\arg\max}\ W(\mathbf{h}_i | \mathbf{h}_k = \mathbf{h}^*, \mathbf{s}_j = \mathbf{s}^*, k \in \mathcal{N}_i, j \in \mathcal{M}). \tag{A-50}$$

Suppose not, then there exists $\mathbf{h}' \in [0,\Gamma]^{2M+1}$ such that

$$W(\mathbf{h}' | \mathbf{h}_k = \mathbf{h}^*, \mathbf{s}_j = \mathbf{s}^*, k \in \mathcal{N}_i, j \in \mathcal{M}) \geqslant W(\mathbf{h}^* | \mathbf{h}_k = \mathbf{h}^*, \mathbf{s}_j = \mathbf{s}^*, k \in \mathcal{N}_i, j \in \mathcal{M}),$$

or equivalently

$$W(\mathbf{h}_i = \mathbf{h}', \mathbf{h}_k = \mathbf{h}^*, \mathbf{s}_j = \mathbf{s}^*, k \in \mathcal{N}_i, j \in \mathcal{M}) \geqslant W(\mathbf{h}_i = \mathbf{h}^*, \mathbf{s}_j = \mathbf{s}^*, i \in \mathcal{N}, j \in \mathcal{M}).$$

This contradicts Lemma A-1 proving that $W(\mathbf{h}_i, \mathbf{s}_j, i \in \mathcal{N}, j \in \mathcal{M})$ has a unique maximizer given by $\mathbf{h}_i = \mathbf{h}^*$ for each hospital $i \in \mathcal{N}$ and $\mathbf{s}_j = \mathbf{s}^*$ for each PAC provider $j \in \mathcal{M}$. Similarly, for each PAC provider $j \in \mathcal{M}$, we have

$$\mathbf{s}^* = \underset{\mathbf{s}_j \in [0,\Gamma]^{2N}}{\arg\max}\ W(\mathbf{s}_j | \mathbf{h}_i = \mathbf{h}^*, \mathbf{s}_k = \mathbf{s}^*, i \in \mathcal{N}, k \in \mathcal{M}_j), \tag{A-51}$$

i.e., each PAC provider j's problem given by (A-48) has a unique solution $\mathbf{s}_j = \mathbf{s}^*$, when other hospitals and PAC providers choose first-best actions. By (A-50), each hospital i's problem given by (A-47) has a unique solution $\mathbf{h}_i = \mathbf{h}^*$, when other hospitals and PAC providers choose first-best actions. Thus, no hospital or PAC provider can profitably deviate from the first-best action profile, i.e., $\mathbf{h}_i = \mathbf{h}^*$ for each hospital $i \in \mathcal{N}$ and $\mathbf{s}_j = \mathbf{s}^*$ for each PAC provider $j \in \mathcal{M}$; thus it is an equilibrium. Plugging the first-best actions in (A-45)-(A-46) one can verify that all hospitals and PAC providers break even in this equilibrium.

Now we prove by contradiction that there exist no other equilibria other than the first best. Suppose there exists another equilibrium in which the actions of hospital $i \in \mathcal{N}$ and PAC provider $j \in \mathcal{M}$ are

$$\check{\mathbf{h}}_i = (\check{a}_i^h, \check{b}_{i1}^h, \ldots, \check{b}_{iM}^h, \check{e}_{i1}^h, \ldots, \check{e}_{iM}^h) \text{ and } \check{\mathbf{s}}_j = (\check{b}_{1j}^s, \ldots, \check{b}_{Nj}^s, \check{e}_{1j}^s, \ldots, \check{e}_{Nj}^s).$$

Thus, in this proposed equilibrium, $\mathbf{h}_i = \check{\mathbf{h}}_i$ is a solution for each hospital i's problem (A-47), and $\mathbf{s}_j = \check{\mathbf{s}}_j$ is a solution for each PAC provider j's problem (A-48), i.e.,

$$\check{\mathbf{h}}_i \in \underset{\mathbf{h}_i \in [0,\Gamma]^{2M+1}}{\arg\max}\ W(\mathbf{h}_i | \check{\mathbf{h}}_k, \check{\mathbf{s}}_j, k \in \mathcal{N}_i, j \in \mathcal{M}), \tag{A-52}$$

$$\check{\mathbf{s}}_j \in \underset{\mathbf{s}_j \in [0,\Gamma]^{2N}}{\arg\max}\ W(\mathbf{s}_j | \check{\mathbf{h}}_i, \check{\mathbf{s}}_k, i \in \mathcal{N}, k \in \mathcal{M}_j). \tag{A-53}$$

By (A-4), we have $\check{a}_i^h = a_h^*$ for each hospital $i \in \mathcal{N}$. Below we prove that

$$\check{b}_{ij}^h = b_h^* \text{ and } \check{b}_{ij}^s = b_s^*, \text{ for each } i \in \mathcal{N} \text{ and } j \in \mathcal{M}. \tag{A-54}$$

Since (A-54) is assumed to hold when $p_{ij} = 0$ WLOG (see the last paragraph of Section 3 for details), it suffices to consider the case of $p_{ij} > 0$. By (A-52)-(A-53), we have

$$\check{b}_{ij}^h \in \underset{b_{ij}^h \in [0,\Gamma]}{\arg\max}\, W(b_{ij}^h | \check{a}_i^h, \check{e}_{ij}^h, \check{\mathbf{h}}_k, \check{\mathbf{s}}_j, k \in \mathcal{N}_i, j \in \mathcal{M}) = \underset{b_{ij}^h \in [0,\Gamma]}{\arg\min}\, p_{ij} \left[C^s(b_{ij}^h, \check{b}_{ij}^s) + b_{ij}^h + \check{b}_{ij}^s \right], \tag{A-55}$$

$$\check{b}_{ij}^s \in \underset{b_{ij}^s \in [0,\Gamma]}{\arg\max}\, W(b_{ij}^s | \check{e}_{ij}^s, \check{\mathbf{h}}_i, \check{\mathbf{s}}_k, i \in \mathcal{N}, k \in \mathcal{M}_j) = \underset{b_{ij}^s \in [0,\Gamma]}{\arg\min}\, p_{ij} \left[C^s(\check{b}_{ij}^h, b_{ij}^s) + \check{b}_{ij}^h + b_{ij}^s \right], \tag{A-56}$$

where the equalities follow by plugging in the expression of W from (A-3). In the proof of Lemma A-1, we have solved for the optimization problem in (A-56) and obtained a unique best response $\check{b}_{ij}^s = g(\check{b}_{ij}^h) \in (0, \Gamma)$, where $g(\cdot)$ is given by (A-5). Now we solve the optimization problem in (A-55). For any fixed $\check{b}_{ij}^s \in [0, \Gamma]$, we have

$$\begin{aligned}
&\frac{\partial^2 \left[C^s(b_{ij}^h, \check{b}_{ij}^s) + b_{ij}^h + \check{b}_{ij}^s \right]}{\partial (b_{ij}^h)^2} = -\frac{\partial^2 C^s(b_{ij}^h, \check{b}_{ij}^s)}{\partial (b^h)^2} < 0, \\
&\lim_{b_{ij}^h \downarrow 0} \frac{\partial \left[C^s(b_{ij}^h, \check{b}_{ij}^s) + b_{ij}^h + \check{b}_{ij}^s \right]}{\partial b_{ij}^h} = \lim_{b_{ij}^h \downarrow 0} \left[\frac{\partial C^s(b_{ij}^h, \check{b}_{ij}^s)}{\partial b^h} + 1 \right] > 0, \\
&\lim_{b_{ij}^h \uparrow \Gamma} \frac{\partial \left[C^s(b_{ij}^h, \check{b}_{ij}^s) + b_{ij}^h + \check{b}_{ij}^s \right]}{\partial b_{ij}^h} = \lim_{b_{ij}^h \uparrow \Gamma} \left[\frac{\partial C^s(b_{ij}^h, \check{b}_{ij}^s)}{\partial b^h} + 1 \right] < 0,
\end{aligned}$$

where all inequalities follow from Assumption A-1(ii). Thus, the optimization problem in (A-55) has a unique solution $\check{b}_{ij}^h$ and is determined by the FOC, i.e., $\partial C^s(\check{b}_{ij}^h, \check{b}_{ij}^s)/\partial b^h + 1 = 0$. Plugging in $\check{b}_{ij}^s = g(\check{b}_{ij}^h)$, we obtain

$$\frac{\partial C^s(\check{b}_{ij}^h, g(\check{b}_{ij}^h))}{\partial b^h} + 1 = 0. \tag{A-57}$$

In the proof of Lemma A-1, we have proven that (A-57) has a unique solution given by b_h^*; see (A-8). Thus, for all hospital $i \in \mathcal{N}$ and PAC provider $j \in \mathcal{M}$, we have $\check{b}_{ij}^h = b_h^*$ and $\check{b}_{ij}^s = g(\check{b}_{ij}^h) = g(b_h^*) = b_s^*$, where the last equality follows from (A-9). Following this procedure, one can verify that $\check{e}_{ij}^h = e_h^*$ and $\check{e}_{ij}^s = e_s^*$ for each hospital $i \in \mathcal{N}$ and PAC provider $j \in \mathcal{M}$. It then follows that $\check{\mathbf{h}}_i = \mathbf{h}^*$ and $\check{\mathbf{s}}_j = \mathbf{s}^*$ for each hospital $i \in \mathcal{N}$ and PAC provider $j \in \mathcal{M}$, contradicting our assumption that $(\check{\mathbf{h}}_i, \check{\mathbf{s}}_j, i \in \mathcal{N}, j \in \mathcal{M})$ is different from first best. □

E. Endogenous discharge decisions

In our original model presented in Section 3, we assumed that a certain percentage of patients are discharged to PAC, and that this proportion is fixed. However, there is no consensus on the optimal PAC setting for patients being discharged from the hospital, as highlighted by Li et al. (2020). Generally, there are two options to consider:

- PAC institutions: These facilities, such as skilled nursing facilities (SNFs), inpatient rehabilitation centers, or long-term hospital care, typically offer more intensive care, potentially reducing unnecessary readmissions. However, they also come with higher costs.
- Home: Patients can also receive PAC through visits from in-home healthcare providers. This option is typically less costly than PAC in an institution, but it may not offer the same level of care (Werner et al. 2019).

Hospitals need to optimize their discharge decisions, taking into account this trade-off among different PAC settings, while also making investments to coordinate care with all types of PAC providers. In this section, we extend our model to examine this additional decision and demonstrate that our payment model can be applied (using the same underlying principles outlined in Section 5) to incentivize hospitals to make socially optimal decisions when a patient can be discharged to these different settings.

Model: To incorporate hospitals' decisions regarding patients' discharge destinations, we introduce the assumption that there are L different types of PAC settings, including home. Each hospital i determines the proportion, $\rho_i^\ell \in [0,1]$, of their patients discharged to PAC providers of type ℓ, where $\ell \in \mathcal{L}$. We let $\vec{\rho}_i = \left\{\rho_i^\ell, \ell \in \mathcal{L}\right\}$ denote the vector of these proportions for hospital i. We represent the proportion of patients discharged from hospital i to type-ℓ PAC provider j as p_{ij}^ℓ. Therefore, we have $\sum_{\ell \in \mathcal{L}} \rho_i^\ell = 1$ and $\sum_{j \in \mathcal{M}^\ell} p_{ij}^\ell / p_i = \rho_i^\ell$, where $\mathcal{M}^\ell = \{1, ..., M^\ell\}$ and M^ℓ denotes the number of PAC providers of type ℓ.

Additional notation and terminology required in this section are based on the ones introduced in Section 3. We use C^ℓ to denote the cost of PAC for type-ℓ providers, and R^ℓ to represent the readmission probability for patients discharged to a type-s PAC setting. Similar to our previous model, we assume that $C^\ell : [0,\Gamma] \times [0,\Gamma] \to \mathbb{R}^+$ is dependent on the investments made by the

hospital, denoted by $b^{h,\ell}$, and by the PAC provider, denoted by b^{ℓ}. The readmission probability from a type-ℓ PAC setting, $R^{\ell}:[0,\Gamma]\times[0,\Gamma]\times[0,1]\to[0,1]$, depends on the investments of the hospital, denoted by $e^{h,\ell}$, and the PAC provider, denoted by e^{ℓ}, in reducing readmissions, as well as the discharge decisions $\vec{\rho}$. The cost of treating a readmitted patient is denoted by ξ^{h} for hospital care and ξ^{ℓ} for PAC care at a type-ℓ provider.

In contrast to the approach outlined in Section 3, our current assumption takes into account the interdependence between readmission probabilities and the characteristics denoted by $\vec{\rho}$ for each type of PAC providers. This consideration is essential for encompassing the variation in care intensity across diverse PAC settings. Notably, hospitals tend to direct more critically ill patients towards PAC facilities that offer more concentrated and intensive care. Rather than directly modeling these intricate allocation decisions, we leverage the influence of $\vec{\rho}$ to encapsulate the effects of patient distribution on readmission probabilities.

Objective functions: In the current setting, the objective of hospital i is defined as similar to (1)

$$\Pi_i^h(\mathbf{h}_i) = T_i^h - \mathcal{C}_i^h(\mathbf{h}_i), \tag{A-58}$$

where

$$\mathcal{C}_i^h(\mathbf{h}_i) = p_i\left[C^h\left(a_i^h\right)+a_i^h\right]+\sum_{\ell\in\mathcal{L}}\sum_{j\in\mathcal{M}^{\ell}} p_{ij}^{\ell}\left[R^{\ell}\left(e_{ij}^{h,\ell},e_{ij}^{\ell},\vec{\rho}_i\right)\xi^h+b_{ij}^{h,\ell}+e_{ij}^{h,\ell}\right] \tag{A-59}$$

is the total cost of the hospital. We use $\mathbf{h}_i=(a_i^h,b_{ij}^{h,\ell},e_{ij}^{h,\ell},\vec{\rho}_i,\ell\in\mathcal{L},j\in\mathcal{M}^{\ell})$ to denote the actions of hospital i.

Similarly, the objective of type-ℓ PAC provider j is defined similarly to (3)

$$\Pi_j^{\ell}(\mathbf{v}_j) = T_j^{\ell} - \mathcal{C}_j^{\ell}(\mathbf{v}_j), \tag{A-60}$$

where

$$\mathcal{C}_j^{\ell}(\mathbf{v}_j) = \sum_{i\in\mathcal{N}} p_{ij}^{\ell}\left[C^{\ell}(b_{ij}^{h,\ell},b_{ij}^{\ell})+R^{\ell}(e_{ij}^{h,\ell},e_{ij}^{\ell},\vec{\rho}_i)\xi^{\ell}+b_{ij}^{\ell}+e_{ij}^{\ell}\right]. \tag{A-61}$$

is the total cost of for this provider. Here $\mathbf{v}_j=\left(b_{ij}^{\ell},e_{ij}^{\ell},i\in\mathcal{N}\right)$ denotes the actions of PAC provider j. For simplicity, we again assume that all readmitted patients receive PAC from the provider that treated them during their initial visit.

Similar to (5), the total welfare W in this case is then given by:

$$W = \upsilon - \sum_{i\in\mathcal{N}}\mathcal{C}_i^h(\mathbf{h}_i) - \sum_{s\in\mathcal{S}}\sum_{j\in\mathcal{M}^{\ell}}\mathcal{C}_j^{\ell}(\mathbf{v}_j). \tag{A-62}$$

The total welfare consists of: (i) patient utility; (ii) the total cost of hospital care; and (iii) the total cost of all PAC providers.

For any fixed discharge decisions $\vec{\rho}_i$ for all $i \in \mathcal{N}$, we assume that the socially optimal actions are uniquely determined by the FOCs of total welfare, as in the original model; see Appendix E. Moreover, in case of multiple discharge decisions being socially optimal, the regulator chooses one of them using a certain tie-breaking rule. We assume, for simplicity, that hospitals follow the same tie-breaking rule. Under these assumptions, the socially optimal actions for all hospitals, denoted by $(a_h^*, b_{h,\ell}^*, e_{h,\ell}^*, \vec{\rho}^*)$ with a slight abuse of notation, are identical, where $\vec{\rho}^* = \{\rho_\ell^*, \ell \in \mathcal{L}\}$ and it is possible that $\rho_\ell^* = 0$ for some ℓ. Similarly, the socially optimal actions for type-ℓ PAC providers, denoted by (b_ℓ^*, e_ℓ^*), are identical, for each $\ell \in \mathcal{L}$.

Payment scheme: We now present an extension of our proposed payment model and demonstrate that it continues to incentivize hospitals and PAC providers to make socially optimal decisions. Before delving into the technical details, we first explain the underlying concept.

The payment scheme outlined in Section 5 aims to improve care coordination by incentivizing hospitals to reduce both their total care costs and the overall cost of PAC for their patients, which includes costs associated with readmissions. Similarly, the payment scheme encourages PAC providers to consider the cost of hospital care for readmitted patients in their decision-making process. This is achieved by first setting hospital and PAC provider-specific benchmarks for these costs, based on the average costs of other hospitals and PAC providers. The payments of hospitals and PAC providers are then linked to their performance relative to these benchmarks. In the current context, we apply the same concept, but we need to modify how the benchmarks are determined.

We will first present the payment scheme for the PAC providers, as it closely resembles the payment scheme (21) in our original model. The payment for type-ℓ PAC provider j is given by:

$$T_j^\ell = \underbrace{\left[\hat{C}_j^\ell + \hat{b}_j^\ell + \hat{e}_j^\ell + \hat{R}_j^\ell \xi^\ell\right] \sum_{i \in \mathcal{N}} p_{ij}^\ell}_{\text{Cost of care}} + \underbrace{\sum_{i \in \mathcal{N}} p_{ij}^\ell \left[\hat{R}_j^\ell - R^\ell(e_{ij}^{h,\ell}, e_{ij}^\ell, \vec{\rho}_i)\right] \xi^h}_{\text{Outcome-based adjustment}}, \tag{A-63}$$

where

$$\hat{C}_j^\ell = \frac{\sum_{i \in \mathcal{N}} \sum_{k \in \mathcal{M}_j^\ell} p_{ik}^\ell C^\ell\left(b_{ik}^{h,\ell}, b_{ik}^\ell\right)}{\sum_{i \in \mathcal{N}} \sum_{k \in \mathcal{M}_j^\ell} p_{ik}^\ell} \text{ and } \hat{R}_j^\ell = \frac{\sum_{i \in \mathcal{N}} \sum_{k \in \mathcal{M}_j^\ell} p_{ik}^\ell R^\ell(e_{ik}^{h,\ell}, e_{ik}^\ell, \vec{\rho}_i)}{\sum_{i \in \mathcal{N}} \sum_{k \in \mathcal{M}_j^\ell} p_{ik}^\ell}$$

represent the average cost and the readmission likelihood for patients who are treated by all type-ℓ PAC providers, excluding provider j. Additionally,

$$\hat{b}_j^\ell = \frac{\sum_{i \in \mathcal{N}} \sum_{k \in \mathcal{M}_j^\ell} p_{ik}^\ell b_{ik}^\ell}{\sum_{i \in \mathcal{N}} \sum_{k \in \mathcal{M}_j^\ell} p_{ik}^\ell} \text{ and } \hat{e}_j^\ell = \frac{\sum_{i \in \mathcal{N}} \sum_{k \in \mathcal{M}_j^\ell} p_{ik}^\ell e_{ik}^\ell}{\sum_{i \in \mathcal{N}} \sum_{k \in \mathcal{M}_j^\ell} p_{ik}^\ell}$$

are the average investments to reduce costs and readmissions by all type-ℓ PAC providers, excluding provider j. It is worth noting that (A-63) follows the same structure as (21).

The payment scheme for hospitals is slightly different from that in Section 5 because of the additional decision the hospitals need to make regarding the discharge destination of patients. The payment scheme is modified to make hospitals internalize the cost of PAC in general. The hospital payment scheme consists of two main parts: (i) cost of care payments to cover the costs of the hospital, $T^{h,0}$; and (ii) outcome-based payment based on the performances of type-s PAC providers that the hospital discharged patients to, $T^{h,\ell}$ for $\ell \in \mathcal{L}$.

We start with the cost of care component. As in (20), the hospital is compensated for the cost of care as well as the investments to reduce costs and readmissions of the PAC providers that it discharges patients to, as follows

$$\underbrace{T_i^{h,0}}_{\text{Cost of care payment}} = p_i\left[\bar{C}_i^h + \bar{a}_i^h + \hat{R}_i^h \xi^h\right] + \sum_{l\in\mathcal{L}} p_i \bar{\rho}_i^\ell \left(\hat{b}_i^{h,\ell} + \hat{e}_i^{h,\ell}\right), \tag{A-64}$$

where

$$\bar{\rho}_i^\ell = \frac{\sum_{k\in\mathcal{N}_i} p_k \rho_k^\ell}{\sum_{k\in\mathcal{N}_i} p_k} \tag{A-65}$$

represents the average fraction of patients discharged to type-ℓ PAC providers, excluding patients discharged from hospital i. Additionally, $\hat{b}_i^{h,\ell}$ and $\hat{e}_i^{h,\ell}$ represent the average costs incurred by hospitals to improve care (in terms of cost and readmission probability, respectively) in type-ℓ PAC providers, excluding patients discharged from hospital i, defined as follows (similar to $\bar{b}_i^h$ and $\bar{e}_i^h$ in Section 4.1)

$$\hat{b}_i^{h,\ell} = \frac{\sum\limits_{k\in\mathcal{N}_i}\sum\limits_{j\in\mathcal{M}^\ell} p_{kj}^\ell b_{kj}^{h,\ell}}{\sum\limits_{k\in\mathcal{N}_i}\sum\limits_{j\in\mathcal{M}^\ell} p_{kj}^\ell}, \;\; \hat{e}_i^{h,\ell} = \frac{\sum\limits_{k\in\mathcal{N}_i}\sum\limits_{j\in\mathcal{M}^\ell} p_{kj}^\ell e_{kj}^{h,\ell}}{\sum\limits_{k\in\mathcal{N}_i}\sum\limits_{j\in\mathcal{M}^\ell} p_{kj}^\ell}, \ell \in \mathcal{L},$$

and (similar to $\bar{R}_i^h$ in Section 4.1)

$$\hat{R}_i^h = \frac{\sum\limits_{k\in\mathcal{N}_i}\sum\limits_{\ell\in\mathcal{L}}\sum\limits_{j\in\mathcal{M}^\ell} p_{kj}^\ell R^\ell(e_{kj}^{h,\ell}, e_{kj}^\ell, \vec{\rho}_k)}{\sum\limits_{k\in\mathcal{N}_i}\sum\limits_{\ell\in\mathcal{L}}\sum\limits_{j\in\mathcal{M}^\ell} p_{kj}^\ell}$$

is the proportion of readmitted patients, excluding patients discharged from hospital i.

The outcome-based payment component, based on the performance of type-ℓ PAC providers, is determined as follows:

$$\underbrace{T_i^{h,\ell}}_{\text{PAC cost component}} = p_i\bar{\rho}_i^\ell\left(\hat{C}_i^{\ell,h} + \hat{R}_i^{h,\ell}\xi^\ell\right) - \sum_{j\in\mathcal{M}^\ell} p_{ij}^s\left[C^\ell\left(b_{ij}^{h,\ell}, b_{ij}^\ell\right) + R^\ell\left(e_{ij}^{h,\ell}, e_{ij}^\ell, \vec{\rho}_i\right)\xi^\ell\right]$$

$$+p_i\bar{\rho}_i^{\ell}(\hat{b}_i^{\ell,h}+\hat{e}_i^{\ell,h})-\sum_{j\in\mathcal{M}^{\ell}}p_{ij}^{\ell}\Big[b_{ij}^{\ell}+e_{ij}^{\ell}\Big], \tag{A-66}$$

where

$$\hat{R}_i^{h,\ell}=\frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^{\ell}}p_{kj}^{\ell}R^{\ell}(e_{kj}^{h,\ell},e_{kj}^{\ell},\vec{\rho}_k)}{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^{\ell}}p_{kj}^{\ell}},\ \text{and}\ \hat{C}_i^{\ell,h}=\frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^{\ell}}p_{kj}^{\ell}C^{\ell}(b_{kj}^{h,\ell},b_{kj}^{\ell})}{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^{\ell}}p_{kj}^{\ell}}, \tag{A-67}$$

denote the proportion of readmitted patients and the average cost of type-ℓ PAC providers, excluding patients discharged from hospital i, and

$$\hat{b}_i^{\ell,h}=\frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^{\ell}}p_{kj}^{\ell}b_{kj}^{\ell}}{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^{\ell}}p_{kj}^{\ell}},\ \text{and}\ \hat{e}_i^{\ell,h}=\frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^{\ell}}p_{kj}^{\ell}e_{kj}^{\ell}}{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^{\ell}}p_{kj}^{\ell}} \tag{A-68}$$

are the average investments to reduce type-ℓ PAC providers' costs and readmissions, respectively, excluding patients discharged from hospital i.

The total payment amount for hospital i is calculated by summing up these components:

$$T_i^h=T_i^{h,0}+\sum_{\ell\in\mathcal{L}}T_i^{h,\ell}. \tag{A-69}$$

To highlight the intuition behind the payment scheme for hospitals, we first note that $T_i^{h,o}$ and the "cost of care" component in (20) are almost identical in principle: both consider the total cost incurred by the hospital in providing care and making improvement investments. Additionally, $T_i^{h,\ell}$ is similar to the "Outcome-based adjustment for care coordination" component in (20) with a subtle difference: there is an additional term in the second line of (A-66). This term incentivizes hospitals to consider the costs associated with different types of PAC providers' investment in reducing costs and readmissions. It does not appear in (20) because the discharge destination is assumed to be exogenous in that section. However, this component could be incorporated in the original payment scheme, as outlined in Remark 6.1.

We next prove that this payment scheme induces first-best actions from all providers.[7]

Proposition A-1. *If the regulator uses* (A-63) *to reimburse hospitals and* (A-69) *to reimburse PAC providers, then the unique Nash equilibrium is for each hospital $i\in\mathcal{N}$ and type-ℓ PAC provider $j\in\mathcal{M}^{\ell}$, $\ell\in\mathcal{L}$, to pick first-best actions $a_i^h=a_h^*, b_{ij}^{h,\ell}=b_{h,\ell}^*, e_{ij}^{h,\ell}=e_{h,\ell}^*, \rho_i^{\ell}=\rho_{\ell}^*$, and $b_{ij}^{\ell}=b_{\ell}^*, e_{ij}^{\ell}=e_{\ell}^*$, respectively. In addition, all providers break even in this equilibrium.*

[7] Without loss of generality and, as in our original model, we assume that hospital i and type-ℓ PAC provider j choose first-best actions when $p_{ij}^{\ell}=0$; see the last paragraph of Section 3 for details.

Proof: We continue to adopt Assumption A-1, with functions C^s and R adapted into sets of functions C^ℓ and R^ℓ for all $\ell \in \mathcal{L}$, and Assumption A-1(iii) applied to R^ℓ for any given $\vec{\rho}$. These assumptions ensure that for any given $\vec{\rho}$ the first-best actions are uniquely determined by the FOCs of total welfare; see the proof of Lemma A-2 below. We will also prove that the total welfare with first-best actions (as functions of $\vec{\rho}$) plugged in achieves maximum at some discharge decisions which are the same for all hospitals. Notably, we do not impose additional conditions to ensure unique first-best discharge decisions and will prove in Proposition A-1 that our payment scheme restores first best provided that hospitals follow the same break-even rule as the regulator when multiple discharge decisions are optimal.

Lemma A-2 (First-best benchmark). *The regulator's objective in* (A-62) *has a maximizer in which* $a_i^h = a_h^*$ *and* $\vec{\rho}_i = \vec{\rho}^*$ *for each hospital* $i \in \mathcal{N}$*, and for each type-*ℓ*,* $\ell \in \mathcal{L}$*, PAC provider* $j \in \mathcal{M}^\ell$ *such that* $p_{ij}^\ell > 0$*,* $b_{ij}^{h,\ell} = b_{h,\ell}^*, b_{ij}^\ell = b_\ell^*, e_{ij}^{h,\ell} = e_{h,\ell}^*$*, and* $e_{ij}^\ell = e_\ell^*$*, where* $a_h^*, b_{h,\ell}^*, e_{h,\ell}^*, b_\ell^*, e_\ell^* \in (0, \Gamma)$ *are unique and satisfy the following FOCs:*

$$\frac{dC^h(a_h^*)}{da_h} + 1 = 0, \tag{A-70}$$

$$\frac{\partial C^\ell(b_{h,\ell}^*, b_\ell^*)}{\partial b^h} + 1 = 0, \tag{A-71}$$

$$\frac{\partial C^\ell(b_{h,\ell}^*, b_\ell^*)}{\partial b^\ell} + 1 = 0, \tag{A-72}$$

$$\frac{\partial R^\ell(e_{h,\ell}^*, e_\ell^*, \vec{\rho}^*)}{\partial e^h}(\xi^h + \xi^\ell) + 1 = 0, \tag{A-73}$$

$$\frac{\partial R^\ell(e_{h,\ell}^*, e_\ell^*, \vec{\rho}^*)}{\partial e^\ell}(\xi^h + \xi^\ell) + 1 = 0. \tag{A-74}$$

Proof of Lemma A-2: Let $\vec{\mathbf{h}} = \{\mathbf{h}_i, i \in \mathcal{N}\}$ and $\vec{\mathbf{v}} = \{\mathbf{v}_j, j \in \mathcal{M}^\ell, \ell \in \mathcal{L}\}$ denote the actions of all hospitals and all PAC providers, respectively. Thus, total welfare W is a function of $\vec{\mathbf{h}}$ and $\vec{\mathbf{v}}$ and by (A-59), (A-61), and (A-62), is given by

$$\begin{aligned} W =& \upsilon - \sum_{i \in \mathcal{N}} p_i \left(C^h(a_i^h) + a_i^h\right) \\ & - \sum_{i \in \mathcal{N}} \sum_{\ell \in \mathcal{L}} \sum_{j \in \mathcal{M}^\ell} p_{ij}^\ell \left[C^\ell(b_{ij}^{h,\ell}, b_{ij}^\ell) + b_{ij}^{h,\ell} + b_{ij}^\ell + R^\ell(e_{ij}^{h,\ell}, e_{ij}^\ell, \vec{\rho}_i)(\xi^h + \xi^\ell) + e_{ij}^{h,\ell} + e_{ij}^\ell\right]. \end{aligned} \tag{A-75}$$

For notational simplicity, we will drop the arguments when it is clear from the context. It is straightforward to verify that (A-4) holds and thus W is maximized at $a_i^h = a_h^*$ for all $i \in \mathcal{N}$; the proof is identical to that in Lemma A-1. In addition, for any fixed $\vec{\rho}_i$, $i \in \mathcal{N}$, by (A-75) we have: (i) when $p_{ij}^\ell = 0$, W is independent of hospital i's investments $(b_{ij}^{h,\ell}, e_{ij}^{h,\ell})$ and PAC provider j's investments $(b_{ij}^\ell, e_{ij}^\ell)$. WLOG we assume that first-best actions are taken (see the last paragraph of Section 3 for details), i.e., $b_{ij}^{h,\ell} = b_{h,\ell}^*, e_{ij}^{h,\ell} = \tilde{e}_{h,\ell}(\vec{\rho}_i), b_{ij}^\ell = b_\ell^*, e_{ij}^\ell = \tilde{e}_\ell(\vec{\rho}_i)$ as determined by (A-71), (A-72), (A-76),

and (A-77), respectively; (ii) when $p_{ij}^{\ell} > 0$, the optimal investments are characterized by $b_{h,\ell}^*, b_{\ell}^*$ defined as in (A-71)-(A-72), and $\tilde{e}_{h,\ell}(\vec{\rho}_i), \tilde{e}_{\ell}(\vec{\rho}_i)$ defined as follows (this proof is omitted as it is similar to that in Lemma A-1):

$$\frac{\partial R^{\ell}(\tilde{e}_{h,\ell}(\vec{\rho}_i), \tilde{e}_{\ell}(\vec{\rho}_i), \vec{\rho}_i)}{\partial e^h}(\xi^h + \xi^{\ell}) + 1 = 0, \tag{A-76}$$

$$\frac{\partial R^{\ell}(\tilde{e}_{h,\ell}(\vec{\rho}_i), \tilde{e}_{\ell}(\vec{\rho}_i), \vec{\rho}_i)}{\partial e^{\ell}}(\xi^h + \xi^{\ell}) + 1 = 0. \tag{A-77}$$

Plugging in (A-75) and noting that $\sum_{j \in \mathcal{M}^{\ell}} p_{ij}^{\ell}/p_i = \rho_i^{\ell}$, we obtain

$$\begin{aligned} W =& v - C^h(a_h^*) - a_h^* \\ & - \sum_{i \in \mathcal{N}} \sum_{\ell \in \mathcal{L}} p_i \rho_i^{\ell} \left[C^{\ell}(b_{h,\ell}^*, b_{\ell}^*) + b_{h,\ell}^* + b_{\ell}^* + R^{\ell}(\tilde{e}_{h,\ell}(\vec{\rho}_i), \tilde{e}_{\ell}(\vec{\rho}_i), \vec{\rho}_i)(\xi^h + \xi^{\ell}) + \tilde{e}_{h,\ell}(\vec{\rho}_i) + \tilde{e}_{\ell}(\vec{\rho}_i) \right]. \end{aligned}$$

The welfare-maximizing patient discharge problem can thus be expressed as, for each $i \in \mathcal{N}$,

$$\underset{\vec{\rho}_i}{\text{minimize}} \ \sum_{\ell \in \mathcal{L}} \rho_i^{\ell} \left[C^{\ell}(b_{h,\ell}^*, b_{\ell}^*) + b_{h,\ell}^* + b_{\ell}^* + R^{\ell}(\tilde{e}_{h,\ell}(\vec{\rho}_i), \tilde{e}_{\ell}(\vec{\rho}_i), \vec{\rho}_i)(\xi^h + \xi^{\ell}) + \tilde{e}_{h,\ell}(\vec{\rho}_i) + \tilde{e}_{\ell}(\vec{\rho}_i) \right] \tag{A-78}$$

$$\text{s.t. } \vec{\rho}_i \in [0,1]^L, \sum_{\ell \in \mathcal{L}} \rho_i^{\ell} = 1. \tag{A-79}$$

Since $\tilde{e}_{h,\ell}(\vec{\rho}_i)$ and $\tilde{e}_{\ell}(\vec{\rho}_i)$ defined as in (A-76)-(A-77) are the unique unconstrained minimizer of $R^{\ell}(e_{h,\ell}, e_{\ell}, \vec{\rho}_i)(\xi^h + \xi^{\ell}) + e_{h,\ell} + e_{\ell}$, the minimum value $R^{\ell}(\tilde{e}_{h,\ell}(\vec{\rho}_i), \tilde{e}_{\ell}(\vec{\rho}_i), \vec{\rho}_i)(\xi^h + \xi^{\ell}) + \tilde{e}_{h,\ell}(\vec{\rho}_i) + \tilde{e}_{\ell}(\vec{\rho}_i)$ is continuous in $\vec{\rho}_i$; this establishes continuity of objective (A-78) in $\vec{\rho}_i$. Since the constraint set defined by (A-79) is compact, the minimization problem (A-78)-(A-79) has at least one solution. Moreover, each solution is independent of hospital index i because it does not appear in the objective function in (A-78) and the constraint set defined in (A-79). The proof is complete by defining $\vec{\rho}^*$ as the first-best discharge decisions the regulator chooses for each hospital $i \in \mathcal{N}$. □

Proof of Proposition A-1: The proof is based on the observation that under the proposed payment scheme, the difference between a hospital's objective and the regulator's objective is independent of that hospital's actions, and the difference between a PAC provider's objective and the regulator's objective is independent of that PAC provider's actions. More precisely, given the actions of all other hospitals and PAC providers, by (A-58)-(A-59) and (A-64)-(A-69), hospital i's objective is

$$\begin{aligned} \Pi_i^h(\mathbf{h}_i) =& p_i \left[\bar{C}_i^h + \bar{a}_i^h - C^h(a_i^h) - a_i^h \right] + p_i \hat{R}_i^h \xi^h - \sum_{\ell \in \mathcal{L}} \sum_{j \in \mathcal{M}^{\ell}} p_{ij}^{\ell} R^{\ell}\left(e_{ij}^{h,\ell}, e_{ij}^{\ell}, \vec{\rho}_i\right) \xi^h \\ & + \sum_{\ell \in \mathcal{L}} \left[p_i \bar{\rho}_i^{\ell} \left(\hat{b}_i^{h,\ell} + \hat{e}_i^{h,\ell} \right) - \sum_{j \in \mathcal{M}^{\ell}} p_{ij}^{\ell} \left(b_{ij}^{h,\ell} + e_{ij}^{h,\ell} \right) \right] + \sum_{\ell \in \mathcal{L}} \left[p_i \bar{\rho}_i^{\ell} (\hat{b}_i^{\ell,h} + \hat{e}_i^{\ell,h}) - \sum_{j \in \mathcal{M}^{\ell}} p_{ij}^{\ell} \left(b_{ij}^{\ell} + e_{ij}^{\ell} \right) \right] \\ & + \sum_{\ell \in \mathcal{L}} \left\{ p_i \bar{\rho}_i^{\ell} \left(\hat{C}_i^{\ell,h} + \hat{R}_i^{h,\ell} \xi^{\ell} \right) - \sum_{j \in \mathcal{M}^{\ell}} p_{ij}^{\ell} \left[C^{\ell}\left(b_{ij}^{h,\ell}, b_{ij}^{\ell} \right) + R^{\ell}\left(e_{ij}^{h,\ell}, e_{ij}^{\ell}, \vec{\rho}_i \right) \xi^{\ell} \right] \right\}. \end{aligned}$$

By (A-60)-(A-61) and (A-63), PAC provider j's objective is

$$\Pi_j^\ell(\mathbf{v}_j) = \sum_{i\in\mathcal{N}} p_{ij}^\ell \Big[\hat{C}_j^\ell + \hat{b}_j^\ell + \hat{e}_j^\ell - C^\ell(b_{ij}^{h,\ell}, b_{ij}^\ell) - b_{ij}^\ell - e_{ij}^\ell + \Big(\hat{R}_j^\ell - R^s(e_{ij}^{h,\ell}, e_{ij}^\ell, \vec{\rho}_i)\Big)(\xi^h + \xi^\ell)\Big].$$

Subtracting each objective by W from (A-75), we obtain

$$\begin{aligned}\Pi_i^h(\mathbf{h}_i) - W(\vec{\mathbf{h}}, \vec{\mathbf{v}}) =& p_i\left(\bar{C}_i^h + \bar{a}_i^h + \hat{R}_i^h \xi^h\right) + \sum_{k\in\mathcal{N}_i}\Big[C^h\left(a_k^h\right) + a_k^h\Big] - v \\ &+ \sum_{\ell\in\mathcal{L}} p_i \bar{\rho}_i^\ell \Big[\hat{b}_i^{h,\ell} + \hat{e}_i^{h,\ell} + \hat{b}_i^{\ell,h} + \hat{e}_i^{\ell,h} + \hat{C}_i^{\ell,h} + \hat{R}_i^{h,\ell}\xi^\ell\Big] \\ &+ \sum_{k\in\mathcal{N}_i}\sum_{\ell\in\mathcal{L}}\sum_{j\in\mathcal{M}^\ell} p_{kj}^\ell \Big[C^\ell(b_{kj}^{h,\ell}, b_{kj}^\ell) + R^s\left(e_{kj}^{h,\ell}, e_{kj}^\ell, \vec{\rho}_k\right)(\xi^h + \xi^\ell) + b_{kj}^{h,\ell} + b_{kj}^\ell + e_{kj}^{h,\ell} + e_{kj}^\ell\Big],\end{aligned}$$

and

$$\begin{aligned}\Pi_j^\ell(\mathbf{v}_j) - W(\vec{\mathbf{h}}, \vec{\mathbf{v}}) =& \sum_{i\in\mathcal{N}} p_{ij}^s \Big[\hat{C}_j^\ell + \hat{b}_j^\ell + \hat{e}_j^\ell + \hat{R}_j^\ell(\xi^h + \xi^\ell)\Big] - v + \sum_{i\in\mathcal{N}}\Big(p_i\Big[C^h\left(a_i^h\right) + a_i^h\Big] + p_{ij}^\ell\left[b_{ij}^{h,\ell} + e_{ij}^{h,\ell}\right]\Big) \\ &+ \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{L}_\ell}\sum_{m\in\mathcal{M}^k} p_{im}^k\Big[C^k(b_{im}^{h,k}, b_{im}^k) + R^k\left(e_{im}^{h,k}, e_{im}^k, \vec{\rho}_i\right)(\xi^h + \xi^k) + b_{im}^{h,k} + b_{im}^k + e_{im}^{h,k} + e_{im}^k\Big] \\ &+ \sum_{i\in\mathcal{N}}\sum_{m\in\mathcal{M}_j^\ell} p_{im}^\ell\Big[C^\ell(b_{im}^{h,\ell}, b_{im}^\ell) + R^s\left(e_{im}^{h,\ell}, e_{im}^\ell, \vec{\rho}_i\right)(\xi^h + \xi^\ell) + b_{im}^{h,\ell} + b_{im}^\ell + e_{im}^{h,\ell} + e_{im}^\ell\Big].\end{aligned}$$

Therefore, the difference between objectives of the regulator and any hospital i does not depend on the hospital's actions $\mathbf{h}_i$, and the difference between objectives of the regulator and any PAC provider j does not depend on the PAC provider's actions $\mathbf{v}_j$. This implies that the equilibrium actions are equal to the first-best actions; we omit the proof as it is similar to that for Theorem 1. Plugging in the first-best actions one can verify that all hospitals and PAC providers break even in equilibrium. □

F. Endogenous readmission cost

We initially assumed that the treatment costs for readmitted patients, denoted by ξ^h for hospitals and ξ^s for PAC providers, are exogenous. However, in practice, hospitals and PAC providers may invest to reduce treatment costs, which can affect costs of treating readmitted patients. In this section, we extend our model to show that our payment schemes induce first-best actions by assuming that the readmission cost is the same as the cost for the initial (index) admission. Specifically, we define C^h and C^s as the treatment costs for hospitals and PAC providers, respectively, for both the initial admission and readmission.

In this case, the objective of hospital i is given by:

$$\Pi_i^h\left(\mathbf{h}_i\right) = T_i^h - \mathcal{C}_i^h(\mathbf{h}_i), \tag{A-80}$$

where $\mathcal{C}_i^h(\mathbf{h}_i)$ represents the total cost of the hospital, defined as:

$$\mathcal{C}_i^h(\mathbf{h}_i) = p_i\Big[C^h\left(a_i^h\right) + a_i^h\Big] + \sum_{j\in\mathcal{M}} p_{ij}\Big[b_{ij}^h + e_{ij}^h + R\left(e_{ij}^h, e_{ij}^s\right)\left(C^h\left(a_i^h\right) + a_i^h + b_{ij}^h\right)\Big] \tag{A-81}$$

and, as in Section 3, a_i^h, b_{ij}^h, and e_{ij}^h represent the investments of hospital i for cost reduction, coordination, and readmission reduction, respectively. Similarly, the objective of PAC provider j is given by:

$$\Pi_j^s(\mathbf{s}) = T_j^s - \mathcal{C}_j^s(\mathbf{s}_j), \tag{A-82}$$

where $\mathcal{C}_j^s(\mathbf{s}_j)$ represents the total cost of the PAC provider, defined as:

$$\mathcal{C}_j^s(\mathbf{s}_j) = \sum_{i\in\mathcal{N}} p_{ij}\Big[e_{ij}^s + \left(1 + R\left(e_{ij}^h, e_{ij}^s\right)\right)\left(C^s\left(b_{ij}^h, b_{ij}^s\right) + b_{ij}^s\right)\Big] \tag{A-83}$$

and, as in Section 3, b_{ij}^s and e_{ij}^s represent the investments of PAC provider j for cost reduction and readmission reduction, respectively. The main difference between (1) and (A-81) is that we use $\left(C^h\left(a_i^h\right) + a_i^h + b_{ij}^h\right)$ to capture the total cost of readmitted patients instead of ξ^h. Similarly, $C^s\left(b_{ij}^h, b_{ij}^s\right) + b_{ij}^s$ replaces ξ^s in (3) to obtain (A-83).

To update the proposed payment model to account for the cost of readmitted patients, the payment amount to hospital i is determined by:

$$\begin{aligned} T_i^h = & \underbrace{p_i\Big[\left(1+\bar{R}_i^h\right)\left(\bar{C}_i^h + \bar{a}_i^h + \bar{b}_i^h\right) + \bar{e}_i^h\Big]}_{\text{Cost of care}} \\ & + \underbrace{\sum_{j\in\mathcal{M}} p_{ij}\Big[\left(1+\bar{R}_i^h\right)\left(\bar{C}_i^{sh} + \bar{b}_j^s\right) - \left(1 + R(e_{ij}^h, e_{ij}^s)\right)\left(C^s\left(b_{ij}^h, b_{ij}^s\right) + b_{ij}^s\right)\Big]}_{\text{Outcome-based adjustment}}, \end{aligned} \tag{A-84}$$

The term $\bar{C}_i^h + \bar{a}_i^h + \bar{b}_i^h$ in the first component ("Cost of care") above is the payment to cover the cost of treatment for readmitted patients in the hospital and the outcome-based payment reflects the PAC cost of treating readmitted patients. Similar to the difference between objective functions in this section and those in Section 5 as explained above (see (1) and (A-81)), the main difference in the current payment amount is that the cost of treatment for readmitted patients (ξ^h and ξ^s) in (20) are replaced by the corresponding costs in the current model in (A-84). Similarly, for PAC providers, we modify the payment as follows

$$\begin{aligned} T_j^s = & \underbrace{\sum_{i\in\mathcal{N}} p_{ij}\Big[\left(1+\bar{R}_j^s\right)\left(\bar{C}_j^s + \bar{b}_j^s\right) + \bar{e}_j^s\Big]}_{\text{Cost of care}} \\ & + \underbrace{\sum_{i\in\mathcal{N}} p_{ij}\Big[\left(1+\bar{R}_j^s\right)\left(\bar{C}_i^h + \bar{a}_i^h + \bar{b}_i^h\right) - \left(1 + R(e_{ij}^h, e_{ij}^s)\right)\left(C^h(a_i^h) + a_i^h + b_{ij}^h\right)\Big]}_{\text{Outcome-based adjustment}}, \end{aligned} \tag{A-85}$$

As for hospitals, $\bar{C}_j^s + \bar{b}_j^s$ in the first component covers the cost of treatment for readmitted patients in PAC providers and the outcome-based payment is based on the cost of treating readmitted patients in a hospital, whereas in (21) these costs are captured by ξ^h for hospitals and ξ^s for PAC providers.

The objective of the regulator remains the same as in (5), where $\mathcal{C}_i^h$ and $\mathcal{C}_j^s$ are defined as in (A-81) and (A-83), respectively, for all $i \in \mathcal{N}$ and $j \in \mathcal{M}$. Under the assumption that the regulator's objective has unique optimal actions (denoted again by (a_h^*, b_h^*, e_h^*) for hospitals and by (b_s^*, e_s^*) for PAC providers) and assuming these actions satisfy FOCs, we show that the payment scheme leads to first-best actions.

Proposition A-2. *If the regulator uses* (A-84) *to reimburse hospitals and* (A-85) *to reimburse PAC providers, then the unique Nash equilibrium is for each each hospital $i \in \mathcal{N}$ and PAC provider $j \in \mathcal{M}$ to pick first-best actions $a_i^h = a_h^*, b_{ij}^h = b_h^*, e_{ij}^h = e_h^*$, and $b_{ij}^s = b_s^*, e_{ij}^s = e_s^*$, respectively. In addition, all providers break even in this equilibrium.*

Moreover, our model can be extended to accommodate scenarios where the cost of readmitted patients deviates from that of index admissions, and where patients may need multiple readmissions, as discussed in Section 5.3 of Arifoğlu et al. (2021).

Proof: We continue to adopt Assumption A-1 with (A-1)-(A-2) revised into

$$\lim_{e^h \downarrow 0} \frac{\partial R\left(e^h, e^s\right)}{\partial e^h} < -\frac{1}{C^h\left(a_h^*\right) + C^s\left(b_h^*, b_s^*\right) + a_h^* + b_h^* + b_s^*} < \lim_{e^h \uparrow \Gamma} \frac{\partial R\left(e^h, e^s\right)}{\partial e^h} \text{ for any } e^s \in [0, \Gamma], \quad \text{(A-86)}$$

$$\lim_{e^s \downarrow 0} \frac{\partial R\left(e^h, e^s\right)}{\partial e^s} < -\frac{1}{C^h\left(a_h^*\right) + C^s\left(b_h^*, b_s^*\right) + a_h^* + b_h^* + b_s^*} < \lim_{e^s \uparrow \Gamma} \frac{\partial R\left(e^h, e^s\right)}{\partial e^s} \text{ for any } e^h \in [0, \Gamma]. \quad \text{(A-87)}$$

Under these conditions, the socially optimal actions uniquely exist and are determined by the FOCs of total welfare W, as shown below.

Lemma A-3 (First-best benchmark). *The regulator's objective in* (5) *has a unique maximizer in which $a_i^h = a_h^*$ for each hospital $i \in \mathcal{N}$, and for each PAC provider $j \in \mathcal{M}$ such that $p_{ij} > 0$, $b_{ij}^h = b_h^*, b_{ij}^s = b_s^*, e_{ij}^h = e_h^*$, and $e_{ij}^s = e_s^*$, where $a_h^*, b_h^*, b_s^* \in (0, \Gamma)$ are defined in* (6)-(8)*, $e_h^*, e_s^* \in (0, \Gamma)$ satisfy the following FOCs:*

$$\frac{\partial R\left(e_h^*, e_s^*\right)}{\partial e^h}\left[C^h\left(a_h^*\right) + C^s\left(b_h^*, b_s^*\right) + a_h^* + b_h^* + b_s^*\right] + 1 = 0, \quad \text{(A-88)}$$

$$\frac{\partial R\left(e_h^*, e_s^*\right)}{\partial e^s}\left[C^h\left(a_h^*\right) + C^s\left(b_h^*, b_s^*\right) + a_h^* + b_h^* + b_s^*\right] + 1 = 0. \quad \text{(A-89)}$$

Proof of Lemma A-3: Plugging (A-81) and (A-83) in (5), we obtain

$$W = v - \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{M}} p_{ij}\Big[\left(1 + R(e_{ij}^h, e_{ij}^s)\right)\left(C^h(a_i^h) + a_i^h + C^s(b_{ij}^h, b_{ij}^s) + b_{ij}^h + b_{ij}^s\right) + e_{ij}^h + e_{ij}^s\Big]. \quad \text{(A-90)}$$

It is straightforward to verify that, for any fixed e^h_{ij} and e^s_{ij}, $i\in\mathcal{N}$, $j\in\mathcal{M}$, the regulator's objective in (A-90) has a unique maximizer in which $a^h_i = a^*_h$ for each hospital $i\in\mathcal{N}$, and for each PAC provider $j\in\mathcal{M}$ such that $p_{ij}>0$, $b^h_{ij}=b^*_h$ and $b^s_{ij}=b^*_s$; the proof is identical to that in Lemma A-1. Below we characterize (e^*_h, e^*_s), i.e., first-best investments hospitals and PAC providers make to reduce readmissions. Let $W^*(e^h_{ij}, e^s_{ij}, i\in\mathcal{N}, j\in\mathcal{M}) = W|_{\left\{a^h_{ij}=a^*_h, b^h_{ij}=b^*_h, b^s_{ij}=b^*_s, i\in\mathcal{N}, j\in\mathcal{M}\right\}}$ for notational simplicity. When $p_{ij}=0$, W^* is independent of e^h_{ij} and e^s_{ij}. WLOG we assume that first-best actions are taken (see the last paragraph of Section 3 for details), i.e., $e^h_{ij}=e^*_h$ and $e^s_{ij}=e^*_s$, where e^*_h and e^*_s are given by (A-88)-(A-89). When $p_{ij}>0$, we take the first and second partial derivatives of W^* with respect to e^s_{ij} and obtain

$$\frac{\partial W^*}{\partial e^s_{ij}} = -p_{ij}\left\{\frac{\partial R(e^h_{ij}, e^s_{ij})}{\partial e^s}\Big[C^h(a^*_h)+C^s(b^*_h,b^*_s)+a^*_h+b^*_h+b^*_s\Big]+1\right\},$$
$$\frac{\partial^2 W^*}{\partial (e^s_{ij})^2} = -p_{ij}\frac{\partial^2 R(e^h_{ij}, e^s_{ij})}{\partial (e^s)^2}\Big[C^h(a^*_h)+C^s(b^*_h,b^*_s)+a^*_h+b^*_h+b^*_s\Big].$$

For any fixed $e^h_{ij}\in[0,\Gamma]$, we have $\partial^2 W^*/\partial(e^s_{ij})^2<0$, $\lim_{e^s_{ij}\downarrow 0}\partial W^*/\partial e^s_{ij}>0$, and $\lim_{e_s\uparrow\Gamma}\partial W^*/\partial e^s_{ij}<0$ by Assumption A-1(iii) and (A-87). Hence there exists a unique $z(e^h_{ij})\in(0,\Gamma)$ that satisfies

$$\frac{\partial R(e^h_{ij}, z(e^h_{ij}))}{\partial e^s}\Big[C^h(a^*_h)+C^s(b^*_h,b^*_s)+a^*_h+b^*_h+b^*_s\Big]+1=0. \tag{A-91}$$

Applying the Implicit Function Theorem, we obtain

$$\frac{dz(e^h_{ij})}{de^h_{ij}} = -\frac{\partial^2 R(e^h_{ij}, z(e^h_{ij}))/\partial e^h \partial e^s}{\partial^2 R\left(e^h_{ij}, z(e^h_{ij})\right)/\partial (e^s)^2}. \tag{A-92}$$

Since W^* is concave in e^s_{ij} by Assumption A-1(iii), we have

$$W^*|_{e^s_{ij}=z(e^h_{ij})} = \sup_{e^s_{ij}\in[0,\Gamma]} W^*.$$

Next we show that for any given $(\vec{\mathbf{h}},\vec{\mathbf{s}})\setminus\{e^h_{ij}, e^s_{ij}\}$, there exists a unique $e^*_h\in(0,\Gamma)$ that satisfies

$$W^*|_{\left\{e^h_{ij}=e^*_h, e^s_{ij}=e^*_s\right\}} = \sup_{e^h_{ij}\in[0,\Gamma]} W^*|_{e^s_{ij}=z(e^h_{ij})}.$$

Let $W^*(e^h_{ij}) = W^*|_{e^s_{ij}=z(e^h_{ij})}$ for notational simplicity. Then,

$$\begin{aligned}\frac{dW^*(e^h_{ij})}{de^h_{ij}} &= -p_{ij}\left\{\left(\frac{\partial R(e^h_{ij}, z(e^h_{ij}))}{\partial e^h}+\frac{\partial R(e^h_{ij}, z(e^h_{ij}))}{\partial e^s}\frac{dz(e^h_{ij})}{de^h_{ij}}\right)\Big[C^h(a^*_h)+C^s(b^*_h,b^*_s)+a^*_h+b^*_h+b^*_s\Big]+1+\frac{dz(e^h_{ij})}{de^h_{ij}}\right\}\\ &= -p_{ij}\left\{\frac{\partial R(e^h_{ij}, z(e^h_{ij}))}{\partial e^h}\Big[C^h(a^*_h)+C^s(b^*_h,b^*_s)+a^*_h+b^*_h+b^*_s\Big]+1\right\},\end{aligned}$$

where the second equality follows from (A-91).

$$\frac{d^2W^*(e^h_{ij})}{d(e^h_{ij})^2} = -p_{ij}\left[\frac{\partial^2 R(e^h_{ij}, z(e^h_{ij}))}{\partial e^h\partial e^s}\frac{dz(e^h_{ij})}{de^h_{ij}}+\frac{\partial^2 R(e^h_{ij}, z(e^h_{ij}))}{\partial (e^h)^2}\right]\Big[C^h(a^*_h)+C^s(b^*_h,b^*_s)+a^*_h+b^*_h+b^*_s\Big]$$

$$=p_{ij}\left[\frac{\left(\frac{\partial^2 R(e^h_{ij},z(e^h_{ij}))}{\partial e^h \partial e^s}\right)^2}{\frac{\partial^2 R(e^h_{ij},z(e^h_{ij}))}{\partial (e^s)^2}}-\frac{\partial^2 R(e^h_{ij},z(e^h_{ij}))}{\partial (e^h)^2}\right]\left[C^h(a^*_h)+C^s(b^*_h,b^*_s)+a^*_h+b^*_h+b^*_s\right]<0,$$

where the second equality follows by plugging in $dz(e^h_{ij})/de^h_{ij}$ from (A-92), and the inequality follows from Assumption A-1(iii). Moreover, we have $\lim_{e^h_{ij}\downarrow 0} dW^*(e^h_{ij})/de^h_{ij} > 0$ and $\lim_{e^h_{ij}\uparrow \Gamma} dW^*(e^h_{ij})/de^h_{ij} < 0$ by (A-86). Thus there exists a unique $e^*_h \in (0,\Gamma)$ that satisfies (A-88) with $e^*_s = z(e^*_h)$; (A-89) follows by substituting $e^h_{ij} = e^*_h$ into (A-91). □

Proof of Proposition A-2: The proof is based on the observation that under the proposed payment scheme, the difference between a hospital's objective and the regulator's objective is independent of that hospital's actions, and the difference between a PAC provider's objective and the regulator's objective is independent of that PAC provider's actions. More precisely, given the actions of all other hospitals and PAC providers, by (A-80)-(A-81) and (A-84), hospital i's objective is

$$\begin{aligned}\Pi^h_i(\mathbf{h}_i)=&\sum_{j\in\mathcal{M}} p_{ij}\left[\left(1+\bar{R}^h_i\right)\left(\bar{C}^h_i+\bar{a}^h_i+\bar{C}^{sh}_i+\bar{b}^h_i+\bar{b}^s_j\right)+\bar{e}^h_i\right]\\&-\sum_{j\in\mathcal{M}} p_{ij}\left[\left(1+R(e^h_{ij},e^s_{ij})\right)\left(C^h(a^h_i)+a^h_i+C^s(b^h_{ij},b^s_{ij})+b^h_{ij}+b^s_{ij}\right)+e^h_{ij}\right].\end{aligned}$$

By (A-82)-(A-83) and (A-85), PAC provider j's objective is

$$\begin{aligned}\Pi^s_j(\mathbf{v}_j)=&\sum_{i\in\mathcal{N}} p_{ij}\left[\left(1+\bar{R}^s_j\right)\left(\bar{C}^h_i+\bar{a}^h_i+\bar{C}^s_j+\bar{b}^h_i+\bar{b}^s_j\right)+\bar{e}^s_j\right]\\&-\sum_{i\in\mathcal{N}} p_{ij}\left[\left(1+R(e^h_{ij},e^s_{ij})\right)\left(C^h(a^h_i)+a^h_i+C^s\left(b^h_{ij},b^s_{ij}\right)+b^h_{ij}+b^s_{ij}\right)+e^s_{ij}\right].\end{aligned}$$

Subtracting each objective by W from (A-90), we obtain

$$\begin{aligned}\Pi^h_i(\mathbf{h}_i)-W(\vec{\mathbf{h}},\vec{\mathbf{v}})=&\sum_{j\in\mathcal{M}} p_{ij}\left[\left(1+\bar{R}^h_i\right)\left(\bar{C}^h_i+\bar{a}^h_i+\bar{C}^{sh}_i+\bar{b}^h_i+\bar{b}^s_j\right)+\bar{e}^h_i\right]+\sum_{j\in\mathcal{M}} p_{ij}e^s_{ij}-\upsilon\\&+\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}} p_{kj}\Big\{\left[1+R(e^h_{kj},e^s_{kj})\right]\left[C^h(a^h_k)+a^h_k+C^s(b^h_{kj},b^s_{kj})+b^h_{kj}+b^s_{kj}\right]+e^h_{kj}+e^s_{kj}\Big\}.\end{aligned}$$

and

$$\begin{aligned}\Pi^s_j(\mathbf{v}_j)-W(\vec{\mathbf{h}},\vec{\mathbf{v}})=&\sum_{i\in\mathcal{N}} p_{ij}\left[\left(1+\bar{R}^s_j\right)\left(\bar{C}^h_i+\bar{a}^h_i+\bar{b}^h_i+\bar{C}^s_j+\bar{b}^s_j\right)+\bar{e}^s_j\right]+\sum_{i\in\mathcal{N}} p_{ij}e^h_{ij}-\upsilon\\&+\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} p_{ik}\Big\{\left[1+R(e^h_{ik},e^s_{ik})\right]\left[C^h(a^h_i)+a^h_i+C^s(b^h_{ik},b^s_{ik})+b^h_{ik}+b^s_{ik}\right]+e^h_{ik}+e^s_{ik}\Big\}.\end{aligned}$$

Therefore, the difference between objectives of the regulator and any hospital i does not depend on the hospital's actions $\mathbf{h}_i$, and the difference between objectives of the regulator and any PAC provider j does not depend on the PAC provider's actions $\mathbf{v}_j$. This implies that the equilibrium actions are equal to the first-best actions; we omit the proof as it is similar to that for Theorem 1. Plugging in the first-best actions one can verify that all hospitals and PAC providers break even in equilibrium. □

G. Uniform Investments

In our original model in Section 5, we assume that each hospital makes different *PAC provider dependent* investments at cost represented by b^h_{ij} there, to reduce PAC treatment costs with each PAC provider (vis-a-vis, we assumed PAC providers makes hospital-dependent investments at cost represented by b^s_{ij} there). However, some investments, such as installing an integrated IT system, could be considered fixed costs that impact the collaboration of a hospital with all PAC providers who are willing to participate in cost-reduction investments.

To model the impact of uniform (non-PAC/hospital-dependent) investments, assume that each hospital makes investment $H_i \in [0, \Gamma]$, $i \in \mathcal{N}$, and each PAC provider makes investment $F_j \in [0, \Gamma]$, $j \in \mathcal{M}$, to reduce PAC treatment costs. Additionally, assume that the cost of PAC treatment $C^s : [0, \Gamma] \times [0, \Gamma] \to \mathbb{R}_+$ is a function of hospital's action H_i and PAC provider's action F_j. All other components of the model remain identical to those introduced in Section 3.

In this case, the objective of hospital i is

$$\Pi^h_i(\mathbf{h}_i) = T^h_i - \mathcal{C}^h_i(\mathbf{h}_i), \tag{A-93}$$

where $\mathcal{C}^h_i(\mathbf{h}_i)$ represents the total cost of the hospital, defined as:

$$\mathcal{C}^h_i(\mathbf{h}_i) = p_i \left[C^h(a^h_i) + a^h_i + H_i\right] + \sum_{j \in \mathcal{M}} p_{ij} \left[R(e^h_{ij}, e^s_{ij})\xi^h + e^h_{ij}\right]. \tag{A-94}$$

Similarly, the objective of PAC provider j is

$$\Pi^s_j(\mathbf{s}) = T^s_j - \mathcal{C}^s_j(\mathbf{s}_j), \tag{A-95}$$

where $\mathcal{C}^s_j(\mathbf{s}_j)$ represents the total cost of the PAC provider, defined as:

$$\mathcal{C}^s_j(\mathbf{s}_j) = \tilde{p}_j F_j + \sum_{i \in \mathcal{N}} p_{ij} \Big[C^s(H_i, F_j) + R(e^h_{ij}, e^s_{ij})\xi^s + e^s_{ij}\Big]. \tag{A-96}$$

The main difference between our original model (1) and (A-93) is that now we use H_i to capture the total cost of hospital i's investments to reduce PAC treatment cost instead of b^h_{ij} in (1). Similarly, we use F_j in (A-95) to capture the total cost of PAC j's investments instead of b^s_{ij} in (3).

To incorporate this change into the payment model, we introduce average investments $\bar{H}_i$ and $\bar{F}_j$ as benchmarks. Let

$$\bar{H}_i = \frac{\sum\limits_{k \in \mathcal{N}_i} p_k H_k}{1 - p_i}, \tag{A-97}$$

denote the average investment for PAC treatment cost reduction of all hospitals, excluding hospital i and

$$\bar{F}_j = \frac{\sum\limits_{k \in \mathcal{M}_j} \tilde{p}_k F_k}{1 - \tilde{p}_j}, \tag{A-98}$$

denote the average investment for PAC treatment cost reduction of all the PAC providers, excluding provider j. The modified payment model for hospitals and PAC providers is as follows:

$$T_i^h = p_i\left[\bar{C}_i^h + \bar{a}_i^h + \bar{H}_i + \bar{e}_i^h + \bar{R}_i^h\xi^h\right] + \sum_{j\in\mathcal{M}} p_{ij}\left[\left(\bar{C}_i^{sh} - C^s(H_i,F_j)\right) + \left(\bar{R}_i^h - R(e_{ij}^h,e_{ij}^s)\right)\xi^s\right], \quad \text{(A-99)}$$

$$T_j^s = \sum_{i\in\mathcal{N}} p_{ij}\left[\bar{C}_j^s + \bar{F}_j + \bar{e}_j^s + \bar{R}_j\xi^s\right] + \sum_{i\in\mathcal{N}} p_{ij}\left(\bar{R}_j - R(e_{ij}^h,e_{ij}^s)\right)\xi^h, \quad \text{(A-100)}$$

where benchmarks $\bar{H}_i$ and $\bar{F}_j$ are defined as in (A-97) and (A-98), respectively, and other benchmark parameters (i.e., $\bar{a}_i^h, \bar{e}_i^h, \bar{C}_i^{sh}, \bar{R}_i^h, \bar{C}_j^s, \bar{R}_j, \bar{e}_j^s$) are defined as in Section 4.1.

The payment model in this case is similar to that in Section 5, see (20) and (21), with the only difference being the way hospitals and PAC providers are compensated for their cost reduction investments. In (20) hospital i receives $p_i\bar{b}_i^h$ to recoup the cost of their investments to reduce PAC costs (since it is assumed to be variable cost there), whereas in (A-99) they receive $p_i\bar{H}_i$. For PAC providers, they receive $\sum_{i\in\mathcal{N}} p_{ij}\bar{b}_j^s$ for the cost of investments in (21) which becomes $\sum_{i\in\mathcal{N}} p_{ij}\bar{F}_j$ in (A-100).

The objective of the regulator in this case is given by (5), where $\mathcal{C}_i^h$ and $\mathcal{C}_j^s$ are defined as in (A-94) and (A-96), respectively, for each $i\in\mathcal{N}$ and $j\in\mathcal{M}$. Assuming that the regulator's objective has a unique optimal solution and that the optimal actions satisfy the FOCs, we find that the first-best actions for hospitals are identical and denoted by (a_h^*, H^*, e_h^*) for each hospital, while the first-best actions for PAC providers are identical and denoted by (F^*, e_s^*) for each PAC provider. We next demonstrate that this payment scheme induces first-best actions.

Proposition A-3. *If the regulator uses* (A-99) *to reimburse hospitals and* (A-100) *to reimburse PAC providers, then the unique Nash equilibrium is for each each hospital $i\in\mathcal{N}$ and PAC provider $j\in\mathcal{M}$ to pick first-best actions $a_i^h = a_h^*, H_i = H^*, e_{ij}^h = e_h^*$, and $F_j = F^*, e_{ij}^s = e_s^*$, respectively. In addition, all providers break even in this equilibrium.*

Similarly, we can demonstrate that when hospitals and PAC providers make uniform investments for reducing readmissions, our proposed payment model, with appropriate adjustments, continues to induce first-best actions.

Proof: We continue to adopt Assumption A-1 with conditions for PAC cost in Assumption A-1 adapted to: (i) PAC cost $C^s(H,F)$ is decreasing and convex in hospital and PAC provider actions:

$$\frac{\partial C^s}{\partial H} < 0, \frac{\partial^2 C^s}{\partial H^2} > 0, \frac{\partial C^s}{\partial F} < 0, \frac{\partial^2 C^s}{\partial F^2} > 0, \frac{\partial^2 C^s}{\partial H^2}\frac{\partial^2 C^s}{\partial F^2} > \left(\frac{\partial^2 C^s}{\partial H\partial F}\right)^2, \quad \text{(A-101)}$$

and (ii) the following boundary conditions hold

$$\lim_{H\downarrow 0}\frac{\partial C^s}{\partial H} < -1 < \lim_{H\uparrow\Gamma}\frac{\partial C^s}{\partial H} \text{ for any } F\in[0,\Gamma], \quad \text{(A-102)}$$

$$\lim_{F\downarrow 0}\frac{\partial C^s}{\partial F}<-1<\lim_{F\uparrow\Gamma}\frac{\partial C^s}{\partial F}\text{ for any }H\in[0,\Gamma]. \tag{A-103}$$

Under these conditions, the socially optimal actions uniquely exist and are determined by the FOCs of total welfare W, as shown below.

Lemma A-4 (First-best benchmark). *The regulator's objective in* (5) *has a unique maximizer in which* $a_i^h=a_h^*$ *and* $H_i=H^*$ *for each hospital* $i\in\mathcal{N}$, $F_j=F^*$ *for each PAC provider* $j\in\mathcal{M}$, *and when* $p_{ij}>0$, $e_{ij}^h=e_h^*$ *and* $e_{ij}^s=e_s^*$, *where* $a_h^*,e_h^*,e_s^*\in(0,\Gamma)$ *are defined in* (6), (9), (10), $H^*,F^*\in(0,\Gamma)$ *satisfy the following FOCs:*

$$\frac{\partial C^s\left(H^*,F^*\right)}{\partial H}+1=0, \tag{A-104}$$

$$\frac{\partial C^s\left(H^*,F^*\right)}{\partial F}+1=0. \tag{A-105}$$

Proof of Lemma A-3: Plugging (A-94) and (A-96) in (5), we obtain

$$W=v-\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}}p_{ij}\Big[C^h\left(a_i^h\right)+a_i^h+C^s(H_i,F_j)+H_i+F_j+R(e_{ij}^h,e_{ij}^s)(\xi^h+\xi^s)+e_{ij}^h+e_{ij}^s\Big]. \tag{A-106}$$

It is straightforward to verify that, for any fixed H_i and F_j, $i\in\mathcal{N}$, $j\in\mathcal{M}$, the regulator's objective in (A-106) has a unique maximizer in which $a_i^h=a_h^*$ for each hospital $i\in\mathcal{N}$, and for each PAC provider $j\in\mathcal{M}$ such that $p_{ij}>0$, $e_{ij}^h=e_h^*$ and $e_{ij}^s=e_s^*$; the proof is identical to that in Lemma A-1. Below we analyze the first-best investments hospitals and PAC providers make to reduce PAC costs, i.e.,

$$\underset{\vec{H}\in[0,\Gamma]^N,\vec{F}\in[0,\Gamma]^M}{\text{maximize}}\ W, \tag{A-107}$$

where we denote $\vec{H}=\{H_i,i\in\mathcal{N}\}$ and $\vec{F}=\{F_j,j\in\mathcal{M}\}$. The objective W given by (A-106) is concave in $(\vec{F},\vec{H})$ because the Hessian matrix $D^2W(\vec{F},\vec{H})$ is negative semi-definite due to $\partial^2C^s/\partial H^2>0,\partial^2C^s/\partial F^2>0$, and

$$\sum_{i\in\mathcal{N}}p_{ij}\frac{\partial^2C^s(H_i,F_j)}{\partial F^2}>\sum_{i\in\mathcal{N}}\frac{\left(p_{ij}\frac{\partial^2C^s(H_i,F_j)}{\partial H\partial F}\right)^2}{\sum_{k\in\mathcal{M}}p_{ik}\frac{\partial^2C^s(H_i,F_k)}{\partial H^2}}\text{ for each }j\in\mathcal{M}, \tag{A-108}$$

which follows from $(\partial^2C^s/\partial H^2)(\partial^2C^s/\partial F^2)>(\partial^2C^s/\partial H\partial F)^2$. In addition, the choice set of (A-107) is compact. Thus, S has a unique maximizer denoted by H_i^* and F_j^*, $i\in\mathcal{N}$, $j\in\mathcal{M}$. By (A-106),

$$\frac{\partial W}{\partial H_i}=-\sum_{j\in\mathcal{M}}p_{ij}\Big[\frac{\partial C^s(H_i,F_j)}{\partial H}+1\Big], \tag{A-109}$$

$$\frac{\partial^2 W}{\partial H_i^2} = -\sum_{j\in\mathcal{M}} p_{ij}\frac{\partial^2 C^s(H_i,F_j)}{\partial H^2}, \tag{A-110}$$

$$\frac{\partial W}{\partial F_j} = -\sum_{i\in\mathcal{N}} p_{ij}\Big[\frac{\partial C^s(H_i,F_j)}{\partial F}+1\Big], \tag{A-111}$$

$$\frac{\partial^2 W}{\partial F_j^2} = -\sum_{i\in\mathcal{N}} p_{ij}\frac{\partial^2 C^s(H_i,F_j)}{\partial F^2}. \tag{A-112}$$

For any fixed $\vec{F}$ and each $i\in\mathcal{N}$, we have $\partial^2 W/\partial H_i^2<0$ by (A-110) and (A-101), $\lim_{H_i\downarrow 0}\partial W/\partial H_i>0$ and $\lim_{H_i\uparrow\Gamma}\partial W/\partial H_i<0$ by (A-109) and (A-102). For any fixed $\vec{H}$ and each $j\in\mathcal{M}$, we have $\partial^2 W/\partial F_j^2<0$ by (A-112) and (A-101), $\lim_{F_j\downarrow 0}\partial W/\partial F_j>0$ and $\lim_{F_j\uparrow\Gamma}\partial W/\partial F_j<0$ by (A-111) and (A-103). Thus, the first-best investments H_i^* and F_j^*, $i\in\mathcal{N}$, $j\in\mathcal{M}$, are determined by FOCs:

$$\sum_{j=1}^{M} p_{ij}\Big[\frac{\partial C^s\left(H_i^*,F_j^*\right)}{\partial H}+1\Big]=0 \text{ for all } i\in\mathcal{N}, \tag{A-113}$$

$$\sum_{i=1}^{N} p_{ij}\Big[\frac{\partial C^s\left(H_i^*,F_j^*\right)}{\partial F}+1\Big]=0 \text{ for all } j\in\mathcal{M}. \tag{A-114}$$

Let H^* and F^* be determined by (A-104)-(A-105). The existence and uniqueness of H^* and F^* are ensured by (A-101)-(A-103). Moreover, one can verify that $H_i=H^*$ and $F_j=F^*$ for each $i\in\mathcal{N}$ and $j\in\mathcal{M}$ is a solution of (A-113)-(A-114). Thus, first-best actions, as uniquely determined by (A-113)-(A-114), are given by $H_i^*=H^*$ and $F_j^*=F^*$ for each $i\in\mathcal{N}$ and $j\in\mathcal{M}$. □

Proof of Proposition A-3: The proof is based on the observation that under the proposed payment scheme, the difference between a hospital's objective and the regulator's objective is independent of that hospital's actions, and the difference between a PAC provider's objective and the regulator's objective is independent of that PAC provider's actions. More precisely, given the actions of all other hospitals and PAC providers, by (A-93)-(A-94) and (A-99), hospital i's objective is

$$\begin{aligned}\Pi_i^h(\mathbf{h}_i)=&\sum_{j\in\mathcal{M}} p_{ij}\Big[\bar{C}_i^h+\bar{a}_i^h+\bar{H}_i+\bar{R}_i^h(\xi^h+\xi^s)+\bar{C}_i^{sh}+\bar{e}_i^h\Big]\\&-\sum_{j\in\mathcal{M}} p_{ij}\Big[C^h(a_i^h)+a_i^h+H_i+R(e_{ij}^h,e_{ij}^s)(\xi^h+\xi^s)+C^s(H_i,F_j)+e_{ij}^h\Big].\end{aligned}$$

By (A-95)-(A-96) and (A-100), PAC provider j's objective is

$$\Pi_j^s(\mathbf{v}_j)=\sum_{i\in\mathcal{N}} p_{ij}\Big[\bar{C}_j^s+\bar{F}_j+\bar{R}_j(\xi^h+\xi^s)+\bar{e}_j^s\Big]-\sum_{i\in\mathcal{N}} p_{ij}\Big[C^s(H_i,F_j)+F_j+R(e_{ij}^h,e_{ij}^s)(\xi^h+\xi^s)+e_{ij}^s\Big].$$

By (5), (A-94), and (A-96), total welfare is

$$W(\vec{\mathbf{h}},\vec{\mathbf{v}})=\upsilon-\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} p_{ij}\Big[C^h(a_i^h)+a_i^h+C^s(H_i,F_j)+H_i+F_j+R(e_{ij}^h,e_{ij}^s)(\xi^h+\xi^s)+e_{ij}^h+e_{ij}^s\Big].$$

Subtracting each objective by W from (A-106), we obtain

$$\Pi_i^h(\mathbf{h}_i) - W(\vec{\mathbf{h}}, \vec{\mathbf{v}}) = \sum_{j\in\mathcal{M}} p_{ij}\Big[\bar{C}_i^h + \bar{a}_i^h + \bar{H}_i + \bar{R}_i^h(\xi^h+\xi^s) + \bar{C}_i^{sh} + \bar{e}_i^h\Big] + \sum_{j\in\mathcal{M}} p_{ij}(F_j + e_{ij}^s) - \upsilon$$
$$+ \sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}} p_{kj}\Big[C^h(a_k^h) + a_k^h + C^s(H_k, F_j) + H_k + F_j + R(e_{kj}^h, e_{kj}^s)(\xi^h+\xi^s) + e_{kj}^h + e_{kj}^s\Big].$$

and

$$\Pi_j^s(\mathbf{v}_j) - W(\vec{\mathbf{h}}, \vec{\mathbf{v}}) = \sum_{i\in\mathcal{N}} p_{ij}\Big[\bar{C}_j^s + \bar{F}_j + \bar{R}_j(\xi^h+\xi^s) + \bar{e}_j^s\Big] + \sum_{i\in\mathcal{N}} p_{ij}\Big[C^h(a_i^h) + a_i^h + H_i + e_{ij}^h\Big] - \upsilon$$
$$+ \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} p_{ik}\Big[C^h(a_i^h) + a_i^h + C^s(H_i, F_k) + H_i + F_k + R(e_{ik}^h, e_{ik}^s)(\xi^h+\xi^s) + e_{ik}^h + e_{ik}^s\Big].$$

Therefore, the difference between objectives of the regulator and any hospital i does not depend on the hospital's actions $\mathbf{h}_i$, and the difference between objectives of the regulator and any PAC provider j does not depend on the PAC provider's actions $\mathbf{v}_j$. This implies that the equilibrium actions are equal to the first-best actions; we omit the proof as it is similar to that for Theorem 1. Plugging in the first-best actions one can verify that all hospitals and PAC providers break even in equilibrium. □

H. Fixed Costs of Investments

In our original model presented in Section 3, we assume variable costs of investments which are accounted on a per-patient basis; see, e.g., b_{ij}^h and e_{ij}^h in (2). However, investments such as staff training and process re-engineering, may primarily require a lump-sum investment independent from patient volume. In this extension, we incorporate fixed costs of investments and show that our proposed payment model can still induce the first-best outcome. We consider two separate case in which the hospital/PAC provider lump-sum investment is uniform or specific to PAC provider/hospital.

Uniform investments. Assume that each hospital i makes a *total* cost of investments $H_i \in [0,\Gamma]$, $i\in\mathcal{N}$, and each PAC provider j makes a *total* cost of investments $F_j \in [0,\Gamma]$, $j\in\mathcal{M}$, to reduce the per-patient PAC treatment cost $C^s(H_i,F_j)\in\mathbb{R}_+$; all other model components remain identical to those introduced in Section 3. In this case, the objective of hospital i is given by (1), where

$$\mathcal{C}_i^h(\mathbf{h}_i) = p_i\left[C^h(a_i^h) + a_i^h\right] + H_i + \sum_{j\in\mathcal{M}} p_{ij}\left[R(e_{ij}^h, e_{ij}^s)\xi^h + e_{ij}^h\right] \tag{A-115}$$

represents the total cost of the hospital. The objective of PAC provider j is given by (3), where

$$\mathcal{C}_j^s(\mathbf{s}_j) = F_j + \sum_{i\in\mathcal{N}} p_{ij}\Big[C^s(H_i,F_j) + R(e_{ij}^h, e_{ij}^s)\xi^s + e_{ij}^s\Big] \tag{A-116}$$

represents the total cost of the PAC provider. The only difference from our original model in (2) and (4) is that, to reduce the PAC cost, hospital i and PAC provider j respectively make *uniform* investments at lump-sum costs H_i and F_j independent from the patient volume.

To incorporate this change into the payment model, we define

$$\bar{H}_i = \frac{\sum_{k\in\mathcal{N}_i} H_k}{N-1}, \tag{A-117}$$

as the average investment to reduce the PAC cost of all hospitals, excluding hospital i, and

$$\bar{F}_j = \frac{\sum_{k\in\mathcal{M}_j} F_k}{M-1}, \tag{A-118}$$

the average investment of all PAC providers, excluding provider j. The payment amounts to hospital i and PAC provider j are given respectively by:

$$T_i^h = p_i\left[\bar{C}_i^h + \bar{a}_i^h + \bar{e}_i^h + \bar{R}_i^h\xi^h\right] + \bar{H}_i + \sum_{j\in\mathcal{M}} p_{ij}\left[\left(\bar{C}_i^{sh} - C^s(H_i,F_j)\right) + \left(\bar{R}_i^h - R(e_{ij}^h,e_{ij}^s)\right)\xi^s\right], \tag{A-119}$$

$$T_j^s = \sum_{i\in\mathcal{N}} p_{ij}\left[\bar{C}_j^s + \bar{e}_j^s + \bar{R}_j\xi^s\right] + \bar{F}_j + \sum_{i\in\mathcal{N}} p_{ij}\left(\bar{R}_j - R(e_{ij}^h,e_{ij}^s)\right)\xi^h, \tag{A-120}$$

where benchmarks $\bar{H}_i$ and $\bar{F}_j$ are defined as in (A-117)-(A-118), respectively, and other benchmark parameters (i.e., $\bar{a}_i^h, \bar{e}_i^h, \bar{C}_i^{sh}, \bar{R}_i^h, \bar{C}_j^s, \bar{R}_j, \bar{e}_j^s$) are defined as in Section 4.1. Compared to the payment model in Section 5, here we modify provider reimbursement for the costs of investments to reduce the PAC cost, i.e., from $\sum_{j\in\mathcal{M}} p_{ij}b_i^h$ for hospital i and $\sum_{i\in\mathcal{N}} p_{ij}\bar{b}_j^s$ for PAC provider j to $\bar{H}_i$ and $\bar{F}_j$, respectively, to reflect the change in cost structure (i.e., variable versus fixed).

The total social welfare W is given by (5), where $\mathcal{C}_i^h$ and $\mathcal{C}_j^s$ are defined as in (A-115)-(A-116), respectively, for each $i\in\mathcal{N}$ and $j\in\mathcal{M}$. We adapt Assumption A-1 to ensure the existence of a unique set of first-best actions determined by the FOCs of W. Next, we demonstrate that our payment model induces the first-best actions.

Proposition A-4. *If the regulator uses* (A-119) *to reimburse hospitals and* (A-120) *to reimburse PAC providers, then the unique Nash equilibrium is for each each hospital $i\in\mathcal{N}$ and PAC provider $j\in\mathcal{M}$ to pick first-best actions $a_i^h = a_h^*, H_i = H^*, e_{ij}^h = e_h^*$, and $F_j = F^*, e_{ij}^s = e_s^*$, respectively. In addition, all providers break even in this equilibrium.*

Similarly, we can demonstrate that when hospitals and PAC providers incure fixed costs of investments for reducing readmissions, our proposed payment model, with appropriate adjustments, continues to induce first-best actions.

Proof: We continue to adopt Assumption A-1 with part (ii) adapted to (A-101)-(A-103). Plugging (A-115)-(A-116) in (5), we obtain

$$W = v - \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} p_{ij}\left[C^h\left(a_i^h\right) + a_i^h + C^s(H_i,F_j) + R(e_{ij}^h,e_{ij}^s)(\xi^h+\xi^s) + e_{ij}^h + e_{ij}^s\right] - \sum_{i\in\mathcal{N}} H_i - \sum_{j\in\mathcal{M}} F_j.$$

We have verified that the first-best actions uniquely exist and are determined by the FOCs of total welfare W, i.e., $a_i^h = a_h^*$ and $H_i = H^*$ for each hospital $i \in \mathcal{N}$, $F_j = F^*$ for each PAC provider $j \in \mathcal{M}$, and when $p_{ij} > 0$, $e_{ij}^h = e_h^*$ and $e_{ij}^s = e_s^*$, where $a_h^*, e_h^*, e_s^* \in (0, \Gamma)$ are defined in (6), (9), (10), and $H^*, F^* \in (0, \Gamma)$ are defined in (A-104)-(A-105). The proof is identical to that for Lemma A-4 except that the first-best investments to reduce the PAC cost, i.e., H_i^* and F_j^*, $i \in \mathcal{N}$, $j \in \mathcal{M}$, are determined by the following FOCs:

$$1 + \sum_{j=1}^{M} p_{ij} \frac{\partial C^s\left(H_i^*, F_j^*\right)}{\partial H} = 0 \text{ for all } i \in \mathcal{N}, \tag{A-121}$$

$$1 + \sum_{i=1}^{N} p_{ij} \frac{\partial C^s\left(H_i^*, F_j^*\right)}{\partial F} = 0 \text{ for all } j \in \mathcal{M}. \tag{A-122}$$

It is straightforward to verify that $H_i^* = H^*$ and $F_j^* = F^*$ for each $i \in \mathcal{N}$ and $j \in \mathcal{M}$ is a solution of (A-121)-(A-122). In addition, the first-best actions uniquely exist due to concavity of W. Thus, in the first-best actions, $H_i = H^*$ for each hospital $i \in \mathcal{N}$, and $F_j = F^*$ for each PAC provider $j \in \mathcal{M}$.

Next, we prove that, under our proposed payment model, a unique equilibrium exists in which providers choose the first-best actions. By (1), (A-115), and (A-119), hospital i's objective is

$$\begin{aligned}\Pi_i^h(\mathbf{h}_i) = &\sum_{j\in\mathcal{M}} p_{ij}\Big[\bar{C}_i^h + \bar{a}_i^h + \bar{R}_i^h(\xi^h + \xi^s) + \bar{C}_i^{sh} + \bar{e}_i^h\Big] + \bar{H}_i - H_i \\ &- \sum_{j\in\mathcal{M}} p_{ij}\Big[C^h(a_i^h) + a_i^h + R(e_{ij}^h, e_{ij}^s)(\xi^h + \xi^s) + C^s(H_i, F_j) + e_{ij}^h\Big],\end{aligned}$$

where $\mathbf{h}_i = (a_i^h, H_i, e_{ij}^h, j \in \mathcal{M})$. By (3), (A-116) and (A-120), PAC provider j's objective is

$$\Pi_j^s(\mathbf{v}_j) = \sum_{i\in\mathcal{N}} p_{ij}\Big[\bar{C}_j^s + \bar{R}_j(\xi^h + \xi^s) + \bar{e}_j^s\Big] + \bar{F}_j - F_j - \sum_{i\in\mathcal{N}} p_{ij}\Big[C^s(H_i, F_j) + R(e_{ij}^h, e_{ij}^s)(\xi^h + \xi^s) + e_{ij}^s\Big],$$

where $\mathbf{v}_j = (F_j, e_{ij}^s, i \in \mathcal{N})$. Subtracting each objective by W, we obtain

$$\begin{aligned}\Pi_i^h(\mathbf{h}_i) - W(\vec{\mathbf{h}}, \vec{\mathbf{v}}) = &\sum_{j\in\mathcal{M}} p_{ij}\Big[\bar{C}_i^h + \bar{a}_i^h + \bar{R}_i^h(\xi^h + \xi^s) + \bar{C}_i^{sh} + \bar{e}_i^h\Big] + \bar{H}_i + \sum_{j\in\mathcal{M}} p_{ij}e_{ij}^s - \upsilon + \sum_{k\in\mathcal{N}_i} H_k + \sum_{j\in\mathcal{M}} F_j \\ &+ \sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}} p_{kj}\Big[C^h(a_k^h) + a_k^h + C^s(H_k, F_j) + R(e_{kj}^h, e_{kj}^s)(\xi^h + \xi^s) + e_{kj}^h + e_{kj}^s\Big],\end{aligned}$$

and

$$\begin{aligned}\Pi_j^s(\mathbf{v}_j) - W(\vec{\mathbf{h}}, \vec{\mathbf{v}}) = &\sum_{i\in\mathcal{N}} p_{ij}\Big[\bar{C}_j^s + \bar{R}_j(\xi^h + \xi^s) + \bar{e}_j^s\Big] + \bar{F}_j + \sum_{i\in\mathcal{N}} p_{ij}\Big[C^h(a_i^h) + a_i^h + e_{ij}^h\Big] - \upsilon + \sum_{i\in\mathcal{N}} H_i + \sum_{k\in\mathcal{M}_j} F_k \\ &+ \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} p_{ik}\Big[C^h(a_i^h) + a_i^h + C^s(H_i, F_k) + R(e_{ik}^h, e_{ik}^s)(\xi^h + \xi^s) + e_{ik}^h + e_{ik}^s\Big],\end{aligned}$$

where $\vec{\mathbf{h}} = \{\mathbf{h}_i, i \in \mathcal{N}\}$ and $\vec{\mathbf{v}} = \{\mathbf{v}_j, j \in \mathcal{M}\}$. Therefore, the difference between objectives of the regulator and any hospital i does not depend on the hospital's actions $\mathbf{h}_i$, and the difference between

objectives of the regulator and any PAC provider j does not depend on the PAC provider's actions $\mathbf{v}_j$. This implies that the equilibrium actions are equal to the first-best actions; we omit the proof as it is similar to that for Theorem 1. Plugging in the first-best actions one can verify that all hospitals and PAC providers break even in equilibrium. $\square$

Provider-specific investments. Assume that each hospital $i \in \mathcal{N}$ makes a *total* cost of investments $H_{ij} \in [0, \Gamma]$, and each PAC provider $j \in \mathcal{N}$ makes a *total* cost of investments $F_{ij} \in [0, \Gamma]$, to reduce the PAC treatment cost per patient discharged from hospital i to PAC provider j, as denoted by $C^s(H_{ij}, F_{ij}) \in \mathbb{R}_+$; all other model components remain identical to those introduced in Section 3. In this case, the objective of hospital i is given by (1), where

$$\mathcal{C}_i^h(\mathbf{h}_i) = p_i \left[C^h(a_i^h) + a_i^h\right] + \sum_{j \in \mathcal{M}} H_{ij} + \sum_{j \in \mathcal{M}} p_{ij} \left[R(e_{ij}^h, e_{ij}^s)\xi^h + e_{ij}^h\right] \tag{A-123}$$

represents the total cost of the hospital. The objective of PAC provider j is given by (3), where

$$\mathcal{C}_j^s(\mathbf{s}_j) = \sum_{i \in \mathcal{N}} F_{ij} + \sum_{i \in \mathcal{N}} p_{ij} \Big[C^s(H_{ij}, F_{ij}) + R(e_{ij}^h, e_{ij}^s)\xi^s + e_{ij}^s\Big] \tag{A-124}$$

represents the PAC provider's total cost. The only difference from our original model in (2) and (4) is that, to reduce the PAC cost, hospital i and PAC provider j respectively make *provider-specific* investments at lump-sum costs H_{ij} and F_{ij} independent from the patient volume.

To incorporate this change into the payment model, we define

$$\bar{H}_i = \frac{\sum\limits_{k \in \mathcal{N}_i} \sum\limits_{j \in \mathcal{M}} H_{kj}}{N-1}, \tag{A-125}$$

as the average investment to reduce the PAC cost of all hospitals, excluding hospital i, and

$$\bar{F}_j = \frac{\sum\limits_{i \in \mathcal{N}} \sum\limits_{k \in \mathcal{M}_j} F_{ik}}{M-1}, \tag{A-126}$$

the average investment of all PAC providers, excluding provider j. The payment amounts to hospital i and PAC provider j are given respectively by:

$$T_i^h = p_i \Big[\bar{C}_i^h + \bar{a}_i^h + \bar{e}_i^h + \bar{R}_i^h \xi^h\Big] + \bar{H}_i + \sum_{j \in \mathcal{M}} p_{ij} \left[\left(\bar{C}_i^{sh} - C^s(H_{ij}, F_{ij})\right) + \left(\bar{R}_i^h - R(e_{ij}^h, e_{ij}^s)\right)\xi^s\right], \tag{A-127}$$

$$T_j^s = \sum_{i \in \mathcal{N}} p_{ij} \Big[\bar{C}_j^s + \bar{e}_j^s + \bar{R}_j \xi^s\Big] + \bar{F}_j + \sum_{i \in \mathcal{N}} p_{ij} \left(\bar{R}_j - R(e_{ij}^h, e_{ij}^s)\right)\xi^h, \tag{A-128}$$

where benchmarks $\bar{H}_i$ and $\bar{F}_j$ are defined as in (A-125)-(A-126), respectively, and other benchmark parameters (i.e., $\bar{a}_i^h, \bar{e}_i^h, \bar{C}_i^{sh}, \bar{R}_i^h, \bar{C}_j^s, \bar{R}_j, \bar{e}_j^s$) are defined as in Section 4.1. Compared to the payment model in Section 5, here we modify provider reimbursement for the costs of investments to reduce

the PAC cost, i.e., from $\sum_{j\in\mathcal{M}} p_{ij} b_i^h$ for hospital i and $\sum_{i\in\mathcal{N}} p_{ij}\bar{b}_j^s$ for PAC provider j to $\bar{H}_i$ and $\bar{F}_j$, respectively, to reflect the change in cost structure (i.e., variable versus fixed).

The total social welfare W is given by (5), where $\mathcal{C}_i^h$ and $\mathcal{C}_j^s$ are defined as in (A-123)-(A-124), respectively, for each $i\in\mathcal{N}$ and $j\in\mathcal{M}$. We adapt Assumption A-1 to ensure that the first-best actions and each provider's best response are uniquely determined by the FOCs of W and provider objective, respectively. Next, we demonstrate that our payment model induces the first-best actions.

Proposition A-5. *If the regulator uses* (A-127) *to reimburse hospitals and* (A-128) *to reimburse PAC providers, then the unique Nash equilibrium is for each each hospital $i\in\mathcal{N}$ and PAC provider $j\in\mathcal{M}$ to pick first-best actions $a_i^h = a_h^*, H_{ij} = H_{ij}^*, e_{ij}^h = e_h^*$, and $F_{ij} = F_{ij}^*, e_{ij}^s = e_s^*$, respectively.*

Similarly, we can demonstrate that when hospitals and PAC providers incur fixed costs of investments for reducing readmissions, our proposed payment model, with appropriate adjustments, continues to induce first-best actions.

Proof: We continue to adopt Assumption A-1 with part (ii) adapted to (A-101) and for each i,j such that $p_{ij}>0$,

$$\lim_{H\downarrow 0}\frac{\partial C^s}{\partial H} < -\frac{1}{p_{ij}} < \lim_{H\uparrow\Gamma}\frac{\partial C^s}{\partial H} \text{ for any } F\in[0,\Gamma], \tag{A-129}$$

$$\lim_{F\downarrow 0}\frac{\partial C^s}{\partial F} < -\frac{1}{p_{ij}} < \lim_{F\uparrow\Gamma}\frac{\partial C^s}{\partial F} \text{ for any } H\in[0,\Gamma]. \tag{A-130}$$

Plugging (A-123)-(A-124) in (5), we obtain

$$W = v - \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} p_{ij}\Big[C^h\left(a_i^h\right) + a_i^h + C^s(H_{ij},F_{ij}) + R(e_{ij}^h,e_{ij}^s)(\xi^h+\xi^s) + e_{ij}^h + e_{ij}^s\Big] - \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}}\Big[H_{ij}+F_{ij}\Big].$$

Similar to the proof of Lemma A-1, it is straightforward to verify that the first-best actions uniquely exist and are determined by the FOCs of total welfare W, i.e., $a_i^h = a_h^*$ for each hospital $i\in\mathcal{N}$, and for each PAC provider $j\in\mathcal{M}$ such that $p_{ij}>0$, $H_{ij}=H_{ij}^*$, $e_{ij}^h = e_h^*$, $F_{ij}=F_{ij}^*$, and $e_{ij}^s = e_s^*$, where $a_h^*, e_h^*, e_s^*\in(0,\Gamma)$ are defined in (6), (9), (10), and $H_{ij}^*, F_{ij}^*\in(0,\Gamma)$ are as follows:

$$1 + p_{ij}\frac{\partial C^s\left(H_{ij}^*,F_{ij}^*\right)}{\partial H} = 0, \tag{A-131}$$

$$1 + p_{ij}\frac{\partial C^s\left(H_{ij}^*,F_{ij}^*\right)}{\partial F} = 0. \tag{A-132}$$

Next, we prove that, under our proposed payment model, a unique equilibrium exists in which providers choose the first-best actions. By (1), (A-123), and (A-127), hospital i's objective is

$$\Pi_i^h(\mathbf{h}_i) = \sum_{j\in\mathcal{M}} p_{ij}\Big[\bar{C}_i^h + \bar{a}_i^h + \bar{R}_i^h(\xi^h+\xi^s) + \bar{C}_i^{sh} + \bar{e}_i^h\Big] + \bar{H}_i - \sum_{j\in\mathcal{M}} H_{ij}$$

$$- \sum_{j\in\mathcal{M}} p_{ij}\Big[C^h(a_i^h) + a_i^h + R(e_{ij}^h, e_{ij}^s)(\xi^h+\xi^s) + C^s(H_{ij}, F_{ij}) + e_{ij}^h\Big],$$

where $\mathbf{h}_i = (a_i^h, H_{ij}, e_{ij}^h, j\in\mathcal{M})$. By (3), (A-124) and (A-128), PAC provider j's objective is

$$\Pi_j^s(\mathbf{v}_j) = \sum_{i\in\mathcal{N}} p_{ij}\Big[\bar{C}_j^s + \bar{R}_j(\xi^h+\xi^s) + \bar{e}_j^s\Big] + \bar{F}_j - \sum_{i\in\mathcal{N}} F_{ij} - \sum_{i\in\mathcal{N}} p_{ij}\Big[C^s(H_{ij}, F_{ij}) + R(e_{ij}^h, e_{ij}^s)(\xi^h+\xi^s) + e_{ij}^s\Big],$$

where $\mathbf{v}_j = (F_{ij}, e_{ij}^s, i\in\mathcal{N})$. Subtracting each objective by W, we obtain

$$\begin{aligned}\Pi_i^h(\mathbf{h}_i) - W(\vec{\mathbf{h}}, \vec{\mathbf{v}}) = &\sum_{j\in\mathcal{M}} p_{ij}\Big[\bar{C}_i^h + \bar{a}_i^h + \bar{R}_i^h(\xi^h+\xi^s) + \bar{C}_i^{sh} + \bar{e}_i^h\Big] + \bar{H}_i + \sum_{j\in\mathcal{M}} p_{ij}e_{ij}^s - \upsilon + \sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}} H_{kj} \\ &+ \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} F_{ij} + \sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}} p_{kj}\Big[C^h(a_k^h) + a_k^h + C^s(H_{kj}, F_{kj}) + R(e_{kj}^h, e_{kj}^s)(\xi^h+\xi^s) + e_{kj}^h + e_{kj}^s\Big],\end{aligned}$$

and

$$\begin{aligned}\Pi_j^s(\mathbf{v}_j) - W(\vec{\mathbf{h}}, \vec{\mathbf{v}}) = &\sum_{i\in\mathcal{N}} p_{ij}\Big[\bar{C}_j^s + \bar{R}_j(\xi^h+\xi^s) + \bar{e}_j^s\Big] + \bar{F}_j + \sum_{i\in\mathcal{N}} p_{ij}\Big[C^h(a_i^h) + a_i^h + e_{ij}^h\Big] - \upsilon + \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} H_{ij} \\ &+ \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} F_{ik} + \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} p_{ik}\Big[C^h(a_i^h) + a_i^h + C^s(H_{ik}, F_{ik}) + R(e_{ik}^h, e_{ik}^s)(\xi^h+\xi^s) + e_{ik}^h + e_{ik}^s\Big],\end{aligned}$$

where $\vec{\mathbf{h}} = \{\mathbf{h}_i, i\in\mathcal{N}\}$ and $\vec{\mathbf{v}} = \{\mathbf{v}_j, j\in\mathcal{M}\}$. Therefore, the difference between objectives of the regulator and any hospital i does not depend on the hospital's actions $\mathbf{h}_i$, and the difference between objectives of the regulator and any PAC provider j does not depend on the PAC provider's actions $\mathbf{v}_j$. This implies that the equilibrium actions are equal to the first-best actions; we omit the proof as it is similar to that for Theorem 1. □

I. Non-identical providers and risk adjustment

To implement coordinating reimbursement schemes in real-world scenarios, it is essential to account for heterogeneity among providers and patients across various dimensions, such as geographic and demographic factors. While we previously assumed that the regulator could identify identical providers or pairs of identical providers, and that all patients were identical, this assumption may not hold in practice. Nevertheless, if the heterogeneity factors can be observed by the regulator and are exogenous to the providers, the proposed scheme can be modified to accommodate this heterogeneity. This approach follows the framework outlined in Shleifer (1985), Savva et al. (2019), and Arifoğlu et al. (2021).

The CJR model also employs a risk-adjustment procedure based on a linear regression model with variables on patient characteristics and health status to adjust the target price for different patients. This risk-adjustment approach aims to account for variations in expected costs associated with patient complexity and comorbidities. In the following discussion, we illustrate this concept

by considering the case where each type of provider differs along one characteristic. However, it is important to note that all the results can be generalized to incorporate multiple characteristics per provider, as discussed in Shleifer (1985) and Savva et al. (2019).

To demonstrate, assume that the readmission probability R is a function of the investments of the hospital e_i^h, $i \in \mathcal{N}$, and the PAC provider e_j^s, $j \in \mathcal{M}$, (as in Section 5) as well as the observable exogenous characteristics of the hospital β_h, of the PAC provider β_s, and the patient β_p. Consequently, the first-best outcomes are dependent on the specific characteristics β_h, β_s, and β_p, see (6)–(10)).

In this modified approach, instead of using average values, as shown in Section 4.1, the regulator estimates $\bar{R}_i^h$ and $\bar{R}_j^s$, for all $i \in \mathcal{N}$ and $j \in \mathcal{M}$. These estimates are obtained through an estimation procedure, such as linear regression, based on observed readmission probabilities and the corresponding observable characteristics of hospitals β_i^h, $i \in \mathcal{N}$, the PAC provider characteristics β_j^s, $j \in \mathcal{M}$, and the patient β^p. Following the proof provided for Theorem 1, it can be shown that all providers will take first-best actions under this revised scheme.

Moreover, if the estimation procedure accurately captures the true values, the targets set at the estimated $\bar{R}_i^h$'s and $\bar{R}_j^s$'s will result in all providers achieving a break-even outcome. This implies that the reimbursement scheme aligns with the actual costs incurred by providers, ensuring a fair and balanced outcome.

J. Alternative cost modeling

Assume that the cost of acute care $C^h : [0,\Gamma]^N \to \mathbb{R}_+$, is a function of the hospital's multidimensional effort τ^h, and the dollar cost of this effort is $a^h(\tau^h) : [0,\Gamma]^N \to [0,\Gamma]$. Similarly, assume that the PAC cost $C^s : [0,\Gamma]^N \times [0,\Gamma]^N \to \mathbb{R}_+$ depends on the multidimensional effort made by the hospital ψ^h with dollar cost $b^h(\psi^h) : [0,\Gamma]^N \to [0,\Gamma]$, and the multidimensional effort of the PAC provider ψ^s with dollar cost $b^s(\psi^s) : [0,\Gamma]^N \to [0,\Gamma]$. The probability of patient readmission $R : [0,\Gamma]^N \times [0,\Gamma]^N \to [0,1]$ is a function of the multidimensional effort made by the hospital κ^h with dollar cost $e^h(\kappa^h) : [0,\Gamma]^N \to [0,\Gamma]$, and the multidimensional effort made by the PAC provider κ^s with dollar cost $e^s(\kappa^s) : [0,\Gamma]^N \to [0,\Gamma]$. We subscript each effort by the corresponding provider index(es), e.g., τ_i^h for hospital i's effort to reduce the cost of acute care and ψ_{ij}^h for hospital i's effort to reduce PAC cost of patients discharged to PAC provider j. We use $\mathbf{h}_i = (\tau_i^h, \psi_{ij}^h, \kappa_{ij}^h, j \in \mathcal{M})$ to denote the actions of hospital i and $\mathbf{s}_j = (\psi_{ij}^s, \kappa_{ij}^s, i \in \mathcal{N})$ to denote the actions of PAC provider j.

Remark A-1. In keeping with the notation in our original model, we use C^a, C^s, R to represent the costs and readmission functions. It is important to note, however, that their arguments are unobservable efforts to reduce costs or the readmission probability as opposed to observable effort costs in our original model.

In this case, the objective of hospital i is

$$\Pi_i^h(\mathbf{h}_i) = T_i^h - \mathcal{C}_i^h(\mathbf{h}_i), \tag{A-133}$$

where $\mathcal{C}_i^h(\mathbf{h}_i)$ represents the total cost of the hospital, defined as:

$$\mathcal{C}_i^h(\mathbf{h}_i) = p_i \left[C^h(\tau_i^h) + a^h(\tau_i^h)\right] + \sum_{j\in\mathcal{M}} p_{ij} \left[R(\kappa_{ij}^h, \kappa_{ij}^s)\xi^h + b^h(\psi_{ij}^h) + e^h(\kappa_{ij}^h)\right], \tag{A-134}$$

Similarly, the objective of PAC provider j is

$$\Pi_j^s(\mathbf{s}_j) = T_j^s - \mathcal{C}_j^s(\mathbf{s}_j), \tag{A-135}$$

where $\mathcal{C}_j^s(\mathbf{s}_j)$ represents the total cost of the PAC provider, defined as:

$$\mathcal{C}_j^s(\mathbf{s}_j) = \sum_{i\in\mathcal{N}} p_{ij} \left[C^s(\psi_{ij}^h, \psi_{ij}^s) + R(\kappa_{ij}^h, \kappa_{ij}^s)\xi^s + b^s(\psi_{ij}^s) + e^s(\kappa_{ij}^s)\right]. \tag{A-136}$$

Transfer payments to hospital i and PAC provider j, i.e., T_i^h and T_j^s, under all payment models previously studied remain unchanged. They are specified in (14)-(15) for bundled payment, (20) and (21) for our proposed payment model, and (23)-(22) for the CJR-type payment model, respectively, with the understanding that costs (care and effort) and the readmission probability are functions of unobservable efforts as in (A-134) and (A-136) above.

In what follows, we show that this alternative cost model is equivalent to our original model in two steps: (i) we prove that any set of first-best efforts in the alternative cost model yield the first-best costs/investments of efforts in our original model; (ii) we prove that, under each payment model (i.e., the bundled payment model, our proposed payment model, and CJR-type payment model), in any equilibrium of the alternative cost model, each provider's optimal efforts induce a cost of effort equal to the optimal investment (potentially a best response function) in our original model.

(i) Plugging (A-134) and (A-136) into (5), we obtain the following expression of total welfare

$$\begin{aligned} W = &\upsilon - \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} p_{ij}\Big[C^h(\tau_i^h) + a^h(\tau_i^h) + C^s(\psi_{ij}^h, \psi_{ij}^s) + b^h(\psi_{ij}^h) + b^s(\psi_{ij}^s) \\ &+ R(\kappa_{ij}^h, \kappa_{ij}^s)(\xi^h + \xi^s) + e^h(\kappa_{ij}^h) + e^s(\kappa_{ij}^h)\Big]. \end{aligned} \tag{A-137}$$

Since W is separable in $\tau_i^h, (\psi_{ij}^h, \psi_{ij}^s), (\kappa_{ij}^h, \kappa_{ij}^s)$, $i \in \mathcal{N}$, $j \in \mathcal{M}$, the socially optimal (i.e., first-best) efforts are determined by solving the following problems separately:

$$\begin{aligned} &\underset{\tau_i^h\in[0,\Gamma]^N}{\text{minimize}}\ C^h(\tau_i^h) + a^h(\tau_i^h),\ i \in \mathcal{N}, \\ &\underset{\psi_{ij}^h, \psi_{ij}^s\in[0,\Gamma]^N}{\text{minimize}}\ C^s(\psi_{ij}^h, \psi_{ij}^s) + b^h(\psi_{ij}^h) + b^s(\psi_{ij}^s),\ i \in \mathcal{N}, j \in \mathcal{M},\ \text{s.t. } p_{ij} > 0, \end{aligned}$$

$$\underset{\kappa_{ij}^h,\kappa_{ij}^s\in[0,\Gamma]^N}{\text{minimize}} R(\kappa_{ij}^h,\kappa_{ij}^s)(\xi^h+\xi^s)+e^h(\kappa_{ij}^h)+e^s(\kappa_{ij}^h),\ i\in\mathcal{N},j\in\mathcal{M},\ \text{s.t. } p_{ij}>0.$$

Since the objective function and constraint set in each problem do not depend on hospital and PAC provider indexes (i,j), the first-best efforts can be derived by solving these generic problems:

$$\underset{\tau^h\in[0,\Gamma]^N}{\text{minimize}} C^h(\tau^h)+a^h(\tau^h), \tag{A-138}$$

$$\underset{\psi^h,\psi^s\in[0,\Gamma]^N}{\text{minimize}} C^s(\psi^h,\psi^s)+b^h(\psi^h)+b^s(\psi^s), \tag{A-139}$$

$$\underset{\kappa^h,\kappa^s\in[0,\Gamma]^N}{\text{minimize}} R(\kappa^h,\kappa^s)(\xi^h+\xi^s)+e^h(\kappa^h)+e^s(\kappa^h). \tag{A-140}$$

We assume that all functions in the objectives are twice differentiable. This and compactness of the constraint sets ensure the existence of a solution (i.e., a set of first-best efforts) to each of the generic problem above. Next, we prove (i) for problem (A-139), i.e., any first-best efforts (ψ^h,ψ^s) yield $b^h(\psi^h)=b_h^*$ and $b^s(\psi^s)=b_s^*$; the proof for problems (A-138) and (A-140) is similar and omitted for brevity.

We assume that $b^h(\psi^h)$ and $b^s(\psi^s)$ are onto, i.e., the range consists of all points in $[0,\Gamma]$, and construct the following minimization problem at each fixed $(B^h,B^s)\in[0,\Gamma]^2$.

$$\underset{\psi^h,\psi^s\in[0,\Gamma]^N}{\text{minimize}} C^s(\psi^h,\psi^s), \tag{A-141}$$

$$\text{s.t. } b^h(\psi^h)\leqslant B^h, b^s(\psi^s)\leqslant B^s. \tag{A-142}$$

Since $b^h(\psi^h)$ and $b^s(\psi^s)$ are continuous, increasing, and onto functions, the constraint set

$$(\psi^h,\psi^s)\in\mathbf{\Psi}_{\mathsf{w}}(B^h,B^s)=\{\psi^h,\psi^s\in[0,\Gamma]^N|b^h(\psi^h)\leqslant B^h,b^s(\psi^s)\leqslant B^s\}$$

is a compact-valued correspondence $\mathbf{\Psi}_{\mathsf{w}}:[0,\Gamma]^2\rightrightarrows[0,\Gamma]^{2N}$ satisfying the closed graph property and hence is upper hemicontinuous. In Lemma A-5 below, we prove that $\mathbf{\Psi}_{\mathsf{w}}(B^h,B^s)$ is lower hemicontinuous, and hence continuous on $[0,\Gamma]^2$. By the Maximum Theorem, the minimized value of problem (A-141)-(A-142) as denoted by $\hat{C}^s(B^h,B^s)$, is a continuous function on $[0,\Gamma]^2$. Thus, problem

$$\underset{B^h,B^s\in[0,\Gamma]}{\text{minimize}} \hat{C}^s(B^h,B^s)+B^h+B^s \tag{A-143}$$

has a solution denoted by B_h^*,B_s^*. We assume the solution to be uniquely determined by FOCs, which is ensured by convexity and boundary conditions on $\hat{C}^s$ identical to those for C^s in our original model.[8] By Lemma A-1, $(B_h^*,B_s^*)=(b_h^*,b_s^*)$ given by (7)-(8) wherein $C^s(b^h,b^s)=\hat{C}^s(b^h,b^s)$.

To complete the proof for this part, it suffices to show that any first-best efforts (ψ^h,ψ^s) yield $b^h(\psi^h)=B_h^*$ and $b^s(\psi^s)=B_s^*$. Suppose not, we have $(b^h(\psi^h),b^s(\psi^s))\neq(B_h^*,B_s^*)$ and there are two

[8] Strict convexity of $\hat{C}^s(B^h,B^s)$ is ensured by strict convexity of $C^s(\psi^h,\psi^s)$; see Theorem 9.17 of Sundaram (1996).

cases: i) (ψ^h, ψ^s) is a solution to problem (A-141)-(A-142) for some $(B^{h'}, B^{s'}) \neq (B_h^*, B_s^*)$. In this case,

$$\begin{aligned} &C^s(\psi^h, \psi^s) + b^h(\psi^h) + b^s(\psi^s) = \hat{C}^s(B^{h'}, B^{s'}) + B^{h'} + B^{s'} \\ >&\hat{C}^s(B_h^*, B_s^*) + B_h^* + B_s^* = C^s(\psi_h^*, \psi_s^*) + b^h(\psi_h^*) + b^s(\psi_s^*), \end{aligned} \tag{A-144}$$

where the inequality holds because problem (A-143) has a unique solution (B_h^*, B_s^*) at which problem (A-141)-(A-142) is solved by (ψ_h^*, ψ_s^*). Furthermore, $(\psi^h, \psi^s) \neq (\psi_h^*, \psi_s^*)$ because

$$(b^h(\psi^h), b^s(\psi^s)) = (B^{h'}, B^{s'}) \neq (B_h^*, B_s^*) = (b^h(\psi_h^*), b^s(\psi_s^*)),$$

where the equalities hold because constraints in (A-142) are binding at at any solution of problem (A-141)-(A-142) for each $(B^h, B^s) \in [0, \Gamma]^2$, since $C^s(\psi^h, \psi^s), b^h(\psi^h), b^s(\psi^s)$ are strictly increasing in each argument. By (A-144) and $(\psi^h, \psi^s) \neq (\psi_h^*, \psi_s^*)$, (ψ^h, ψ^s) is not a solution of problem (A-139), contradicting our assumption that (ψ^h, ψ^s) is a first-best solution.

ii) (ψ^h, ψ^s) is not a solution to problem (A-141)-(A-142) for all $(B^h, B^s) \in [0, \Gamma]^2$. Let $(\psi^{h'}, \psi^{s'})$ denote a solution of problem (A-141)-(A-142) at $(B^h, B^s) = (b^h(\psi^h), b^s(\psi^s))$. We have $b^h(\psi^{h'}) = b^h(\psi^h)$ and $b^s(\psi^{s'}) = b^s(\psi^s)$ since constraints in (A-142) are binding at any solution of problem (A-141)-(A-142).

$$C^s(\psi^h, \psi^s) + b^h(\psi^h) + b^s(\psi^s) = C^s(\psi^h, \psi^s) + b^h(\psi^{h'}) + b^s(\psi^{s'}) > C^s(\psi^{h'}, \psi^{s'}) + b^h(\psi^{h'}) + b^s(\psi^{s'}),$$

where the inequality holds because $(\psi^{h'}, \psi^{s'})$ is a solution of problem (A-141)-(A-142) at $(B^h, B^s) = (b^h(\psi^h), b^s(\psi^s)) = (b^h(\psi^{h'}), b^s(\psi^{s'}))$ whereas (ψ^h, ψ^s) is not. Thus, (ψ^h, ψ^s) is not a solution of problem (A-139), contradicting our assumption that (ψ^h, ψ^s) is a first-best solution.

Lemma A-5. *$\Psi_{\mathsf{w}}(B^h, B^s)$ is lower hemicontinuous at each fixed $B^h, B^s \in [0, \Gamma]$.*

Proof of Proposition A-5: We will invoke the sequential characterization of lower hemicontinuity; see Proposition 4 on p. 299 of Ok (2007). That is, for each $(B^h, B^s) \in [0, \Gamma]^2$ and any $(\psi^h, \psi^s) \in \Psi_{\mathsf{w}}(B^h, B^s)$, we will, for any sequence $(B_m^h, B_m^s) \in [0, \Gamma]^2$ converging to (B^h, B^s) as $m \to \infty$, construct a sequence $(\psi_m^h, \psi_m^s) \in \Psi_{\mathsf{w}}(B_m^h, B_m^s)$ that converges to (ψ^h, ψ^s).

We define ψ_m^h as the solution to the following minimization problem

$$\underset{\psi \in [0,\Gamma]^N}{\text{minimize}} \ |\psi - \psi^h|, \tag{A-145}$$

$$\text{s.t. } b^h(\psi) \leqslant \min\{B_m^h, b^h(\psi^h)\}. \tag{A-146}$$

A unique solution exists because the objective is strictly convex and the constraint set is compact and non-empty. Also we have (i) $b^h(\psi_m^h) \leqslant B_m^h$ by (A-146) and $\min\{B_m^h, b^h(\psi^h)\} \leqslant B_m^h$, and (ii)

$\psi_m^h \to \psi^h$ as $m \to \infty$ as proven in the next paragraph. Similarly, we can construct $\psi_m^s \in \{\psi \in [0,\Gamma]^N | b^s(\psi) \leqslant B_m^s\}$ that converges to ψ^s as $m \to \infty$.

When $b^h(\psi^h) < B^h$, there exists $M > 0$ such that $\min\{B_m^h, b^h(\psi^h)\} = b^h(\psi^h)$ for all $m > M$. Plugging into (A-146), we have $\psi_m^h = \psi^h$ for all $m > M$, so (ii) trivially holds. When $b^h(\psi^h) = B^h$, we prove (ii) by contradiction. Suppose that ψ_m^h does not converge to ψ^h, then there exists $M > 0$ such that for all $m > M$, $|\psi_m^h - \psi^h| > \epsilon > 0$. By continuity of $b^h(\cdot)$, we have $|b^h(\psi_m^h) - b^h(\psi^h)| > \epsilon_b > 0$ for all $m > M$. Since $b^h(\psi_m^h) \leqslant b^h(\psi^h)$ for each m by (A-146), we have $b^h(\psi_m^h) < b^h(\psi^h) - \epsilon_b = B^h - \epsilon_b$ for all $m > M$. By convergence of B_m^h to B^h, there exists $M' > 0$ such that $B_m^h > B^h - \epsilon_b/2$ for all $m > M'$. We now prove that ψ_m^h is not the solution to problem (A-145)-(A-146) for all $m > \max\{M, M'\}$. Define $\psi_m^{h'} = \psi_m^h + \epsilon_h(\psi^h - \psi_m^h)$, where $\epsilon_h > 0$ and is small enough such that $b^h(\psi_m^{h'}) < B^h - \epsilon_b/2$; existence of ϵ_h is ensured by continuity of $b^h(\cdot)$ and $b^h(\psi_m^h) < B^h - \epsilon_b$. We also have $\psi_m^{h'} = \epsilon_h \psi^h + (1-\epsilon_h)\psi_m^h \in [0,\Gamma]^N$ for all $\epsilon_h \in (0,1)$ due to $\psi_m^h, \psi^h \in [0,\Gamma]^N$. Thus, all constraints in problem (A-145)-(A-146) are satisfied at $\psi_m^{h'}$. In addition, we have $|\psi_m^{h'} - \psi^h| = (1-\epsilon_h)|\psi_m^h - \psi^h| < |(\psi_m^h - \psi^h)|$ by $\epsilon_h \in (0,1)$, so $\psi_m^{h'}$ is a strict improvement over ψ_m^h which thus is sub-optimal. □

(ii) Expanded provider objectives under the bundled payment model and the CJR-type payment model (which mathematically includes our proposed payment model at $\theta = 1$) are given by (A-14)-(A-15) and (A-28)-(A-29), respectively, with the understanding that here costs (care and effort) and readmission probability are functions of unobservable efforts. Since each provider's objective is strictly decreasing in her own costs of care and efforts and readmission probability, hospital i's equilibrium actions minimize $\Pi_i^h(\mathbf{h}_i)$ with ψ_j^s and κ_j^s solving the following problems:

$$\underset{\psi^s \in [0,\Gamma]^N}{\text{minimize}}\ C^s(\psi^h, \psi^s), \tag{A-147}$$

$$\text{s.t. } b^s(\psi^s) = B^s \text{ for each } B^s \in [0,\Gamma], \tag{A-148}$$

$$\underset{\kappa^s \in [0,\Gamma]^N}{\text{minimize}}\ R(\kappa^h, \kappa^s), \tag{A-149}$$

$$\text{s.t. } e^s(\kappa^s) = K^s \text{ for each } K^s \in [0,\Gamma]. \tag{A-150}$$

Any ψ^s that is not a solution to problem (A-147)-(A-148) cannot emerge in equilibrium because PAC provider j can strictly increase $\Pi_j^s(\mathbf{s}_j)$ by choosing a minimizer of problem (A-147)-(A-148), which strictly decreases $C^s(\psi^h, \psi^s)$ and does not change $b^s(\psi^s)$. Similarly, in equilibrium κ^s is a solution to problem (A-149)-(A-150).

Denote the minimized value functions of problems (A-147)-(A-148) and (A-149)-(A-150) by $\hat{C}^{sh}(\psi^h, B^s)$ and $\hat{R}^h(\kappa^h, K^s)$, respectively, which are continuous with proof similar to that for $\hat{C}^s(B^h, B^s)$. In what follows, we prove $e^h(\kappa_i^h) = z_h(K^s)$ for each $K^s \in [0,\Gamma]$ in equilibrium under the CJR-type payment model, where z_h is the hospital's best response function in our original model

and is determined by (A-34) with $\partial R(e^h, e^s)/\partial e^h$ replaced by $\partial \hat{R}(K^h, K^s)/\partial K^h$, where $\hat{R}(K^h, K^s)$ is the minimized value function of the following problem:

$$\underset{\kappa^h \in [0,\Gamma]^N}{\text{minimize}} \ \hat{R}^h(\kappa^h, K^s), \tag{A-151}$$

$$\text{s.t. } e^h(\kappa^h) = K^h. \tag{A-152}$$

Since hospital objective (A-28) is separable in τ^h, ψ^h, κ^h. the equilibrium κ^h can be derived by solving the following generic problem for each PAC provider with cost of effort $K_s \in [0, \Gamma]$:

$$\underset{\kappa^h \in [0,\Gamma]^N}{\text{minimize}} \ \hat{R}^h(\kappa^h, K^s)((2-\theta)\xi^h + \xi^s) + e^h(\kappa^h). \tag{A-153}$$

Any minimizer κ^h must be a solution to problem (A-151)-(A-152). If not, the hospital can strictly lower the objective in (A-153) by choosing a solution to problem (A-151)-(A-152) which decreases the readmission probability $\hat{R}^h$ and does not affect the cost of effort e^h. Thus, problem (A-153) is equivalent to

$$\underset{K^h \in [0,\Gamma]}{\text{minimize}} \ \hat{R}(K^h, K^s)((2-\theta)\xi^h + \xi^s) + K^h. \tag{A-154}$$

There exists a solution to this problem, because the constraint set is compact and the objective is continuous. In particular, continuity of $\hat{R}(K^h, K^s)$ follows by continuity of $R(\kappa^h, \kappa^s)$ and applying the Maximum Theorem to problems (A-148)-(A-150) and (A-151)-(A-152). We assume that the solution is unique and determined by FOC, i.e.,

$$\frac{\partial \hat{R}(K^h, K^s)}{\partial K^h}((2-\theta)\xi^h + \xi^s) + 1 = 0.$$

Since this is identical to the definition of z_h in (A-34), in equilibrium we have $e^h(\kappa^h) = K^h = z_h(K^s)$.

In a similar proof, one can show that all other best response PAC costs of efforts, for hospitals and PAC providers under the CJR-type payment model and the bundled payment model, are identical to the corresponding best response functions in our original model with PAC cost and readmission functions defined appropriately; the hospital decision on the effort to reduce acute care cost is identical to that in the first-best problem (A-138). We omit technical details to avoid repetition.