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

Tolga Tezcan
Jones Graduate School of Business, Rice University, 6100 Main Street, Houston, TX 77005, USA
tolga.tezcan@rice.edu

Diagnosis-Related Group (DRG) and bundled payment models are widely used in healthcare reimbursement by entities like the Centers for Medicare & Medicaid Services (CMS) and insurance companies. However, these models were primarily designed for conditions managed by a single healthcare provider in a centralized manner, often overlooking the complexities of cases requiring post-acute care (PAC) following an initial hospital stay. This can result in inadequate incentives for effective care coordination between hospitals and PAC providers, especially when treatment decisions are decentralized.

Motivated by the Comprehensive Care for Joint Replacement (CJR) payment model recently introduced by CMS, which holds hospitals accountable for the quality and cost of the entire CJR episode, including the cost of PAC, we propose simple payment models that incentivize hospitals and PAC providers to collaboratively enhance the cost efficiency and quality of care for such conditions. Our approach extends traditional payment models by introducing performance targets for all providers, encompassing the entire care episode. Using a game-theoretical model, we demonstrate that the proposed payment model elicits socially optimal actions from all providers, under various assumptions. Importantly, our models do not require detailed knowledge of the hospital-PAC network structure but rely solely on observed cost and quality outcomes within the entire system.

Furthermore, while the CJR payment model represents a positive step forward, our analysis reveals potential areas for improvement. Specifically, we suggest that holding both hospitals and PAC providers financially accountable, instead of solely focusing on hospitals, would yield further enhancements in the care delivery model.

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 cost-effective 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. This payment amount serves as a benchmark, fostering indirect competition among healthcare providers to improve cost efficiency. Consequently, providers who manage to operate more efficiently than the average are able to generate positive margins from their services. The implementation of DRG-based payment systems, particularly through the prospective payment system initiated by CMS in 1983, has resulted in reduced hospital spending growth for Medicare (Davis and Rhodes 1988). Furthermore, according to theoretical evidence, these payment models elicit socially optimal efforts from providers towards cost reduction (Shleifer 1985).

However, despite the success in incentivizing cost reduction, there is a legitimate concern that DRG-based payment models may potentially encourage providers to prioritize cost savings over the quality of care. To address this concern, regulators have introduced outcome-based DRG 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. The Affordable Care Act (ACA) played a pivotal role in promoting the implementation of outcome-based payment models by CMS, aiming to align payment with the quality of care provided (Chernew et al. 2020). 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 (2023b). 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: DRG 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 post-acute care (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). 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 (EHRs) 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 DRG-based payments 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. 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) payment 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 their spending during the performance year is below the target price, while repayments may be required if spending exceeds the target price (CMS 2018, Department of Health and Human Services 2021). However, PAC providers continue to be paid using a DRG- and outcome-based payment model.

The CJR program draws on concepts from bundled-payment models but incorporates unique features. While traditional bundled payment models involve a single payment to cover the entire episode of care (a feature CJR model adopted) the CJR model utilizes separate payments to hospitals and PAC providers, recognizing them as distinct entities. However, this separation of payments also introduces challenges in determining responsibility for high costs and adverse quality outcomes throughout the entire episode of care, as opposed to cases where a single entity oversees the entire process. The complexity is further compounded by the collaborative relationships between hospitals and multiple PAC providers, making it challenging to attribute costs and quality outcomes accurately. Addressing these issues becomes crucial to prevent potential free-riding problems in multi-provider settings, as discussed in Chapter 5.3.8 of Salanie (1997).

In an effort to further incentivize care coordination, the CJR program allows hospitals to establish "gainsharing agreements" with PAC providers to share rewards and penalties. However, designing effective gainsharing contracts can be challenging, especially when there are informational asymmetries between the two providers (Gupta et al. 2021, Ghamat et al. 2021). Additionally, complications arise due to the hospital's limited ability to mandate which PAC provider a patient should choose (McGarry and Grabowski 2017). Thus, despite the potential benefits of bundled payment models and gainsharing agreements, their implementation is complex in practice.

New payment models for care coordination: In light of the challenges faced by existing payment models, we propose new payment models that provide incentives for hospitals and PAC providers to contain costs and improve quality through coordinated efforts. We develop a stylized model where the quality and cost of care provided by a PAC provider can be improved by additional effort from the hospital that provided acute care to the patient. Using a game theoretical approach, we demonstrate that our proposed payment model achieves socially optimal (first-best) outcomes in terms of both cost and quality.

One notable advantage of our payment model is its simplicity, as as it does not impose additional informational burdens on payers beyond what existing DRG and bundled payment models require. Our payment model utilizes benchmarks derived from the average cost and quality performance of all providers involved, aligning our payment models with the recent payment reforms implemented by CMS. Providers are incentivized to collaborate and enhance care quality in a cost-effective manner by comparing their joint performance to these carefully chosen benchmarks.

Another significant benefit is that our payment model does not rely on the network structure of hospitals and PAC providers, making it adaptable to various healthcare environments and conditions without necessitating modifications. Moreover, we demonstrate the model's flexibility by extending it to accommodate various modifications, such as different types of PAC providers, different coordination and cost structures, and heterogeneous hospitals and PAC entities.

The main distinction between our proposed payment model and the existing single-entity DRG and outcome-based payment models lies in the fact that each entity is held responsible for the cost and quality of the entire episode of treatment, rather than only their individual aspects. This extension builds on prior research on moral hazard in teams (Holmstrom 1982).

Furthermore, we compare the equilibrium outcomes under our proposed payment model to those under a CJR-type payment model. While CJR is a step in the right direction and has led to cost reductions (CMS 2022), we show that it does not provide sufficient incentives for achieving socially optimal outcomes. This limitation arises from the fact that only hospitals are held accountable for overall costs, while PAC providers are solely responsible for their own cost and quality. While gainsharing agreements have the potential to improve care coordination under CJR, designing effective agreements remains an ongoing challenge (Gupta et al. 2021, Ghamat et al. 2021). In contrast, our proposed payment model eliminates the reliance on such agreements, offering a promising and pragmatic alternative.

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 growing literature on designing payment models for healthcare provision in various settings (see for example So and Tang (2000), Ata et al. (2013), Jiang et al. (2012), Bavafa et al. (2021), Bastani et al. (2016), Goodman and Dai (2023)). Specifically, we focus on relative performance-based payment models, similar to those recently implemented by Centers for Medicare & Medicaid Services (CMS), and our work complements the existing literature on this topic.

Savva et al. (2019) demonstrate the effectiveness of combining endogenous benchmarks with average cost-based DRG-based payment models to reduce waiting times for emergency departments in a cost-effective manner. Arifoğlu et al. (2021), motivated by the Hospital Readmissions Reduction Program, concentrate on hospital readmissions and provide similar insights on reducing them. Additionally, Debo et al. (2021) highlights the vulnerability of DRG-based payment models to upcoding (i.e., coding and billing for more expensive services or procedures than were actually performed or rendered) incentives and offers solutions to eliminate such behavior.

The main distinction between our paper and this literature stream is the latter's primary focus on care settings where a single provider delivers care in a centralized manner, whereas our study investigates settings where care is provided by multiple entities. By investigating these multi-entity settings, we demonstrate that relative performance-based payment models can effectively achieve first-best (or socially optimal) outcomes in terms of quality and cost.

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 et al. (2016) compare providers' actions under fee-for-service with bundled payments in a model where hospitals choose the treatment intensity for heterogeneous patients considering potential treatment failures. Andritsos and Tang (2018) compare fee-for-service and bundled payment schemes in a model with readmissions in a single-entity setting. Guo et al. (2019) extend Andritsos and Tang (2018) by explicitly modeling hospitals' capacity and its impact on patient waiting times. 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. Vlachy et al. (2023) explore the impact of bundled payments on the coordination of hospitals' and physicians' decisions. Gupta and Mehrotra (2015) examine how bundled payment contract selection processes should be designed by regulators, drawing inspiration from the Bundled Payments for Care Improvement Initiative.

The primary distinction between this literature and our approach lies in our focus on relative performance-based payment models, aligning with CMS's predominant reliance on such models. 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 within the context of a CJR-type bundled payment model. 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 the CJR payment model. 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 requires additional effort 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 moral hazard in teams: Our paper makes a contribution to the broader theory of moral hazard in teams. Holmstrom (1982) demonstrates that, in team settings, individual members may be less motivated to exert their best effort due to the challenges of measuring and attributing individual performance, leading to issues of free-riding. He also establishes that certain contracts can achieve optimal effort from each team member when the regulator possesses complete information about the costs and benefits associated with individual effort.

Similarly, Mookherjee (1984) provides necessary and sufficient conditions for the optimality of rank-order based incentive schemes in a similar team setting. Although our model shares similarities with Holmstrom (1982) and Mookherjee (1984), a crucial distinction exists: the regulator has the ability to observe the costs of each entity within the care team in their models. Instead, we demonstrate that a contract based on the relative performance of similar entities induces first-best effort in equilibrium, thereby extending the main finding of Shleifer (1985) from the context where entities operate individually, to situations where entities collaborate as part of a team.

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: acute care in a hospital and subsequent PAC provided by another entity, such as skilled nursing facilities (SNFs). 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. 2019)).

To assess the influence of payment models, our model involves three key decision-makers: the regulator, hospitals, and PAC providers. The regulator oversees multiple providers in different

settings and has the authority to implement a payment model that determines the reimbursement for each provider's care services. The cost and readmission likelihood in each treatment episode depends on the efforts of the hospital, the PAC provider, and the hospital's coordination efforts with the PAC provider. With the objective of maximizing social welfare (i.e., achieving the first-best outcomes), the regulator determines the payment model. In response to the payment model, the providers select their actions to maximize their expected profit from delivering care. We establish the (pure-strategy) Nash equilibrium in a one-round game in which each player holds correct expectations about the other players' actions, with the aim of assessing the long-term impact of reimbursement schemes and comparing these outcomes to the first-best actions.

Next, we present the specific details of the objective functions for each party and establish the optimal effort levels that represent the first-best outcomes.

Care model: We define the cost of an episode of care as the sum of the acute care cost, denoted by C^h, and the PAC cost, denoted by C^s. We also introduce the readmission probability, denoted by R, which represents the likelihood of a patient being readmitted. To examine the impact of payment models on hospitals and PAC providers, we assume that these quantities are functions of the providers' efforts, measured in monetary terms.

Initially, we consider a fixed (i.e., exogenous) probability for patients needing PAC and focus solely on those patients. It is worth noting that, for patients who do not require PAC, single-entity payment models yield first-best efforts because the whole episode of care is managed by the hospital; see for example Arifoğlu et al. (2021). To further explore the model, we extend it to incorporate endogenous discharge decisions in §5.1.

In our model, the cost of acute care, denoted by $C^h : [0, \Gamma] \to \mathbb{R}^+$, is a function of the hospital's effort, a^h, to reduce costs, where $\Gamma > 0$. Similarly, the cost of PAC, denoted by $C^s : [0, \Gamma] \times [0, \Gamma] \to \mathbb{R}^+$, depends on the efforts made by the hospital, b^h, and the PAC provider, b^s, such as improved coordination through shared electronic medical record systems. Throughout, we also assume that all provider actions are bounded by $\Gamma > 0$ without loss of generality.

Additionally, the probability of patient readmission is denoted by $R : [0, \Gamma] \times [0, \Gamma] \to [0, 1]$ and is a function of the efforts of both the hospital, e^h, and the PAC provider, e^s, to minimize readmissions. We assume fixed expected costs for treating readmitted patients in both the hospital, denoted by ξ^h, and the PAC setting, denoted by ξ^s. Our model can be extended to incorporate potential provider actions to reduce the costs associated with treating readmitted patients; see §5.2. We assume that the functions C^h, C^s, and R are twice differentiable.

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 $j = 1, \ldots, M$. The probability of a patient being discharged from hospital i to PAC provider j is denoted by $p_{ij} \geqslant 0$. Throughout

the paper, we set $\mathcal{K} = \{1, \ldots, K\}$ and $\mathcal{K}_i = \mathcal{K} \setminus i$ for any given integer K and for $i = 1, \ldots, K$, for notational simplicity.

We use p_i to represent the fraction of patients receiving acute care from hospital i, where $p_i \equiv \sum_{j \in \mathcal{M}} p_{ij}$. We assume that each hospital provides care to some patients, i.e., $p_i > 0$ for all $i \in \mathcal{N}$, and we normalize the total patient population to one, i.e., $\sum_{i \in \mathcal{N}} p_i = 1$, without loss of generality. Additionally, we assume that PAC providers in the network treat some patients, denoted by $\tilde{p}_j \equiv \sum_{i \in \mathcal{N}} p_{ij} > 0$ for all $j \in \mathcal{M}$.

The effort exerted by hospital i to reduce the cost of acute care is denoted by a_i^h. To capture the efforts made by hospitals and PAC providers for care coordination, we denote PAC provider-specific efforts for cost reduction by b_{ij}^h and efforts to reduce readmissions by e_{ij}^h for hospitals. Similarly, we use e_{ij}^s to represent the effort of PAC provider j in coordinating care with hospital i. For notational simplicity, we use $\mathbf{h}_i = (a_i^h, b_{ij}^h, e_{ij}^h, j \in \mathcal{M})$ to denote the actions of hospital i and $\mathbf{s}_j = (b_{ij}^s, e_{ij}^s, i \in \mathcal{N})$ to denote the actions of PAC provider j.

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 exerts the same effort (i.e, a_i^h) for cost reduction in acute care regardless of the PAC destination, as hospitals are unlikely to make different treatment decisions based on the anticipated PAC provider. However, efforts to reduce readmissions and the cost of PAC depend on the specific PAC provider (e_{ij}^h and e_{ij}^s), as these require coordination between the two parties, such as through shared healthcare records or coordinated care in the PAC setting.

Similarly, the objective of PAC provider j, denoted by Π_j^s, for each treatment episode 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^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

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

Regulator: The regulator aims to maximize the total welfare W from each treatment episode, calculated as the patient surplus minus the total cost. The total welfare W is given by:

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

Here, υ represents the patient surplus from receiving treatment, and the remaining terms constitute the 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 extended to incorporate patient disutility from readmissions, as discussed in a similar manner in §5.3 of Arifoğlu et al. (2021). As υ is fixed, the regulator's objective function is equivalent to minimizing the total expected cost of a care episode.

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 these effort levels can be identified through the following first-order conditions (FOCs) obtained from the regulator's objective function (5):

$$(C^h(a_h^*))' + 1 = 0, \tag{6}$$
$$\frac{\partial C^s(b_h^*, b_s^*)}{\partial b^h} + 1 = 0, \tag{7}$$
$$\frac{\partial C^s(b_h^*, b_s^*)}{\partial b^s} + 1 = 0, \tag{8}$$
$$\frac{\partial R(e_h^*, e_s^*)}{\partial e^h}(\xi^h + \xi^s) + 1 = 0, \tag{9}$$
$$\frac{\partial R(e_h^*, e_s^*)}{\partial e^s}(\xi^h + \xi^s) + 1 = 0. \tag{10}$$

In Appendix A, we provide sufficient conditions on the cost 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 socially optimal 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).

4. Coordinating reimbursement schemes

The proposed payment model for both provider types follows a similar structure and consists of two parts: (i) a payment to cover the costs of care and efforts; and (ii) an outcome-based payment that promotes coordination, both of which are calculated based on the performances of other providers.

To formulate the payment scheme, we introduce the following notation. Let

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

represent the per-patient average cost of acute care for patients treated by all hospitals, excluding hospital i, and

$$\bar{C}_i^{sh} = \frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}} p_{kj} C^s(b_{kj}^h, b_{kj}^s)}{1-p_i}, \tag{11}$$

denote the per-patient average cost of PAC for all patients discharged from all hospitals, excluding hospital i. Similarly, let

$$\bar{R}_i^h = \frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}} p_{kj} R(e_{kj}^h, e_{kj}^s)}{1-p_i}, \tag{12}$$

represent the proportion of readmitted patients who are treated by all hospitals, excluding hospital i. We also define

$$\bar{a}_i^h = \frac{\sum_{k\in\mathcal{N}_i} p_k a_k^h}{1-p_i},\ \bar{b}_i^h = \frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}} p_{kj} b_{kj}^h}{1-p_i},\ \text{and}\ \bar{e}_i^h = \frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}} p_{kj} e_{kj}^h}{1-p_i}, \tag{13}$$

as the weighted average effort levels of all hospitals, excluding hospital i. The payment amount for hospital i is given by:

$$T_i^h = \underbrace{p_i\left[\bar{C}_i^h + \bar{a}_i^h + \bar{b}_i^h + \bar{e}_i^h + \bar{R}_i^h \xi^h\right]}_{\text{Cost of care}} + \underbrace{\sum_{j\in\mathcal{M}} p_{ij}\left[\bar{C}_i^{sh} - C^s(b_{ij}^h, b_{ij}^s) + \left(\bar{R}_i^h - R(e_{ij}^h, e_{ij}^s)\right)\xi^s\right]}_{\text{Outcome-based adjustment for care coordination}}. \tag{14}$$

The payment for hospitals consists of two components: the "Cost of care" component covers the costs of providing care and exerting effort, while the "Outcome-based adjustment for care coordination" component rewards or penalizes hospitals based on the average cost of PAC for their discharged patients and the average cost of treating readmitted patients in PAC, relative to other hospitals and PAC providers. The first component aligns with payment models used for single-entity DRG payments, encouraging cost efficiency by setting relative performance benchmarks, while the second component provides additional incentives for care coordination with PAC providers (we argue below that the removal of this component will compromise care coordination, see Remark 3

below for more details). In addition, our proposed payment method resembles bundled-type payments, involving a single payment for the entire episode of care, including potential readmissions (see Arifoğlu et al. (2021) for a detailed discussion).

The payment model for PAC providers follows a similar principle. Let

$$\bar{C}_j^s = \frac{\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} p_{ik} C^s(b_{ik}^h, b_{ik}^s)}{1-\tilde{p}_j}, \tag{15}$$

denote the per-patient average cost of PAC for patients treated by all PAC providers, excluding provider j and

$$\bar{R}_j^s = \frac{\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} p_{ik} R(e_{ik}^h, e_{ik}^s)}{1-\tilde{p}_j}. \tag{16}$$

represent the proportion of readmitted patients who are treated by all PAC providers, excluding provider j.

Furthermore

$$\bar{b}_j^s = \frac{\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} p_{ik} b_{ik}^s}{1-\tilde{p}_j}, \text{ and } \bar{e}_j^s = \frac{\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{M}_j} p_{ik} e_{ik}^s}{1-\tilde{p}_j}, \tag{17}$$

are the weighted average effort level of all the PAC providers, except provider j. The payment for provider j is given by:

$$T_j^s = \underbrace{\tilde{p}_j\left[\bar{C}_j^s + \bar{b}_j^s + \bar{e}_j^s + \bar{R}_j^s\xi^s\right]}_{\text{Cost of care}} + \underbrace{\sum_{i\in\mathcal{N}} p_{ij}\left[\bar{R}_j^s - R(e_{ij}^h, e_{ij}^s)\right]\xi^h}_{\text{Outcome-based adjustment for care coordination}}. \tag{18}$$

The payment model for PAC providers also includes a "Cost of care" component and an "Outcome-based adjustment for care coordination" component, encouraging cost efficiency and incentivizing coordination with hospitals, respectively.

Theorem 1. *If the regulator uses* (14) *to reimburse hospitals and* (18) *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.*

The proof is presented in Appendix B. Several key aspects regarding our main result merit emphasis.

Remark 1. The proposed payment model is highly appealing due to its simplicity. First, it eliminates the need for regulators to accurately estimate the cost structures or underlying cost functions of hospitals and PAC providers. Instead, regulators only need to observe the eventual

costs after these institutions make efforts to reduce their costs and readmissions. This approach resemblances the single-entity DRG-based payment models discussed in Shleifer (1985), Savva et al. (2019), Arifoğlu et al. (2021). Second, the payment model is independent of the network structure of hospitals and PAC providers, as well as the proportion of patients transferred between them. As a result, it can be seamlessly implemented in healthcare environments of varying structures without requiring any modifications.

Remark 2. We can further simplify the payment model by reducing the number of cost items that the regulator needs to monitor. Specifically, the regulator only needs to observe the overall costs of hospitals and PAC providers for index admissions, along with readmission costs. This eliminates the requirement to monitor each individual component of the cost items separately.

To determine T_j^s as defined in (18), the regulator only needs to observe the aggregated cost of PAC care for each patient (i.e., $C_{ij}^s + b_{ij}^s + e_{ij}^s$), along with the readmission probabilities and costs. Similarly, we can simplify the hospital payment defined in (14) as follows. Define the *total cost* of PAC provider j's treatment cost for those patients discharged from hospital i by

$$\mathscr{C}_{ij}^{sh} = C^s(b_{ij}^h, b_{ij}^s) + b_{ij}^s + e_{ij}^s, \tag{19}$$

and let

$$\bar{\mathscr{C}}_i^{sh} = \frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}} p_{kj}\mathscr{C}_{kj}^{sh}}{1-p_i}. \tag{20}$$

The payment amount for hospital i is then given by

$$\mathscr{T}_i^h = p_i\Big[\bar{C}_i^h + \bar{a}_i^h + \bar{b}_i^h + \bar{e}_i^h + \bar{R}_i^h\xi^h\Big] + \sum_{j\in\mathcal{M}} p_{ij}\Big[\bar{\mathscr{C}}_i^{sh} - \mathscr{C}_{ij}^{sh}\Big] + \sum_{j\in\mathcal{M}} p_{ij}\Big[\bar{R}_i^h - R(e_{ij}^h, e_{ij}^s)\Big]\xi^s. \tag{21}$$

We can show that Theorem 1 remains valid when T_j^s defined in (18) used along with $\mathscr{T}_i^h$.

Remark 3. The conventional bundled payment models do not provide adequate incentives for care coordination between acute and PAC providers. To demonstrate, we note that our payment model reduces to a conventional bundled payment model if we remove the "Outcome-based adjustment for care coordination" components in (14) and (18). In single-entity healthcare systems, this payment model restores optimal outcomes (see §5.1 of Arifoğlu et al. (2021)). However, we can show that hospitals will not exert any effort to reduce treatment costs in PAC with the conventional bundled payment model, i.e. $b_{ij}^h = 0$ and this will result in higher readmission costs in the system, i.e., $R(e_{ij}^h, e_{ij}^s)(\xi^h + \xi^s) + e_{ij}^h + e_{ij}^s > R(e_h^*, e_s^*)(\xi^h + \xi^s) + e_h^* + e_s^*$ in equilibrium for all $i \in \mathcal{N}$ and $j \in \mathcal{M}$.

Remark 4. Comprehensive Joint Replacement (CJR) program is a regionalized payment model (Department of Health and Human Services 2021), meaning that the average payment (referred to as "target price" by CMS) is determined for each region. This takes into account regional differences in costs.

Our payment model can be adapted to accommodate this characteristic by allowing flexibility in calculating payment amounts and benchmark parameters. For example, the payment amount for a hospital could be calculated as a convex combination of the peer performance of hospitals in the same region and the average performance of a selected group of providers. This would allow for more flexibility in setting payment amounts and performance targets, while still taking into account the regional variation in costs of care.

5. Extensions

In this section, we aim to validate the fundamental concept underlying our payment model by demonstrating its adaptability to different settings and its ability to consistently elicit socially optimal actions. We achieve this by adjusting the specific modeling assumptions discussed in the previous section.We focus on the following scenarios:

i. Endogenous discharge decisions: Hospitals make decisions about which PAC type to discharge patients to.
ii. Endogenous readmission treatment costs: The cost of treating a patient who is readmitted depends on the providers' actions.
iii. Uniform effort: Hospitals exert the same level of effort across all the PAC providers it discharges patients to, and similarly, the PAC providers across all hospitals it receives patients from.
iv. Non-identical providers: Providers have different characteristics, such as their quality or cost structure.

The proofs of the results in this section are presented in Appendix C.

5.1. Endogenous discharge decisions

In our original model presented in §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 efforts 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 §4) 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 S different types of PAC settings, including home. Each hospital i determines the proportion, $\rho_i^s \in [0,1]$, of their patients discharged to PAC providers of type s, where $s \in \mathcal{S}$. We let $\vec{\rho}_i = \{\rho_i^s, s \in \mathcal{S}\}$ denote the vector of these proportions for hospital i. We represent the proportion of patients discharged from hospital i to type-s PAC provider j as p_{ij}^s. Therefore, we have $\sum_{s\in\mathcal{S}} \rho_i^s = 1$ and $\sum_{j\in\mathcal{M}^s} p_{ij}^s/p_i = \rho_i^s$, where $\mathcal{M}^s = \{1, ..., M^s\}$ and M^s denotes the number of PAC providers of type s.

Additional notation and terminology required in this section are based on the ones introduced in §3. We use C^s to denote the cost of PAC for type-s providers, and R^s to represent the readmission probability for patients discharged to a type-s PAC setting. Similar to our previous model, we assume that $C^s : [0,\Gamma] \times [0,\Gamma] \to \mathbb{R}^+$ is dependent on the efforts made by the hospital, denoted by $b^{h,s}$, and by the PAC provider, denoted by b^s. The readmission probability from a type-s PAC setting, $R^s : [0,\Gamma] \times [0,\Gamma] \times [0,1] \to [0,1]$, depends on the efforts of the hospital, denoted by $e^{h,s}$, and the PAC provider, denoted by e^s, 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^s for PAC care at a type-s provider.

In contrast to the approach outlined in §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)

$$\Pi_i^h(\mathbf{h}_i) = T_i^h - \mathcal{C}_i^h(\mathbf{h}_i), \tag{22}$$

where

$$\mathcal{C}_i^h(\mathbf{h}_i) = p_i\left[C^h\left(a_i^h\right) + a_i^h\right] + \sum_{s\in\mathcal{S}}\sum_{j\in\mathcal{M}^s} p_{ij}^s\left[R^s\left(e_{ij}^{h,s}, e_{ij}^s, \vec{\rho}_i\right)\xi^h + b_{ij}^{h,s} + e_{ij}^{h,s}\right] \tag{23}$$

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

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

$$\Pi_j^s(\mathbf{v}_j) = T_j^s - \mathcal{C}_j^s(\mathbf{v}_j), \tag{24}$$

where

$$\mathcal{C}_j^s(\mathbf{v}_j) = \sum_{i \in \mathcal{N}} p_{ij}^s \left[C^s(b_{ij}^{h,s}, b_{ij}^s) + R^s(e_{ij}^{h,s}, e_{ij}^s, \vec{\rho}_i)\xi^s + b_{ij}^s + e_{ij}^s \right]. \tag{25}$$

is the total cost of for this provider. Here $\mathbf{v}_j = \left(b_{ij}^s, e_{ij}^s, i \in \mathcal{N}\right)$ 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:

$$W = \upsilon - \sum_{i \in \mathcal{N}} \mathcal{C}_i^h(\mathbf{h}_i) - \sum_{s \in \mathcal{S}} \sum_{j \in \mathcal{M}^s} \mathcal{C}_j^s(\mathbf{v}_j). \tag{26}$$

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 efforts are uniquely determined by the FOCs of total welfare, as in the original model; see Appendix C.1. 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 $(a_h^*, b_{h,s}^*, e_{h,s}^*, \vec{\rho}^*)$ with a slight abuse of notation, are identical, where $\vec{\rho}^* = \{\rho_s^*, s \in \mathcal{S}\}$ and it is possible that $\rho_s^* = 0$ for some s. Similarly, the socially optimal actions for type-s PAC providers, denoted by (b_s^*, e_s^*), are identical, for each $s \in \mathcal{S}$.

Payment scheme: We now present the 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 §4 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 (18) in our original model. The payment for type-s PAC provider j is given by:

$$T_j^s = \underbrace{\left[\hat{C}_j^s + \hat{b}_j^s + \hat{e}_j^s + \hat{R}_j^s \xi^s\right] \sum_{i \in \mathcal{N}} p_{ij}^s}_{\text{Cost of care}} + \underbrace{\sum_{i \in \mathcal{N}} p_{ij}^s \left[\hat{R}_j^s - R^s(e_{ij}^{h,s}, e_{ij}^s, \vec{\rho}_i)\right] \xi^h}_{\text{Outcome-based adjustment}}, \tag{27}$$

where

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

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

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

are the average cost of efforts to reduce costs and readmissions by all type-s PAC providers, excluding provider j. It is worth noting that (27) follows the same structure as (18).

The payment scheme for hospitals is slightly different from that in §4 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, $T^{h,s}$ for $s \in \mathcal{S}$.

We start with the cost of care component. As in (14), the hospital is compensated for the cost of care as well as the effort it is expected to exert to reduce costs and readmissions of the PAC providers that it discharges patients to, as follows

$$\underbrace{T_i^{h,0}}_{\text{Cost of care payment}} = p_i \left[\bar{C}_i^h + \bar{a}_i^h + \hat{R}_i^h \xi^h\right] + \sum_{s \in \mathcal{S}} p_i \bar{\rho}_i^s \left(\hat{b}_i^{h,s} + \hat{e}_i^{h,s}\right), \tag{28}$$

where

$$\bar{\rho}_i^s = \frac{\sum_{k \in \mathcal{N}_i} p_k \rho_k^s}{\sum_{k \in \mathcal{N}_i} p_k} \tag{29}$$

represents the average fraction of patients discharged to type-s PAC providers, excluding patients discharged from hospital i. Additionally, $\hat{b}_i^{h,s}$ and $\hat{e}_i^{h,s}$ represent the average costs incurred by

hospitals to improve care (in terms of cost and readmission probability, respectively) in type-s PAC providers, excluding patients discharged from hospital i, defined as follows (similar to (13))

$$\hat{b}_i^{h,s} = \frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^s} p_{kj}^s b_{kj}^{h,s}}{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^s} p_{kj}^s},\ \hat{e}_i^{h,s} = \frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^s} p_{kj}^s e_{kj}^{h,s}}{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^s} p_{kj}^s}, s\in\mathcal{S},$$

and (similar to (12))

$$\hat{R}_i^h = \frac{\sum_{k\in\mathcal{N}_i}\sum_{s\in\mathcal{S}}\sum_{j\in\mathcal{M}^s} p_{kj}^s R^s(e_{kj}^{h,s}, e_{kj}^s, \vec{\rho}_k)}{\sum_{k\in\mathcal{N}_i}\sum_{s\in\mathcal{S}}\sum_{j\in\mathcal{M}^s} p_{kj}^s}$$

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

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

$$\begin{aligned}\underbrace{T_i^{h,s}}_{\text{PAC cost component}} &= p_i\bar{\rho}_i^s\left(\hat{C}_i^{s,h} + \hat{R}_i^{h,s}\xi^s\right) - \sum_{j\in\mathcal{M}^s} p_{ij}^s\left[C^s\left(b_{ij}^{h,s}, b_{ij}^s\right) + R^s\left(e_{ij}^{h,s}, e_{ij}^s, \vec{\rho}_i\right)\xi^s\right] \\ &\quad + p_i\bar{\rho}_i^s(\hat{b}_i^{s,h} + \hat{e}_i^{s,h}) - \sum_{j\in\mathcal{M}^s} p_{ij}^s\left[b_{ij}^s + e_{ij}^s\right], \end{aligned} \tag{30}$$

where

$$\hat{R}_i^{h,s} = \frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^s} p_{kj}^s R^s(e_{kj}^{h,s}, e_{kj}^s, \vec{\rho}_k)}{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^s} p_{kj}^s},\ \text{and}\ \hat{C}_i^{s,h} = \frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^s} p_{kj}^s C^s(b_{kj}^{h,s}, b_{kj}^s)}{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^s} p_{kj}^s}, \tag{31}$$

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

$$\hat{b}_i^{s,h} = \frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^s} p_{kj}^s b_{kj}^s}{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^s} p_{kj}^s},\ \text{and}\ \hat{e}_i^{s,h} = \frac{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^s} p_{kj}^s e_{kj}^s}{\sum_{k\in\mathcal{N}_i}\sum_{j\in\mathcal{M}^s} p_{kj}^s} \tag{32}$$

are the average efforts to reduce type-s 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:

$$T_i^h = T_i^{h,0} + \sum_{s\in\mathcal{S}} T_i^{h,s}. \tag{33}$$

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 (14) are almost identical in principle: both consider the total cost incurred by the hospital in providing care and making improvement efforts. Additionally, $T_i^{h,s}$ is similar to the "Outcome-based adjustment for care coordination" component in Equation (14) with

a subtle difference: there is an additional term in the second line of (30). 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 (14) 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 2.

We next prove that this payment scheme induces first-best actions from all providers.[1]

Proposition 1. *If the regulator uses* (27) *to reimburse hospitals and* (33) *to reimburse PAC providers, then the unique Nash equilibrium is for each each hospital* $i \in \mathcal{N}$ *and type-s PAC provider* $j \in \mathcal{M}^s$, $s \in \mathcal{S}$, *to pick first-best actions* $a_i^h = a_h^*, b_{ij}^{h,s} = b_{h,s}^*, e_{ij}^{h,s} = e_{h,s}^*, \rho_i^s = \rho_s^*$, *and* $b_{ij}^s = b_s^*, e_{ij}^s = e_s^*$, *respectively. In addition, all providers break even in this equilibrium.*

5.2. 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 make efforts 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:

$$\Pi_i^h(\mathbf{h}_i) = T_i^h - \mathcal{C}_i^h(\mathbf{h}_i), \tag{34}$$

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

$$\mathcal{C}_i^h(\mathbf{h}_i) = p_i\left[C^h\left(a_i^h\right) + a_i^h\right] + \sum_{j\in\mathcal{M}} p_{ij}\left[b_{ij}^h + e_{ij}^h + R\left(e_{ij}^h, e_{ij}^s\right)\left(C^h\left(a_i^h\right) + a_i^h + b_{ij}^h\right)\right] \tag{35}$$

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

$$\Pi_j^s(\mathbf{s}) = T_j^s - \mathcal{C}_j^s(\mathbf{s}_j), \tag{36}$$

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

$$\mathcal{C}_j^s(\mathbf{s}_j) = \sum_{i\in\mathcal{N}} p_{ij}\left[e_{ij}^s + \left(1 + R\left(e_{ij}^h, e_{ij}^s\right)\right)\left(C^s\left(b_{ij}^h, b_{ij}^s\right) + b_{ij}^s\right)\right] \tag{37}$$

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

and, as in §3, b_{ij}^s and e_{ij}^s represent the effort levels of PAC provider j for cost reduction and readmission reduction, respectively. The main difference between (1) and (35) 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 (37).

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

$$T_i^h = \underbrace{p_i\left[\left(1+\bar{R}_i^h\right)\left(\bar{C}_i^h+\bar{a}_i^h+\bar{b}_i^h\right)+\bar{e}_i^h\right]}_{\text{Cost of care}} + \underbrace{\sum_{j\in\mathcal{M}} p_{ij}\left[\left(1+\bar{R}_i^h\right)\left(\bar{C}_i^{sh}+\bar{b}_j^s\right)-\left(1+R(e_{ij}^h,e_{ij}^s)\right)\left(C^s\left(b_{ij}^h,b_{ij}^s\right)+b_{ij}^s\right)\right]}_{\text{Outcome-based adjustment}}, \tag{38}$$

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 expected cost of treating readmitted patients. Similar to the difference between objective functions in this section and those in §4 as explained above (see (1) and (35)), the main difference in the current payment amount is that the cost of treatment for readmitted patients (ξ^h and ξ^s) in (14) are replaced by the corresponding costs in the current model in (38). Similarly, for PAC providers, we modify the payment as follows

$$T_j^s = \underbrace{\sum_{i\in\mathcal{N}} p_{ij}\left[\left(1+\bar{R}_j^s\right)\left(\bar{C}_j^s+\bar{b}_j^s\right)+\bar{e}_j^s\right]}_{\text{Cost of care}} + \underbrace{\sum_{i\in\mathcal{N}} p_{ij}\left[\left(1+\bar{R}_j^s\right)\left(\bar{C}_i^h+\bar{a}_i^h+\bar{b}_i^h\right)-\left(1+R(e_{ij}^h,e_{ij}^s)\right)\left(C^h(a_i^h)+a_i^h+b_{ij}^h\right)\right]}_{\text{Outcome-based adjustment}}, \tag{39}$$

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 (18) 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 (35) and (37), 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 efforts.

Proposition 2. *If the regulator uses* (38) *to reimburse hospitals and* (39) *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 §5.3 of Arifoğlu et al. (2021).

5.3. Uniform efforts

In our original model in §4, we assume that each hospital exerts a different *PAC provider dependent* effort, see term b_{ij}^h there, to reduce PAC treatment costs with each PAC provider (vis-a-vis, we assumed PAC providers make hospital-dependent efforts, see term b_{ij}^s 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 efforts.

To model the impact of uniform (non-PAC/hospital-dependent) efforts, assume that each hospital makes an effort $H_i \in [0, \Gamma]$, $i \in \mathcal{N}$, and each PAC provider makes an effort $F_j \in [0, \Gamma]$, $j \in \mathcal{M}$, 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 effort H_i and PAC provider's effort F_j. All other components of the model remain identical to those introduced in §3.

In this case, the objective of hospital i is

$$\Pi_i^h(\mathbf{h}_i) = T_i^h - \mathcal{C}_i^h(\mathbf{h}_i), \tag{40}$$

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

$$\mathcal{C}_i^h(\mathbf{h}_i) = p_i \left[C^h(a_i^h) + a_i^h + H_i\right] + \sum_{j \in \mathcal{M}} p_{ij} \left[R(e_{ij}^h, e_{ij}^s)\xi^h + e_{ij}^h\right]. \tag{41}$$

Similarly, the objective of PAC provider j is

$$\Pi_j^s(\mathbf{s}) = T_j^s - \mathcal{C}_j^s(\mathbf{s}_j), \tag{42}$$

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

$$\mathcal{C}_j^s(\mathbf{s}_j) = \tilde{p}_j F_j + \sum_{i \in \mathcal{N}} p_{ij} \Big[C^s(H_i, F_j) + R(e_{ij}^h, e_{ij}^s)\xi^s + e_{ij}^s\Big]. \tag{43}$$

The main difference between our original model (1) and (40) is that now we use H_i to capture the total cost of hospital i's effort to reduce PAC treatment cost instead of b_{ij}^h in (1). Similarly, we use F_j in (42) to capture the effort cost of PAC j instead of b_{ij}^s in (3).

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

$$\bar{H}_i = \frac{\sum\limits_{k\in\mathcal{N}_i} p_k H_k}{1-p_i}, \tag{44}$$

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

$$\bar{F}_j = \frac{\sum\limits_{k\in\mathcal{M}_j} \tilde{p}_k F_k}{1-\tilde{p}_j}, \tag{45}$$

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

$$T_i^h = p_i\left[\bar{C}_i^h + \bar{a}_i^h + \bar{H}_i + \bar{e}_i^h + \bar{R}_i^h\xi^h\right] + \sum_{j\in\mathcal{M}} p_{ij}\left[\left(\bar{C}_i^{sh} - C^s(H_i,F_j)\right) + \left(\bar{R}_i^h - R(e_{ij}^h, e_{ij}^s)\right)\xi^s\right], \tag{46}$$

$$T_j^s = \sum_{i\in\mathcal{N}} p_{ij}\left[\bar{C}_j^s + \bar{F}_j + \bar{e}_j^s + \bar{R}_j\xi^s\right] + \sum_{i\in\mathcal{N}} p_{ij}\left(\bar{R}_j - R(e_{ij}^h, e_{ij}^s)\right)\xi^h, \tag{47}$$

where benchmarks $\bar{H}_i$ and $\bar{F}_j$ are defined as in (44) and (45), 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 (15)–(17).

The payment model in this case is similar to that in §4, see (14) and (18), with the only difference being the way hospitals and PAC providers are compensated for their cost reduction efforts. In (14) hospital i receives $p_i\bar{b}_i^h$ to recoup the cost of their effort to reduce PAC costs (since it is assumed to be variable cost there), whereas in (46) they receive $p_i\bar{H}_i$. For PAC providers, they receive $\sum_{i\in\mathcal{N}} p_{ij}\bar{b}_j^s$ for their effort in (18) which becomes $\sum_{i\in\mathcal{N}} p_{ij}\bar{F}_j$ in (47).

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 (41) and (43), 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 efforts.

Proposition 3. *If the regulator uses* (46) *to reimburse hospitals and* (47) *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.*

The robustness of our results considering both variable and fixed efforts to reduce the readmissions can be demonstrated in a similar manner.

5.4. Non-identical providers

To implement coordinating reimbursement schemes, it was previously assumed that the regulator could identify identical providers or at least pairs of identical providers. However, in real-world scenarios, providers often exhibit heterogeneity across various dimensions, such as geographical location and demographic factors. Nevertheless, if these factors can be observed by the regulator and are exogenous to the providers, then 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). 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 efforts of the hospital e_i^h, $i \in \mathcal{N}$, and the PAC provider e_j^s, $j \in \mathcal{M}$, (as in §4) as well as the observable exogenous characteristics of the hospital β_h and of the PAC provider β_s. Therefore, the first-best outcomes are dependent on the specific characteristics β_h and β_s, see (6)–(10)).

In this modified approach, instead of using average values, as shown in equations (12) and (16), 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, which is based on observed readmission probabilities and the corresponding observable characteristics of hospitals β_i^h, $i \in \mathcal{N}$, and the PAC provider characteristics β_j^s, $j \in \mathcal{M}$. By 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.

6. Equilibrium under a CJR-type payment model

We have demonstrated that, when both hospitals and PAC providers are held accountable for the costs and quality of the entire episode of care, it elicits socially optimal actions from both entities. However, in the case of the CJR program, only hospitals are directly held financially accountable, while gainsharing agreements are allowed between providers. It is crucial to examine the implications of this arrangement on providers' behavior.

Modeling the actions of hospitals and PAC providers under CJR and gainsharing agreements presents challenges for two reasons. Firstly, while hospitals can share both the risk and reconciliation (i.e, performance-based) payments with PAC providers through gainsharing agreements, CMS

has not provided clear guidelines on how to design such agreements. Consequently, the specific agreement reached between a hospital and a PAC provider is likely to depend on their relative bargaining power. Secondly, hospitals cannot mandate patients to seek care exclusively from their preferred PAC providers, further complicating matters.

Although exploring the intricacies of this interaction between hospitals and PAC providers is beyond the scope of this paper (see (Gupta et al. 2021, Ghamat et al. 2021)), we utilize a simplified model to capture its essence. In this model, we assume that PAC providers bear only a portion of the hospital treatment cost of readmitted patients, denoted by the parameter $\theta \in [0,1]$, while the hospital assumes the remaining cost $(1-\theta)$. Additionally, we assume that PAC providers are responsible for the entire PAC treatment cost of readmitted patients because they are subject to other outcome-based payment programs such as the Skilled Nursing Facility Value-Based Purchasing program (see CMS (2023b)).

The payment amount received by PAC providers takes the following form:

$$T_j^s = \tilde{p}_j \Big[\bar{C}_j^s + \bar{b}_j^s + \bar{e}_j^s + \bar{R}_j^s \xi^s\Big] + \sum_{i\in\mathcal{N}} p_{ij} \Big[\bar{R}_j^s - R(e_{ij}^h, e_{ij}^s)\Big]\theta\xi^h. \tag{48}$$

This payment model aligns with our proposed payment model for PACs (see (18)) when $\theta = 1$. As for hospitals, the payment amount is as follows:

$$\begin{aligned} T_i^h = {} & p_i \Big[\bar{C}_i^h + \bar{a}_i^h + \bar{b}_i^h + \bar{e}_i^h + \bar{R}_i^h \xi^h\Big] \\ & + \sum_{j\in\mathcal{M}} p_{ij} \Big[\bar{C}_i^{sh} - C_s(b_{ij}^h, b_{ij}^s) + \big(\bar{R}_i^h - R(e_{ij}^h, e_{ij}^s)\big)\left((1-\theta)\xi^h + \xi^s\right)\Big]. \end{aligned} \tag{49}$$

The key distinction between this payment model and our proposed payment model for PAC providers (see (14)) is that the hospital is held responsible for the excessive readmission cost associated with providing acute care, which the PAC provider is no longer accountable for. When $\theta = 1$, the payment amount in equation (49) becomes identical to (14) in our proposed payment model.

We next establish the equilibrium under the payment model (48)–(49); we relegate the technical details to Appendix D.

Proposition 4. *If the regulator uses* (48) *to reimburse hospitals and* (49) *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. In addition, if* $\partial^2 R(e^h, e^s)/\partial e^h \partial e^s \geqslant 0$ *for all* $e^h, e^s \in [0, \Gamma]$*, we have* $\tilde{e}^h > e_h^*$ *and* $\tilde{e}^s < e_s^*$ *for any* $\theta \in [0,1)$*.*

Assumption $\partial^2 R(e^h, e^s)/\partial e^h \partial e^s \geqslant 0$ implies that hospital investment is less effective at reducing the readmissions at higher PAC provider investments. With this condition, Proposition 4 establishes that hospitals invest more, while the PAC providers invest less, in reducing readmissions

relative to their respective first-best levels. This is intuitive considering Theorem 1 because now the hospitals bear part of the PAC providers reimbursement adjustment for readmission performance in the hospital setting. As such, hospitals have stronger incentives to reduce readmissions and PAC providers have reduced incentives.

This result demonstrates that the CJR program incentivizes hospitals and PAC providers to make efforts to coordinate, which was not present under separate Diagnosis-Related Group (DRG) based payments, see Remark 3. Furthermore, because this payment model is identical to our proposed payment model when $\theta = 1$, it can potentially lead to first-best effort levels with carefully designed gainsharing agreements between hospitals and PAC providers. However, this presents significant challenges as the hospital may lack the resources or information on PAC provider costs to thoroughly assess the impact of a gainsharing agreement. Additionally, similar gainsharing agreements would need to be established with all PAC providers, as hospitals cannot mandate which PAC provider a patient should choose. Our proposed payment model eliminates the reliance on such agreements, offering a promising alternative.

7. Conclusions

Payment models in healthcare play a vital role in incentivizing healthcare providers to deliver high-quality and cost-effective care. DRG- and performance-based payment models have demonstrated particular effectiveness in eliciting efficient healthcare delivery. While these models are well suited for medical conditions managed by a single entity, challenges arise when dealing with certain conditions, such as joint replacement, which involve multiple independent providers operating in different settings and decentralized treatment decisions. In response to these challenges, bundled payment models have emerged to encompass the costs and services associated with an entire episode of care, offering a more holistic approach to payment.

However, effectively dividing bundled payments among independent providers to ensure comparable incentives as in single-entity settings remains uncertain and requires further research. The CJR payment model implemented by CMS highlights these challenges, as it primarily holds hospitals accountable for the entire joint replacement episode while allowing them to form gainsharing agreements with other providers to improve healthcare delivery collaboratively. However, the design of these gainsharing agreements lacks a consensus and poses implementation issues.

Our contributions: In this paper, we propose an innovative payment model and show that it elicits socially optimal actions from independent healthcare entities responsible for different parts of a specific condition's episode of care. Our proposed payment model utilizes benchmarks derived from the performance of all providers involved and incentivizes them to collaboratively enhance the quality of care in a cost-effective manner by comparing their joint performance to these benchmarks.

The appeal of our method lies in its simplicity and efficiency. Firstly, it eliminates the need for regulators to accurately estimate the cost structures or underlying cost functions of hospitals and PAC providers. This streamlines the implementation process and reduces administrative burdens. Secondly, the payment model is independent of the network structure of hospitals and PAC providers, as well as the proportion of patients transferred between them. Consequently, it can be seamlessly used in healthcare environments of varying structures and for different conditions without requiring modifications. An additional advantage of our proposed payment model is its elimination of the reliance on gainsharing agreements, which can be challenging to design and implement in practical settings. By removing this complexity, our model becomes more straightforward and feasible for adoption.

Overall, our novel payment model presents a promising approach to drive collaboration among healthcare providers and improve the overall quality of care while maintaining cost effectiveness. The model's adaptability and simplicity make it a feasible and attractive option for implementation in a wide range of healthcare settings, appealing to providers and regulators alike.

Potential implementation issues: Our payment model, while not burdening the regulator with additional informational requirements beyond current models, does present the potential for increased risk exposure for PAC providers compared to the CJR payment model. This is primarily due to our model holding PAC providers financially responsible for excess acute-care costs related to readmitted patients. However, we believe that CMS (or other regulators) can address and mitigate these concerns through several strategies.

Firstly, it is essential to recognize that gainsharing agreements, an alternative approach as described above, also entail increased risk for PAC providers, making our proposed payment model not inherently unique in this aspect.

Secondly, many PAC providers, including home healthcare providers, currently enjoy healthy profit margins from treating Medicare patients (MedPAC 2022), suggesting that there is room to increase their share in penalties and rewards without compromising their financial viability.

Thirdly, a cautious and gradual implementation approach can be taken, initially capping the share of PAC providers in imposed rewards and penalties and progressively increasing these caps over several years. This allows PAC providers sufficient time to implement more effective collaboration tools with acute-care providers, reducing risks and enhancing coordination. However, in the long run, incentive payments should not be capped, as caps could impact the effectiveness of the model, as discussed in Arifoğlu et al. (2021).

By carefully considering these factors and implementing the payment model thoughtfully, CMS can address the risk concerns for PAC providers while fostering collaboration and incentivizing improved patient outcomes across independent providers.

Limitations: Our study has several limitations that warrant consideration. Firstly, a more detailed and nuanced model could provide a better understanding of the specific costs incurred by each party involved in improving collaborative care. While we have focused on two simplified cases (variable or fixed costs), a more comprehensive understanding of cost structures would enhance the implementation of our suggested payment schemes. For example, by requiring healthcare providers to adopt electronic medical records software with specific capabilities, similar to the effective facilitation of performance-based payment models by the Electronic Health Record Incentive programs (CMS 2023a), and then using appropriate reimbursement models, the proposed payment models could be further improved in practice.

Additionally, we have assumed that patient volume is sufficiently high to keep the variance of the performance estimates, such as readmission probability, low. However, smaller institutions may face higher variability, impacting the effectiveness of the payment model. For such providers, aggregating data over a longer time period or exploring alternative payment models might be more appropriate.

Moreover, our study employs readmissions as a proxy for a quality measure, while CJR considers a richer set of quality measures. These measures can potentially influence patient choice, and the presence of multiple PAC providers in a region offers patients options for selection. To address these complexities, a more detailed analysis (similar to those in Savva et al. (2019) and Arifoğlu et al. (2021)) would shed further light on the impact of quality outcomes on patient decision-making and how payment models can be extended to accommodate these factors.

References

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

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

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.

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.

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. PMID: 30601709.

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

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.

Britton, M. C., G. M. Ouellet, K. E. Minges, M. Gawel, B. Hodshon, and S. I. Chaudhry (2019). Care transitions between hospitals and skilled nursing facilities: Perspectives of sending and receiving providers. *Journal of Applied Gerontology 38*(6), 785–794.

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

Chernew, M. E., P. H. Conway, and A. B. Frakt (2020). Transforming Medicare's payment systems: Progress shaped by the ACA. *Health Affairs 39*(3), 413–420.

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 (2022). CMS comprehensive care for joint replacement model: Performance year 4 evaluation report. https://innovation.cms.gov/data-and-reports/2021/cjr-py4-annual-report last accessed on July 14, 2023.

CMS (2023a). Promoting interoperability programs. https://www.cms.gov/regulations-and-guidance/legislation/ehrincentiveprograms last accessed on July 14, 2023.

CMS (2023b). Value-based programs. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/Value-Based-Programs last accessed on July 14, 2023.

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.

Debo, L., N. Savva, and R. Shumsky (2021). Hospital reimbursement in the presence of cherry picking and upcoding. Technical report, Dartmouth University.

Department of Health and Human Services (2017). Medicare program; cancellation of advancing care coordination through episode payment and cardiac rehabilitation incentive payment models; changes to comprehensive care for joint replacement payment model: Extreme and uncontrollable circumstances policy for the comprehensive care for joint replacement payment model. *Federal Register (December 1, 2017) 82*(230), 57066 –57104. 83 FR 128.

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.

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 (2023). Impact of physician payment scheme on diagnostic effort and testing. *To appear in Management Science*.

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

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

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

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

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

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://www.healthaffairs.org/do/10.1377/hblog20170829.061481/full/ last accessed on July 14, 2023.

MedPAC (2022). Report to the congress: Medicare payment policy. https://www.medpac.gov/wp-content/uploads/2022/03/Mar22_MedPAC_ReportToCongress_v3_SEC.pdf. 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.

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

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.

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.

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*.

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) Acute care cost is strictly decreasing and strictly convex in hospital investment, i.e.,* $(C^h(a^h))' < 0$ *and* $(C^h(a^h))'' > 0$*, and the following boundary conditions hold.*

$$\lim_{a^h \downarrow 0} (C^h(a^h))' < -1 < \lim_{a^h \uparrow \Gamma} (C^h(a^h))'.$$

(ii) 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) investments, 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

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

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 effort to reduce acute care cost. Taking the first and second partial derivatives of W in (A-3) with respect to a_i^h, we have

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

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

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

Second, we characterize (b_h^*, b_s^*), i.e., first-best efforts made by the hospital and PAC provider to reduce 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 efforts are taken (see the last paragraph of §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

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

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

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

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

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

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

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

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

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

where the second equality follows by plugging in $g'(b_{ij}^h)$ 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 effort 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 efforts are taken (see the last paragraph of §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 $e_{ij}^h \in [0, \Gamma]$, we have $\partial^2 W/\partial (e_{ij}^s)^2 < 0$, $\lim_{e_{ij}^s \downarrow 0} \partial W/\partial e_{ij}^s > 0$, and $\lim_{e_s \uparrow \Gamma} \partial W/\partial e_{ij}^s < 0$ by Assumption A-1(iii). Hence there exists a unique $z(e_{ij}^h) \in (0, \Gamma)$ that satisfies

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

Applying the Implicit Function Theorem, we obtain

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

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

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

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

$$W|_{\{e_{ij}^h = e_h^*, e_{ij}^s = e_s^*\}} = \sup_{e_{ij}^h \in [0,\Gamma]} W|_{e_{ij}^s = z(e_{ij}^h)}.$$

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

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

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

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

where the second equality follows by plugging in $z'(e_{ij}^h)$ from (A-11), and the inequality follows from Assumption A-1(iii). Moreover, we have $\lim_{e_{ij}^h \downarrow 0} dW(e_{ij}^h)/de_{ij}^h > 0$ and $\lim_{e_{ij}^h \uparrow \Gamma} dW(e_{ij}^h)/de_{ij}^h < 0$ 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 $e_{ij}^h = e_h^*$ into (A-10). □

B. 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), (14), and (18), hospital i's objective is

$$\begin{aligned}
\Pi_i^h(\mathbf{h}_i) =& p_i\left[\bar{C}_i^h + \bar{a}_i^h - C^h(a_i^h) - a_i^h\right] \\
&+ \sum_{j \in \mathcal{M}} p_{ij}\left[\bar{C}_i^{sh} - C^s(b_{ij}^h, b_{ij}^s) + (\bar{R}_i^h - R(e_{ij}^h, e_{ij}^s))(\xi^h + \xi^s) + \bar{b}_i^h - b_{ij}^h + \bar{e}_i^h - e_{ij}^h\right],
\end{aligned} \tag{A-12}$$

and PAC provider j's objective is

$$\Pi_j^s(\mathbf{s}_j) = \sum_{i\in\mathcal{N}} p_{ij}\left[\bar{C}_j^s - C^s(b_{ij}^h, b_{ij}^s) + (\bar{R}_j^s - R(e_{ij}^h, e_{ij}^s))(\xi^h+\xi^s) + \bar{b}_j^s - b_{ij}^s + \bar{e}_j^s - e_{ij}^s\right]. \tag{A-13}$$

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

$$\underset{\mathbf{h}_i\in[0,\Gamma]^{2M+1}}{\text{maximize}}\ W(\mathbf{h}_i|\mathbf{h}_k, \mathbf{s}_j, k\in\mathcal{N}_i, j\in\mathcal{M}) \tag{A-14}$$

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

$$\underset{\mathbf{s}_j\in[0,\Gamma]^{2N}}{\text{maximize}}\ W(\mathbf{s}_j|\mathbf{h}_i, \mathbf{s}_k, i\in\mathcal{N}, k\in\mathcal{M}_j). \tag{A-15}$$

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.,

$$\underset{\mathbf{h}_i\in[0,\Gamma]^{2M+1}, \mathbf{s}_j\in[0,\Gamma]^{2N}, i\in\mathcal{N}, j\in\mathcal{M}}{\text{maximize}}\ W(\mathbf{h}_i, \mathbf{s}_j, i\in\mathcal{N}, j\in\mathcal{M}), \tag{A-16}$$

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 $i \in \mathcal{N}$,

$$\mathbf{h}^* = \underset{\mathbf{h}_i \in [0,\Gamma]^{2M+1}}{\arg\max}\ W(\mathbf{h}_i | \mathbf{h}_k = \mathbf{h}^*, \mathbf{s}_j = \mathbf{s}^*, k \in \mathcal{N}_i, j \in \mathcal{M}). \tag{A-17}$$

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

$$\mathbf{s}^* = \underset{\mathbf{s}_j \in [0,\Gamma]^{2N}}{\arg\max}\ W(\mathbf{s}_j | \mathbf{h}_i = \mathbf{h}^*, \mathbf{s}_k = \mathbf{s}^*, i \in \mathcal{N}, k \in \mathcal{M}_j), \tag{A-18}$$

i.e., each PAC provider j's problem given by (A-15) has a unique solution $\mathbf{s}_j = \mathbf{s}^*$, when other hospitals and PAC providers choose first-best actions. By (A-17), each hospital i's problem given by (A-14) 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-12)-(A-13) 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-14), and $\mathbf{s}_j = \check{\mathbf{s}}_j$ is a solution for each PAC provider j's problem (A-15), i.e.,

$$\check{\mathbf{h}}_i \in \underset{\mathbf{h}_i \in [0,\Gamma]^{2M+1}}{\arg\max}\ W(\mathbf{h}_i | \check{\mathbf{h}}_k, \check{\mathbf{s}}_j, k \in \mathcal{N}_i, j \in \mathcal{M}), \tag{A-19}$$

$$\check{\mathbf{s}}_j \in \underset{\mathbf{s}_j \in [0,\Gamma]^{2N}}{\arg\max}\ W(\mathbf{s}_j | \check{\mathbf{h}}_i, \check{\mathbf{s}}_k, i \in \mathcal{N}, k \in \mathcal{M}_j). \tag{A-20}$$

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

$$\check{b}_{ij}^h = b_h^* \text{ and } \check{b}_{ij}^s = b_s^*, \text{ for each } i \in \mathcal{N} \text{ and } j \in \mathcal{M}. \tag{A-21}$$

Since (A-21) is assumed to hold when $p_{ij}=0$ WLOG (see the last paragraph of §3 for details), it suffices to consider the case of $p_{ij}>0$. By (A-19)-(A-20), we have

$$\check{b}^h_{ij} \in \arg\max_{b^h_{ij}\in[0,\Gamma]} W(b^h_{ij}|\check{a}^h_i, \check{e}^h_{ij}, \check{\mathbf{h}}_k, \check{\mathbf{s}}_j, k\in\mathcal{N}_i, j\in\mathcal{M}) = \arg\min_{b^h_{ij}\in[0,\Gamma]} p_{ij}\left[C^s(b^h_{ij},\check{b}^s_{ij})+b^h_{ij}+\check{b}^s_{ij}\right], \tag{A-22}$$

$$\check{b}^s_{ij} \in \arg\max_{b^s_{ij}\in[0,\Gamma]} W(b^s_{ij}|\check{e}^s_{ij}, \check{\mathbf{h}}_i, \check{\mathbf{s}}_k, i\in\mathcal{N}, k\in\mathcal{M}_j) = \arg\min_{b^s_{ij}\in[0,\Gamma]} p_{ij}\left[C^s(\check{b}^h_{ij},b^s_{ij})+\check{b}^h_{ij}+b^s_{ij}\right], \tag{A-23}$$

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-23) and obtained a unique best response $\check{b}^s_{ij}=g(\check{b}^h_{ij})\in(0,\Gamma)$, where $g(\cdot)$ is given by (A-5). Now we solve the optimization problem in (A-22). For any fixed $\check{b}^s_{ij}\in[0,\Gamma]$, we have

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

where all inequalities follow from Assumption A-1(ii). Thus, the optimization problem in (A-22) has a unique solution $\check{b}^h_{ij}$ and is determined by the FOC, i.e., $\partial C^s(\check{b}^h_{ij},\check{b}^s_{ij})/\partial b^h+1=0$. Plugging in $\check{b}^s_{ij}=g(\check{b}^h_{ij})$, we obtain

$$\frac{\partial C^s(\check{b}^h_{ij},g(\check{b}^h_{ij}))}{\partial b^h}+1=0. \tag{A-24}$$

In the proof of Lemma A-1, we have proven that (A-24) 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 $\check{b}^h_{ij}=b^*_h$ and $\check{b}^s_{ij}=g(\check{b}^h_{ij})=g(b^*_h)=b^*_s$, where the last equality follows from (A-9). Following this procedure, one can verify that $\check{e}^h_{ij}=e^*_h$ and $\check{e}^s_{ij}=e^*_s$ 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. □

C. Proof of the results in §5

C.1. Endogenous discharge decisions

We continue to adopt Assumption A-1, with functions C^s and R adapted into sets of functions C^s and R^s for all $s\in\mathcal{S}$, and Assumption A-1(iii) applied to R^s for any given $\vec{\rho}$. These assumptions ensure that for any given $\vec{\rho}$ the first-best efforts 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 efforts (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 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* (26) *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-*$s \in \mathcal{S}$ *PAC provider* $j \in \mathcal{M}^s$ *such that* $p_{ij}^s > 0$*,* $b_{ij}^{h,s} = b_{h,s}^*, b_{ij}^s = b_s^*, e_{ij}^{h,s} = e_{h,s}^*$*, and* $e_{ij}^s = e_s^*$*, where* $a_h^*, b_{h,s}^*, e_{h,s}^*, b_s^*, e_s^* \in (0, \Gamma)$ *are unique and satisfy the following FOCs:*

$$(C^h(a_h^*))' + 1 = 0, \tag{A-25}$$

$$\frac{\partial C^s(b_{h,s}^*, b_s^*)}{\partial b^h} + 1 = 0, \tag{A-26}$$

$$\frac{\partial C^s(b_{h,s}^*, b_s^*)}{\partial b^s} + 1 = 0, \tag{A-27}$$

$$\frac{\partial R^s(e_{h,s}^*, e_s^*, \vec{\rho}^*)}{\partial e^h}(\xi^h + \xi^s) + 1 = 0, \tag{A-28}$$

$$\frac{\partial R^s(e_{h,s}^*, e_s^*, \vec{\rho}^*)}{\partial e^s}(\xi^h + \xi^s) + 1 = 0. \tag{A-29}$$

Proof of Lemma A-2: Let $\vec{\mathbf{h}} = \{\mathbf{h}_i, i \in \mathcal{N}\}$ and $\vec{\mathbf{v}} = \{\mathbf{v}_j, j \in \mathcal{M}^s, s \in \mathcal{S}\}$ 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 (23), (25), and (26), is given by

$$\begin{aligned} W =& \upsilon - C^h(a_i^h) - a_i^h \\ & - \sum_{i \in \mathcal{N}} \sum_{s \in \mathcal{S}} \sum_{j \in \mathcal{M}^s} p_{ij}^s \left[C^s(b_{ij}^{h,s}, b_{ij}^s) + b_{ij}^{h,s} + b_{ij}^s + R^s(e_{ij}^{h,s}, e_{ij}^s, \vec{\rho}_i)(\xi^h + \xi^s) + e_{ij}^{h,s} + e_{ij}^s \right]. \end{aligned} \tag{A-30}$$

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-30) we have: (i) when $p_{ij}^s = 0$, W is independent of hospital i's efforts $(b_{ij}^{h,s}, e_{ij}^{h,s})$ and PAC provider j's efforts (b_{ij}^s, e_{ij}^s). WLOG we assume that first-best efforts are taken (see the last paragraph of §3 for details), i.e., $b_{ij}^{h,s} = b_{h,s}^*, e_{ij}^{h,s} = \tilde{e}_{h,s}(\vec{\rho}_i), b_{ij}^s = b_s^*, e_{ij}^s = \tilde{e}_s(\vec{\rho}_i)$ as determined by (A-26), (A-27), (A-31), and (A-32), respectively; (ii) when $p_{ij}^s > 0$, the optimal efforts are characterized by $b_{h,s}^*, b_s^*$ defined as in (A-26)-(A-27), and $\tilde{e}_{h,s}(\vec{\rho}_i), \tilde{e}_s(\vec{\rho}_i)$ defined as follows (this proof is omitted as it is similar to that in Lemma A-1):

$$\frac{\partial R^s(\tilde{e}_{h,s}(\vec{\rho}_i), \tilde{e}_s(\vec{\rho}_i), \vec{\rho}_i)}{\partial e^h}(\xi^h + \xi^s) + 1 = 0, \tag{A-31}$$

$$\frac{\partial R^s(\tilde{e}_{h,s}(\vec{\rho}_i), \tilde{e}_s(\vec{\rho}_i), \vec{\rho}_i)}{\partial e^s}(\xi^h + \xi^s) + 1 = 0. \tag{A-32}$$

Plugging in (A-30) and noting that $\sum_{j\in\mathcal{M}^s} p_{ij}^s/p_i = \rho_i^s$, we obtain

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

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

$$\underset{\vec{\rho}_i}{\text{minimize}} \sum_{s\in\mathcal{S}} \rho_i^s \left[C^s(b_{h,s}^*, b_s^*) + b_{h,s}^* + b_s^* + R^s(\tilde{e}_{h,s}(\vec{\rho}_i), \tilde{e}_s(\vec{\rho}_i), \vec{\rho}_i)(\xi^h + \xi^s) + \tilde{e}_{h,s}(\vec{\rho}_i) + \tilde{e}_s(\vec{\rho}_i)\right] \quad \text{(A-33)}$$

$$\text{s.t. } \vec{\rho}_i \in [0,1]^S, \sum_{s\in\mathcal{S}} \rho_i^s = 1. \quad \text{(A-34)}$$

Since $\tilde{e}_{h,s}(\vec{\rho}_i)$ and $\tilde{e}_s(\vec{\rho}_i)$ defined as in (A-31)-(A-32) are the unique unconstrained minimizer of $R^s(e_{h,s}, e_s, \vec{\rho}_i)(\xi^h + \xi^s) + e_{h,s} + e_s$, the minimum value $R^s(\tilde{e}_{h,s}(\vec{\rho}_i), \tilde{e}_s(\vec{\rho}_i), \vec{\rho}_i)(\xi^h + \xi^s) + \tilde{e}_{h,s}(\vec{\rho}_i) + \tilde{e}_s(\vec{\rho}_i)$ is continuous in $\vec{\rho}_i$; this establishes continuity of objective (A-33) in $\vec{\rho}_i$. Since the constraint set defined by (A-34) is compact, the minimization problem (A-33)-(A-34) 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-33) and the constraint set defined in (A-34). 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 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 (22)-(23) and (28)-(33), hospital i's objective is

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

By (24)-(25) and (27), PAC provider j's objective is

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

Subtracting each objective by W from (A-30), we obtain

$$\begin{aligned} \Pi_i^h(\mathbf{h}_i) - W(\vec{\mathbf{h}}, \vec{\mathbf{v}}) =& p_i\left(\bar{C}_i^h + \bar{a}_i^h + \hat{R}_i^h\xi^h\right) + \sum_{k\in\mathcal{N}_i}\left[C^h\left(a_k^h\right) + a_k^h\right] - v \\ & + \sum_{s\in\mathcal{S}} p_i\bar{\rho}_i^s\left[\hat{b}_i^{h,s} + \hat{e}_i^{h,s} + \hat{b}_i^{s,h} + \hat{e}_i^{s,h} + \hat{C}_i^{s,h} + \hat{R}_i^{h,s}\xi^s\right] \end{aligned}$$

$$+\sum_{k\in\mathcal{N}_i}\sum_{s\in\mathcal{S}}\sum_{j\in\mathcal{M}^s} p^s_{kj}\Big[C^s(b^{h,s}_{kj},b^s_{kj})+R^s\left(e^{h,s}_{kj},e^s_{kj},\vec{\rho}_k\right)(\xi^h+\xi^s)+b^{h,s}_{kj}+b^s_{kj}+e^{h,s}_{kj}+e^s_{kj}\Big],$$

and

$$\begin{aligned}\Pi^s_j(\mathbf{v}_j)-W(\vec{\mathbf{h}},\vec{\mathbf{v}})=&\sum_{i\in\mathcal{N}} p^s_{ij}\left[\hat{C}^s_j+\hat{b}^s_j+\hat{e}^s_j+\hat{R}^s_j(\xi^h+\xi^s)\right]-v+\sum_{i\in\mathcal{N}}\Big\{p_i\Big[C^h\left(a^h_i\right)+a^h_i\Big]+p^s_{ij}\left[b^{h,s}_{ij}+e^{h,s}_{ij}\right]\Big\}\\&+\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{S}_s}\sum_{l\in\mathcal{M}^k} p^k_{il}\Big[C^k(b^{h,k}_{il},b^k_{il})+R^k\left(e^{h,k}_{il},e^k_{il},\vec{\rho}_i\right)(\xi^h+\xi^k)+b^{h,k}_{il}+b^k_{il}+e^{h,k}_{il}+e^k_{il}\Big]\\&+\sum_{i\in\mathcal{N}}\sum_{l\in\mathcal{M}^s_s} p^s_{il}\Big[C^s(b^{h,s}_{il},b^s_{il})+R^s\left(e^{h,s}_{il},e^s_{il},\vec{\rho}_i\right)(\xi^h+\xi^s)+b^{h,s}_{il}+b^s_{il}+e^{h,s}_{il}+e^s_{il}\Big].\end{aligned}$$

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. □

C.2. Endogenous readmission cost

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

$$\lim_{e^h\downarrow 0}\frac{\partial R\left(e^h,e^s\right)}{\partial e^h}<-\frac{1}{C^h\left(a^*_h\right)+C^s\left(b^*_h,b^*_s\right)+a^*_h+b^*_h+b^*_s}<\lim_{e^h\uparrow\Gamma}\frac{\partial R\left(e^h,e^s\right)}{\partial e^h}\text{ for any }e^s\in[0,\Gamma],\quad\text{(A-35)}$$

$$\lim_{e^s\downarrow 0}\frac{\partial R\left(e^h,e^s\right)}{\partial e^s}<-\frac{1}{C^h\left(a^*_h\right)+C^s\left(b^*_h,b^*_s\right)+a^*_h+b^*_h+b^*_s}<\lim_{e^s\uparrow\Gamma}\frac{\partial R\left(e^h,e^s\right)}{\partial e^s}\text{ for any }e^h\in[0,\Gamma].\quad\text{(A-36)}$$

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* $a^h_i=a^*_h$ *for each hospital* $i\in\mathcal{N}$*, and for each PAC provider* $j\in\mathcal{M}$ *such that* $p_{ij}>0$*,* $b^h_{ij}=b^*_h, b^s_{ij}=b^*_s, e^h_{ij}=e^*_h$*, and* $e^s_{ij}=e^*_s$*, 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:*

$$\frac{\partial R\left(e^*_h,e^*_s\right)}{\partial e^h}\Big[C^h\left(a^*_h\right)+C^s\left(b^*_h,b^*_s\right)+a^*_h+b^*_h+b^*_s\Big]+1=0,\quad\text{(A-37)}$$

$$\frac{\partial R\left(e^*_h,e^*_s\right)}{\partial e^s}\Big[C^h\left(a^*_h\right)+C^s\left(b^*_h,b^*_s\right)+a^*_h+b^*_h+b^*_s\Big]+1=0.\quad\text{(A-38)}$$

Proof of Lemma A-3: Plugging (35) and (37) in (5), we obtain

$$W=v-\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} p_{ij}\Big[\left(1+R(e^h_{ij},e^s_{ij})\right)\left(C^h(a^h_i)+a^h_i+C^s(b^h_{ij},b^s_{ij})+b^h_{ij}+b^s_{ij}\right)+e^h_{ij}+e^s_{ij}\Big].\quad\text{(A-39)}$$

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-39) 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 efforts 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 efforts are taken (see the last paragraph of §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-37)-(A-38). 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-36). Hence there exists a unique $z(e^h_{ij}) \in (0, \Gamma)$ that satisfies

$$\frac{\partial R(e^h_{ij}, z(e^h_{ij}))}{\partial e^s}\Big[C^h(a^*_h) + C^s(b^*_h, b^*_s) + a^*_h + b^*_h + b^*_s\Big] + 1 = 0. \tag{A-40}$$

Applying the Implicit Function Theorem, we obtain

$$z'(e^h_{ij}) = -\frac{\partial^2 R(e^h_{ij}, z(e^h_{ij}))/\partial e^h \partial e^s}{\partial^2 R\left(e^h_{ij}, z(e^h_{ij})\right)/\partial (e^s)^2}. \tag{A-41}$$

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,

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

where the second equality follows from (A-40).

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

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

where the second equality follows by plugging in $z'(e^h_{ij})$ from (A-41), 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-35). Thus there exists a unique $e^*_h\in(0,\Gamma)$ that satisfies (A-37) with $e^*_s=z(e^*_h)$; (A-38) follows by substituting $e^h_{ij}=e^*_h$ into (A-40). □

Proof of Proposition 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 (34)-(35) and (38), 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 (36)-(37) and (39), 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-39), 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. □

C.3. Uniform Efforts

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 investments, i.e.,

$$\frac{\partial C^s}{\partial H}<0,\frac{\partial^2 C^s}{\partial H^2}>0,\frac{\partial C^s}{\partial F}<0,\frac{\partial^2 C^s}{\partial F^2}>0,\frac{\partial^2 C^s}{\partial H^2}\frac{\partial^2 C^s}{\partial F^2}>\left(\frac{\partial^2 C^s}{\partial H\partial F}\right)^2, \tag{A-42}$$

and (ii) the following boundary conditions hold

$$\lim_{H\downarrow 0}\frac{\partial C^s}{\partial H}<-1<\lim_{H\uparrow \Gamma}\frac{\partial C^s}{\partial H} \text{ for any } F\in[0,\Gamma], \tag{A-43}$$

$$\lim_{F\downarrow 0}\frac{\partial C^s}{\partial F}<-1<\lim_{F\uparrow \Gamma}\frac{\partial C^s}{\partial F} \text{ for any } H\in[0,\Gamma]. \tag{A-44}$$

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:*

$$\frac{\partial C^s\left(H^*,F^*\right)}{\partial H}+1=0, \tag{A-45}$$

$$\frac{\partial C^s\left(H^*,F^*\right)}{\partial F}+1=0. \tag{A-46}$$

Proof of Lemma A-3: Plugging (41) and (43) in (5), we obtain

$$W=v-\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}}p_{ij}\Big[C^h\left(a_i^h\right)+a_i^h+C^s(H_i,F_j)+H_i+F_j+R(e_{ij}^h,e_{ij}^s)(\xi^h+\xi^s)+e_{ij}^h+e_{ij}^s\Big]. \tag{A-47}$$

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-47) 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 efforts hospitals and PAC providers make to reduce PAC costs, i.e.,

$$\underset{\vec{H}\in[0,\Gamma]^N,\vec{F}\in[0,\Gamma]^M}{\text{maximize}}\ W, \tag{A-48}$$

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-47) 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

$$\sum_{i\in\mathcal{N}}p_{ij}\frac{\partial^2C^s(H_i,F_j)}{\partial F^2}>\sum_{i\in\mathcal{N}}\frac{\left(p_{ij}\frac{\partial^2C^s(H_i,F_j)}{\partial H\partial F}\right)^2}{\sum_{k\in\mathcal{M}}p_{ik}\frac{\partial^2C^s(H_i,F_k)}{\partial H^2}} \text{ for each } j\in\mathcal{M}, \tag{A-49}$$

which follows from $(\partial^2 C^s/\partial H^2)(\partial^2 C^s/\partial F^2) > (\partial^2 C^s/\partial H \partial F)^2$. In addition, the choice set of (A-48) 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-47),

$$\frac{\partial W}{\partial H_i} = -\sum_{j\in\mathcal{M}} p_{ij}\Big[\frac{\partial C^s(H_i,F_j)}{\partial H}+1\Big], \tag{A-50}$$

$$\frac{\partial^2 W}{\partial H_i^2} = -\sum_{j\in\mathcal{M}} p_{ij}\frac{\partial^2 C^s(H_i,F_j)}{\partial H^2}, \tag{A-51}$$

$$\frac{\partial W}{\partial F_j} = -\sum_{i\in\mathcal{N}} p_{ij}\Big[\frac{\partial C^s(H_i,F_j)}{\partial F}+1\Big], \tag{A-52}$$

$$\frac{\partial^2 W}{\partial F_j^2} = -\sum_{i\in\mathcal{N}} p_{ij}\frac{\partial^2 C^s(H_i,F_j)}{\partial F^2}. \tag{A-53}$$

For any fixed $\vec{F}$ and each $i \in \mathcal{N}$, we have $\partial^2 W/\partial H_i^2 < 0$ by (A-51) and (A-42), $\lim_{H_i\downarrow 0}\partial W/\partial H_i > 0$ and $\lim_{H_i\uparrow\Gamma}\partial W/\partial H_i < 0$ by (A-50) and (A-43). For any fixed $\vec{H}$ and each $j \in \mathcal{M}$, we have $\partial^2 W/\partial F_j^2 < 0$ by (A-53) and (A-42), $\lim_{F_j\downarrow 0}\partial W/\partial F_j > 0$ and $\lim_{F_j\uparrow\Gamma}\partial W/\partial F_j < 0$ by (A-52) and (A-44). Thus, the first-best investments H_i^* and F_j^*, $i \in \mathcal{N}$, $j \in \mathcal{M}$, are determined by FOCs:

$$\sum_{j=1}^{M} p_{ij}\Big[\frac{\partial C^s\left(H_i^*,F_j^*\right)}{\partial H}+1\Big] = 0 \text{ for all } i \in \mathcal{N}, \tag{A-54}$$

$$\sum_{i=1}^{N} p_{ij}\Big[\frac{\partial C^s\left(H_i^*,F_j^*\right)}{\partial F}+1\Big] = 0 \text{ for all } j \in \mathcal{M}. \tag{A-55}$$

Let H^* and F^* be determined by (A-45)-(A-46). The existence and uniqueness of H^* and F^* are ensured by (A-42)-(A-44). 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-54)-(A-55). Thus, first best actions, as uniquely determined by (A-54)-(A-55), are given by $H_i^* = H^*$ and $F_j^* = F^*$ for each $i \in \mathcal{N}$ and $j \in \mathcal{M}$. □

Proof of Proposition 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 (40)-(41) and (46), 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 (42)-(43) and (47), 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), (41), and (43), 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-47), we obtain

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

and

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

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. □

C.4. Non-identical providers

We continue to adopt Assumption A-1 with conditions in Assumption A-1(iii) assumed to hold for any hospital and PAC provider characteristic pair (β_h,β_s). By Lemma A-1, the first best actions are $a_h^*, b_h^*, e_h^*(\beta_i^h,\beta_j^s)$ for hospital i and $b_s^*, e_s^*(\beta_i^h,\beta_j^s)$ for PAC provider j, where $a_h^*, b_h^*, b_s^* \in (0,\Gamma)$ are determined by (6)-(8), and $e_h^*(\beta_i^h,\beta_j^s), e_s^*(\beta_i^h,\beta_j^s) \in (0,\Gamma)$ are determined by the following FOCs:

$$\frac{\partial R(e_h^*(\beta_i^h,\beta_i^s), e_s^*(\beta_i^h,\beta_i^s), \beta_i^h, \beta_i^s)}{\partial e^h}(\xi^h+\xi^s) + 1 = 0, \tag{A-56}$$

$$\frac{\partial R(e_h^*(\beta_i^h,\beta_i^s), e_s^*(\beta_i^h,\beta_i^s), \beta_i^h, \beta_i^s)}{\partial e^s}(\xi^h+\xi^s) + 1 = 0. \tag{A-57}$$

In what follows, we redefine $\bar{R}_i^h$ and $\bar{R}_j^s$ using linear regression based on hospital and PAC provider characteristics and show that the reimbursement scheme (14) and (18) restores first best outcomes and all providers break even in equilibrium.

Consider any given hospital and a PAC provider the hospital discharges patients to. Let β_h^m and β_s^m denote the hospital's and PAC provider's observable characteristics, respectively. Under the proposed reimbursement scheme with any benchmark parameters exogenous to their actions, equilibrium outcome coincides with first best and is given by $a_h^*, b_h^*, b_s^*, e_h^m = e_h^*(\beta_h^m,\beta_s^m), e_s^m = e_s^*(\beta_h^m,\beta_s^m)$;

we omit the proof as it is similar to that for Theorem 1. We next use these *ex post* actions and linear regression based on observable provider characteristics to derive benchmark parameters $\bar{R}_i^h$ and $\bar{R}_j^s$, which approximate the first-best outcomes for any hospital $i \in \mathcal{N}$ and PAC provider $j \in \mathcal{M}$.

Expanding (A-56)-(A-57) around $(e_h^m, e_s^m, \beta_h^m, \beta_s^m)$, we obtain

$$\frac{\partial^2 R^m}{\partial (e^h)^2}(e_i^h - e_h^m) + \frac{\partial^2 R^m}{\partial e^s \partial e^h}(e_i^s - e_s^m) + \frac{\partial^2 R^m}{\partial \beta^h \partial e^h}(\beta_i^h - \beta_h^m) + \frac{\partial^2 R^m}{\partial \beta^s \partial e^h}(\beta_i^s - \beta_s^m) \approx 0,$$
$$\frac{\partial^2 R^m}{\partial e^h \partial e^s}(e_i^h - e_h^m) + \frac{\partial^2 R^m}{\partial (e^s)^2}(e_i^s - e_s^m) + \frac{\partial^2 R^m}{\partial \beta_h \partial e^s}(\beta_i^h - \beta_h^m) + \frac{\partial^2 R^m}{\partial \beta_s \partial e^s}(\beta_i^s - \beta_s^m) \approx 0,$$

where $R^m = R(e_h^m, e_s^m, \beta_h^m, \beta_s^m)$. Solving for e_i^h and e_i^s, we obtain estimates of first best actions for any hospital $i \in \mathcal{N}$ and PAC provider $j \in \mathcal{M}$ as

$$\bar{e}^h(\beta_i^h, \beta_i^s) = e_h^m - \frac{\left(\frac{\partial^2 R^m}{\partial (e^s)^2}\frac{\partial^2 R^m}{\partial \beta_h \partial e^h} - \frac{\partial^2 R^m}{\partial e^s \partial e^h}\frac{\partial^2 R^m}{\partial \beta_h \partial e^s}\right)(\beta_i^h - \beta_h^m) + \left(\frac{\partial^2 R^m}{\partial (e^s)^2}\frac{\partial^2 R^m}{\partial \beta_s \partial e^h} - \frac{\partial^2 R^m}{\partial e^s \partial e^h}\frac{\partial^2 R^m}{\partial \beta_s \partial e^s}\right)(\beta_i^s - \beta_s^m)}{\frac{\partial^2 R^m}{\partial (e^h)^2}\frac{\partial^2 R^m}{\partial (e^s)^2} - \left(\frac{\partial^2 R^m}{\partial e^s \partial e^h}\right)^2},$$
$$\bar{e}^s(\beta_i^h, \beta_i^s) = e_s^m - \frac{\left(\frac{\partial^2 R^m}{\partial (e^h)^2}\frac{\partial^2 R^m}{\partial \beta_h \partial e^s} - \frac{\partial^2 R^m}{\partial e^h \partial e^s}\frac{\partial^2 R^m}{\partial \beta_h \partial e^h}\right)(\beta_i^h - \beta_h^m) + \left(\frac{\partial^2 R^m}{\partial (e^h)^2}\frac{\partial^2 R^m}{\partial \beta_s \partial e^s} - \frac{\partial^2 R^m}{\partial e^h \partial e^s}\frac{\partial^2 R^m}{\partial \beta_s \partial e^h}\right)(\beta_i^s - \beta_s^m)}{\frac{\partial^2 R^m}{\partial (e^h)^2}\frac{\partial^2 R^m}{\partial (e^s)^2} - \left(\frac{\partial^2 R^m}{\partial e^s \partial e^h}\right)^2}.$$

Approximating $R(e_i^h, e_j^s, \beta_i^h, \beta_j^s)$ by the first-order Taylor series around $(e_h^m, e_s^m, \beta_h^m, \beta_s^m)$ yields

$$\bar{R}(\beta_i^h, \beta_j^s) = R^m + \frac{\partial R^m}{\partial e^h}\left(\bar{e}^h(\beta_i^h, \beta_j^s) - e_h^m\right) + \frac{\partial R^m}{\partial e^s}\left(\bar{e}^s(\beta_i^h, \beta_j^s) - e_s^m\right) + \frac{\partial R^m}{\partial \beta_h}(\beta_i^h - \beta_h^m) + \frac{\partial R^m}{\partial \beta_s}(\beta_i^s - \beta_s^m).$$

We use $\bar{R}(\beta_i^h, \beta_j^s)$ in place of $\bar{R}_i^h$ and $\bar{R}_j^s$, i.e., hospital and PAC provider payment amounts are

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

By definition $\bar{R}(\beta_i^h, \beta_j^s)$ is independent of hospital i and PAC provider j actions, and approximates the first best outcome when provider heterogeneity is small.[2] Thus, the proposed reimbursement scheme restores first best outcomes with all providers break even.

D. Proof of the results in §6

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

$$\lim_{e^h\downarrow 0}\frac{\partial R(e^h, e^s)}{\partial e^h} < -\frac{1}{\xi^h + \xi^s} \text{ and } \lim_{e^h\uparrow \Gamma}\frac{\partial R(e^h, e^s)}{\partial e^h} > -\frac{1}{2\xi^h + \xi^s} \text{ for all } e^s \in [0, \Gamma], \tag{A-58}$$
$$\lim_{e_s\downarrow 0}\frac{\partial R(e^h, e^s)}{\partial e^s} < -\frac{1}{\xi^s} \text{ and } \lim_{e^s\uparrow \Gamma}\frac{\partial R(e^h, e^s)}{\partial e^s} > -\frac{1}{\xi^h + \xi^s} \text{ for all } e^h \in [0, \Gamma], \tag{A-59}$$

[2] This can be achieved by grouping similar provider pairs and picking a representative pair to define β_h^m and β_s^m.

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

Proof of Proposition 4: By (1), (2), and (49), hospital i's objective is

$$\begin{aligned}\Pi_i^h(\mathbf{h}_i) =& p_i\left[\bar{C}_i^h + \bar{a}_i^h - C^h(a_i^h) - a_i^h\right] \\ &+ \sum_{j\in\mathcal{M}} p_{ij}\left[\bar{C}_i^{sh} - C_s(b_{ij}^h, b_{ij}^s) + \left(\bar{R}_i^h - R(e_{ij}^h, e_{ij}^s)\right)\left((2-\theta)\xi^h + \xi^s\right) + \bar{b}_i^h - b_{ij}^h + \bar{e}_i^h - e_{ij}^h\right].\end{aligned}$$

By (3), (4), and (48), PAC provider j's objective is

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

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' efforts 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 efforts are taken (see the last paragraph of §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

$$\frac{\partial \Pi_i^h}{\partial e_{ij}^h} = -p_{ij}\left[\frac{\partial R(e_{ij}^h, e_{ij}^s)}{\partial e^h}((2-\theta)\xi^h + \xi^s) + 1\right], \tag{A-60}$$

$$\frac{\partial \Pi_j^s}{\partial e_{ij}^s} = -p_{ij}\left[\frac{\partial R(e_{ij}^h, e_{ij}^s)}{\partial e^s}(\theta\xi^h + \xi^s) + 1\right]. \tag{A-61}$$

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 $\theta \in [0,1]$. 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 $\partial^2 R(e^h, e^s)/\partial e^h \partial e^s \gtrless 0$.

(i) At any fixed $e_{ij}^s \in [0, \Gamma]$, by (A-60) 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

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

by (A-58). 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.,

$$\frac{\partial R(z_h(e_{ij}^s), e_{ij}^s)}{\partial e^h}((2-\theta)\xi^h + \xi^s) + 1 = 0. \tag{A-62}$$

At any fixed $e^h_{ij} \in [0,\Gamma]$, by (A-59), (A-61), and Assumption A-1(iii), we have $\partial^2 \Pi^s_j / \partial (e^s_{ij})^2 < 0$, $\lim_{e^s \uparrow \Gamma} \partial \Pi^s_j / \partial e^s_{ij} < 0$, and $\lim_{e^s \downarrow 0} \partial \Pi^s_j / \partial e^s_{ij} > 0$ for any $\theta \in (\underline{\theta}_s, \bar{\theta}_s)$, where

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

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

by (A-59). Thus, the PAC provider has a unique best response $z_s(e^h_{ij}) \in (0,\Gamma)$ and is determined by the FOC of Π^s_j, i.e.,

$$\frac{\partial R(e^h_{ij}, z_s(e^h_{ij}))}{\partial e^s} (\theta \xi^h + \xi^s) + 1 = 0. \tag{A-63}$$

(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 $e^s_{ij} = z_s(e^h_{ij})$ into (A-62), we can characterize the hospital's equilibrium readmission-reduction effort $\tilde{e}_h$ by $\Psi(\tilde{e}_h) = 0$, where

$$\Psi(e^h) = \frac{\partial R(e^h, z_s(e^h))}{\partial e^h} ((2-\theta)\xi^h + \xi^s) + 1. \tag{A-64}$$

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

$$\frac{d\Psi(e^h)}{de^h} = ((2-\theta)\xi^h + \xi^s) \left(\frac{\partial^2 R(e^h, z_s(e^h))}{\partial (e^h)^2} + \frac{\partial^2 R(e^h, z_s(e^h))}{\partial e^s \partial e^h} \frac{dz_s(e^h)}{de^h} \right), \tag{A-65}$$

where

$$\frac{dz_s(e^h)}{de^h} = -\frac{\partial^2 R(e^h, z_s(e^h)) / \partial e^h \partial e^s}{\partial^2 R(e^h, z_s(e^h)) / \partial (e^s)^2} \tag{A-66}$$

by taking the derivative of (A-63) with respect to e^h_{ij}. Plugging (A-66) into (A-65), we have

$$\frac{d\Psi(e^h)}{de^h} = ((2-\theta)\xi^h + \xi^s) \left(\frac{\partial^2 R(e^h, z_s(e^h))}{\partial (e^h)^2} - \frac{\left(\partial^2 R(e^h, z_s(e^h)) / \partial e^h \partial e^s \right)^2}{\partial^2 R(e^h, z_s(e^h)) / \partial (e^s)^2} \right) > 0, \tag{A-67}$$

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 SNF chooses $e_s = \tilde{e}_s$.

(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-63)-(A-64). 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-63) is continuous in $e^h \in (0,\Gamma)$ and $\theta \in \Theta$, so as $\Psi(e^h, \theta)$

defined as in (A-64). 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-67), $\tilde{e}_h$ 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

$$\frac{d\tilde{e}_h}{d\theta} = -\frac{\partial\Psi(\tilde{e}_h, \theta)/\partial\theta}{\partial\Psi(\tilde{e}_h, \theta)/\partial e^h} = -\frac{((2-\theta)\xi^h + \xi^s)\frac{\partial^2 R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h \partial e^s}\frac{\partial z_s(\tilde{e}_h, \theta)}{\partial\theta} - \xi^h \frac{\partial R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h}}{((2-\theta)\xi^h + \xi^s)\left[\frac{\partial^2 R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial (e^h)^2} + \frac{\partial^2 R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h \partial e^s}\frac{\partial z_s(\tilde{e}_h, \theta)}{\partial e^h}\right]}. \tag{A-68}$$

By (A-66) and Assumption A-1(iii), the denominator on the right-hand side of (A-68) is positive for all $\theta \in [0,1]$, thus

$$\operatorname{sgn}\left(\frac{d\tilde{e}_h}{d\theta}\right) = -\operatorname{sgn}\left(((2-\theta)\xi^h + \xi^s)\frac{\partial^2 R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h \partial e^s}\frac{\partial z_s(\tilde{e}_h, \theta)}{\partial\theta} - \xi^h \frac{\partial R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h}\right). \tag{A-69}$$

Taking the derivative of (A-63) with respect to θ, we have

$$\frac{\partial z_s(e^h, \theta)}{\partial\theta} = -\frac{\xi^h \frac{\partial R(e^h, z_s(e^h, \theta))}{\partial e^s}}{(\theta\xi^h + \xi^s)\frac{\partial^2 R(e^h, z_s(e^h, \theta))}{\partial (e^s)^2}}. \tag{A-70}$$

Plugging it into (A-69), we obtain

$$\begin{aligned}
\operatorname{sgn}\left(\frac{d\tilde{e}_h}{d\theta}\right) &= \operatorname{sgn}\left(\frac{(2-\theta)\xi^h + \xi^s}{\theta\xi^h + \xi^s}\frac{\partial^2 R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h \partial e^s}\frac{\frac{\partial R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^s}}{\frac{\partial^2 R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial (e^s)^2}} + \frac{\partial R(\tilde{e}_h, z_s(\tilde{e}_h, \theta))}{\partial e^h}\right) \\
&= \operatorname{sgn}\left(\frac{(2-\theta)\xi^h + \xi^s}{\theta\xi^h + \xi^s}\frac{\partial^2 R(\tilde{e}_h, \tilde{e}_s)}{\partial e^h \partial e^s}\frac{\frac{\partial R(\tilde{e}_h, \tilde{e}_s)}{\partial e^s}}{\frac{\partial^2 R(\tilde{e}_h, \tilde{e}_s)}{\partial (e^s)^2}} + \frac{\partial R(\tilde{e}_h, \tilde{e}_s)}{\partial e^h}\right) \\
&= \operatorname{sgn}\left(\frac{(2-\theta)\xi^h + \xi^s}{\theta\xi^h + \xi^s}\frac{\partial^2 R(\tilde{e}_h, \tilde{e}_s)}{\partial e^h \partial e^s}\frac{\partial R(\tilde{e}_h, \tilde{e}_s)}{\partial e^s} + \frac{\partial R(\tilde{e}_h, \tilde{e}_s)}{\partial e^h}\frac{\partial^2 R(\tilde{e}_h, \tilde{e}_s)}{\partial (e^s)^2}\right) < 0,
\end{aligned} \tag{A-71}$$

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

$$\frac{d\tilde{e}_s}{d\theta} = \frac{dz_s(\tilde{e}_h, \theta)}{d\theta} = \frac{\partial z_s(\tilde{e}_h, \theta)}{\partial e^h}\frac{d\tilde{e}_h}{d\theta} + \frac{\partial z_s(\tilde{e}_h, \theta)}{\partial\theta} = -\frac{\frac{\partial^2 R(\tilde{e}_h, \tilde{z}_s)}{\partial e^h \partial e^s}}{\frac{\partial^2 R(\tilde{e}_h, \tilde{z}_s)}{\partial (e^s)^2}}\frac{d\tilde{e}_h}{d\theta} - \frac{\xi^h \frac{\partial R(\tilde{e}_h, \tilde{z}_s)}{\partial e^s}}{(\theta\xi^h + \xi^s)\frac{\partial^2 R(\tilde{e}_h, \tilde{z}_s)}{\partial (e^s)^2}} > 0, \tag{A-72}$$

where the second equality follows from differentiation by parts, the third follows from (A-66) and (A-70), and the inequality follows from $d\tilde{e}_h/d\theta < 0$ by (A-71), $\partial^2 R(e^h, e^s)/(\partial e^h \partial e^s) \geqslant 0$, $\partial R(e^h, e^s)/\partial e^h < 0$, $\partial R(e^h, e^s)/\partial e^s < 0$, and $\partial^2 R(e^h, e^s)/\partial (e^s)^2 > 0$ by Assumption A-1(iii). $\square$