Provider Payment Models for Generative AI in Healthcare

Elodie Adida* Tinglong Dai†

*School of Business, University of California, Riverside, Riverside, California 92521, elodieg@ucr.edu
†Carey Business School, Johns Hopkins University, Baltimore, Maryland 21202;
Hopkins Business of Health Initiative, Johns Hopkins University, Washington, District of Columbia 20001, dai@jhu.edu

Generative AI (GenAI) tools, such as ambient listening, hold promise for transforming medical practice by integrating diverse data formats and delivering multimodal outputs. Yet, their uptake in healthcare is limited by the need for customization and the lack of suitable provider payment models. In current U.S. practice, the costs associated with using these tools are classified as indirect, with no direct mechanisms for reimbursing providers. This paper examines how different provider payment models affect the quality and uptake of GenAI tools in clinical settings. We develop a theoretical framework that captures the strategic interplay between a developer, who sets the quality and usage fee of a GenAI tool, and a healthcare provider, who decides on usage in response to the reimbursement structure. We show that without reimbursement, providers restrict GenAI usage to more complex cases, leading to suboptimal quality and limited uptake. Fee-for-service models, while encouraging widespread use at high reimbursement rates, can lead developers to compromise on quality. Conversely, lower reimbursement rates may incentivize higher quality but still fall short of the socially optimal level. We propose a hybrid payment model that integrates fee-for-service with value-based payments, showing that aligning developer and provider incentives requires defining value metrics based on the quality of the GenAI tool itself, rather than the benefits it delivers. Interestingly, in this model, as development costs rise, the minimum fee-for-service needed to align incentives decreases. Our paper demonstrates how the interplay between the downstream and upstream dynamics of GenAI tools in healthcare influences their quality and drives their uptake.

Key words: Physician payment, healthcare operations management, artificial intelligence.
History: This version is dated January 14, 2025

1. Introduction

The ongoing boom in generative artificial intelligence (GenAI) began with OpenAI's ChatGPT, which launched on November 30, 2022, and reached 100 million users in just two months (Wachter and Brynjolfsson 2024). Originally a text-only tool, ChatGPT has evolved to support a wide range of input and output formats. It has also spurred the development of a variety of multimodal GenAI tools, including Anthropic's Claude, Google's Gemini, and Meta's LLaMA, along with GenAI-based enterprise solutions from companies such as Microsoft, Oracle, and Salesforce. Applications of GenAI

in routine healthcare include interpreting various types of medical data—such as text, voice, medical images, lab results, and electronic health records—and generating outputs such as visit summaries, scripts, assessments and plans, and image annotations (Omiye et al. 2024). According to a March 2024 survey by McKinsey, more than 29% of healthcare organizations have incorporated AI-based GenAI tools into their workflows (Lamb et al. 2024).

The widespread adoption of GenAI tools in medical practice raises questions about how to compensate providers for using such tools. Addressing this question presents novel challenges for policymakers and payers for two reasons. First, the quality of GenAI tools can vary depending on the developer's level of customization, which stands in stark contrast to pre-GenAI tools. Most pre-GenAI tools—often based on computer vision models (Dai and Abràmoff 2023)—are designed for specific medical conditions or patient populations, with inflexible inputs and outputs, marketed as black-box devices that generally cannot be modified once approved by the FDA (Lai et al. 2024). By contrast, GenAI tools often require customization for real-world application, are rarely subject to FDA clearance, and exhibit varying quality levels.[1] Conceivably, the payment model for using these tools can impact the customization process and, in turn, the quality of these tools. Second, unlike pre-GenAI tools for which the variable cost of using them is often negligible, the variable costs of using Gen-AI tools are significant (Goldman Sachs 2024), necessitating usage-based pricing (Rand 2023).

Compensating physicians for acquiring and using AI under the Medicare Physician Fee Schedule (PFS) presents a multifaceted challenge. The Centers for Medicare & Medicaid Services (CMS) has historically treated the costs associated with AI tools not related to medical decision-making as indirect, effectively excluding them from the PFS methodology (CMS 2021). This classification has faced pushback, prompting CMS to implement temporary workarounds to align payment rates for AI-related services with those for similar services. These temporary measures, however, are not viable as long-term solutions. CMS is actively working to adapt the PFS to reflect the growing role of AI tools as part of routine healthcare delivery (Zink et al. 2024).

In this paper, we examine how payment structures influence both the uptake of GenAI tools (downstream) and the quality of these tools (upstream). Several research questions guide our analysis: Is the status quo—in which providers receive no direct payment for using AI tools—sufficient? If not, how do commonly used payment systems, such as fee-for-service and value-based payment models, compare? Finally, can the physician payment model be designed to achieve the first-best outcome that maximizes social welfare?

[1] In the rest of the paper, we use "AI" and "GenAI" interchangeably unless otherwise noted. In other words, the AI tools we refer to exclude those influencing medical decision-making (e.g., diagnosis, treatment, and cure)—an important topic but beyond the scope of this paper.

To address these questions, we develop a parsimonious model to investigate the effects of provider payment structures on the quality and usage of GenAI tools. The model captures the strategic interaction between an AI developer and a healthcare provider, bridging upstream and downstream perspectives. The AI developer sets the tool's quality and price based on expected usage and development costs; the healthcare provider, typically a physician, decides on its use for a case by weighing clinical complexity, the cost of using AI, reimbursement (zero under the status quo), and the expected benefit from AI. By examining different payment models—including the status quo, fee-for-service, and hybrid scenarios—we analyze how economic incentives influence both the quality of the AI tool and the extent of its clinical use.

We begin by analyzing the status quo scenario, in which the provider is not reimbursed for using GenAI in medical practice. In this setting, we examine the decisions made by both the AI developer and the healthcare provider. Specifically, the developer determines the quality of the AI system and sets the price per use, whereas the provider decides whether and for which patients to use the AI tool, based on the associated costs and expected benefits. Our analysis reveals that without reimbursement, the provider uses the AI tool only for more complex patient cases, where the benefits justify the costs. The developer sets a price and quality level that balances the cost of developing a higher-quality product with the provider's willingness to pay. At equilibrium, the developer selects a price that maximizes profit, resulting in a quality level below the socially optimal standard and leaving only a proportion of the patients who could benefit from the AI receiving it. The intuition is that without reimbursement, the financial burden of using the AI tool rests solely on the provider, limiting its widespread use even when it could benefit more patients. As a result, developers have little incentive to invest in higher-quality AI, because the limited provider usage does not justify the additional costs of developing higher-quality tools.

Next, we examine the fee-for-service payment system, a prevalent reimbursement model for pre-GenAI tools (Parikh and Helmchen 2022; Wu et al. 2023) and one considered instrumental in promoting the use of AI in routine healthcare practice due to its simplicity and financial predictability (Abràmoff et al. 2024). Under this model, the provider receives a fixed reimbursement rate for each patient on whom AI is used. Our analysis reveals several key insights. The provider's AI usage depends on the reimbursement rate, ranging from no usage to partial or full adoption across all patients. The developer's optimal strategy is similarly contingent on the reimbursement rate. If the fee is sufficiently high, the developer minimizes quality to reduce costs while setting the price equal to the fee, leading to AI usage for all patients. At a lower reimbursement rate, the developer increases quality to justify the price, but the quality still falls short of the socially optimal level, and AI uptake is limited to a subset of patients. Whereas the fee-for-service model promotes broad AI usage at high reimbursement levels, it falls short of achieving first-best outcomes. Specifically, it incentivizes

developers to tailor quality to reimbursement rates rather than optimizing for social welfare, resulting in inefficiencies in both quality and usage.

Our analysis of the fee-for-service payment system aligns with perspectives in health economics, such as those by Zink et al. (2024) in a recent *JAMA Internal Medicine* article, on the implications of reimbursing healthcare providers for the full cost of acquiring AI tools. Specifically, Zink et al. (2024) argue such reimbursement can incentivize providers to use AI even when unnecessary, leading to wasteful healthcare spending. However, their argument assumes the quality of AI tools is exogenous and independent of the provider payment model. By accounting for the *endogenous quality* of AI tools, our analysis enriches this perspective to reveal the upstream effects of provider payment schemes. Full cost coverage can distort the incentive structure for AI developers. When reimbursed for the full costs of acquiring and using AI, providers often extend its application broadly across all patients. This usage pattern dampens developers' incentives to improve the quality of their AI tools. Our finding highlights the importance of jointly evaluating the usage and quality implications of AI tools within a fee-for-service payment framework.

Finally, given the inefficiencies inherent in both the status quo of no reimbursement and the fee-for-service system, we propose a hybrid payment model to better align the incentives of both providers and AI developers. This model integrates elements of the fee-for-service approach with those of a value-based payment system. However, the concept of *value* in healthcare is notoriously ambiguous (Reinhardt 2016). We explore two potential definitions of value in the context of AI in healthcare: (1) value as the *benefit* derived from the use of the AI tool, and (2) value as the *quality* of the AI tool itself. Under the benefit-based definition, we find a value-based payment system tends to favor the treatment of more complex patients, providing limited incentives for providers to use AI in less-complex cases. This incentive can lead to suboptimal outcomes where AI is underutilized in less complex cases, even when it could be beneficial to those patients. By contrast, the quality-based definition incentivizes providers to prioritize the use of higher-quality AI tools, with these incentives independent of patient complexity.

We find a quality-based payment system works best when physician altruism is relatively low. Although more altruistic physicians leading to better outcomes may seem intuitive, we show that when altruism is high, physicians become less responsive to financial incentives, resulting in selective, rather than broad, use of AI. This reduces the effectiveness of the payment system in ensuring the socially optimal full patient coverage. On the other hand, when altruism is lower, the payment model is more successful in encouraging physicians to use AI across all patients.

When the quality-based payment system proves ineffective under high levels of physician altruism, we propose a hybrid payment system that compensates providers through a combination of fee-for-service and quality-based payments. We show that, as long as the fee-for-service payment is

sufficiently high, this hybrid model induces providers to choose the first-best quality level. At the same time, it incentivizes providers to use the AI system for the same patient population as in the first-best scenario. This hybrid approach offers a promising solution to balance the dual goals of promoting appropriate use and ensuring high quality of AI tools.

Interestingly, as the cost of developing the AI system increases, we find the minimum required fee per service *decreases*. Although this may seem counterintuitive, the reason is that higher development costs lead to a lower optimal quality level in the first-best scenario, thereby requiring less intensive incentives to induce the first-best outcome.

Our paper represents a first attempt to understand the tension between upstream development efforts and downstream deployment decisions for medical AI. As noted above, CMS currently considers the costs of using GenAI tools by healthcare providers to be indirect and, therefore, does not reimburse them for using AI. However, CMS has demonstrated a willingness to experiment with innovative payment models to encourage the adoption of medical AI tools, particularly in light of the ongoing boom in GenAI, which is poised to transform much of healthcare (Zink et al. 2024). Our paper provides a theoretical foundation for policymakers as they navigate the trade-offs of various provider payment models. The proposed coordinated payment model, which combines fee-for-service with value-based payments, aligns with industry trends toward payment reforms designed to facilitate the integration of medical AI tools into routine care (Abràmoff et al. 2024).

2. Literature

Our paper contributes to several streams of literature: (1) physician payment models, (2) the economics and operations of AI, and (3) human-AI interaction. Our paper is also conceptually related to the new technology adoption literature.

First, provider payment models have emerged as an important topic in the healthcare operations management literature (Betcheva et al. 2021; Dai and Tayur 2020; Keskinocak and Savva 2020). The literature stream has explored prospective payment systems (Dada and White 1999), bundled payment models (Adida et al. 2017; Andritsos and Tang 2018; Guo et al. 2019; Vlachy et al. 2023), hospital readmissions reduction programs (Andritsos and Tang 2018; Arifoğlu et al. 2021; Zhang et al. 2016), reference pricing (Nassiri et al. 2022; Savva et al. 2019), out-of-pocket expenses (Dai et al. 2017), referral services (Adida and Bravo 2019), outcomes-based reimbursement (Adida 2021; Xu et al. 2022; Zorc et al. 2023), and payment models for diagnostic services (Adida and Dai 2024). To our knowledge, no prior work has examined provider payment models in the context of AI tools, where both upstream (AI development) and downstream (AI deployment) decisions are intertwined.

Our work also connects to another strand of the health care operations management literature that examines the misalignment of incentives between an upstream producer (e.g., a pharmaceutical

manufacturer) and a downstream user (e.g., a healthcare provider); see, for example, Chick et al. (2008), Taylor and Xiao (2014), Xu et al. (2022), and Zhang et al. (2020). Departing from this literature, we study a situation where the downstream user's actions benefit the end users (patients) but do not directly benefit themselves in that the healthcare provider is reimbursed by a third party.[2] By comparison, in the literature, the downstream users' actions are usually directly tied to their revenue. Because of this key difference, the design of the provider payment model is key to aligning upstream development and downstream deployment decisions.

Second, our paper contributes to the growing literature on the economics and operations of AI, addressing both upstream and downstream considerations. For instance, Gurkan and de Véricourt (2022) examine the AI flywheel effect, where predictive models improve as larger training datasets lead to more accurate predictions, which in turn generate additional data from broader use. They show firms often provide datasets larger than optimal, fostering effective monitoring of developer effort and initiating a virtuous cycle of product improvement. In the context of medical AI devices, Dai and Tayur (2022) emphasize the importance of physician buy-in and patient acceptance for successful integration, outlining service-design principles to facilitate both, building on earlier research that explores how physicians' use of AI impacts their professional reputation (Dai and Singh 2020) and how the sequencing of service activities shapes customer experiences (Das Gupta et al. 2016; Li et al. 2022). Agrawal et al. (2024) highlight the need for system-wide organizational changes to support AI adoption, particularly in settings with interdependent decisions, where modular structures reduce disruptions and coordinated processes enhance AI's effectiveness. Perhaps most relevant to our work, recent commentaries in medical journals, such as those by Abràmoff et al. (2024) and Parikh and Helmchen (2022), address the challenges of compensating providers for using predictive AI tools without formalizing their ideas. Different from existing work, our paper focuses on GenAI tools in healthcare, which present unique challenges, including extensive customization requirements that drive the quality of such tools.

Third, our work also contributes to the literature on human-AI interaction. de Véricourt and Gurkan (2023) study the case where a decision-maker is uncertain about the accuracy of an AI algorithm and updates his belief about its accuracy by interacting with it over time. Grand-Clément and Pauphilet (2023), in a continuous, general-prediction framework, contend the optimal AI algorithms are not necessarily those with the highest levels of predictive accuracy, but those that take into account the downstream adherence behavior of users. They develop an adherence-aware optimization framework to address the accuracy-adherence tradeoff. Recent research has increasingly focused on the effectiveness of integrating AI with human judgment to improve decision quality in

[2] Requiring patients to pay for AI solutions out of pocket, although possible, is a source of health inequity and not a sustainable solution (Abràmoff et al. 2024).

areas such as healthcare (Mullainathan and Obermeyer 2021; Orfanoudaki et al. 2022) and retail (Karlinsky-Shichor and Netzer 2024). One possible mechanism of synergy may be that AI improves predictive accuracy, whereas humans provide interpretation and contextual insights (Agrawal et al. 2018). Similarly, Boyacı et al. (2024) document accuracy gains from human-AI collaboration, albeit at some cognitive cost to human agents. Other studies similarly highlight that human oversight of AI decision-making improves outcomes and can support cognitive growth in the workplace (Chen et al. 2022; Kim et al. 2024). To our knowledge, none of these papers examines how the incentive structure for downstream use of AI affects upstream AI development. By modeling the impact and optimal design of provider payment models, our work has broader implications for aligning AI development and deployment with societal goals.

Finally, our paper is conceptually related to the new technology adoption literature (e.g., Cho and McCardle 2009; Cohen et al. 2016; Kundu and Ramdas 2022; Uppari et al. 2019; Zhang and Lee 2022), which has examined the interaction between a technology supplier's quality decision and a buyer's technology adoption decision in the face of a heterogeneous patient population. Unlike these papers, which study the decision to adopt an exogenous technology, our paper examines how to pay the healthcare provider for using the technology, which involves a third party (the payer); the payment system itself may influence the endogenous decision about the quality of the technology as well as how the technology is used. Put differently, our paper advances this literature by examining the interaction between the development and deployment of technology as shaped by the incentive environment designed by a third party.

3. Model

In this section, we describe our model setup and notation system. Our model captures the interaction between an AI developer and a healthcare provider serving patients. Accordingly, it is divided into two phases: (1) AI development and (2) AI deployment.

In the first stage (AI development), the AI developer decides on the quality and price of the AI system such as an ambient listening AI tool (e.g., Tierney et al. 2024). Specifically, the developer determines the quality level, q, and sets a price, p, for using the AI system on a per-use (i.e., per-patient) basis.[3] The quality of the AI system, denoted by q, is directly tied to the resources invested by the developer, such as time, staffing, and computational power; substantial fine-tuning activities are required to ensure quality before deploying an AI model for clinical use (Yaraghi 2024). The cost to the developer for choosing a quality level q is given by cq^2, where c represents the quality cost coefficient. This quadratic cost function reflects the increasing marginal cost of improving quality, a

[3] Charging healthcare providers for the use of GenAI on a per-use basis is a common practice (Cohen and Toubiana 2024), in no small part because of the non-trivial variable costs associated with using AI (Rand 2023).

standard assumption in the literature (see, e.g., Lahiri and Dey 2013; Moorthy 1988). The developer's objective is to choose a quality level $q > 0$ and a price $p > 0$ that balance the cost of development with the expected revenue from selling the AI tool to the provider. The trade-off involves setting a price that maximizes profit while ensuring the quality is adequate for the provider to use the AI tool. Alternatively, the developer can opt out, resulting in no participation and no profit. We use a tie-breaking rule such that if indifferent, the developer does not participate.

In the second stage (AI development), the healthcare provider decides whether and how to use the AI tool in each patient case, based on the quality and price of the AI tool. Patients are heterogeneous with respect to their complexity level, which we denote by $x \in [0, 1]$. For simplicity of analysis, the complexity level is assumed to follow a uniform distribution over the interval $[0, 1]$, and the total number of patients is normalized to one. The provider selects a subset of patients, denoted by the interval $[x_1, x_2] \subseteq [0, 1]$, on whom to apply the AI tool. The provider's objective consists of both the costs incurred and the benefits derived from using the AI tool. Specifically, the provider's utility function incorporates the per-use fee charged by the developer and the patient welfare. The welfare term reflects the provider's altruism, represented by a constant δ, which weighs the patient benefit relative to the provider's costs. This mixed-incentive model aligns with the literature on physician agency (Bester and Dahm 2017; Gaynor et al. 2023; Jelovac 2001; McGuire 2000), reflecting that providers balance their financial considerations with the quality of care they deliver. We use a tie-breaking rule such that if indifferent, the provider does not use the AI.

A patient with complexity x can be treated with or without the GenAI tool, where the benefit of using the tool is modeled as $B(x, q) = b \cdot q \cdot x$, with b capturing the economic value of enhanced care. This benefit increases with both the quality of the GenAI tool (q) and the complexity of the patient's condition (x). For example, ambient listening systems, which automate electronic health record (EHR) documentation, enable physicians to pay undivided attention to patients, particularly in complex cases where documentation burdens are high (Topol 2019). For simpler cases with minimal documentation needs, the incremental benefit of such tools is smaller, because these cases can be effectively managed even without AI. The value of the GenAI tool depends critically on its quality. High-quality tools integrate seamlessly into clinical workflows, improving documentation accuracy and reducing physician workload (Cohen and Toubiana 2024). Providers must balance the cost of using these tools against the anticipated benefits, weighing whether its benefits justify the price of high-quality AI. This balancing creates a trade-off for providers: high-quality tools enhance care for complex cases but come with higher costs. Providers optimize this trade-off by selectively using AI tools in scenarios in which the benefits outweigh the costs.

Our model captures the strategic decisions of both the AI developer and the healthcare provider, reflecting the interplay between AI development and deployment. In addition, our model incorporates

the role of a payer who must determine whether and how to reimburse providers for using GenAI tools. We first analyze two provider payment models: (1) no reimbursement, reflecting the status quo; and (2) fee-for-service reimbursement, where providers receive a per-use payment for using GenAI tools. Our analysis evaluates the impact of these payment models on the quality, pricing, and usage of GenAI tools. By comparing these outcomes with the first-best scenario, in which social welfare is maximized, we highlight inefficiencies inherent in the considered models. Furthermore, we explore whether reimbursement rules can be redesigned to achieve first-best outcomes. In doing so, we explore novel payment models that align the incentives of AI developers and healthcare providers.

3.1. Discussion on Modeling Assumptions

We make several simplifying assumptions to be aligned with the health economics and healthcare operations management literature, and, in certain cases, to make our model tractable.

First, we conceptualize the physician's goal as maximizing a weighted sum of her direct financial payoff (compensation from the payer minus the fee paid to the AI developer) and patient welfare. Accounting for financial incentives in physician decision-making is prevalent in both the health economics (e.g., Bester and Dahm 2017; Jelovac 2001) and healthcare operations management (e.g., Adida and Dai 2024; Guo et al. 2019) literature. Our model incorporates this consideration by including a term proportional to the patient's utility in the physician's objective function. This approach internalizes the physician's concern for the patient's welfare.

Second, we posit that the benefit to the patient of using the AI tool increases linearly with both the quality of the tool and the complexity of the case. This linear assumption is made primarily for tractability. In real-world scenarios, the benefits of AI tools may exhibit non-linear characteristics due to various factors, such as diminishing returns to quality improvements. Nevertheless, we expect our main results to remain robust as long as the relationship between utility, quality, and patient complexity is monotonic and concave.

Third, we study a pricing scenario where the price of the AI tool is fully endogenous and set by the AI developer. This endogenous pricing assumption allows us to analyze the strategic behavior of the AI developer in setting the optimal price that balances potential revenues and development costs. However, scenarios exist in which prices may be exogenously determined due to factors such as the market power of the provider or budget constraints. Nevertheless, our core findings are robust and applicable even under such exogenous pricing conditions. We can show the qualitative nature of our results remains unchanged when prices are externally imposed, as the fundamental trade-offs between cost, quality, and adoption persist.

Fourth, we focus on the physician's decision to use AI and treat it as independent from the patient-specific decision-making process. The reason is that the decision to adopt the AI system is often made

before the patient's precise condition is revealed. In this way, the AI adoption is seen as a general choice made upfront based on a basic understanding of the case complexity.

Finally, we assume the payer determines the provider payment model before the AI developer sets the price. This assumption allows us to examine how different reimbursement models, such as fee-for-service or value-based payments, affect the developer's pricing and quality decisions and the subsequent provider's AI use decisions. Even if in some cases the price may be set before the reimbursement policy, the fundamental relationships between quality, price, and reimbursement in our model remain valid across different timing structures.

3.2. Benchmark Equilibrium: First-Best Scenario

In the first-best scenario, the goal of a social planner is to select both the quality level of the AI tool and the set of patients on whom the tool is used to maximize total welfare. This welfare is defined as the combination of patient benefits and quality costs, where the price represents an internal cash-flow exchange and does not affect the first-best outcome.

The social planner has the option of either not participating (resulting in an objective value of zero) or participating if a positive objective value can be achieved by selecting an optimal quality level and identifying the appropriate set of patients to benefit from the AI tool. If the social planner chooses to participate, the goal is to maximize the following objective function:

$$-cq^2 + bq \int_{x_1}^{x_2} x \; dx = -cq^2 + bq \cdot \frac{x_2^2 - x_1^2}{2},$$

where q represents the quality level of the AI tool, c is the cost coefficient associated with quality, b is the benefit scaling parameter, and x_1 and x_2 denote the range of patient complexity levels. The integral term $\int_{x_1}^{x_2} x \; dx$ reflects the aggregate complexity of the patient population, indicating the social planner prioritizes using AI on higher-complexity patients due to their greater potential benefit from AI assistance.

LEMMA 1. *In the first-best scenario, the social planner chooses to participate, and the socially optimal quality level is given by* $q^{FB} = \frac{b}{4c}$. *Under this optimal quality level, the social planner uses the AI tool for all eligible patients within the specified complexity range.*

Lemma 1 highlights that the socially optimal quality level balances the marginal benefits of increasing quality with the marginal costs. The optimal quality $q^{FB} = \frac{b}{4c}$ ensures the AI tool is used effectively across all eligible patients, maximizing social welfare. This benchmark equilibrium serves as a reference point for comparing outcomes under different reimbursement scenarios throughout the rest of the paper.

4. Analysis of Provider Payment Models

In this section, we examine two payment models for providers using the GenAI tool. The first model represents the status quo in the U.S., where the provider receives no reimbursement for using these technologies. The second model aligns with the fee-for-service approach, where the provider is reimbursed on a per-use basis for using the GenAI tool.

4.1. Status Quo: No Reimbursement

Under the current reimbursement framework, the U.S. Centers for Medicare & Medicaid Services (CMS) consider the costs associated with the acquisition and use of GenAI to be indirect, treating them similarly to office equipment and software (CMS 2021). As a result, physicians do not receive direct reimbursement for incorporating AI into clinical practice. To analyze this scenario, we investigate the physician's decision to use AI and the developer's decision regarding the quality of AI in the absence of reimbursement.

In the first stage, the AI developer aims to maximize their profit by solving:

$$\max_{p,q>0} \quad -cq^2 + p(x_2 - x_1),$$

where the parameters x_1 and x_2 are determined in the second stage and may depend on the selected price p and quality q.

In the second stage, given the price p and quality q decisions made by the developer, the provider solves:

$$\begin{aligned}\max_{x_1,x_2} \quad & -p(x_2 - x_1) + \delta bq \int_{x_1}^{x_2} x\,dx \\ & = -p(x_2 - x_1) + \delta bq \frac{x_2^2 - x_1^2}{2}.\end{aligned}$$

Before determining the developer's decisions, we first derive the provider's optimal decision in response to the developer's given price and quality choices:

Lemma 2. *In the second stage, given the price $p > 0$ and quality $q > 0$ selected by the developer, the provider uses the AI on some patients if and only if $p < \delta bq$. In this case, the provider uses the AI for patients with complexity $x \in \left[\frac{p}{\delta bq}, 1\right]$.*

The intuition behind this result is that AI will be used only when its quality is sufficiently high or its price is low relative to the benefits derived from patient welfare. Because the benefits of using AI are greater for more complex patients, whereas the costs remain uniform across patients, AI is used for the most complex cases.

Proposition 1. *In the first stage, the developer opts to participate and selects a price $p^* = \frac{(\delta b)^2}{16c}$ and quality $q^* = \frac{\delta b}{8c}$. In the second stage, the provider uses the AI on half of all patients, specifically those with complexity $x \in \left[\frac{1}{2}, 1\right]$.*

Proposition 1 has several implications. First, as long as the altruism parameter δ is less than 2, the equilibrium quality of the AI tool is lower than the socially optimal quality level. This result reveals a source of inefficiency in the absence of reimbursements, where the quality of the AI tool does not reach its potential socially optimal value, due to insufficient financial incentives for the developer under the status quo. Second, the proposition means that, regardless of the parameter values, only half of the patients receive AI treatment in the equilibrium scenario, in contrast to the first-best scenario (see Lemma 1) where all patients would benefit from the AI tool. This disparity highlights a gap between the equilibrium and the socially optimal outcome. Third, efforts aimed at influencing the provider's degree of altruism (δ)—also referred to as patient-centeredness (Bergeson and Dean 2006)—or modifying the development cost (c) are insufficient to fully align the equilibrium outcome with the first-best scenario. Although increasing the altruism parameter to a high level would align the quality level with first-best, it would not address the discrepancy in the proportion of patients receiving AI treatment.

In conclusion, the absence of reimbursement for AI tools under the status quo hinders the attainment of socially optimal outcomes. This realization motivates the need to explore alternative reimbursement models that better align the incentives of the AI developer and the healthcare provider.

4.2. Fee-for-Service Reimbursement

We now examine a fee-for-service payment system, which has been the primary method of reimbursing providers for the use of FDA-approved AI devices (Parikh and Helmchen 2022; Wu et al. 2023). The simplicity of the fee-for-service model makes it an attractive option for reimbursing physicians for the use of AI tools (Abràmoff et al. 2024).

Under the fee-for-service payment system, the provider receives a reimbursement for each patient on whom AI was used. This reimbursement f represents the fixed amount paid to the provider per use of the AI tool.

In the second stage, given the quality decision q and the price decision p chosen by the developer in the first stage, the provider aims to maximize his net benefit from using the AI tool. The provider's problem can be formulated as

$$\begin{aligned}\max_{x_1,x_2} \quad & (f-p)(x_2-x_1)+\delta bq\int_{x_1}^{x_2} x\,dx \\ & =(f-p)(x_2-x_1)+\delta bq\frac{x_2^2-x_1^2}{2},\end{aligned}$$

where x_1 and x_2 denote the range of patient complexity, δ is the provider's degree of altruism, and b is the utility scaling parameter.

In essence, the provider's problem under the fee-for-service model mirrors the problem without reimbursement, but with one important modification: the price p is adjusted by the reimbursement

f. Specifically, the effective price considered by the provider is $p - f$. This adjusted price can be positive or negative, depending on the magnitude of f. If f is large enough to exceed p, the effective price becomes negative, indicating a net gain to the provider per use of the AI tool.

This adjustment reflects the economic rationale that higher reimbursements foster broader adoption of AI tools, because they reduce the net-cost burden on providers. Conversely, if reimbursement is inadequate, the net cost may still deter providers from using the AI tool, particularly for less complex patients for whom the perceived benefit does not outweigh the cost. The provider's decision is thus influenced by the balance between the reimbursement received and the price paid, which ultimately affects the range of patient complexity x for which the AI tool is used. We characterize the provider's decision in the following lemma.

LEMMA 3. *In the second stage, given the quality decision $q > 0$ and price decision $p > 0$ selected by the developer in the first stage,*

(i) if $f \geq p$, the provider uses AI on all patients;

(ii) if $p - \delta bq \leq f < p$, the provider uses the AI on patients with complexity $x \in \left[\frac{p-f}{\delta bq}, 1\right]$;

(iii) else, the provider does not use AI.

The intuition behind Lemma 3 is straightforward: a high fee motivates the provider to use AI for all patients. At intermediate fee levels, the provider limits AI use to more complex cases, with usage increasing as the fee increases. When the fee is too low, the provider avoids using AI entirely.

This lemma aligns with the economic arguments presented in the recent commentary by Zink et al. (2024), which emphasizes the need to avoid full cost coverage of AI by payers to mitigate inefficiencies in the healthcare system. According to Zink et al. (2024), reimbursing the full cost of AI services may result in overutilization and excessive healthcare spending, because providers have little incentive to limit the use of AI tools when the financial burden is entirely borne by the payer. Their argument, however, assumes AI quality is exogenous and unaffected by the reimbursement model. Our analysis enriches their argument by highlighting that covering the full cost of AI can also undermine the incentive structure for AI developers. Specifically, as we see next, when payers cover the entire cost of AI services, developers are less motivated to invest in enhancing the quality of their AI systems. This guaranteed reimbursement creates a financial cushion that diminishes the pressure to improve AI performance.

The proposition below characterizes the developer's decision in the first stage. For ease of presentation, we define

$$\varphi(q, f) = 8\delta bcq^3 - \delta^2 b^2 q^2 + f^2.$$

PROPOSITION 2. *In the first stage, the developer's optimal decisions are as follows:*

(i) If $f \geq \frac{\delta^2 b^2}{27c}$*, the developer sets the quality at* $q = \epsilon$ *(i.e., as small as possible) and the price at* $p = f$*. In this scenario, the provider uses AI on all patients.*

(ii) If $f < \frac{\delta^2 b^2}{27c}$*, the optimal decisions are to set the quality at* $\bar{q}_2$*, the unique solution in* q *of* $\varphi(q, f) = 0$ *on* $[\delta b/(12c), \infty)$*, and the price at* $\bar{p}_2 \equiv (\delta b \bar{q}_2 + f)/2$*. In this case, the provider uses AI on patients with complexity* $[1/2 - f/(2\delta b \bar{q}_2), 1]$*.*

Proposition 2 can be interpreted as follows. When the fee per patient using AI is high, the developer anticipates AI to be used for all patients. As a result, the developer selects the quality level to be low, minimizing costs, and sets the price equal to the reimbursement to maximize revenue while ensuring AI usage. When the fee is less high, quality is intermediate and AI is used only on a fraction of patients.

Note $\partial\varphi(q, f)/\partial q = 2\delta b q(12cq - \delta b)$ is positive on $[\delta b/(12c), \infty)$. It follows that the quality level $\bar{q}_2$, the unique root in q of $\varphi(q, f)$ on $[\delta b/(12c), \infty)$, must increase as f decreases to maintain $\varphi(q, f) = 0$. This property indicates an inverse relationship between the fee and the quality: lower fees necessitate higher quality to ensure equilibrium conditions are met. Hence, we have the following corollary:

COROLLARY 1. *(i) The quality of the AI tool* q^* *weakly decreases as the reimbursement rate* f *increases.*

(ii) The price of the AI tool p^* *increases with the reimbursement rate* f*.*

Conventional health-policy discourse contends that low reimbursement rates are responsible for poor quality in medications (Hernandez 2023) and limited access to healthcare services (Alexander and Schnell 2024). However, Corollary 1(i) reveals this intuition may not apply to physicians' use of AI tools: higher reimbursement rates can, seemingly paradoxically, lead to *lower* quality. The mechanism driving this result lies in the behavior of AI developers. Increased reimbursement expands the provider's use of the AI tool across a broader range of patients. Anticipating this wider market, developers may strategically reduce the quality of their tools to minimize development costs while maintaining substantial market share.

Importantly, this finding does not negate the idea that higher reimbursement enhances patient access. Instead, it highlights the trade-off between access and quality in designing provider payment models. To achieve both goals, policymakers must carefully consider how reimbursement structures influence developer incentives.

In light of Lemma 3, Corollary 1 implies that although lower reimbursement rates may encourage higher quality by limiting the financial pressures on developers to reduce costs, they also restrict the provider's use of AI to a narrower subset of patients. This finding reflects a constrained market environment where both quality and price are adjusted downward. Consequently, setting reimbursement rates too low risks undermining the broader uptake of AI tools, reducing their potential societal

benefits. Policymakers, therefore, face the challenge of striking a balance between ensuring quality and accessibility in the reimbursement design, aiming to achieve both in tandem.

COROLLARY 2. *The fee-for-service payment model does not achieve coordination with the first-best outcome. Moreover, when $\delta < 1.5$, the quality level under fee-for-service is strictly lower than the first-best quality.*

Corollary 2 builds on Corollary 1 by highlighting the consequence of the tension between accessibility and quality of AI tools inherent within the fee-for-service payment system. Specifically, setting the fee-for-service rate f high enough may align the coverage of AI services with the first-best scenario, ensuring all patients have access to the AI tool. However, this full coverage comes at the cost of significantly diminished quality—the quality level would approach zero, far below the first-best benchmark. Conversely, reducing the fee could lead to improvements in quality, but these improvements would still fall short of the first-best standard, and the reduction in fee would simultaneously decrease the number of patients who receive AI services. This trade-off reveals the fee-for-service model's limitations in balancing quality and accessibility, calling for a better-designed payment scheme, which we examine in the next section.

5. Value-Based Provider Payment: Benefit vs. Quality

Given the inefficiencies inherent in both the no-reimbursement status quo and the traditional fee-for-service system, we now explore alternative payment models that can more effectively align the incentives of providers and AI developers.

One potential remedy is value-based payment models, which have gained traction as alternatives to traditional fee-for-service systems by aiming to incentivize providers based on the value they deliver, rather than the volume of services provided (Abràmoff et al. 2024; Adida and Bravo 2019). However, despite its popularity, value-based payment is not without challenges—chief among them is the difficulty in precisely defining what "value" means in the context of healthcare (Reinhardt 2016). In this section, we examine two potential definitions of value in the context of AI in healthcare: (1) value as the *benefit* derived from the use of the AI tool, and (2) value as the *quality* of the AI tool itself. These definitions guide our analysis of how different payment models can impact the adoption and quality of AI technologies.

5.1. Benefit-Based Payment

We start with considering a payment structure whereby reimbursement is tied directly to the benefit provided by the AI tool in improving patient outcomes, formalized as $f + \gamma B$, where f and γ are constants, and B represents the value delivered to the patient.

Given the developer's quality decision q and price decision p in the first stage, the provider solves:

$$\max_{x_1,x_2} \int_{x_1}^{x_2} (f+\gamma bqx-p)\,dx + \delta bq \int_{x_1}^{x_2} x\,dx = (f-p)(x_2-x_1) + (\gamma+\delta)\,bq\frac{x_2^2-x_1^2}{2}.$$

This formulation shows the benefit-based payment model intensifies the provider's incentives to deliver value, as captured by $\delta+\gamma$. However, this model is inherently biased toward more complex cases, because they generate higher measurable benefits. Consequently, providers are incentivized to prioritize more complex patients, resulting in a misalignment with the first-best scenario, where care would be optimally distributed across all patient types.

LEMMA 4. *No benefit-based payment can achieve the first-best outcome.*

The intuition behind Lemma 4 is that although the benefit-based payment model strengthens incentives for high-quality care by tying reimbursement to the benefit generated by the use of AI, it inherently favors more complex cases. Under this model, providers receive higher reimbursement for using AI on complex patients because these cases yield greater measurable value.

Paying providers based on the benefits delivered, although theoretically more appealing than the status quo or the fee-for-service model, disproportionately incentivizes care for more complex patients. For this reason, this intuitive design can end up *widening* disparities in access to AI tools, leaving less complex cases underserved.

5.2. Quality-Based Payment

Given the limitations of the benefit-based model, we next explore a payment structure that aligns provider incentives with the intrinsic quality of the AI tools themselves. Specifically, we consider a reimbursement scheme of the form γq, where q denotes the quality of the AI tool and γ is a constant.

This approach is designed to align both upstream and downstream incentives by encouraging developers to invest in higher-quality AI systems while ensuring providers are motivated to use these tools appropriately across different patient complexities. Under this model, the provider's decision-making process, given the quality decision q and price decision p by the developer, is captured by:

$$\max_{x_1,x_2} \int_{x_1}^{x_2} (\gamma q-p)\,dx + \delta bq \int_{x_1}^{x_2} x\,dx = (\gamma q-p)(x_2-x_1) + \delta bq\frac{x_2^2-x_1^2}{2}.$$

Here, the provider's use of AI is contingent on the following conditions:

- If $p-\gamma q \le 0$, the provider uses AI for all patients.
- If $p-(\gamma+\delta b)q \le 0 < p-\gamma q$, AI is used only for patients with complexity $x \in \left[\frac{p-\gamma q}{\delta bq}, 1\right]$.
- Otherwise, the provider does not use AI.

To determine the quality, q, and the price, p, the developer faces two potential optimization problems: one that results in full coverage and another that leads to partial coverage. The solution entails

comparing the payoffs associated with each scenario. Coordination is achieved only when the full-coverage case dominates the partial-coverage case, as the social optimum corresponds to full coverage.

The following proposition provides the necessary and sufficient condition for the quality-based payment system to coordinate the system:

PROPOSITION 3. *Setting $\gamma = \frac{b}{2}$ enables the quality-based payment scheme to align decisions with the first-best outcome when $\delta \leq \frac{1}{2}$. However, when $\delta > \frac{1}{2}$, the provider uses AI on only a subset of patients, thereby making coordination unattainable.*

Proposition 3 show that, under appropriate conditions, a quality-based payment model can align the incentives of the AI developer and healthcare provider, achieving a first-best outcome. The conditions reveal a counterintuitive relationship between physician altruism, denoted by δ, and the effectiveness of quality-based payment systems in achieving the social optimum. Conventional wisdom in health policy suggests inefficiencies arise primarily from physicians being insufficiently patient-centric (i.e., δ is too low). This perspective is behind many policy reforms aimed at fostering patient-centered care by increasing the emphasis on patient welfare (i.e., increasing δ), thereby potentially reducing the role of financial incentives in shaping physician behavior. However, the above proposition presents a paradox: for a quality-based payment system to align physician actions with the first-best outcome, δ must be sufficiently small. Specifically, when $\delta \leq \frac{1}{2}$, setting the quality coefficient γ at $b/2$ optimally balances the trade-offs between quality care and cost efficiency, incentivizing physicians to make decisions that align with the socially optimal outcome.

As δ increases, the physician's intrinsic motivation to prioritize patient outcomes renders financial incentives less effective. In the case in which $\delta > \frac{1}{2}$, the second optimization problem becomes dominant (with an interior stationary point), inducing the provider to use AI on only a subset of patients, thereby making coordination unattainable. This result aligns with recent literature highlighting the complexities of designing payment systems in healthcare. Studies suggest that although altruistic motivations are desirable, they can weaken the influence of incentive structures designed to improve efficiency and quality outcomes. For instance, Prendergast (2007) argues high intrinsic motivation may limit the ability of external rewards to influence behavior. Echoing these findings, Proposition 3 reveals the nuanced role of δ in shaping healthcare efficiency, meaning policies solely focused on increasing patient-centeredness may overlook the role of financial incentives in shaping physician behavior.

6. A Hybrid Payment System

We have shown in the preceding section that the quality-based payment system proves ineffective under high levels of physician altruism. We now shift our focus to a hybrid payment system that

merges aspects of both fee-for-service and quality-based payment mechanisms. In this system, the provider is compensated with a per-visit fee of f in addition to the quality-based payment (γq). Because coordination is achievable using a quality-based payment when $\delta \leq 1/2$ (see Section 5.2), we focus in this section on the case of $\delta > 1/2$.

6.1. Coordinating Hybrid Model

Under this hybrid payment model, the provider's decision-making process in the second stage, given the quality decision q and price decision p by the developer, is captured by:

$$\max_{x_1,x_2} \int_{x_1}^{x_2} (f + \gamma q - p)\, dx + \delta b q \int_{x_1}^{x_2} x\, dx = (f + \gamma q - p)(x_2 - x_1) + \delta b q \frac{x_2^2 - x_1^2}{2}.$$

Under a hybrid payment structure combining fee-for-service and quality-based incentives, the developer's decision-making differs significantly depending on the extent of market coverage. We examine two cases: full and partial market coverage, each reflecting distinct economic and operational conditions for optimal quality and pricing decisions.

Case 1: Full Market Coverage. When the developer anticipates full market coverage, it maximizes utility by selecting quality (q) and price (p) to achieve optimal revenue subject to a minimum fixed payment (f):

$$\max_{p,q>0} -cq^2 + p \quad \text{s.t.} \quad p - \gamma q \leq f.$$

Solving this yields an optimal quality level $q = \gamma/(2c)$ and a price $p = f + \gamma^2/(2c)$, resulting in an objective value of $f + \gamma^2/(4c)$.

Case 2: Partial Market Coverage. Under partial market coverage, the developer's utility depends not only on fee-for-service payments but also on the quality-sensitive revenue from covered patients. This scenario's objective function is:

$$\max_{p,q>0} \psi(p,q) = -cq^2 + p\left(1 - \frac{p - f - \gamma q}{\delta b q}\right),$$

subject to $p - \gamma q > f$ and $p - (\gamma + \delta b)q \leq f$. In this scenario, quality incentives are linked directly to coverage decisions. The solution yields an optimal quality level $q = (\delta b + \gamma)^2/(12c\delta b)$.

The comparison between full and partial coverage gives rise to the existence of a threshold payment $\bar{f}$ above which the provider's and developer's incentives are aligned under a hybrid model. A reimbursement amount above this threshold level guarantees the provider's financial incentives are sufficient for full AI usage, whereas the developer commits to a quality level that aligns with first-best outcomes. The following proposition provides the condition for the quality-based payment scheme to coordinate the AI developer's and the provider's incentives:

PROPOSITION 4. *Suppose $\delta \geq 1/2$, and let*

$$\tilde{f} = \begin{cases} \frac{\left(\delta+\frac{1}{2}\right)^3 b^2}{12\delta c\sqrt{3}} & \text{if } \delta > 1+\sqrt{3}/2 \\ \frac{b^2\left(\delta+\frac{1}{2}\right)^2\left(\delta-\frac{1}{2}\right)}{12\delta c} & \text{else.} \end{cases}$$

- *If $f \geq \tilde{f}$, setting $\gamma = \frac{b}{2}$ enables the quality-based payment scheme to align decisions with the first-best outcome.*
- *$f^* \in (0, \tilde{f}]$ exists such that coordination is achieved if and only if $\gamma = \frac{b}{2}$ and $f \geq f^*$.*

Proposition 4 first provides a sufficient condition leading to a coordinating payment scheme. Namely, it obtains a closed-form expression for $\tilde{f}$, the minimum fixed payment to the provider that ensures the first-best outcome arises. As long as the payment meets this minimum value, the provider has sufficient incentives to use AI on all patients, and the quality level the developer selects matches the first-best as well as long as γ is set adequately.

Proposition 4 also gives a necessary and sufficient condition for coordination, in the form of a proof of the existence of a threshold f^* on this fixed payment. Due to a lack of tractability, we do not have a closed-form expression on this threshold. However, an extensive numerical study, described below, makes clear that f^* is well approximated by $\tilde{f}$, and has the same monotonicity properties.

6.2. Numerical Illustration

To illustrate how the proposed hybrid payment scheme coordinates both upstream and downstream activities, we use ambient listening solutions in healthcare as a concrete example. These technologies have the potential to significantly reduce the time physicians spend on electronic health record (EHR) documentation, enabling greater focus on direct patient care. For instance, first-year internal medicine residents allocate approximately 87% of their work hours away from patients, with a significant portion dedicated to EHR tasks (Chaiyachati et al. 2019). Similarly, primary care physicians spend a median of 5.9 hours daily on EHR activities, including 1.4 hours after clinic hours (Tai-Seale et al. 2019). Ambient listening technologies automate EHR documentation, streamlining workflows and improving physician efficiency, while increasing patient-facing time (Topol 2019).

To calibrate the parameters used in our analysis, we focus on two key quantities: the scaling factor for benefit (b) and the cost coefficient (c). The parameter b quantifies the economic value of additional patient-facing time enabled by ambient listening technologies. The quality of the AI tool (q) is normalized such that $q = 1$ represents the maximum achievable reduction in EHR documentation time, and $q = 0$ reflects no reduction. Recent findings indicate AI scribes reduce documentation time for primary care physicians by 13.8%, from 10.9 to 9.4 minutes per appointment (Rotenstein et al. 2024). Assuming a value of \$100 for additional patient-facing time per patient, we estimate $b = \$100 \times 0.138 = \$13.8$ per unit of quality level per unit of complexity.

The parameter c captures the incremental cost of improving the quality of ambient listening systems, modeled quadratically to reflect diminishing returns. Using average cost data for AI scribe tools ($0.20–$8 per visit, with a midpoint of $4.10; Cohen and Toubiana 2024), we estimate the imputed unit development cost as $4.1. The cost function is modeled as $cq^2 = \$4.1$ at $q = 1$, so we obtain $c = \$4.1$.

Table 1 Parameter Values for Sensitivity Analysis

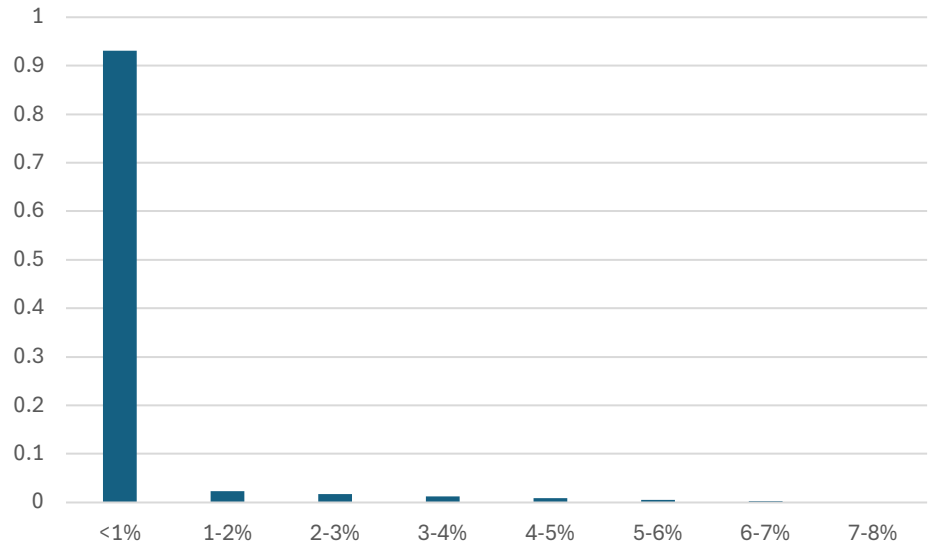
Parameter	Base-Case Estimate	Sensitivity Range	References
Benefit per unit quality level, b ($/unit)	13.80	10–20	Rotenstein et al. (2024)
Cost scaling factor, c ($/unit)	4.10	1–10	Cohen and Toubiana (2024)
Altruism parameter, δ	1.0	0.5–5.0	Adida and Dai (2024)

To conduct some sensitivity analysis when the key coefficients vary around their base-case values, we consider a wide range of scenarios of parameters as listed in Table 1. Using increments of 0.1, 0.1, and 0.05 for these parameters results in 810,000 parameter scenarios. In each scenario, we evaluate the gap between f^* and $\tilde{f}$, calculated as $(\tilde{f} - f^*)/\tilde{f}$. We also test in each scenario the monotonicity of f^* with respect to respectively c, b, and δ within the range described above. Our findings are the following: the maximum gap between f^* and $\tilde{f}$ across all scenarios equals 7.06%, but the average gap is much lower, at 0.22% (see Figure 1). In addition, in 93.13% of cases, the gap was smaller than 1%. This findings confirms $\tilde{f}$ is a very good approximation for f^*. Moreover, we obtained that in 100% of scenarios, f^* is monotonically decreasing in c, monotonically increasing in b and monotonically increasing in δ. Using the closed-form expression of $\tilde{f}$, it is straightforward to verify $\tilde{f}$ is monotonically decreasing in c, monotonically increasing in b, and monotonically increasing in δ. Hence, we find numerically that f^* has the same monotonicity properties as $\tilde{f}$.

Remark 1. We observe that the minimum fixed payment increases in δ and b and decreases in c.

An interesting result from the above remark is that the minimum required fee per service to enable coordination to the first-best decreases as the cost of developing the AI system increases. At first glance, this finding seems counterintuitive, because one might expect higher development costs to require higher per-service fees in the coordinating payment model. However, the intuition lies in how higher costs affect the optimal quality level in the first-best scenario. As development costs increase, the optimal quality level decreases in the first-best outcome. This reduction in the optimal quality level simplifies the coordination problem, because the incentives required to motivate the developer to achieve this reduced quality target are lower. Consequently, the fee per service decreases, even in a more expensive development environment. This observation reveals the intricate relationship between development costs, quality targets, and incentives in aligning developer and provider incentives.

Figure 1 **Histogram of the gap between f^* and $\tilde{f}$ across all scenarios**



7. Extensions

In this section, we explore several extensions to our main model, each addressing specific dimensions of the problem to test the robustness of our findings and assess the implications of alternative payment mechanisms. Section 7.1 examines the introduction of a usage cost, capturing operational expenses incurred by the provider. Section 7.2 analyzes a subscription-based pricing model, contrasting it with fee-for-service structures. Section 7.3 investigates the impact of provider-facing benefits. Finally, Section 7.4 evaluates the potential for subsidies to address misalignments between private and social incentives.

7.1. Usage Cost

We extend our analysis to include a model with a usage cost, representing a fixed cost u incurred by the provider for each patient treated with the AI system. This cost reflects operational expenses such as setting up the AI system, inputting and verifying patient-specific information, or managing technical requirements such as software updates and data integration. These setup and operational efforts are well-documented as integral to the deployment of AI tools in healthcare settings, particularly for systems that require customization or preprocessing for effective use.

In the first stage, the AI developer maximizes profit by solving

$$\max_{p,q>0} \quad -cq^2 + (p+u)(x_2 - x_1),$$

where x_1 and x_2 are determined in the second stage as functions of the price p and quality q. In the second stage, given p and q, the provider solves:

$$\max_{x_1,x_2} \quad -(p+u)(x_2 - x_1) + \delta bq \int_{x_1}^{x_2} x, dx.$$

The provider uses AI for cases with complexity levels in the range $x \in \left[\frac{p+u}{\delta bq}, 1\right]$, provided $p+u < \delta bq$. In the first stage, the developer thus solves:

$$\begin{aligned} \max_{p,q>0} \quad & -cq^2 + p\left(1 - \frac{p+u}{\delta bq}\right) \\ \text{s.t.} \quad & p+u \le \delta bq. \end{aligned}$$

The optimal price p satisfies:

$$\frac{(p+u)^2}{(2p+u)^3} = \frac{2c}{\delta^2 b^2},$$

and the optimal quality is given by:

$$q = \frac{2p+u}{\delta b}.$$

This solution ensures the constraints, including the second-order condition $q > \frac{\delta b}{12c}$, are satisfied. Participation by the developer depends on the profitability of this solution.

Unlike in the main model, introducing a usage cost changes the first-best outcome: full patient coverage may no longer be socially optimal. The fixed usage cost introduces a threshold complexity level below which the cost of AI usage outweighs its benefits.

Numerical experiments confirm that in the case in which the provider incurs a usage cost for using AI, the first-best outcome cannot be achieved under the status quo or a fee-for-service payment model. The results align with our findings from the main model. These limitations stem from the inherent misalignment of incentives in such payment structures. Our further numerical analysis shows a hybrid payment model, which combines fee-for-service with quality-based payments, can overcome these challenges even in the presence of a usage cost. Specifically, the hybrid model aligns the incentives of developers and providers, fostering the development and adoption of higher-quality AI systems. Also, it restores patient coverage to the same level as in the first-best outcome, ensuring equitable access across the population.

7.2. Subscription-based Pricing

An alternative pricing mechanism is subscription based (Cohen and Toubiana 2024). Under subscription-based pricing, the developer specifies a subscription fee per time unit and a quality level q. Upon acceptance of the contract, the provider gains unrestricted access to use AI for an unlimited number of cases without additional charges.[4] We denote p as the present value of the expected time-discounted subscription-fee revenue.

[4] Patient panel sizes are typically calculated based on the number of available appointment slots, clinician workdays, and the average number of patient visits per year, reflecting a capacity-driven approach to panel management (Paige et al. 2020). As a result, panel size is largely fixed, meaning that under a subscription model, the provider's decision focuses on whether to adopt the AI tool rather than selecting specific patients based on complexity.

In the first stage, the developer solves

$$\max_{p,q>0} \quad -cq^2 + p.$$

In the second stage, given the price p and quality q decisions made by the developer, if participating, the provider solves

$$\max_{x_1,x_2} \quad -p + \delta bq \int_{x_1}^{x_2} x\,dx.$$

If participating, the provider uses AI on all patients (i.e., $x_1 = 0,\ x_2 = 1$). As a result, the provider participates if and only if $p < \delta bq/2$. Hence, in the first stage, the developer sets $p = \delta bq/2 - \epsilon$ and selects the quality level q to maximize $-cq^2 + \delta bq/2$. It follows that the developer sets $q = \delta b/(4c)$ and its profit is positive; thus, the developer elects to participate. We summarize these results in the following proposition.

PROPOSITION 5. *Under a subscription model, the developer sets the quality level at $q = \delta b/(4c)$; the provider participates and uses AI on all patients.*

Although a subscription-based payment scheme between the provider and the developer addresses a separate issue from how the provider is reimbursed by the payer for AI usage, it is worth noting that such a model achieves coordination to the first-best outcome if and only if $\delta = 1$. For this level of altruism, both the quality level and the range of patients for whom AI is used are aligned to the first-best. However, the quality decision is misaligned if $\delta \neq 1$. A direct subsidy to the provider would not help address this misalignment. Specifically, a subsidy s granted to the provider would simply increase the provider's willingness to pay for the subscription, which would raise the subscription price by s, but would not change the developer's quality decision. Essentially, such a fixed subsidy would be transferred to the developer without enabling coordination of the quality decision to the first-best.

7.3. Note Time versus "Pajama Time"

In our main analysis, we posit that the benefit from using the AI tool accrues entirely to the patient, with no direct utility to the provider beyond altruistic considerations. In practice, this assumption may not hold. For instance, ambient listening tools can provide benefits beyond reducing note time during patient visits. In addition to improving patient engagement during consultations, they can reduce "pajama time"—the after-hours documentation burden often borne by providers (Lohr 2023). Such additional benefits could directly enhance the provider's utility from adopting the AI tool.

From a modeling perspective, this extension can be accommodated by reinterpreting the altruism parameter, δ, to reflect both patient-related benefits (e.g., improved engagement) and provider-facing

efficiencies (e.g., reduced documentation burden). This reinterpretation broadens the scope of δ to include provider utility derived from the AI tool. This adjustment does not alter our key findings. The parameter δ continues to serve as a composite measure of the alignment between private and social benefits in adopting AI tools. Whether the benefits accrue entirely to patients or are shared between patients and providers, the core trade-offs and policy implications—such as the role of payment models in incentivizing optimal adoption—remain robust.

7.4. Subsidies

Consider a price subsidy, whereby the provider is subsidized for a proportion of the AI system's price. In this framework, the provider would experience an effective price $p' > 0$, which is proportional to the nominal price p. Consequently, in the second stage, the provider's adoption threshold would adjust, enabling the use of AI for a greater subset of patients with complexity levels $x \in [p'/(\delta bq), 1]$. However, achieving first-best coordination requires AI utilization for all patients. A partial price subsidy, although expanding usage, cannot eliminate the misalignment between private and social incentives, thus falling short of the first-best outcome.

Direct subsidies to AI developers are generally not feasible due to practical and political constraints. For the sake of completeness of analysis, we consider a development-cost subsidy to the AI developer such that the experienced development cost is $c' > 0$, proportional to the initial cost c. In this case, the subsidy alters the developer's investment decision. However, as indicated by the equilibrium outcome, the provider would still use AI for only half of all patients, perpetuating the divergence from the first-best scenario. Thus, a partial subsidy for development costs, even if it was feasible, does not achieve coordination.

8. Conclusions

As GenAI becomes increasingly integrated into routine healthcare delivery, the design of appropriate reimbursement models is important to ensure the development and widespread use of high-quality AI tools. Current U.S. payment systems do not adequately address the unique financial and operational needs associated with the adoption of GenAI. Without the right incentives, both AI developers and healthcare providers may face barriers that prevent the effective use of these advanced tools in clinical practice.

Our paper presents a novel analytical framework to capture how various reimbursement models affect the development and use of GenAI in routine healthcare. We show existing U.S. provider payment models—either offering no reimbursement or relying solely on fee-for-service—are insufficient for promoting the development of high-quality AI tools and ensuring their optimal use in clinical practice. Under the status quo of no reimbursement, providers shoulder the full financial burden of

AI use, leading to limited adoption. This, in turn, results in suboptimal quality from developers, who lack the incentives to invest in improving their tools. On the other hand, whereas the fee-for-service model may drive wider AI use, it risks incentivizing the development of lower-quality tools, particularly when high reimbursement rates diminish the pressure on developers to produce high-quality AI. Our findings align with predictions from health economics (see, e.g., Zink et al. 2024) yet introduce a new dimension by examining the upstream effects of AI payment models, thereby addressing the impact on the *quality* of AI systems.

To address these shortcomings of the status quo (no reimbursement) and the fee-for-service payment model, we propose a hybrid payment model that integrates both fee-for-service payments and quality-based incentives. This approach effectively aligns the interests of AI developers and healthcare providers, promoting the development of high-quality AI systems while ensuring these tools are used appropriately across diverse patient populations. Crucially, the value component in this model should be tied to the quality of the AI system itself, rather than solely to the benefit generated by AI. By focusing on the quality of the AI system, this hybrid system ensures providers are incentivized to select and implement AI tools that deliver superior clinical performance, regardless of patient complexity.

One of the more nuanced insights from our analysis is the paradoxical role of physician altruism in quality-based payment systems. Although greater physician altruism may seem beneficial for patient-centered care, we find it can undermine the effectiveness of financial incentives. When physicians are highly motivated by patient welfare, they are less responsive to quality-based financial rewards, leading to selective AI use and limiting the overall coordination between developers and providers. As such, payment models must carefully consider the balance between financial incentives and the intrinsic motivations of healthcare providers to ensure optimal outcomes.

We also highlight a counterintuitive result: higher development costs can actually simplify the coordination problem, because they lead to lower optimal quality levels in the first-best scenario, thereby reducing the need for high-powered incentives. This finding challenges conventional wisdom and underscores the challenges of designing payment systems that balance the trade-offs between cost, quality, and uptake in healthcare.

Our paper serves as a foundation for future studies on AI reimbursement models, which could explore other important factors, such as market competition among AI developers, regulatory frameworks for ensuring safety and efficacy, and the role of bundled or outcome-based payment models in further optimizing AI adoption. As GenAI continues to evolve and reshape the healthcare landscape, thoughtful payment system design will be critical to unlocking the full potential of these technologies for improving clinical outcomes and reducing healthcare costs.

References

Abràmoff MD, Dai T, Zou J (2024) Scaling adoption of medical AI — reimbursement from value-based care and fee-for-service perspectives. *NEJM AI* 1(5):AIpc2400083.

Adida E (2021) Outcome-based pricing for new pharmaceuticals via rebates. *Management Science* 67(2):892–913.

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

Adida E, Dai T (2024) Impact of physician payment scheme on diagnostic effort and testing. *Management Science* 70(8):5408–5425.

Adida E, Mamani H, Nassiri S (2017) Bundled payment vs. fee-for-service: Impact of payment scheme on performance. *Management Science* 63(5):1606–1624.

Agrawal A, Gans J, Goldfarb A (2018) Prediction, judgment, and complexity: A theory of decision-making and artificial intelligence. *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press), 89–110.

Agrawal A, Gans JS, Goldfarb A (2024) Artificial intelligence adoption and system-wide change. *Journal of Economics & Management Strategy* 33(2):327–337.

Alexander D, Schnell M (2024) The impacts of physician payments on patient access, use, and health. *American Economic Journal: Applied Economics* 16(3):142–177.

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

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

Bergeson SC, Dean JD (2006) A systems approach to patient-centered care. *JAMA* 296(23):2848.

Bester H, Dahm M (2017) Credence goods, costly diagnosis and subjective evaluation. *The Economic Journal* 128(611):1367–1394.

Betcheva L, Erhun F, Jiang H (2021) OM Forum–Supply chain thinking in healthcare: Lessons and outlooks. *Manufacturing & Service Operations Management* 23(6):1333–1353.

Boyacı T, Canyakmaz C, de Véricourt F (2024) Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science* 70(2):1258–1275.

Chaiyachati KH, Shea JA, Asch DA, Liu M, Bellini LM, Dine CJ, Sternberg AL, Gitelman Y, Yeager AM, Asch JM, Desai S (2019) Assessment of inpatient time allocation among first-year internal medicine residents using time-motion observations. *JAMA Internal Medicine* 179(6):760–767.

Chen N, Hu M, Li W (2022) Algorithmic decision-making safeguarded by human knowledge. Working Paper, University of Toronto, Toronto.

Chick SE, Mamani H, Simchi-Levi D (2008) Supply chain coordination and influenza vaccination. *Operations Research* 56(6):1493–1506.

Cho SH, McCardle KF (2009) The adoption of multiple dependent technologies. *Operations Research* 57(1):157–169.

CMS (2021) Comment on CMS-2021-0119-0053. Public Comment, URL https://www.regulations.gov/comment/CMS-2021-0119-25800, received on September 9, 2021.

Cohen MC, Lobel R, Perakis G (2016) The impact of demand uncertainty on consumer subsidies for green technology adoption. *Management Science* 62(5):1235–1258.

Cohen MC, Toubiana EM (2024) AI medical scribes: A revolution in healthcare documentation, working Paper, McGill University.

Dada M, White WD (1999) Evaluating financial risk in the Medicare prospective payment system. *Management Science* 45(3):316–329.

Dai T, Abràmoff MD (2023) Incorporating artificial intelligence into healthcare workflows: Models and insights. *Tutorials in Operations Research: Advancing the Frontiers of OR/MS: From Methodologies to Applications*, 133–155 (INFORMS).

Dai T, Akan M, Tayur S (2017) Imaging room and beyond: The underlying economics behind physicians' test-ordering behavior in outpatient services. *Manufacturing & Service Operations Management* 19(1):99–113.

Dai T, Singh S (2020) Conspicuous by its absence: Diagnostic expert testing under uncertainty. *Marketing Science* 39(3):540–563.

Dai T, Tayur S (2020) OM Forum—Healthcare operations management: A snapshot of emerging research. *Manufacturing & Service Operations Management* 22(5):869–887.

Dai T, Tayur S (2022) Designing AI-augmented healthcare delivery systems for physician buy-in and patient acceptance. *Production and Operations Management* 31(12):4443–4451.

Das Gupta A, Karmarkar US, Roels G (2016) The design of experiential services with acclimation and memory decay: Optimal sequence and duration. *Management Science* 62(5):1278–1296, publisher: INFORMS.

de Véricourt F, Gurkan H (2023) Is your machine better than you? You may never know. *Management Science* (in press).

Gaynor M, Mehta N, Richards-Shubik S (2023) Optimal contracting with altruistic agents: Medicare payments for dialysis drugs. *American Economic Review* 113(6):1530–1571.

Goldman Sachs (2024) GenAI: Too much spend, too little benefit? URL https://www.goldmansachs.com/images/migrated/insights/pages/gs-research/gen-ai--too-much-spend,-too-little-benefit-/TOM_AI%202.0_ForRedaction.pdf, published June 25, 2024.

Grand-Clément J, Pauphilet J (2023) The best decisions are not the best advice: Making adherence-aware recommendations. URL https://arxiv.org/abs/2209.01874.

Guo P, Tang CS, Wang Y, Zhao M (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.

Gurkan H, de Véricourt F (2022) Contracting, pricing, and data collection under the AI flywheel effect. *Management Science* 68(12):8791–8808.

Hernandez I (2023) Drug shortages: Examining supply challenges, impacts, and policy solutions from a federal health program perspective (testimony before the U.S. Senate Finance Committee). https://www.finance.senate.gov/imo/media/doc/1205_hernandez_testimony.pdf.

Jelovac I (2001) Physicians' payment contracts, treatment decisions and diagnosis accuracy. *Health Economics* 10(1):9–25.

Karlinsky-Shichor Y, Netzer O (2024) Automating the B2B salesperson pricing decisions: A human-machine hybrid approach. *Marketing Science* 43(1):138–157.

Keskinocak P, Savva N (2020) OM Forum—A review of the healthcare-management (modeling) literature published in Manufacturing & Service Operations Management. *Manufacturing & Service Operations Management* 22(1):59–72.

Kim Y, Knight B, Mitrofanov D, Xu Y (2024) AI and worker learning: Evidence from a large-scale field experiment. Working Paper, Loyola University of Chicago, Chicago.

Kundu A, Ramdas K (2022) Timely after-sales service and technology adoption: Evidence from the off-grid solar market in Uganda. *Manufacturing & Service Operations Management* 24(3):1329–1348.

Lahiri A, Dey D (2013) Effects of piracy on quality of information goods. *Management Science* 59(1):245–264.

Lai J, Xu LL, Fang X, Dai T (2024) Regulating adaptive medical artificial intelligence: Can less oversight lead to greater compliance? Working paper, Johns Hopkins University.

Lamb J, Israelstam G, Agarwal R, Bhasker S (2024) Generative AI in healthcare: Adoption trends and what's next .

Li Y, Dai T, Qi X (2022) A theory of interior peaks: Activity sequencing and selection for service design. *Manufacturing & Service Operations Management* 24(2):993–1001.

Lohr S (2023) A.I. outshines in health care. At paperwork. *The New York Times* (June 27), p. A1.

McGuire TG (2000) Physician agency. *Handbook of Health Economics*, chapter 9, 461–536 (Elsevier).

Moorthy KS (1988) Product and price competition in a duopoly. *Marketing Science* 7(2):141–168.

Mullainathan S, Obermeyer Z (2021) Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics* 137(2):679–727.

Nassiri S, Adida E, Mamani H (2022) Reference pricing for healthcare services. *Manufacturing & Service Operations Management* 24(2):921–937.

Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R (2024) Large language models in medicine: The potentials and pitfalls: A narrative review. *Annals of Internal Medicine* 177(2):210–220.

Orfanoudaki A, Saghafian S, Song K, Chakkera HA, Cook C (2022) Algorithm, human, or the centaur: How to enhance clinical care? Working Paper No. RWP22-027, Harvard Kennedy School, Cambridge, MA.

Paige NM, Apaydin EA, Goldhaber-Fiebert JD, Mak S, Miake-Lye IM, Begashaw MM, Severin JM, Shekelle PG (2020) What is the optimal primary care panel size?: A systematic review. *Annals of Internal Medicine* 172(3):195.

Parikh RB, Helmchen LA (2022) Paying for artificial intelligence in medicine. *npj Digital Medicine* 5(63).

Prendergast C (2007) The motivation and bias of bureaucrats. *American Economic Review* 97(1):180–196.

Rand B (2023) How will the tech titans behind ChatGPT, Bard, and LLaMA make money? URL https://hbswk.hbs.edu/item/how-will-the-tech-titans-behind-chat-gpt-bard-and-llama-make-money.

Reinhardt UE (2016) Sense and nonsense in defining "value"' in health care. https://nihcm.org/assets/articles/sense-nonsense-defining-value-in-health-care-uwe-reinhardt.pdf, presented at the National Institute of Health Care Management, Capitol Hill Briefing on The Future of Health Care in America.

Rotenstein L, Melnick ER, Iannaccone C, Zhang J, Mugal A, Lipsitz SR, Healey MJ, Holland C, Snyder R, Sinsky CA, Ting D, Bates DW (2024) Virtual scribes and physician time spent on electronic health records. *JAMA Network Open* 7(5):e2413140.

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

Tai-Seale M, Olson TB, Lee SG, Chan J, Morikawa S, Durbin D, Li AS, Luft EA, Browne G, Friedman RL (2019) Electronic health record time and patient safety culture among primary care physicians. *JAMA Network Open* 2(4):e194431.

Taylor TA, Xiao W (2014) Subsidizing the distribution channel: Donor funding to improve the availability of malaria drugs. *Management Science* 60(10):2461–2477.

Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, Liu V, Lee K (2024) Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery* 5(3).

Topol E (2019) *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (New York: Basic Books).

Uppari BS, Popescu I, Netessine S (2019) Selling off-grid light to liquidity-constrained consumers. *Manufacturing & Service Operations Management* 21(2):308–326.

Vlachy J, Ayer T, Ayvaci M, Raghunathan S (2023) The business of healthcare: The role of physician integration in bundled payments. *Manufacturing & Service Operations Management* 25(3):996–1012.

Wachter RM, Brynjolfsson E (2024) Will generative artificial intelligence deliver on its promise in health care? *JAMA* 331(1):65.

Wu K, Wu E, Theodorou B, Liang W, Mack C, Glass L, Sun J, Zou J (2023) Characterizing the clinical adoption of medical AI devices through U.S. insurance claims. *NEJM AI* 1(1).

Xu L, Li H, Zhao H (2022) Outcome-based reimbursement: The solution to high drug spending? *Manufacturing & Service Operations Management* 24(4):2029–2047.

Yaraghi N (2024) Generative AI in health care: Opportunities, challenges, and policy. Brookings Institution, URL https://www.brookings.edu/articles/generative-ai-in-health-care-opportunities-challenges-and-policy/.

Zhang C, Atasu A, Ayer T, Toktay LB (2020) Truthful mechanisms for medical surplus product allocation. *Manufacturing & Service Operations Management* 22(4):735–753.

Zhang DJ, Gurvich I, Van Mieghem JA, Park E, Young RS, Williams MV (2016) Hospital readmissions reduction program: An economic and operational analysis. *Management Science* 62(11):3351–3371.

Zhang W, Lee HH (2022) Investment strategies for sourcing a new technology in the presence of a mature technology. *Management Science* 68(6):4631–4644.

Zink A, Chernew ME, Neprash HT (2024) How should medicare pay for artificial intelligence? *JAMA Internal Medicine* .

Zorc S, Chick SE, Hasija S (2023) Choosing outcomes-based reimbursement policies: Should we worry about collusion? Working paper, University of Virginia, Darden School of Business, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2973048.

Online Appendix to "Physician Compensation in the Age of Generative Artificial Intelligence"

Proof of Lemma 1. At the first-best, the social planner decides which patients $[x_1, x_2]$ to use AI on (where $0 \le x_1 \le x_2 \le 1$), to maximize

$$bq\frac{x_2^2 - x_1^2}{2},$$

hence it is clear that it is socially optimal to use AI on all patients. The socially optimal quality is obtained by solving

$$\max_{q>0} \quad -cq^2 + \frac{bq}{2},$$

which leads to $q = b/(4c)$. The social planner's objective then equals

$$-\frac{b^2}{16c} + \frac{b^2}{8c} = \frac{b^2}{16c} > 0,$$

so the social planner opts to participate. *Q.E.D.*

Proof of Lemma 2. In stage 2, the provider decides which patients $[x_1, x_2]$ to use AI on (where $0 \le x_1 \le x_2 \le 1$), to maximize

$$\begin{aligned} & f(x_1, x_2) = -p(x_2 - x_1) + \delta bq\frac{x_2^2 - x_1^2}{2}, \\ \text{where} \quad & \frac{\partial f}{\partial x_1}(x_1, x_2) = p - \delta bq x_1 \\ & \frac{\partial f}{\partial x_2}(x_1, x_2) = -p + \delta bq x_2. \end{aligned}$$

If $p > \delta bq$, then $\frac{\partial f}{\partial x_1} > 0$ and $\frac{\partial f}{\partial x_2} < 0$ for all $x_1, x_2 \in [0,1]$, so $x_1^* = x_2^*$. No patient receives AI.

Otherwise, if $p \le \delta bq$, we have $\frac{\partial f}{\partial x_1} > 0$ if and only if $x_1 < \frac{p}{\delta bq}$ and $\frac{\partial f}{\partial x_2} > 0$ if and only if $x_2 > \frac{p}{\delta bq}$. Hence, f is unimodal in x_1 and reaches its maximum at $\frac{p}{\delta bq}$. Moreover, for $x_2 \ge \frac{p}{\delta bq}$, f is increasing in x_2, so the optimal solution is $(x_1, x_2) = (\frac{p}{\delta bq}, 1)$. *Q.E.D.*

Proof of Proposition 1. If $p > \delta bq$, AI is not used in stage 2, so the developer does not participate as there is no possibility of earning any revenue. In stage 1, the developer solves the following problem (and participates if and only if the optimal objective value is positive):

$$\begin{aligned} \max_{p,q>0} \quad & \varphi(p,q) = -cq^2 + p\left(1 - \frac{p}{\delta bq}\right) = -cq^2 + p - \frac{p^2}{\delta bq} \\ \text{s.t.} \quad & p \le \delta bq. \end{aligned}$$

We have

$$\begin{aligned} \frac{\partial \varphi}{\partial p}(p,q) &= 1 - \frac{2p}{\delta bq} \\ \frac{\partial \varphi}{\partial q}(p,q) &= -2cq + \frac{p^2}{\delta bq^2} \\ \frac{\partial^2 \varphi}{\partial p^2}(p,q) &= -\frac{2}{\delta bq} < 0 \end{aligned}$$

$$\frac{\partial^2 \varphi}{\partial q^2}(p,q) = -2c - 2\frac{p^2}{\delta b q^3} < 0$$
$$\frac{\partial^2 \varphi}{\partial p \partial q}(p,q) = \frac{2p}{\delta b q^2}.$$

The first-order conditions can be written as:

$$2p = \delta b q$$
$$2c\delta b q^3 = p^2,$$

that is,

$$q = \frac{\delta b}{8c}, \quad p = \frac{\delta^2 b^2}{16c}.$$

Note in particular that, at this stationary point solution, the constraint $p \le \delta b q$ is valid.

The second-order condition requires that, at the stationary point,

$$\frac{2}{\delta b q}\left(2c + 2\frac{p^2}{\delta b q^3}\right) - \frac{4p^2}{\delta^2 b^2 q^4} > 0,$$
$$\Leftrightarrow \frac{32c^2}{\delta^2 b^2} > 0 \text{ after simplifications.}$$

Hence, the unique stationary point is the optimal solution. Moreover, at this solution, the objective value equals $\delta^2 b^2/(64c) > 0$, so the developer opts to participate. *Q.E.D.*

Proof of Lemma 3. The proof is similar to the proof of Lemma 2. If $p - f > \delta b q$ (i.e., if $f < p - \delta b q$), AI is not used on any patient. Otherwise, if $f \ge p - \delta b q$, the optimal solution is $(x_1, x_2) = (\frac{p-f}{\delta b q}, 1)$. *Q.E.D.*

Proof of Proposition 2. We start with proving the following lemma:

Lemma A1. *The inequality*

$$\varphi\left(\frac{\delta b + \sqrt{(\delta b)^2 - 24cf}}{12c}, f\right) \le 0$$

is equivalent to the condition

$$\frac{\delta^2 b^2}{27c} \le f \le \frac{\delta^2 b^2}{24c}.$$

Proof of Lemma A1. To establish this equivalence, observe that $f \le \delta^2 b^2/24c$ is required for the square root to exist, and

$$\varphi\left(\frac{\delta b + \sqrt{(\delta b)^2 - 24cf}}{12c}, f\right) = f^2 + \frac{b^4\delta^4 + b\delta\left(b^2\delta^2 - 24cf\right)^{3/2} - 36b^2c\delta^2 f}{216c^2}. \tag{A1}$$

This expression equals zero when $f = \delta^2 b^2/(27c)$. Next, we examine the total derivative of (A1) with respect to f:

$$\frac{d}{df}\varphi\left(\frac{\delta b + \sqrt{(\delta b)^2 - 24cf}}{12c}, f\right) = 2\left[f - \frac{b\delta\left(\delta b + \sqrt{(\delta b)^2 - 24cf}\right)}{12c}\right]. \tag{A2}$$

This derivative is always negative because

$$\frac{b\delta\left(\delta b + \sqrt{(\delta b)^2 - 24cf}\right)}{12c} > \frac{(\delta b)^2}{12c} \ge 2f > f.$$

Because the derivative is negative and φ equals zero at $f = \delta^2 b^2/(27c)$, it follows that

$$\varphi\left(\frac{\delta b + \sqrt{(\delta b)^2 - 24cf}}{12c}, f\right) \leq 0 \tag{A3}$$

if and only if

$$\frac{\delta^2 b^2}{27c} \leq f \leq \frac{\delta^2 b^2}{24c}. \tag{A4}$$

Q.E.D.

To prove Proposition 2, we need to solve two optimization problems, and select the one leading to the higher objective value (provided it is positive, to ensure participation). The first optimization problem is

$$\begin{aligned} \max_{p,q>0} \quad & -cq^2 + p \\ \text{s.t.} \quad & p \leq f. \end{aligned}$$

The optimal solution is $q = \epsilon$ and $p = f$, with an objective value of $f - c\epsilon^2 \approx f$.

The second optimization problem is

$$\begin{aligned} \max_{p,q>0} \quad & \psi(p,q) = -cq^2 + p\left(1 - \frac{p-f}{\delta bq}\right) \\ \text{s.t.} \quad & p > f \\ & p - \delta bq \leq f. \end{aligned}$$

We have

$$\begin{aligned} \frac{\partial \psi}{\partial p}(p,q) &= 1 - \frac{2p-f}{\delta bq} \\ \frac{\partial \psi}{\partial q}(p,q) &= -2cq + \frac{p(p-f)}{\delta bq^2}. \end{aligned}$$

The first-order conditions can be written as:

$$\begin{aligned} 2p - f &= \delta bq \\ 2c\delta bq^3 &= p(p-f). \end{aligned}$$

The former implies $p - \delta bq = f - p \leq f$, consistent with the second constraint of the optimization problem. The latter implies in particular $p - f > 0$, consistent with the first constraint of the optimization problem. Hence, at a stationary point, both constraints are satisfied.

We have

$$\begin{aligned} \frac{\partial^2 \psi}{\partial p^2}(p,q) &= -\frac{2}{\delta bq} < 0 \\ \frac{\partial^2 \psi}{\partial q^2}(p,q) &= -2c - 2\frac{p(p-f)}{\delta bq^3} < 0 \\ \frac{\partial^2 \psi}{\partial p \partial q}(p,q) &= \frac{2p-f}{\delta bq^2}, \end{aligned}$$

where the first two inequalities are trivially satisfied at a stationary point. The second-order conditions also require that, at a stationary point,

$$\frac{2}{\delta bq}\left(2c + 2\frac{p(p-f)}{\delta bq^3}\right) - \left(\frac{2p-f}{\delta bq^2}\right)^2 > 0.$$

Using the first-order condition, the above inequality simplifies to $12c/(\delta bq) - 1/q^2 > 0$, or equivalently, $q > \delta b/(12c)$.

Plugging the first first-order condition into the second, we obtain

$$\frac{\delta^2 b^2 q^2 - f^2}{4} = 2\delta bcq^3,$$

or equivalently

$$\varphi(q, f) \equiv 8\delta bcq^3 - \delta^2 b^2 q^2 + f^2 = 0.$$

We need to solve this equation for q to find the quality at a stationary point for a given f. We have

$$\frac{\partial \varphi(q, f)}{\partial q} = 24\delta bcq^2 - 2\delta^2 b^2 q = 2q\delta b(12cq - \delta b).$$

Hence, $\varphi(q, f)$ is unimodal in q, reaching a minimum when $q = \delta b/(12c)$. At that minimum, $\varphi(q, f)$ takes the value (after simplifications)

$$\varphi\left(\frac{\delta b}{12c}, f\right) = f^2 - \frac{\delta^4 b^4}{3 \times 12^2 c^2}.$$

Therefore, the equation $\varphi(q, f) = 0$ in q has two solutions $\bar{q}_1, \bar{q}_2$ if and only if $f < \delta^2 b^2/(12c\sqrt{3})$. (If $f = \delta^2 b^2/(12c\sqrt{3})$, there is a single solution $q = \delta b/(12c)$ and $\varphi(q, f) \geq 0$ for all q, and if f is above the threshold, there is no solution.) In addition, because $\varphi(0, f) > 0$, both solutions $\bar{q}_1, \bar{q}_2$ are non-negative; they are equidistant to $\delta b/(12c)$. It follows that a unique stationary point $\bar{q}_2$ exists satisfying the second-order conditions if and only if $f < \delta^2 b^2/(12c\sqrt{3})$. Moreover, we have $\delta b/(12c) < \bar{q}_2 < \delta b/(6c)$.

It remains to ensure that at this solution (together with the associated price, $\bar{p}_2 \equiv (\delta b\bar{q}_2 + f)/2$), the objective value is larger than that at the first optimization problem. Namely, we need to ensure that, at this stationary point,

$$\begin{aligned} &-cq^2 + p\left(1 - \frac{p - f}{\delta bq}\right) = -3cq^2 + \frac{\delta bq}{2} + \frac{f}{2} > f, \\ \Leftrightarrow &-3cq^2 + \frac{\delta bq}{2} - \frac{f}{2} > 0. \end{aligned}$$

This degree-2 polynomial is less than or equal to zero if $(\delta b)^2 \leq 24cf$. Otherwise, it has two positive roots that are equidistant to $\delta b/(12c)$. Hence, the objective value is larger than that at the first optimization problem if and only if the stationary point is located in between these two roots, that is, if and only if

$$\bar{q}_2 \leq \frac{\delta b + \sqrt{(\delta b)^2 - 24cf}}{12c},$$

or equivalently, if and only if

$$\varphi\left(\frac{\delta b + \sqrt{(\delta b)^2 - 24cf}}{12c}, f\right) > 0.$$

Using Lemma A2, this inequality is equivalent to

$$\frac{\delta^2 b^2}{27c} \leq f \leq \frac{\delta^2 b^2}{24c}. \tag{A5}$$

Conclusion:

- if $f > \delta^2 b^2/(12c\sqrt{3})$, no stationary point exists in the second optimization problem. At the boundaries/limit, the objective is the same or worse than the first optimization problem;

- else, if $f \geq \delta^2 b^2/(24c)$, the second optimization problem has a worse objective than the first;
- else, if $f \geq \delta^2 b^2/27c$, the second optimization problem has a worse objective than the first.
- else, that is, if $f < \delta^2 b^2/(27c)$, one stationary point exists that is the unique maximizer, given by $\bar{q}_2$ as the larger root in q of $\varphi(q, f)$, and associated price $\bar{p}_2 \equiv (\delta b \bar{q}_2 + f)/2$.

Q.E.D.

Proof of Corollary 1. We start with proving the following lemma:

Lemma A2. *The inequality*

$$\varphi\left(\frac{\delta b + \sqrt{(\delta b)^2 + 48cf}}{24c}, f\right) \leq 0$$

is equivalent to the condition

$$\frac{\delta^2 b^2}{27c} \geq f.$$

Proof of Lemma A2. We start by noting

$$\varphi\left(\frac{\delta b + \sqrt{\delta^2 b^2 + 48cf}}{24c}, f\right) = \frac{b\delta f\sqrt{b^2\delta^2 + 48cf}}{36c} - \frac{b^3\delta^3\left(\sqrt{b^2\delta^2 + 48cf} + b\delta\right)}{864c^2} + f^2,$$

which is equal to zero when $f = \frac{b^2\delta^2}{27c}$. The total derivative of the above quantity with respect to f is given by

$$2f\left(\frac{b\delta}{\sqrt{b^2\delta^2 + 48cf}} + 1\right) > 0.$$

Therefore, it follows that

$$\varphi\left(\frac{\delta b + \sqrt{\delta^2 b^2 + 48cf}}{24c}, f\right) < 0$$

if and only if $f < \frac{b^2\delta^2}{27c}$. *Q.E.D.*

To prove Corollary 1(i), note $\varphi(q^*, f) = 0$ implies

$$\frac{\partial \varphi}{\partial q}\frac{dq^*}{df} + \frac{\partial \varphi}{\partial f} = 0. \tag{A6}$$

Because $\partial\varphi/\partial f = 2f > 0$, it follows that

$$\frac{\partial \varphi}{\partial q}\frac{dq^*}{df} < 0.$$

Moreover, since $\partial\varphi/\partial q > 0$ when $q \in [\delta b/(12c), \infty)$, it follows that $dq^*/df < 0$.

To prove Corollary 1(ii), note

$$\frac{dp^*}{df} = \frac{1}{2}\delta b\frac{dq^*}{df} + \frac{1}{2}.$$

Moreover, from (A6),

$$\frac{dq^*}{df} = -\frac{\partial\varphi/\partial f}{\partial\varphi/\partial q} = -\frac{2f}{2\delta b q^*(12cq^* - \delta b)}.$$

As a result,

$$\frac{dp^*}{df} = \frac{-2f\delta b + 2\delta b q^*(12cq^* - \delta b)}{4\delta b q^*(12cq^* - \delta b)},$$

which has the sign of $12c(q^*)^2 - \delta b q^* - f$. This degree-2 polynomial in q^* has a positive discriminant, a positive root, and a negative root; moreover it gros to infinity as q^* grows large. Hence, on the positive domain, the polynomial is positive if and only if q^* is above the positive root, namely, $(\delta b + \sqrt{(\delta b)^2 + 48cf})/(24c)$. Because φ is increasing in q for $q > \delta b/(12c)$, it follows that $dp^*/df \geq 0$ if and only if

$$\varphi\left(\frac{\delta b + \sqrt{\delta^2 b^2 + 48cf}}{24c}, f\right) \leq 0.$$

By Lemma A2, we have $dp^*/df \geq 0$ if and only if $\delta^2 b^2/27c \geq f$. By part (ii) of Proposition 2, we have $f < \delta^2 b^2/27c$. As a result, $dp^*/df \geq 0$. *Q.E.D.*

Proof of Corollary 2. As noted in the proof of Proposition 2, the quality set with a fee-for-service reimbursement is either $q = \epsilon$ (near zero) with AI used on all patients, or $q = \bar{q}_2$, where $\delta b/(12c) < \bar{q}_2 < \delta b/(6c)$ with AI used on a subset of patients. Because at the first-best, AI is used on all patients with a non-near-zero quality, matching the first-best is impossible.

Furthermore, $\delta b/(6c) < b/(4c)$ when $\delta < 1.5$. Hence, the quality is lower than that at the first-best, given by $b/(4c)$. *Q.E.D.*

Proof of Lemma 4. Coordination would require that the second case of Proposition 2 holds and

$$\frac{1}{2} - \frac{f}{2\delta b \bar{q}_2} = 0,$$

that is, $f = \delta b \bar{q}_2$. Moreover, $\varphi(\bar{q}_2, f) = 0$ implies

$$8\delta bc\bar{q}_2^3 - \delta^2 b^2 \bar{q}_2^2 + \delta^2 b^2 \bar{q}_2^2 = 0,$$

that is, $\delta bc\bar{q}_2^3 = 0$ which contradicts $\delta > 0$ and $\bar{q}_2 > \delta b/(12c)$. *Q.E.D.*

Proof of Proposition 3. We need to solve two optimization problems, and select the one leading to the higher objective value (provided it is positive, to ensure participation). The first optimization problem is

$$\begin{aligned} \max_{p,q>0} \quad & -cq^2 + p \\ \text{s.t.} \quad & p - \gamma q \leq 0. \end{aligned}$$

The optimal solution is $q = \gamma/(2c)$ and $p = \gamma^2/(2c)$, with an objective value of $\gamma^2/(4c)$. With these decisions, AI is used on all patients, so this coordinates to the first-best if and only if the quality decisions match, that is, $\gamma = b/2$.

The second optimization problem is

$$\begin{aligned} \max_{p,q>0} \quad & \psi(p,q) = -cq^2 + p\left(1 - \frac{p - \gamma q}{\delta b q}\right) \\ \text{s.t.} \quad & p - \gamma q > 0 \\ & p - (\gamma + \delta b)q \leq 0. \end{aligned}$$

If this problem dominates the first, AI is used for only a fraction of patients, making coordination impossible, regardless of γ. Hence, to focus on whether coordination is possible, we set $\gamma = b/2$ and we seek to determine whether the optimal objective value of the second optimization problem may dominate that of the first.

We have

$$\frac{\partial \psi}{\partial p}(p,q) = 1 - \frac{2p - \gamma q}{\delta b q} = 1 - \frac{2p}{\delta b q} + \frac{\gamma}{\delta b}$$
$$\frac{\partial \psi}{\partial q}(p,q) = -2cq + \frac{p^2}{\delta b q^2}.$$

The first-order conditions can be written as:

$$p = \frac{\delta b + \gamma}{2} q$$
$$2c\delta b q^3 = p^2.$$

Plugging the first condition into the second, we obtain $p = q = 0$ (leading to an objective value of zero, worse than the first problem) or

$$q = \frac{(\delta b + \gamma)^2}{8c\delta b}, \quad p = \frac{(\delta b + \gamma)^3}{16c\delta b}.$$

Regarding the constraints, we have (using $\gamma = b/2$)

$$p - \gamma q = (\delta b - \gamma) q/2 = b(\delta - 1/2) q/2 > 0 \text{ if and only if } \delta > 1/2$$
$$p - (\gamma + \delta b) q = -p \leq 0.$$

Hence, if $\delta > 1/2$, the stationary point satisfies the constraints; otherwise, no stationary point (other than $(0,0)$) exists in the feasible domain, so the first constraint is tight at the optimum ($p = \gamma q$), and the problem reduces to the first optimization problem. We thus focus on the case $\delta > 1/2$ in the remainder of the proof.

We have

$$\frac{\partial^2 \psi}{\partial p^2}(p,q) = -\frac{2}{\delta b q} < 0$$
$$\frac{\partial^2 \psi}{\partial q^2}(p,q) = -2c - 2\frac{p^2}{\delta b q^3} < 0$$
$$\frac{\partial^2 \psi}{\partial p \partial q}(p,q) = \frac{2p}{\delta b q^2}.$$

The second-order conditions require that, at a stationary point,

$$\frac{2}{\delta b q}\left(2c + 2\frac{p^2}{\delta b q^3}\right) - \left(\frac{2p}{\delta b q^2}\right)^2 > 0.$$

Using the first-order condition, the above inequality simplifies to $4c/(\delta b q) > 0$, which is trivially satisfied. Hence, the stationary point is the optimal solution.

It remains to compare the objective of the second optimization problem at the optimal solution with that of the first one. After simplifications, the objective of the second problem at the unique stationary point is

$$(\gamma + \delta b)^4 / (64 c \delta^2 b^2),$$

which is larger than the objective of the first problem ($\gamma^2/(4c)$) because

$$\frac{(\gamma+\delta b)^4}{64c\delta^2 b^2} \geq \frac{\gamma^2}{4c} \Leftrightarrow (\gamma+\delta b)^2 \geq 4\delta b\gamma$$
$$\Leftrightarrow (\gamma-\delta b)^2 \geq 0.$$

Hence, coordination is possible if $\delta < 1/2$; otherwise, the second optimization problem dominates. *Q.E.D.*

Proof of Proposition 4. We need to solve two optimization problems, and select the one leading to the higher objective value (provided it is positive, to ensure participation). The first optimization problem is

$$\begin{aligned} \max_{p,q>0} \quad & -cq^2+p \\ \text{s.t.} \quad & p-\gamma q \leq f. \end{aligned}$$

The optimal solution is $q=\gamma/(2c)$ and $p=f+\gamma^2/(2c)$, with an objective value of $f+\gamma^2/(4c)$. With these decisions, AI is used on all patients so this coordinates to the first-best if and only if the quality decisions match, that is, $\gamma=b/2$.

The second optimization problem is

$$\begin{aligned} \max_{p,q>0} \quad & \psi(p,q) = -cq^2+p\left(1-\frac{p-f-\gamma q}{\delta bq}\right) \\ \text{s.t.} \quad & p-\gamma q > f \\ & p-(\gamma+\delta b)q \leq f. \end{aligned}$$

If the second optimization problem yields a higher objective value than the first, AI is used only for a subset of patients, making coordination unattainable irrespective of the value of γ. To obtain the conditions under which coordination is feasible, we fix $\gamma=b/2$ and evaluate whether the optimal objective value of the second optimization problem can exceed that of the first.

We have

$$\frac{\partial\psi}{\partial p}(p,q) = 1-\frac{2p-f-\gamma q}{\delta bq} = 1-\frac{2p-f}{\delta bq}+\frac{\gamma}{\delta b}$$
$$\frac{\partial\psi}{\partial q}(p,q) = -2cq+\frac{p(p-f)}{\delta bq^2}.$$

The first-order conditions can be written as:

$$2p-f=(\delta b+\gamma)q$$
$$2c\delta bq^3 = p(p-f).$$

The former implies $p-(\delta b+\gamma)q = f-p \leq f$, consistent with the second constraint of the optimization problem. The first constraint of the optimization problem is $p-\gamma q > f$, i.e., using the first of the FOC, $q > f/(\delta b-\gamma)$. This requires $\delta b > \gamma$, which is valid when $\gamma=b/2$ and $\delta > 1/2$. We still need to check whether this constraint is satisfied at a stationary point. (If not, the constraint is tight, and the second problem reduces to the first optimization problem.)

We have

$$\frac{\partial^2\psi}{\partial p^2}(p,q) = -\frac{2}{\delta bq} < 0$$

$$\frac{\partial^2 \psi}{\partial q^2}(p,q) = -2c - 2\frac{p(p-f)}{\delta b q^3} < 0$$
$$\frac{\partial^2 \psi}{\partial p \partial q}(p,q) = \frac{2p-f}{\delta b q^2},$$

where the first two inequalities are trivially satisfied at a stationary point. The second-order conditions also require that, at a stationary point,

$$\frac{2}{\delta b q}\left(2c + 2\frac{p(p-f)}{\delta b q^3}\right) - \left(\frac{2p-f}{\delta b q^2}\right)^2 > 0.$$

Using the first-order condition, the above inequality simplifies to $12c\delta bq > (\delta b+\gamma)^2$ or, equivalently, $q > (\delta b+\gamma)^2/(12c\delta b)$.

Plugging the first first-order condition into the second, we obtain

$$\frac{(\delta b+\gamma)^2 q^2 - f^2}{4} = 2\delta bcq^3$$

or, equivalently,

$$\varphi(q) \equiv 8\delta bcq^3 - (\delta b+\gamma)^2 q^2 + f^2 = 0.$$

We need to solve this equation for q to find the quality at a stationary point. We have

$$\varphi'(q) = 24\delta bcq^2 - 2(\delta b+\gamma)^2 q = 2q(12\delta bcq - (\delta b+\gamma)^2).$$

Hence, $\varphi(q)$ is unimodal, reaching a minimum when $q = (\delta b+\gamma)^2/(12\delta bc)$. At that minimum, $\varphi(q)$ takes the value (after simplifications)

$$\varphi\left(\frac{(\delta b+\gamma)^2}{12\delta bc}\right) = f^2 - \frac{(\delta b+\gamma)^6}{3\times(12\delta bc)^2}.$$

Therefore, the equation $\varphi(q)=0$ has two solutions $\bar{q}_1, \bar{q}_2$ if and only if $f < (\delta b+\gamma)^3/(12\delta bc\sqrt{3}) = \bar{f}$ (using the fact that $\gamma = b/2$). (If $f = (\delta b+\gamma)^3/(12\delta bc\sqrt{3})$, there is a single solution $q = (\delta b+\gamma)^2/(12\delta bc)$ and $\varphi(q) \geq 0$ for all q, and if f is above the threshold, there is no solution.) In addition, because $\varphi(0) > 0$, both solutions $\bar{q}_1, \bar{q}_2$ are non-negative; they are equidistant to $(\delta b+\gamma)^2/(12\delta bc)$. It follows that a unique stationary point $\bar{q}_2$ exists satisfying the second-order conditions if and only if $f < (\delta b+\gamma)^3/(12\delta bc\sqrt{3})$. Moreover, we have $(\delta b+\gamma)^2/(12\delta bc) < \bar{q}_2 < (\delta b+\gamma)^2/(6\delta bc)$.

Because the first-order condition gives $p = \frac{((\delta b+\gamma)q+f)}{2}$, and because $\gamma = b/2$ with $\delta > 1/2$, the stationary point satisfies the first constraint $p - \gamma q > f$ if and only if $q > \frac{f}{\delta b-\gamma}$. This condition holds if and only if one of the following is true:

$$\varphi\left(\frac{f}{\delta b-\gamma}\right) < 0 \quad \text{or} \quad \bar{q}_1 > \frac{f}{\delta b-\gamma}.$$

Simplifying further, we find that $\varphi\left(\frac{f}{\delta b-\gamma}\right)$ has the same sign as the expression $2cf - \gamma(\delta b-\gamma)$. Thus, the constraint is satisfied if either $2cf - \gamma(\delta b-\gamma) < 0$ or both $2cf - \gamma(\delta b-\gamma) > 0$ and $\frac{f}{\delta b-\gamma} < \frac{(\delta b+\gamma)^2}{12\delta bc}$. Specifically, this implies either:

$$f < \frac{\gamma(\delta b-\gamma)}{2c} \quad \text{or} \quad \frac{\gamma(\delta b-\gamma)}{2c} < f < \frac{(\delta b+\gamma)^2(\delta b-\gamma)}{12\delta bc}.$$

Therefore, the stationary point satisfies the first constraint if and only if:

$$f < \frac{(\delta b+\gamma)^2(\delta b-\gamma)}{12\delta bc} = \frac{b^2(\delta+\frac{1}{2})^2(\delta-\frac{1}{2})}{12\delta c}.$$

If f exceeds this threshold, the stationary point becomes infeasible, leaving the second optimization problem with no interior solution. Observe

$$\frac{b^2(\delta+\frac{1}{2})^2(\delta-\frac{1}{2})}{12\delta c} < \bar{f} \quad \Leftrightarrow \quad \sqrt{3}(\delta-\frac{1}{2}) < \delta+\frac{1}{2}$$
$$\Leftrightarrow \quad \delta < 1+\frac{\sqrt{3}}{2} \approx 1.87.$$

Provided that the stationary point is feasible, it remains to compare the objective value at this solution (together with the associated price, $\bar{p}_2 \equiv ((\delta b+\gamma)\bar{q}_2+f)/2$) verus that at the first optimization problem. Namely, the first optimization problem dominates when, at this stationary point,

$$-cq^2+p\left(1-\frac{p-f-\gamma q}{\delta b q}\right) = -3cq^2+\left(\frac{(\delta b+\gamma)q}{2}+\frac{f}{2}\right)\left(1+\frac{\gamma}{\delta b}\right) < f+\frac{\gamma^2}{4c},$$
$$\Leftrightarrow -3cq^2+\frac{(\delta b+\gamma)^2}{2\delta b}q-\frac{f}{2}+\frac{f\gamma}{2\delta b}-\frac{\gamma^2}{4c}<0.$$

If the optimal solution to the second optimization problem lies on one of the boundaries, two possibilities exist: either the first or the second constraint is tight. If the first constraint in tight, the problem reduces to the first optimization problem. If the second constraint is tight, the optimal solution is $q=0, p=f$ with an objective value of zero, which is worse than the first optimization problem.

Hence,

- if $\delta > 1+\sqrt{3}/2$ and $f > b^2(\delta+1/2)^3/(12\delta c\sqrt{3})$, or if $\delta < 1+\sqrt{3}/2$ and $f > b^2(\delta+1/2)^2(\delta-1/2)/(12\delta c)$, no feasible stationary point exists in the second optimization problem. The solution is that of the first optimization problem. Hence, setting $\gamma = b/2$ coordinates to the first-best;
- otherwise, a feasible stationary point (solution of the second optimization problem) exists. We need to find it and test whether the resulting objective is worse than the objective of the first problem. We first need to find $\bar{q}_2$ the larger root of the equation

  $$\varphi(q) \equiv 8\delta bcq^3-(\delta b+\gamma)^2q^2+f^2=0.$$

  Set $\bar{p}_2 \equiv ((\delta b+\gamma)\bar{q}_2+f)/2$. Then, we test whether

  $$-3cq^2+\frac{(\delta b+\gamma)^2}{2\delta b}q-\frac{f}{2}+\frac{f\gamma}{2\delta b}-\frac{\gamma^2}{4c}<0. \tag{A7}$$

  If so, setting $\gamma = b/2$ coordinates to the first-best. If not, the second optimization problem dominates the first, and no coordination is possible, because only a fraction of patients receives AI at the optimal solution.

Taking the derivative with respect to f of the equation $\varphi(q)=0$, we obtain

$$\frac{\partial \bar{q}_2}{\partial f} = \frac{f}{\bar{q}_2}\frac{1}{(\delta b+\gamma)^2-12\delta bc\bar{q}_2}.$$

Because we have $\bar{q}_2 > \frac{(\delta b+\gamma)^2}{12\delta bc}$, it follows that $\frac{\partial \bar{q}_2}{\partial f}<0$. Furthermore, by taking the partial derivative of the left-hand side of (A7) with respect to f and using the expression for $\frac{\partial \bar{q}_2}{\partial f}$, we obtain:

$$\frac{f-(\delta b-\gamma)\bar{q}_2}{2\delta b\bar{q}_2}<0,$$

where the inequality holds due to the first feasibility constraint. As a result, (A7) is equivalent to requiring f to exceed a certain threshold. Given that coordination is possible for f above $\tilde{f}$ and impossible for $f=0$, it follows that the threshold f^* lies within the interval $(0,\tilde{f}]$. *Q.E.D.*