MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 1523-4614 | EISSN 1526-5498 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Primary Care First Initiative: Impact on care delivery and outcomes

Elodie Adida
School of Business, University of California, Riverside, elodie.goodman@ucr.edu,

Fernanda Bravo
Anderson School of Management, University of California Los Angeles, fernanda.bravo@anderson.ucla.edu,

Problem definition: The Centers for Medicare & Medicaid Services launched the Primary Care First (PCF) initiative in January 2021. The initiative builds upon prior innovative payment models and aims at incentivizing a redesign of primary care delivery, including new modes of delivery such as remote care. To achieve this goal, the initiative blends capitation and fee-for-service (FFS) payments and includes performance-based adjustments linked to service quality and health outcomes. We analyze a model motivated by this new payment system, and its impact on the different stakeholders, and derive insights on how to design it to reach the best possible outcome.

Methodology/Results: We propose an analytical model that captures patient heterogeneity in terms of health complexity, provider choice of care delivery mode (referral to a specialist, in-person visit, or remote care), and quality of service (health outcomes and wait time). We analyze the provider decision on the mode of care delivery under both FFS and PCF and study whether PCF can be designed to yield a socially optimal outcome. We characterize analytically when patients, payer, and providers are better off under PCF and show that in many cases, PCF can be designed to yield a socially optimal outcome. We numerically calibrate our model for 14 states in the US. We observe that the average health status in a state is a source of heterogeneity that crucially drives the performance of PCF. We find that the model motivated by the current PCF implementation results in *too much* adoption of referral care and *too little* adoption of remote care. In addition, states with poor average health status may use more in-person care than socially optimal under a baseline (low) level of capitation. Moreover, relying on high levels of capitation leads to low adoption of in-person care.

Managerial Implications: Our results have health policy implications, by shedding light on how PCF might impact patients, payer, and providers. Under the current performance-based adjustments, low levels of capitation should be preferred. PCF has the potential to be designed to achieve socially optimal outcomes. However, the fee per visit may need to be tailored to the local population's health status.

Key words: Healthcare, Primary Care, Payment system.
History: Submitted 3/30/2022; Accepted 2/21/2023

1. Introduction

Primary care occupies a key function in the US health care system. The Primary Care Physician (or provider for short) is often the first and main point of contact with the healthcare system for patients and plays an important role in managing the care delivered to them. Yet, few doctors select to practice primary care in the US, in part because of the relatively lower compensation compared to other specialties (Burton et al. 2017), causing shortages in parts of the country (Association of American Medical Colleges 2019). Thus, there is a consensus among experts that reform is needed to improve primary care delivery.

Recent technological advances enable innovative modes of care delivery that can be beneficial to the patient experience. For example, telemedicine was made possible in part by the widespread adoption of electronic health records. The COVID-19 pandemic has shown that remote care can be feasible and beneficial to patients. It increases convenience and reduces transportation costs for patients. However, the standard visit-based reimbursement model under the Fee-For-Service (FFS) traditional payment system does not encourage using such innovations. Before the COVID-19 pandemic, the number of Medicare beneficiaries receiving telehealth services accounted for only one-quarter of a percent of the more than 35 million Medicare fee-for-service beneficiaries analyzed (LaPointe 2018, ASPE 2020). This very limited adoption is due mainly to barriers linked to Medicare reimbursement policies stated in the Social Security Act. In particular, normally Medicare only covers telehealth services in rural or shortage areas, when the patient is located in specific sites (*not* including the patient's home), for a limited type of practitioners, of interactions (*e.g.*, not audio-only phone calls), and of services. During the pandemic, many of the policies around remote care were waived. Most notably, some private payers and Medicare programs established parity of payment between clinical care and telehealth (Center for Connected Health Policy 2020). However, most experts anticipate that the policies will be re-instated after the pandemic (Shachar et al. 2020). Primary care reform must thus find ways to incentivize the most appropriate modes of care delivery for primary care in a more permanent fashion.

The Centers for Medicare & Medicaid Services (CMS), as the largest payer in the country, has been at the front line of designing this much-needed primary care reform. For example, back in 1992, CMS introduced the Physician Fee Schedule to try and reduce payment disparities among medical specialties, but large disparities have remained despite this effort. Starting in 2011, CMS has participated in the Multi-Payer Advanced Primary Care Practice Demonstration, aiming at encouraging the adoption of the "patient-centered medical home" model of care by primary care practices to better manage the health of patients with chronic conditions. In 2012, CMS launched

the Comprehensive Primary Care Initiative (CPCI) to encourage providers to redesign their practices with the goals of improving care delivery and patient health outcomes while reducing spending (CMS 2011). CPCI was centered on five comprehensive primary care functions: (1) access and continuity; (2) care management; (3) comprehensiveness and coordination; (4) patient and caregiver engagement; and, (5) planned care and population health. CPCI implemented a different way of paying for primary care services to improve the quality of care and compensate the primary care practice in a more holistic manner than FFS does. It included both a risk-adjusted prospective monthly payment per beneficiary and shared savings bonuses subject to meeting quality targets, in addition to the regular FFS visit-based reimbursements. CPCI took place over four years in seven regions involving a total of 502 medical practices. The results of the initiative were mixed (Ginsburg et al. 2016). On the positive side, CPCI seemed to increase access to primary care services, improve the care management of high-risk patients, and improve coordination of care transitions (Peikes et al. 2018). It also resulted in a 2% reduction in emergency department visits. On the negative side, CMS expenses from care management fees paid to physicians surpassed the spending reduction. Moreover, quality and patient experience remained practically unchanged. Hence, CPCI had no significant impact on most care quality measures and did not generate meaningful savings for CMS.

In 2017, CMS launched a new 5-year demonstration building upon CPCI: Comprehensive Primary Care Plus (CPC+). CPC+ had a broader range, involving 14 payers and 2876 practices spread over 14 regions. CPC+ strengthened the incentives introduced under CPCI: "[t]o support a fundamental change in care delivery, practices require a fundamental change in payment structure" (Sessums et al. 2016). To this end, CPC+ deepened the requirements regarding care delivery; it moved further away from FFS by providing both a higher monthly care management payment per beneficiary and a lower per-visit reimbursement (the latter only in Track 2; Track 1 was similar to CPCI); it replaced the shared savings model with a prospective bonus per beneficiary that must be repaid if performance targets are not met. With CPC+, CMS wanted to "help practices move away from one-size-fits-all, fee-for-service health care" (CMS 2016). Through a blend of FFS and capitation, subject to meeting a quality target, CPC+ hoped both to reduce incentives to unnecessarily inflate the volume of care, and to compensate physicians for tasks that can benefit patients but are not currently reimbursed under FFS. The goal was to allow practices "the flexibility to deliver care in the manner that best meets patients' needs (...). Practices might offer non-face-to-face visits (*e.g.*, electronic or telephone), offer visits in alternative locations" (Sessums et al. 2016). The CPC+ initiative has now concluded and the analysis of its results is still ongoing. However, preliminary analysis of the effect of CPC+ shows that it yielded only modest improvements in quality-of-care measures, together with a small increase in Medicare spending (Peikes et al. 2021).

In 2021, CMS launched the first of two cohorts participating in a new multi-payer payment demonstration: Primary Care First (PCF) (the second cohort, for practices that were enrolled in CPC+, debuted in January 2022) (CMS 2021a). The model will be tested over a six-year period. PCF expands the regional reach of CPC+, with 26 regions across the country. PCF builds upon CPC+, with less administrative burden, more transparency, and stronger performance-based incentives, with the potential for substantial bonuses and penalties. Under PCF, the risk-adjusted fixed upfront payment received by the physician is increased while each primary care visit is paid at the same flat fee set much lower than under CPC+. Finally, performance-based incentives are strengthened, with up to 50% of revenue as a bonus or up to 10% of revenue as a penalty (Thacker 2021). Building upon CPC+, "PCF is intended to test whether advanced primary care can reduce total cost of care while improving or maintaining quality" (McDermott and Roth 2019).

One of the key aspects of PCF is to continue incentivizing providers to deliver care in innovative ways, such as via remote care. Doctors may well deliver standard or health maintenance services remotely (at a lower cost, see Rohrer et al. (2010)) without significantly increasing patients' chance of poor health outcomes, especially for less complex patients. Remote care delivery can also help alleviate some of the health disparities between urban and rural areas, by improving the convenience of accessing primary care services (Douthit et al. 2015).

Supporters of PCF argue that, by deepening care delivery requirements while reinforcing the value-based financial incentives and distancing itself further from FFS, PCF could achieve more gains than CPC+ both in financial savings and in quality of care. Indeed, it was shown that a high level of capitated payment is necessary to shift primary care to non-visit-based care (Basu et al. 2017). PCF, by increasing capitated payments to primary care physicians, could equip them with the tools and incentives to create a real change in primary care delivery and health outcomes. On the other hand, critics note that PCF might financially hurt primary care practices due to the potentially heavy penalties, that the payment system's reliance on the capitated payment might be excessive, and that the performance-based incentives could have unintended consequences (such as curtailing necessary hospitalizations) (Sessums et al. 2019). It is unclear whether a low fixed payment per visit combined with a substantial amount of performance incentives and capitation provides the right incentives to transform care delivery as CMS hopes to.

This paper focuses on the impact of a payment system motivated by PCF on the different stakeholders (physician, patient, and payer). We refer to this payment system as PCF acknowledging that it only captures the main aspects of the real PCF in practice. We seek to answer the following research questions: what is the effect of PCF on each agent of the system? Can PCF yield a socially optimal outcome? How should PCF be designed to yield the best possible outcome? How

do patient population characteristics affect that optimal design? We address these questions by focusing on how PCF affects physicians' choice of care delivery mode — namely, in-office visits, remote visits, or referral to a specialist.

We find that PCF improves incentives to deliver remote care, but the payment terms need to be carefully calibrated to further incentivize the use of remote care and to avoid excessive fragmentation of the care delivered (*e.g.*, too many referrals). Importantly, in general, PCF can yield care delivery modes that align with the social optimum, for appropriate values of the performance-based adjustment and visit fee. Using a numerical implementation calibrated using real data, we obtain policy insights that shed light on the effect of the current PCF payment terms. We find that the current PCF maximizes social welfare when provider compensation is not too heavily based on capitation; with too much capitation, in-person care is underutilized. The current performance-based adjustment has the potential to yield socially optimal outcomes, but the current visit fee of $40.82 should be tailored according to the population's average health status. States with higher health status require a higher visit fee to prevent the over-utilization of specialist care. Overall, PCF appears to be a promising improvement over FFS for primary care delivery. Yet, its design should be differentiated to match the needs of the local population (*e.g.*, overall health status).

2. Related Literature

Since the implementation of the Affordable Care Act in 2010, CMS has launched several care delivery and payment initiatives (*e.g.*, Accountable Care Organization model, Episode-Based Payment model, CPC+, PCF, etc.) with the goal of improving service quality and health outcomes, and reducing care delivery cost. Initiatives involving alternative payment models seek to shape providers' behavior by rewarding quality and penalizing unnecessary expenditures. As CMS rolls out these initiatives, it hopes that some of them will prove to be successful in achieving the above-mentioned goals and sustainable, with the potential to eventually partially replace the traditional fee-for-service model, which is ineffective at driving up quality and reducing costs (Robinson 2001).

There has been an increasing interest in studying performance-based payment schemes in various healthcare settings with the goal of enhancing the design of alternative payment models (*e.g.*, Fainman and Kucukyazici (2020)). Fuloria and Zenios (2001) propose an outcome-based payment mechanism to incentivize providers' optimal treatment decisions in the presence of moral hazard. Jiang et al. (2012) study how providers allocate time slots between open-access and traditional appointments subject to a wait time target. The design and performance of bundled payment models, one of the most extensively promoted CMS initiatives, where the physician and the hospital get paid a fixed amount per episode of care, has also generated interest in the literature (*e.g.*, Gupta and Mehrotra 2015, Adida et al. 2017). The adoption of the Accountable Care Organization (ACO)

delivery model has resulted in new market interactions in the industry. Adida and Bravo (2019) study the quality incentive problem in the contracting of referral services between an ACO and an external provider, while Bravo et al. (2022) analyze care coordination with external providers under a shared-saving program.The hospital setting, which combines issues of quality, cost, capacity, and competition, gives rise to unique performance-based payment design problems. Savva et al. (2019) use a modified yardstick competition model to propose a performance-based payment scheme that incentivizes both cost and wait time reduction in the context of an emergency department. Jiang et al. (2020) consider a payer and two competing hospitals with information asymmetry on cost. They propose a performance-based payment scheme that rewards the hospitals for investing in service quality and capacity and show that patients benefit from stronger competition and from the bonus incentive payment. In contrast, in this work, we consider the framework of the PCF scheme as proposed by CMS for primary care; we focus on understanding the incentives behind its payment structure for physicians to adopt alternative care delivery modes and provide insights on how to calibrate it in order to achieve first-best outcomes. To the best of our knowledge, the existing modeling literature has not yet explored the performance of the PCF payment scheme.

The use of innovative payment models in the presence of alternative delivery modes for primary care has gained traction as healthcare systems realize the value of offering efficient primary care services. Campbell et al. (2009) empirically evaluate the effectiveness of performance-based payment on the quality of primary care in England. They find that quality increased in the short term for some chronic conditions, but that care continuity decreased after the implementation of the model. Using simulation, Basu et al. (2017) find that high capitation rates are needed to incentivize primary care practices to deliver team- and non-visit-based care. Zhong et al. (2016) develop a queuing model to study the performance of scheduling policies for a primary care practice in the presence of web consultations and e-visits in addition to in-office visits. In a follow-up paper, Zhong et al. (2018) consider the time allocation problem faced by a physician who delivers care using both e-visits and in-office visits. They show that operational efficiency is achieved only if the e-visit duration is short enough to compensate for the efficiency loss of incorporating online communications into the schedule. From a practice design perspective, Bavafa et al. (2019) study the impact of delegating care to non-physician providers (*e.g.*, nurse practitioner) on visit frequency, patient population size, physician revenue, and population health status. In a related paper, Bavafa et al. (2021) study the impact of patient revisit frequency on physician earnings, patient population size, and health status, in the presence of e-visits. Customization of visit frequency increases physician revenue but it can reduce the patient population size and patients' health status. The above papers model physician decisions under fee-for-service or capitation (or a mix of both). Our setting differs from this literature in that we study the incentive mechanism behind the PCF payment model. We

analyze the adoption of remote care as an alternative to in-person visits and the resulting effect on health quality and spending. In addition to focusing on a different payment mechanism, our paper also differs from Bavafa et al. (2021) because, in our model, the least sick patients use e-visits while the opposite is true in Bavafa et al. (2021). Furthermore, we explicitly incorporate the likelihood of increased health risks that can result from seeing a patient remotely.

In the delivery of specialty care, Rajan et al. (2018) study the quality-speed trade-off faced by specialists caring for a heterogeneous population of chronic patients. They show that revenue-maximizing providers treat a smaller patient population, spend more time with patients and have shorter wait times than welfare-maximizing providers. The adoption of telemedicine can make providers more productive (*i.e.*, see more patients) and increase social welfare, however, some patients might be worse off. Similar to our setting, heterogeneity arises from patients having different health statuses and traveling cost burdens from visiting the clinic.

3. Model

3.1. Model setup

Our model is primarily motivated by PCF and aims to capture the main drivers of the payment system while making some simplifications for the sake of maintaining tractability. We consider a primary care provider who takes care of a fixed panel of Medicare patients, who are included under the PCF agreement. The provider can also admit new patients, who are not under PCF and whose care is covered under a fee-for-service type of agreement with a different payer (*e.g.*, a private insurer). The notation is summarized in Table A1 in the Appendix.

3.1.1. Patient heterogeneity and modes of care. The yearly arrival rate of existing Medicare patients for primary care visits is λ (we do not model any prevention effort aiming at improving patients' health status). For each incoming visit, the provider selects one of three options for delivering care to the patient. The patient may be referred to an external specialist, seen in person by the provider, or may utilize an alternative delivery method, such as an e-visit. We refer to the latter type of delivery as "remote", and to in-person visits as "face-to-face". Episodes of care with Medicare patients have heterogeneous complexities denoted $x \in [0,1]$, where a higher value of x represents a more complex episode of care (we will refer to x as the patient complexity). We model the distribution of the complexity level as uniform on $[0,1]$. Such a distribution is commonly used in the Health Economics and Healthcare OM literature to model patient heterogeneity (*e.g.*, Mahjoub et al. 2018, Adida 2021, Çakıcı and Mills 2021). The Medicare patients are already under the provider's care, and the provider is familiar with the patient's overall health condition, who can thus determine x in advance of the interaction and use it to select the adequate mode of care.

Indeed, a brief description of why an appointment is needed is usually sufficient to decide on a suitable appointment type in most cases, especially for already-established patients.

We determine the provider's optimal delivery mode decision for each visit based on the patient's complexity. Specifically, the provider decides the complexity thresholds $e_0 \leq e_1 \in [0,1]$, where patients with complexity within $[0, e_0]$ are seen remotely, patients with complexity within $(e_0, e_1]$ are seen face-to-face, and patients with complexity within $(e_1, 1]$ are referred to a specialist (but the three modes of care delivery are not necessarily all utilized). We assume that new patients are always seen face-to-face, thus the choice of care delivery mode only affects existing Medicare patients. We make this assumption because it is common for payers to extend the possibility of remote visits only for established patients, whom the provider already knows and has already physically examined in the past, and thus can more easily treat without a new in-person physical exam. For instance, upon the onset of the COVID pandemic, Medicare initially stated in 2020 that "virtual check-in services can only be reported when the billing practice has an established relationship with the patient" (CMS 2020).

3.1.2. Cost of care. Each of the three modes of care delivery incurs a different cost to the physician and the patients. The provider's care delivery cost is denoted $\bar{c}(x)$ for remote care, and $c(x)$ for face-to-face care; the specialist delivery cost is $\tilde{c}(x)$. We assume $\bar{c}(\cdot)$, $c(\cdot)$ and $\tilde{c}(\cdot)$ are increasing, $\bar{c}(x) < c(x) < \tilde{c}(x) \quad \forall x \in (0,1)$, and $c(x) - \bar{c}(x)$ and $\tilde{c}(x) - c(x)$ are non-increasing in x. These assumptions state that more complex patients incur more costs. Furthermore, for any patient, remote care incurs less cost than face-to-face care, which is itself less costly than specialist care. Finally, the cost differential between remote and face-to-face care decreases with patient complexity, so less complex patients represent "low-hanging fruits" for whom there is a more potential cost-saving opportunity by using remote care. Similarly, the cost differential between face-to-face and specialist care is also non-increasing with patient complexity to capture the fact that there is fewer cost-savings potential for high-complexity patients. These conditions are sufficient to ensure the provider's objective and social welfare are concave. The provider internalizes the negative effects of having to refer a patient to a specialist. Indeed, a cost t is perceived per referred patient as reflecting the loss of the personal relationship, and additional communication and coordination effort in order to maintain continuity of care. The patient pays to the physician a co-payment p for face-to-face care and $\bar{p}$ $(\leq p)$ for remote care and pays a co-payment $\tilde{p}$ $(> p)$ to the specialist in case of a referral.

3.1.3. Result of medical interaction and costs. After the needed care for that visit has been delivered, we model the patient's condition as either resolved or not. While this is a simplification of reality, using a binary health outcome is sufficient to capture the key trade-off of cost

and effectiveness between the different care delivery modes. If the condition is not resolved, we refer to the result of care as a "failure" (*e.g.*, the patient is admitted for inpatient services due to a worsening health condition). The chance of failure depends on the patient's complexity and on the type of care delivery. We denote the chance of failure as $\bar{q}(x)$ for a remote visit, $q(x)$ for a face-to-face visit, and $\tilde{q}(x)$ for a referral. We denote

$$Q(x) = \int_0^x q(t)dt$$

the fraction of failures among patients seen face-to-face with complexities up to $x \in [0,1]$. We assume that $\bar{q}(\cdot),\ q(\cdot),\ \tilde{q}(\cdot)$ are monotonically increasing on $[0,1]$ with $\bar{q}(x) > q(x) > \tilde{q}(x)\ \ \forall x \in [0,1]$. Moreover, $\bar{q}(\cdot) = (1+\beta)q(\cdot)$ and $\tilde{q}(\cdot) = (1-\alpha)q(\cdot)$, where $0 < \alpha < 1,\ 0 < \beta \leq 1/q(1) - 1$. Finally, $q(0) = 0$. These assumptions ensure that a higher complexity level increases the chance of failure in any given delivery mode. They also state that the chances of failures across delivery modes are proportional to each other. We scale the levels of complexity so that at the lowest complexity ($x = 0$), the chance of failure is zero. The constant β (resp. α) captures the potential quality loss (resp. gain) due to remotely seeing (resp. referring) the patient. The bounds on α and β ensure that the chance of failure lies within $[0,1]$ for all delivery modes. In our model, remote care leads to a higher chance of failure than face-to-face care. While for certain types of care, like for mental health services, remote care has achieved effectiveness comparable to in-person care, it is reasonable to assume that for primary care, which often involves using the patient's physical characteristics, remote care would in general tend to be less effective than in-person care (Shigekawa et al. 2018).

Failures incur a cost of z for the provider if the provider delivered the care (face-to-face or remotely). This cost can be viewed as a reputation cost, as patients may attribute a failed outcome to the provider's decision not to refer to a specialist. The cost z can also be viewed as the disutility experienced by an altruistic physician upon failure (*i.e.*, the provider internalizes the patient's poor health outcome). In addition to incurring a direct cost, a failure may have financial repercussions on the provider payment under PCF, as detailed below. Failure also imposes a cost of w to the payer due to the need for further costly treatment. Table 1 lists the parameters for each delivery option.

3.1.4. Patient value. Remote care delivery provides a utility u to the patient due to convenience (*e.g.*, avoiding transportation costs). All patients served by the provider incur a disutility v_W proportional to the average wait time, which is the delay until the actual appointment. We use wait time as a proxy for measuring service quality, as we describe in Section 3.1.6. A failure has a negative impact on the patient; we model as v_H the corresponding patient disutility (*e.g.*, inconvenience and cost of seeking further care, discomfort due to continued presence of the medical condition).

We consider two types of payment systems, described next.

Table 1 Summary of model parameters for the three types of care delivery modes

	Remote visit	Face-to-face visit	Referral
Service rate	$\bar{\mu}$	μ	–
Care delivery cost	$\bar{c}(x)$	$c(x)$	$\tilde{c}(x)$
Chance of failure	$\bar{q}(x)$	$q(x)$	$\tilde{q}(x)$
Provider cost per failure	z	z	–
Provider coordination cost for a referral	–	–	t
Payer failure cost	w	w	w
Patient convenience utility	u	–	–
Patient failure disutility	v_H	v_H	v_H
Patient wait disutility	v_W	v_W	–
Patient co-payment	$\bar{p}$	p	$\tilde{p}$

3.1.5. FFS payment system. Under FFS, the payer reimburses the provider $f(x)$ per patient for a face-to-face visit, with $f(\cdot)$ increasing and $f(x) > c(x) \;\; \forall x \in [0,1]$. As explained in the Introduction, under FFS remote care does not incur any reimbursement, as was largely the case before the COVID pandemic (and is expected to be the case afterward). While the delivery cost and the reimbursement depend on patient complexity, we assume that the provider's profit margin m^F does not, where $m^F \equiv f(x) + p - c(x) \;\forall x \in [0,1]$. Namely, the provider is compensated more when the cost of delivery is higher due to a more complex (*e.g.*, more time-consuming) patient condition, but the added compensation mirrors exactly the extra cost so that the margin remains unchanged. Hence, the provider does not earn more profit from more complex patients. This assumption eliminates any cherry-picking incentives and allows us to focus solely on the financial incentives created by the PCF payment system. As an approximation, we assume that the margin m^F is the same for all face-to-face visits, regardless of whether it is for a new or an existing patient.

The payer compensates the specialist at a rate of $\tilde{f}(x)$ for a patient with complexity x, with $\tilde{f}(x) > f(x) \;\forall x \in [0,1]$, under both FFS and PCF.

3.1.6. PCF payment system. Under PCF, the practice receives a fixed amount R per year. (In practice, the provider is paid per beneficiary and per month; the quantity R corresponds to the aggregate amount for the attributed patient population per year.) The fixed amount of R is meant to help the physician cover upfront fixed costs and invest in activities to improve care delivery and patient experience, *e.g.*, hiring new staff, training employees, etc. Indeed, CMS does not prescribe how the population-based payment should be spent. Moreover, under PCF the provider receives a reduced flat fee per visit, r, whether the visit is delivered face-to-face or remotely and regardless of the visit complexity, with $r < f(x) \;\; \forall x \in [0,1]$. (Currently, the flat fee is set at $40.82, see CMS (2021b).) Note that while remote care does not incur reimbursement under FFS, it does under PCF, see *e.g.*, https://innovation.cms.gov/files/x/pcf-faqs.pdf (point 55), which states

that the flat primary care visit fee applies to most telehealth visits (a minority of visit types may not be compensated, and are not considered as part of the scope covered by the model).

Payment to the provider is adjusted for service quality and health outcomes. The service quality must exceed a minimum threshold to pass a "Quality Gateway." If a practice does not pass the Quality Gateway, the performance adjustment is -10% regardless of health outcomes. If it passes, the health outcomes performance adjustment is determined from -10% to $+34\%$ of revenue ("performance-based adjustment"). Service quality is a complex metric because it not only involves objective measures (*e.g.*, wait time) but also subjective aspects of the service (*e.g.*, staff friendliness). In reality, service quality is measured by a collection of criteria not limited to the wait time. However, in this paper, we consider wait time as a proxy for service quality. Indeed, the CAHPS Clinician & Group Adult Survey 3.0 (Agency for Healthcare Research and Quality 2015), which is used as a performance measure under PCF (Thacker 2021), includes questions (*e.g.*, number 6 and 8) related to promptness in the scheduling of an appointment. Hence, wait time plays a role as a measure of service quality under PCF to evaluate the provider's performance. Health outcomes are measured using the Acute Hospital Utilization metric, which is based on inpatient admission and observation stay discharges during the measurement year (CMS (2021b), chapter 5, p. 61).

While in the practical implementation of PCF the performance-based adjustment may take the form of either a bonus or a penalty, it can equivalently be seen as a penalty-based incentive after re-scaling upwards the payments to the provider. We model the health quality adjustment as a penalty proportional to the rate of failure. We denote p_H as the penalty for each failure. Hence in our model, a PCF contract is described by the parameters (R, r, p_H). We model the Quality Gateway criterion test, based on service quality, as determined by the average wait time. We denote $\bar{W}$ the maximum wait time to pass the Quality Gateway. The parameters of the contract vary depending on whether the provider meets the Quality Gateway qualification target. If the provider care decisions result in an average wait time lower than $\bar{W}$, then they *qualify* and they are paid under (R^q, r^q, p_H^q). Otherwise, they do *not qualify* and are compensated under $(R^{nq}, r^{nq}, p_H^{nq})$.

We note that the PCF system used in practice also includes a continuous improvement bonus incentive of up to 16%, which is not captured in this paper. Studying the impact of this bonus is beyond the scope of this manuscript because it would require keeping track of the practice performance over time. The dynamic nature of the problem would bring tractability issues, due both to the dynamic evolution of the performance and the possible strategic behavior of practices optimizing over time.

3.2. Service Dynamics

3.2.1. Framework. We model the provider practice as an M/M/1 queuing system. The service rate depends on the mode of care used. Zhong et al. (2018) report that e-visits can be managed in half the time (or less) of an in-office visit. Thus we assume that remote care has a shorter service time than face-to-face care. For remote care the service rate is $\bar{\mu}$ and for face-to-face care, the service rate is μ ($<\bar{\mu}$) (whether for new or existing patients). We model the provider as operating at a level of utilization equal to $\rho<1$.

By referring patients to a specialist, or by adopting remote care, the provider releases some of her capacity. The shortage of primary doctors is evidence of a backup of demand for primary care. We thus assume that the provider backfills the released capacity with new patients while maintaining a utilization level of ρ. We assume the new patients are non-Medicare, and thus are not part of the PCF program (*i.e.*, are reimbursed under FFS), because the physician accepts other insurers, and recent evidence shows that physician practices receive more than 70% of their revenue through FFS (Sokol 2020). Hence, in this work, panel size only varies through the addition of new non-Medicare patients. (In theory, the practice could be better off by adjusting upwards or downwards the Medicare patient panel size, but this aspect is not considered in this paper. Papers such as Bavafa et al. (2019) analyze in detail panel size decision-making; we leave the study of how to select panel size within PCF as a future research direction.)

We denote the resulting new patient arrival rate as $\lambda_N(e_0,e_1)$. This rate depends on complexity thresholds (e_0,e_1), defined above since the amount of remote care and referrals determines how much time has been freed to see new patients. Thus, the provider's net arrival rate is $e_1\lambda+\lambda_N(e_0,e_1)$ (*i.e.*, existing Medicare patients seen remotely or face-to-face, in addition to new patients).

The average service rate μ_{net} is given by

$$\mu_{net}(e_0,e_1)\equiv\frac{e_1\lambda+\lambda_N(e_0,e_1)}{e_0\lambda/\bar{\mu}+(e_1-e_0)\lambda/\mu+\lambda_N(e_0,e_1)/\mu}.$$

The rate corresponds to the practice's average service rate (patients per year) considering the average appointment time for remote visits ($1/\bar{\mu}$) and face-to-face ($1/\mu$) visits, and adjusting for the fraction of visits of each type (*e.g.*, the proportion of remote visits is $e_0\lambda/(e_1\lambda+\lambda_N(e_0,e_1))$).

The value of $\lambda_N(e_0,e_1)$ is determined so the utilization, after including new patients, remains at ρ, that is, $\lambda_N(e_0,e_1)$ satisfies the following equation:

$$\rho=\frac{e_1\lambda+\lambda_N(e_0,e_1)}{\mu_{net}(e_0,e_1)},$$

$$\textit{i.e., }\lambda_N(e_0,e_1)=\left(\rho-e_0\frac{\lambda}{\bar{\mu}}-(e_1-e_0)\frac{\lambda}{\mu}\right)\mu. \tag{1}$$

3.2.2. Average wait time. The expected wait time in an M/M/1 queuing system is $W = \frac{\rho/\mu_{net}(e_0,e_1)}{1-\rho}$, which, after some simplifications, leads to

$$W(e_0) = \frac{\bar{\mu}\rho^2}{(1-\rho)(\lambda(\bar{\mu}-\mu)e_0 + \bar{\mu}\mu\rho)}. \tag{2}$$

We note that the expected wait time does not depend on the referral threshold e_1. This is because we assume the service rate of face-to-face care is the same for a new patient as for an existing patient. Hence, any patient who is sent for a referral is exactly replaced by a new patient without affecting the average wait time. We also observe that the expected wait time is decreasing in e_0. As e_0 increases, more patients are seen remotely (and fewer face-to-face), and thus the net average service rate μ_{net} increases. As a result, patients spend less time on average with the physician (because $\frac{1}{\mu_{net}(e_0,e_1)} = \frac{\rho}{\lambda e_0(1-\mu/\bar{\mu})+\mu\rho}$ decreases in e_0). It follows that the overall average wait time decreases. We note that if $\bar{W} \geq \frac{\rho}{(1-\rho)\mu}$, any $e_0 \in [0,1]$ satisfies the Quality Gateway condition $W(e_0) \leq \bar{W}$.

3.3. Analysis under FFS

Under FFS, the yearly expected profit for the provider is given by

$$\Pi^{FFS}_{provider} = m^F\lambda(e_1 - e_0) - \lambda\int_0^{e_0} \bar{c}(x)dx - t\lambda(1-e_1) - z\lambda(\beta Q(e_0) + Q(e_1)) + (m^F - zQ(1))\lambda_N(e_0,e_1). \tag{3}$$

In the above expression, the first term is the profit from face-to-face care for existing patients. Remote patients and referrals bring no revenue but incur costs. The second term represents the cost incurred from remote visits. The third term is the coordination cost due to referrals. The fourth term is the cost due to treatment failures of patients seen either remotely or face-to-face. Finally, the last term is the extra profit due to new patients, where $zQ(1)$ is the expected disutility for failed treatment of new patients since $Q(1)$ is the expected value of the patient complexity.

We assume $t/z + Q(1) < q(1)$. This assumption states that the referral cost (t) is not too high compared to the reputation cost (z) incurred by a failure. In other words, this assumption ensures that some referrals take place under FFS. If $t/z + Q(1) \geq q(1)$, referrals are too costly for the provider, and thus $e_1^F = 1$, which would imply that under FFS the physician sees all patients face-to-face without any referral to a specialist. This situation does not match what is observed in practice. The following proposition characterizes the optimal complexity thresholds under FFS.

Proposition 1. *Under FFS, the provider chooses complexity thresholds:*

$$(e_0^F, e_1^F) = \left(0, q^{-1}\left(\frac{t}{z} + Q(1)\right)\right). \tag{4}$$

Intuitively, FFS does not provide incentives to deliver care remotely to any patient for two reasons. First, there is no compensation for remote patients, even though direct and indirect costs are incurred, via a higher chance of poor health outcomes. Second, while remote care releases some

capacity that can be used to bring in new patients (and thus additional compensation), the newly available capacity would not be sufficient to replace every patient seen remotely with a new patient (because the service time for a remote visit, while lower, is not zero). Hence, the compensation gained from new patients would not fully balance out the lost compensation from the use of remote care. As a result, it is optimal for the physician to opt out of remote care entirely. This is consistent with what has been observed in practice: before the COVID-19 pandemic, the use of remote care was extremely limited under the FFS payment system, as described in the Introduction.

Proposition 1 illustrates that the provider faces a trade-off between referring high-complexity patients to a specialist to avoid failures (which are more likely without a referral), and the added coordination and communication cost of managing those referrals. It follows from this proposition that the provider makes fewer referrals to specialists (*i.e.*, e_1^F is higher) when the provider's referral cost t is high, or the provider's failure cost z is low (that is, the provider's referral cost is high in comparison to the cost incurred by a patient failure). Note that profit margin m^F does not drive this decision since the provider can replace referred patients with new patients' appointments and make the same profit. Hence, referring a patient to a specialist has no financial impact as every referred patient can be replaced with a new patient bringing the same level of compensation.

3.4. Analysis under PCF

Under a PCF contract (R, r, p_H) the yearly expected profit for the provider is given by

$$\begin{aligned}\Pi^{PCF}_{provider} =& R + r\lambda e_1 + \bar{p}\lambda e_0 + p\lambda(e_1 - e_0) - \lambda\int_0^{e_0}\bar{c}(x)dx - \lambda\int_{e_0}^{e_1}c(x)dx - t\lambda(1-e_1) \\ &- z\lambda\left[\beta Q(e_0) + Q(e_1)\right] - p_H\lambda\left[(1-\alpha)Q(1) + \alpha Q(e_1) + \beta Q(e_0)\right] + (m^F - zQ(1))\lambda_N(e_0, e_1).\end{aligned} \tag{5}$$

In the above expression, R is the upfront payment (proportional to the practice's attributed Medicare patient population); the flat visit fee r is received for each face-to-face or remote visit; the copayment $\bar{p}$ is received for each remote visit and p for each face-to-face visit. The cost $\bar{c}(x)$ is incurred for a remote visit and cost $c(x)$ is incurred for a face-to-face visit of a patient with complexity x. A cost t is experienced for treatment referrals for the additional coordination and communication burden on the PCP, and a cost z is incurred for treatment failures when the patient is seen either face-to-face or remotely by the practice. Finally, penalty p_H is incurred as an outcome-based adjustment for poor health outcomes (*i.e.*, high failure rate). New patients, who are non-Medicare, continue to bring in the profit margin m^F and expected disutility for failure $zQ(1)$. Note that the contract parameters (R, r, p_H) depend on whether the practice meets the Quality Gateway qualification criterion or not, as described in Section 3.1.6.

The following proposition characterizes the optimal complexity thresholds (e_0^P, e_1^P) under PCF.

PROPOSITION 2. *Under PCF, there exists a unique optimal selection* (e_0, e_1) *and we provide a sequence of steps to obtain it in the proof (in Appendix B). Moreover, if the Quality Gateway criterion is satisfied, the provider selects care complexity thresholds that lead to either remote-only; remote and referral; or remote, face-to-face, and referral care. The latter case — region* $\mathcal{R}$*, with thresholds* $(max\{e_0^{min}, \bar{e}_0\}, \bar{e}_1)$ *— is the only case with qualification where the three modes of care co-exist. (Quantities* $e_0^{min}, \bar{e}_0, \bar{e}_1$ *are defined in the proof.)*

The proof of the proposition, which can be found in Appendix B, provides the mathematical conditions when each of the cases described above arises. The above result identifies three possible sets of thresholds defining different cases of optimal decisions for the provider under PCF with Quality Gateway qualification (*i.e.*, with a remote care threshold e_0^P that is high enough so the average wait time does not exceed the maximum allowed $\bar{W}$). The case where $e_0^P = e_1^P = 1$ represents the extreme case where a single delivery mode is utilized (remote-only). In the case where $e_0^P = e_1^P < 1$, two delivery modes are utilized: remote and referral, with no patient seen face-to-face. Finally, in the region, $\mathcal{R}$, the three modes of care co-exist, as the provider chooses thresholds $(e_0^P, e_1^P) = (max\{e_0^{min}, \bar{e}_0\}, \bar{e}_1)$ where $0 < max\{e_0^{min}, \bar{e}_0\} < \bar{e}_1 < 1$ and e_0^{min} is the minimum fraction of remote care ensuring Quality Gateway qualification.

A central research question we aim to address in this paper is how PCF should be calibrated. In the above analysis, the contract terms determine which modes of care are being utilized when the physician optimally responds to the incentives set by the contract. To focus on the most realistic setting, we frame our discussion under the scenario where the payer, who designs the contract, sets its terms so that the practice meets the Quality Gateway qualification and all three modes of care are utilized. (In Section 5, we formalize the payer's objective as that of maximizing social welfare and describe how the PCF contract can give rise to socially optimum decisions by the provider. We will also focus on the realistic scenario where it is socially optimal to pass the Quality Gateway and have the three modes of care co-exist at the social optimum.) Namely, we focus on the subset of PCF contracts such that the provider chooses to have the least complex patients seen remotely, the most complex patients referred to specialists, and the rest seen face-to-face, with an average wait time below the threshold ensuring Quality Gateway qualification. This implies that the PCF parameters are set so that the complexity thresholds lie in the region $\mathcal{R}$ (the mathematical conditions for this to be the case are detailed in the proof of Proposition 2). We focus on the case where the practice meets the Quality Gateway qualification criterion because participation in Primary Care First is voluntary, and as such, it is unlikely that a practice would choose to participate without meeting the criterion, which would lead to a penalty of 10% of its revenue. Thanks to the large size of the patient pool, the wait time does not significantly deviate from its expected value and thus the practice can fairly accurately anticipate whether or not it will meet the wait time bound.

4. Discussion

4.1. Comparing FFS and PCF care thresholds

As shown in Proposition 1, there is no remote care under FFS, *i.e.*, $e_0^F = 0$, which implies that in all regions, $e_0^F < e_0^P$. In other words, FFS gives no incentives for remote care, but PCF does (there is remote care in the region $\mathcal{R}$ as well as in the regions where $e_0^P = e_1^P = 1$ or $e_0^P = e_1^P < 1$) and hence PCF gives rise to more remote care than FFS.

The comparison of threshold e_1 between FFS and PCF depends on the input parameters. The next lemma establishes how the threshold compares when the outcome of PCF lies in the region $\mathcal{R}$.

LEMMA 1. *In region $\mathcal{R}$, we have $e_1^P < e_1^F$.*

This result proves that PCF gives rise to more referrals than FFS for any contract within region $\mathcal{R}$. There are two reasons for this effect. First, PCF offers a smaller profit margin for face-to-face visits than FFS, hence face-to-face visits from new patients (replacing referred patients) become economically more appealing to the provider under PCF. Second, a face-to-face visit has a higher chance of failure than a referral, and since failures are penalized under PCF through the performance adjustment, PCF has more incentives to refer to a specialist. Hence, overall, in region $\mathcal{R}$, PCF gives rise to more referrals and more remote care than FFS.

4.2. Other Performance Metrics

In this section, we compare FFS and PCF from the perspective of each agent in the system. To this aim, we first define the patient utility and Medicare payer profit as follows:

Patient utility. Patients experience a positive utility u for the convenience of being seen remotely, a disutility v_H for experiencing further complications, and a disutility v_W for waiting to see the primary care provider. Furthermore, the patient pays co-payments p, $\bar{p}$ and $\tilde{p}$ for a face-to-face, remote, and specialist visit, respectively. Thus, the aggregate patient utility is given by

$$\Pi_{patient}(e_0, e_1) = \lambda \left(u e_0 - v_H \left[\beta Q(e_0) + \alpha Q(e_1) + (1-\alpha) Q(1) \right] - v_W W(e_0) - \bar{p} e_0 - p(e_1 - e_0) - \tilde{p}(1 - e_1) \right). \tag{6}$$

Medicare Payer profit. The payer compensates the provider and the specialist for their services. It also faces an additional cost of w if the treatment fails (*e.g.*, future expected care cost). The payer's profit function depends on the payment system. For a PCF contract (R, r, p_H) (where the terms depend on whether the practice qualifies for Quality Gateway or not), the payer's profit is

$$\Pi_{payer}^{\text{PCF}}(e_0, e_1) = -R - \lambda \left(e_1 r + \int_{e_1}^{1} \tilde{f}(x) dx \right) + \lambda (p_H - w) \left[\beta Q(e_0) + \alpha Q(e_1) + (1-\alpha) Q(1) \right]. \tag{7}$$

In the above expression, the payer pays the provider the fixed amount R, pays for remote and face-to-face visits at the flat rate r, and collects penalty p_H for poor health outcomes. Moreover, each failure yields a cost of w. Finally, the payer compensates a specialist visit by the amount $\tilde{f}(x)$ for a patient with complexity x. The payer's profit function under FFS can be written as

$$\Pi^{\text{FFS}}_{payer}(e_0, e_1) = \lambda \left(- \int_{e_0}^{e_1} f(x)dx - w\left[\beta Q(e_0) + \alpha Q(e_1) + (1-\alpha)Q(1)\right] - \int_{e_1}^{1} \tilde{f}(x)dx \right). \quad (8)$$

Now, we define a quantity that is useful in comparing PCF and FFS. Let

$$\Delta Q \equiv \alpha Q(e_1^F) - \beta Q(e_0^P) - \alpha Q(e_1^P) = \alpha \int_{e_1^P}^{e_1^F} q(t)dt - \beta \int_0^{e_0^P} q(t)dt.$$

We refer to ΔQ as the 'health outcome effect'. Intuitively, ΔQ represents the *long-term* benefits of PCF as measured by its failure rate reduction in comparison to FFS.

PROPOSITION 3. *We have the following comparisons between PCF and FFS for the Medicare patient population under PCF contract within the region $\mathcal{R}$:*

1. *The utilization of in-person services — face-to-face provider visits and referrals to a specialist — is lower under PCF than FFS;*
2. *The failure rate is lower under PCF than FFS iff the health outcome effect is positive;*
3. *The average wait time is lower under PCF than FFS;*
4. *The patient panel size, including Medicare and new patients, is larger under PCF than FFS;*
5. *Patients are better off under PCF than FFS iff*

$$-v_H \Delta Q + (\tilde{p} - p)(e_1^F - e_1^P) \leq (u + p - \bar{p})e_0^P + v_W(W(0) - W(e_0^P)),$$

where e_1^F, e_0^P, e_1^P and $W(\cdot)$ are given in Proposition 1, the proof of Proposition 2, and Eq. (2).

Part 1 of the proposition states that PCF lowers the utilization of in-person services by Medicare patients, regardless of the contract terms, because it encourages remote visits. Part 2 of the proposition links the effect of PCF on the failure rate to the sign of the health outcome effect. Note that the health outcome effect is positive when the quality gain from additional referrals under PCF (as compared to FFS) surpasses the potential quality loss due to adopting some level of remote care. We note that ΔQ is decreasing in r and increasing in p_H and therefore to ensure better health outcomes under PCF (relative to FFS), r should be set not too high, and p_H not too low.

Part 3 shows that any PCF contract lowers the average wait time, thanks to the use of remote care. Part 4 proves that PCF increases the panel size. Since PCF leads to more remote care and referrals than FFS, which brings in more new patients, the net panel size is larger under PCF. Finally, part 5 compares the patient benefit between the two payment systems. For patients, the benefit from PCF depends on the interaction between the failure rate (and associated failure disutility), the co-payments, the disutility from the wait time, and the utility of remote care. Patients benefit in terms of health outcomes when PCF leads to fewer failures (positive health

outcome effect). The fact that PCF has more referrals (with higher co-payment than face-to-face) than FFS hurts the patients financially. However, the fact that PCF has more remote care (with lower co-payment than face-to-face) than FFS benefits the patients financially, due to both co-payments and the extra utility enjoyed by patients for the convenience of remote care (u). Finally, patients benefit in terms of wait time under PCF as more remote visits reduce the average wait time to access care.

We next compare the payer expenditure and benefit under FFS and PCF. We denote

$$\begin{aligned}\Delta X \equiv \int_0^{e_1^F} c(x)dx - \int_0^{e_0^P} \bar{c}(x)dx - \int_{e_0^P}^{e_1^P} c(x)dx + t(e_1^P - e_1^F) + \int_{e_1^F}^{e_1^P} \tilde{f}(x)dx \\ - z[\beta Q(e_0^P) + Q(e_1^P) - Q(e_1^F)] - p(e_1^F - e_1^P + e_0^P) + \bar{p}e_0^P\end{aligned}$$

as the 'expenditure effect'. Intuitively, ΔX represents the *short-term* benefits of PCF, as measured by the payer reimbursement expenditures reduction under PCF compared to FFS.

The payer's expenses under PCF critically depend on the fixed payment of R to the physician, a payment that does not exist under FFS. Clearly, for the payer the performance of PCF compared to FFS depends on R — for example, if R is sufficiently high, the physician prefers PCF while the payer is worse off under PCF, a trivial result that does not yield any interesting insight. Hence, we make an assumption on R (in the following result only) to make a more meaningful comparison.

Proposition 4. *Suppose $R \geq 0$ is set so that the provider is indifferent between PCF and FFS for the Medicare patient population. Then,*

1. *The Medicare payer reimbursement expenditures (paid to the provider) are lower under PCF than FFS iff the expenditure effect is positive;*
2. *The Medicare payer is better off under PCF than FFS iff either (i) $\Delta X \geq 0$ and $\Delta Q \geq 0$, or (ii) $\Delta X \geq 0$ and $\Delta Q < 0$ and $w < -\Delta X/\Delta Q$, or (iii) $\Delta X < 0$ and $\Delta Q \geq 0$ and $w > -\Delta X/\Delta Q$.*

We set R to ensure provider participation in PCF by making the provider indifferent between FFS and PCF for the Medicare patient population. This is consistent with the fact that PCF was not designed to financially penalize primary care physicians (who are already scarce). Essentially, the amount R is calibrated to cover the fixed cost of infrastructure improvements that the physician makes to participate in PCF, which is the intent behind this fixed upfront payment.

Part 1 of Proposition 4 indicates that the payer incurs less reimbursement expenditure under PCF when the expenditure effect is positive. However, the payer does not only value reimbursement expenditures when setting up the payment model: the payer also takes into account the failure cost w incurred with each patient failure. Part 2 of the proposition provides conditions that characterize when the payer is better off overall under PCF, depending on the expenditure effect and the health outcome effect. When $\Delta Q < 0$ and $\Delta X < 0$, there are more failures and more reimbursement

expenditures under PCF, so the payer prefers FFS. Similarly, when $\Delta Q \geq 0$ and $\Delta X \geq 0$, PCF leads to fewer failures and fewer reimbursement expenditures and so it dominates FFS. Otherwise, the payer is facing a trade-off. When $\Delta Q < 0$ and $\Delta X > 0$, there are more failures but fewer reimbursement expenditures under PCF, and so the payer prefers PCF as long as the payer's failure penalty, w, is low enough. Conversely, when $\Delta Q > 0$ and $\Delta X < 0$, FFS incurs more failures than PCF but fewer reimbursement expenditures, and so the payer prefers PCF when the failure cost w is high.

5. Social Welfare

5.1. Social Welfare Formulation

As a benchmark, we now seek to find socially optimal care decisions. Such decisions aim at maximizing social welfare. After establishing this benchmark, it will be possible to determine whether or not PCF can be designed to yield decisions that coincide with the social optimum.

Consistent with the literature (*e.g.*, Dranove 1996, Chap. 4, p. 62), we define social welfare as comprising the utility of all the stakeholders involved in the care delivery process, *i.e.*, provider, payer, patients, and specialist. Equations (3) and (5)-(8) provide the patient utility, the provider benefit, and the Medicare payer profit under FFS and PCF. We next obtain the specialist's profit.

Specialist profit. We assume the specialist provides standard care to referred patients, making no decision affecting either the payments or the health outcome. Thus, the specialist's profit is

$$\Pi_{spc}(e_0, e_1) = \lambda\tilde{p}(1 - e_1) + \lambda \int_{e_1}^{1} (\tilde{f}(x) - \tilde{c}(x))dx.$$

The above expression captures the fact that for each referral, the specialist incurs cost $\tilde{c}(x)$, and receives a co-payment from the patient as well as a payment from the payer. Note that the specialist does not incur a reputation cost in case of failure because the specialist did not make any decision that may be seen as responsible for the failure in our model. In other words, failures following referral care are considered non-avoidable, *i.e.*, solely due to the severity of the health condition. Therefore, social welfare, which is the utility of the system comprising provider, Medicare patients, payer, and specialist, is given by the cumulative costs of delivering care across the different modes of delivery, the cumulative costs of failures, the patient's benefit from remote care and the profit from new patients (all other payments, being internal to the system, balance each other out):

$$\Pi_{social}(e_0, e_1) = \lambda_N(e_0, e_1)(m^F - zQ(1)) + \lambda\Big(ue_0 - v_W W(e_0) - \int_0^{e_0} \bar{c}(x)dx - \int_{e_0}^{e_1} c(x)dx - \int_{e_1}^{1} \tilde{c}(x)dx$$
$$- t(1 - e_1) - (v_H + w + z)\left(\beta Q(e_0) + Q(e_1)\right) - (v_H + w)(1 - \alpha)(Q(1) - Q(e_1))\Big). \quad (9)$$

5.2. Socially optimal care modes

The next proposition describes the socially optimal thresholds (e_0^S, e_1^S).

PROPOSITION 5. *At the social optimum, there are four possible sets of care complexity thresholds either* $(1,1)$, $(\bar{e}^S, \bar{e}^S)$, $(\bar{e}_0^S, 1)$ *or* $(\bar{e}_0^S, \bar{e}_1^S)$. *The latter case — region* $\mathcal{R}^S$ *with thresholds* $(\bar{e}_0^S, \bar{e}_1^S)$ *— is the only case where the three modes of care co-exist.*

The proof of the proposition provides the conditions for each of these four cases to arise as well as the (explicit or implicit) expressions of quantities $\bar{e}^S, \bar{e}_0^S$ and $\bar{e}_1^S$. The four possible cases correspond to care delivered either remotely only; remotely and by referral; remotely and face-to-face; and remotely, face-to-face, and by referral, respectively. Remarkably, remote care is always used to deliver care for the lowest complexity patients under the social optimum.

A relevant question is whether FFS and/or PCF may give rise to the socially optimal modes of care. Using Proposition 1, it follows that FFS may not do so because no remote care exists under FFS, while all socially optimal outcomes lead to a non-zero fraction of remote care.

COROLLARY 1. *FFS cannot coordinate to the socially optimum care thresholds.*

The following section investigates whether and how PCF can lead to the socially optimal outcome.

5.3. PCF Coordination

When designing a payment system such as PCF, CMS, as a public insurer, is generally modeled as aiming to maximize social welfare. Thus, whether it is possible to design PCF to yield the social optimum is an important question. Hence, in this section, we investigate whether a PCF contract may lead the provider's selection of care modes to match the social optimum, that is, to $e_0^P = e_0^S$ and $e_1^P = e_1^S$. Since having three modes of care is the most practical scenario, we study coordination in the case where the socially optimal solution results in three modes of care offered — that is, region $\mathcal{R}^S$. The following result assumes $\bar{W}$ can be adjusted with the constraint $\bar{W} \geq W(e_0^S)$ so that the socially optimum solution satisfies the Quality Gateway criteria; this is arguably the most interesting and practical case.

PROPOSITION 6. *Suppose that we are in region* $\mathcal{R}^S$. *If* $\underline{r} < f(0)$, *there exists a family of PCF contracts, characterized by* $r^q \in [\underline{r}, f(0)]$, p_H^q *given as a function of* r^q *and* $\bar{W} = W(e_0^S)$, *that coordinates the provider care decisions to the socially optimal ones.*

The definition of $\underline{r}$ (this lower bound on r^q stems from ensuring $p_H^q > 0$), and the function of r^q leading to determining p_H^q can be obtained from the proof of the result in Appendix B. It can be seen from this result that the coordinating contracts are characterized by well-calibrated values of r^q and p_H^q; amount R does not affect coordination because it is simply a way to ensure physician

participation but does not affect incentives regarding care mode decisions. Proposition 6 thus indicates that PCF may help align the provider decisions with the socially optimal care modes, via an appropriate choice of penalties and visit fees.

Proposition 6 focuses on the case when the socially optimal outcome fulfills the Quality Gateway qualification criterion. We note that in the family of contracts described in the result, the Quality Gateway constraint is binding, that is, the average wait time is equal to the maximum allowed. In some cases, it may be possible to obtain a coordinating contract that relaxes this constraint, *i.e.*, where the average wait time is strictly below the maximum $\bar{W}$. We describe in the proof of Proposition 6 how to construct such a contract. This would yield a single contract — a specific value of r^q and of p_H^q — rather than a *family* of contracts.

It follows from Proposition 6 that PCF holds great promise to improve care delivery to patients, as long as its parameters are carefully selected. In addition, the flat fee per visit and performance adjustment in the PCF contract that achieves the social optimum may not be unique. Indeed, there may be a continuum of contracts, differing via their penalty parameter and visit flat fee, that align with the social optimum. This property is a desirable feature of the PCF payment system as it allows a high degree of flexibility to the payer for choosing a coordinating contract. This result recalls a standard result in Operations Management, stating that there exist infinitely many revenue-sharing contracts that coordinate supply chain decisions, each contract with a continuum of profit shares allocated to the retailer and the supplier (Cachon and Lariviere 2005).

6. Numerical analysis

In this section, we use state-level data to evaluate the effect of the PCF contract on care mode decisions and on the Quality gateway criterion, and we analyze coordinating contracts.

6.1. Calibration of model parameters

The first cohort of the PCF initiative (approximately 900 primary care practices) started in January 2021. A second cohort (primarily for former CPC+ practices) launched in January 2022 and will run for a 5-year period. Because the transition from CPC+ to PCF is still very recent, we calibrate our model using enrollment data from the CPC+ initiative. The CPC+ initiative ran from January 2017 to December 2021; it included over 2,800 primary practices in 14 U.S. regions and close to 1.8M beneficiaries. Table C9 in Appendix C lists the number of practices and Medicare beneficiaries enrolled under the CPC+ initiative at the state level. In each state, our unit of analysis is an average practice. States can sharply differ in population density (and thus in patients' travel convenience from remote services), from a rural state like Montana (MT) with 6.8 ppl. per sq. mi. to an urban state like New Jersey (NJ) with 1195.5 ppl. per sq. mi. (U.S. Census Bureau 2019). The underlying

health status of the population also varies widely. For instance in Oregon (OR), 50% of Medicare patients suffer from chronic conditions, while the rate is 77% in NJ (Table C7 in Appendix C). Demand for primary care depends on the size of the Medicare population, their health status, and the number of primary care physicians in the state. For instance, Hawaii (HI) has the lowest average number of visits per year per practice (due to the state's small Medicare population), while MT has the largest (due to the small number of practices in the state) (Table C9 in the Appendix C). These observations highlight the heterogeneity among states, which we exploit to obtain insights into the optimal design of PCF.

We next provide a summary of the calibration approach. The estimation details and corresponding data are available in Appendix C. We refer the reader to Table C3 for a summary of the values used for each state.

The yearly average arrival rate λ is obtained from the 2016 "Physician Compare" utilization data (Table C9, column (iii)). The target utilization ρ takes values in $\{0.75, 0.8, 0.85\}$, but we present the results for the case $\rho = 0.8$ only, as no significant differences were observed for other values. For the service rate, we have $\mu = \lambda/\rho$ patients per year. We estimate that face-to-face visits last 30 minutes; remote visits are shorter and we conservatively estimate a 20-minute duration. Thus, the service rate for remote visits is estimated as $\bar{\mu} = \mu \times \frac{30 min/visit}{20 min/visit} = 1.5\ \mu$.

Our analytical model assumes a constant FFS margin $m^F := f(x) + p - c(x)$ independent of the patient complexity $x \in [0, 1]$. We model the in-office reimbursement rate and visit cost as linear in patient complexity. Specifically, the reimbursement rate is defined as $f(x) := f_0 + \texttt{FFS\_rate} \times x$, and we estimate f_0 and `FFS_rate` based on CPT codes-based reimbursement rates for primary care in-office visits (see codes in Table C4). It follows that $c(x)$ is linear in x and we denote $c(x) := c_0 + \texttt{FFS\_rate} \times x$. We assume the cost for the lowest-complexity patient is equal to the co-payment, *i.e.*, $c_0 = p$. The cost of remote care is also modeled as linear in the care complexity x, and for simplicity, we assume it is proportional to the cost of face-to-face care. Specifically, $\bar{c}(x) = 90\% \times c(x)$. The specialist visit cost is estimated as a constant and conservatively calibrated as two times the highest in-office cost, *i.e.*, $\tilde{c}(x) := 2 \times c(1)\ \forall x \in [0, 1]$. The specialist rate $\tilde{f}(x)$ is also assumed to be a constant; we estimate it using the average of in-office visit FFS rates for a subset of medical specialties (see Table C5).

PCF aims to encourage quality of care by rewarding/penalizing practices for excessive utilization of inpatient services. Albeit imperfect, we use the 2018 Medicare proportion of *non-elective* inpatient services (in contrast to *elective* inpatient services) as a proxy for the population's underlying health status, and use this quantity to estimate the probability of treatment failure $q(x)$ (see Table C6). We vary the change in the failure probability due to referral (α) and remote care (β) in the range 10%–30%. For illustration purposes, we only report results for $\alpha = 10\%$ and $\beta = 30\%$.

We estimate the convenience of remote care, u, as the total (two-way) travel time to the nearest hospital (Pew Research Center 2018) valued using the state median hourly wage (Maciag 2017) (we make this approximation despite the fact that Medicare patients are typically not fully employed). Thus, patients in low-density states value remote care more because of the longer travel time to access care. The disutility incurred from treatment failure, v_H, is estimated by assuming the patient is admitted to the hospital for an average length-of-stay, which is valued using the state median wage, plus the inconvenience of having to visit the health care facility, which is estimated at u. The average disutility experienced due to waiting time to see the provider, v_W, is hard to quantify because it may be affected by a wide variety of features. We calibrated it by assuming the failure cost and the wait cost are of similar magnitude under the scenario where all care is delivered in person, *i.e.*, $v_W W(0) = v_H Q(1)$. The cost of access delay v_W thus depends on the population's health status and the state median hourly wage. See Table C7 for details.

Medicare's primary care and specialist visit co-payments range from \$15 to \$25 and \$30 to \$50, respectively (Fay 2019). We set the primary care face-to-face co-payment at $p = \$20$, the remote visit co-payment at $\bar{p} = 90\% \times p = \$18$, and the specialist visit co-payment at $\tilde{p} = \$40$ for all states.

The provider's cost of treatment failure is difficult to estimate. In our numerical experiments, we consider $z \in [\$20, \$60]$ per patient, which ensures that the three modes of care co-exist for a wide range of contracts parameters, but we report the results only for $z = \$30$ as no significant changes in insights were observed with other values. The coordination cost associated with a referral is calibrated as $t \in [\$20, \$80]$ per patient referred and we report the values for $t = \$50$. The payer's cost of treatment failure w is conservatively estimated as the cost of an inpatient visit, which in 2017 reached $\$11,700$ according to AHRQ (Agency for Healthcare Research and Quality 2020).

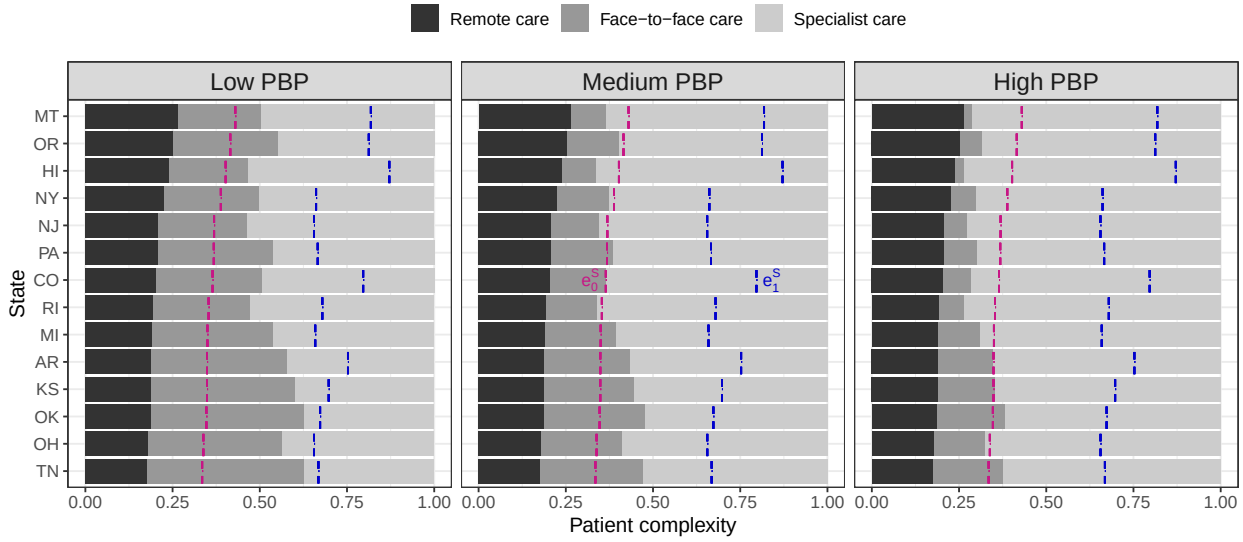
We calibrate the parameters of the baseline PCF contract according to the current implementation of PCF, which has three main components: a population-based payment (PBP), a flat-visit fee (FVF), and a performance-based adjustment (PBA). The PBP corresponds to a capitation payment per beneficiary per month; the exact amount varies depending on the risk profile of the patient population treated by the primary care practice. We will consider three possible levels for the PBP (low, medium, high) for sensitivity analysis: \$28, \$56, and \$84. The FVF is paid to the provider per visit (face-to-face or remote); the current value is \$40.82. The PBA is based on the practice's health outcomes and varies from -10% to $+34\%$ of revenue. The baseline values for these parameters are shown in the lower part of Table C3. We explain the mapping from PCF parameters to our equivalent contract parameters (R, r, p_H) in Appendix C.

6.2. Results

In the first part of this section, we analyze the adoption of the three modes of care under the baseline PCF and compare it to the socially optimal solution. In the second part, we investigate the impact of the Quality Gateway criterion on the modes of care delivery. Finally, in the third part, we study the family of coordinating contracts and discuss what characteristics of the population affect the coordinating PCF contract. Note that, while the insights provided in this section are based on numerical analysis only, we carefully calibrated our model parameters based on realistic values across many states in the US and checked for robustness within reasonable ranges of these parameters. As a result, these insights are valuable for understanding the effect of adjustments to the PCF payment scheme.

6.2.1. Modes of care delivery under the baseline PCF. Figure 1 shows the provider's choice of care mode according to the patient complexity (x-axis) for all states. The figure also shows the care mode thresholds at the social optimum (e_0^S, e_1^S) (dashed red and blue thresholds). The three panels correspond to the three considered possible values of the capitation payment (PBP).

Figure 1 **Care thresholds under the baseline PCF. States are ranked in decreasing value of e_0^S.**



Note. The x-axis corresponds to patient complexity. In each panel, the left dashed threshold corresponds to e_0^S and the right dashed one to e_1^S. The wait time qualification threshold $\bar{W} = (1 + 0.05)W(e_0^S)$, which ensures e_0^S leads to qualification. Low PBP: \$28, Medium PBP = 2×\$28, High PBP = 3×\$28.

Our first observation is that a higher capitation payment (*i.e.*, PBP) results in less face-to-face care. The decrease in face-to-face care is primarily caused by an increase in referrals. This can be explained as follows. Under the baseline PCF the performance penalties are linked to the capitation

payment: higher PBP implies higher penalties (see Appendix C). Thus, the provider is incentivized to adopt more referral care as doing so leads to better health outcomes (due to the lower probability of failure) and hence, lower penalty payments. Furthermore, the provider is able to backfill the capacity released from those referrals with visits from new patients, who are not subject to the performance-based adjustments and generate higher revenue. As a result, a higher PBP can have the unintended consequence of making the provider rely more on specialist care. On the other hand, a higher PBP may lead to an increased patient panel size as new patients are included thanks to the freed-up capacity, which helps improve access to care for new patients. Interestingly, we note that even though remote care can lead to worse health outcomes, the provider still relies on it to care for low-complexity patients in all states even for high values of the capitation amount. This is to ensure that the Quality Gateway qualification criterion is met.

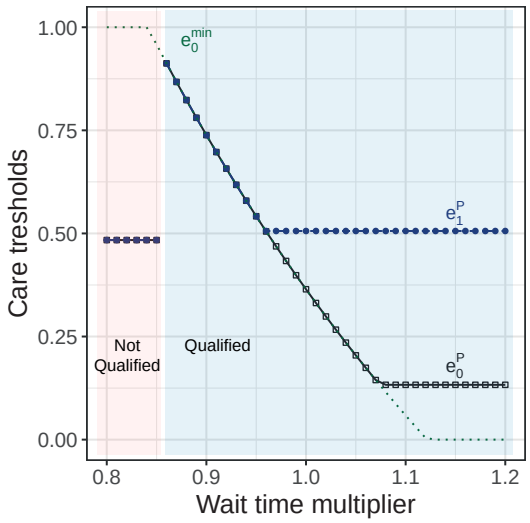
Secondly, we observe that the sub-optimality of the baseline PCF manifests differently for the different states. While all states exhibit less remote care and more referrals than the first-best, at all three PBP levels, the states' "health status" plays a noteworthy role. A state's health status, as measured by the utilization of inpatient services, is inversely related to the failure rate parameter δ. In states with lower health status (*e.g.*, TN, OH, OK, KS), the provider uses less remote care and more in-person care than in states with higher health status. In the case of low PBP, a state with low health status is often associated with more in-person care than prescribed by the first-best. Similarly, a state with high health status (*e.g.*, HI, MT, OR, CO) tends to overuse referrals to a larger extent than states with lower health status. It follows from these observations that how the baseline PCF differs from the first-best varies across states, and depends crucially on how healthy on average the population is in the state.

6.2.2. Effect of Quality Gateway criterion. The Quality Gateway qualification criterion is a minimum service quality level imposed by the payer to be eligible for a performance incentive and avoid a -10% penalty. In our model, this is captured by a maximum average wait time, $\bar{W}$. In designing the PCF payment system, the payer enjoys some flexibility in setting this maximum wait time. In this section, we analyze numerically the impact of adjusting the Quality Gateway condition (via adjusting $\bar{W}$) on the optimal care thresholds chosen by the provider under PCF. To do this we consider a range of $\pm 20\%$ around the value of $W(e_0^S)$ for the maximum wait time $\bar{W}$. Lowering $\bar{W}$ makes it harder for the provider to meet the Quality Gateway qualification condition while increasing it makes it easier.

Figure 2 depicts the optimal thresholds (e_0^P, e_1^P) (y-axis) as the maximum wait time varies in the range $(0.8, 1.2) \times W(e_0^S)$ (the x-axis corresponds to the multiplier in the range from 0.8 to 1.2). Note that the average wait time is decreasing in e_0; therefore meeting the maximum average

wait time constraint imposes a minimum amount of remote care (e_0^{min}). We observe that when the maximum wait time is low, *i.e.*, the Quality Gateway qualification requirement is more stringent, providers may not be able to satisfy the Quality Gateway condition and opt to divert all patients to either a remote setting or to a specialist. This is because, with no qualification, the provider does not make a significant profit from providing face-to-face care to patients reimbursed under PCF. Instead, the provider focuses on maximizing the profit from new non-Medicare patients and thus frees up its in-person capacity to accept as many new patients as possible.

Figure 2 **Impact of Quality Gateway wait time requirement on modes of care delivery (Colorado depicted as an illustrative state)**



For intermediate values of the maximum wait time, the provider adopts just enough remote care to ensure qualification — namely, the provider chooses the care thresholds so that the average wait time equals precisely the maximum allowed wait time $\bar{W}$. Doing so translates into setting the remote care threshold at the minimum ensuring qualification (*i.e.*, $e_0^P = e_0^{min}$). In this region, the maximum wait time is high enough that the qualification requirement can be met, but low enough that it is not profitable to generate an average wait time strictly below the maximum, and thus the constraint is binding. We distinguish two regions in this intermediate range. For low-intermediate maximum wait time, there is no in-person care (*i.e.*, $e_1^P = e_0^P = e_0^{min}$). In this subregion, the maximum waiting time is relatively low, and thus the minimum remote care threshold is relatively high and exceeds the value of e_1 that would be optimal for the provider. Since e_1 needs to be above e_0, the provider selects e_1 as close to its optimal value as it can, that is, at e_0. For high-intermediate maximum wait time, the minimum remote care threshold is below the optimal value of e_1 and thus the three modes of care exist (*i.e.*, $0 < e_0^P = e_0^{min} < e_1^P < 1$).

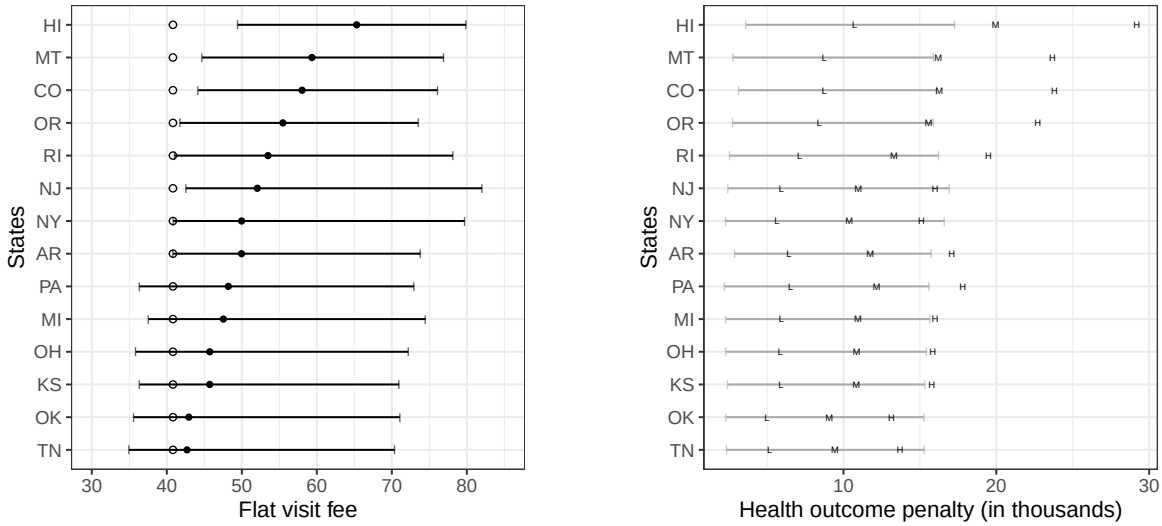
Finally, for large values of the maximum wait time, where the requirement is more lenient, the Quality Gateway qualification condition does not force the provider to adopt remote care beyond the amount that maximizes the provider's profit. The provider implements three modes of care and adopts more remote care than the minimum required by the Quality Gateway condition because it is in their best interest to do so.

To summarize, we see that the maximum wait time in the Quality Gateway qualification condition can impact the provider's choice of care delivery modes and therefore should be carefully selected, in conjunction with the other contract parameters. A too stringent requirement may induce the provider to fail qualification or to qualify while limiting the use of in-person services by PCF beneficiaries and instead reserve it for newly admitted patients.

6.2.3. Coordinating PCF contracts. We next numerically assess the family of coordinating contracts characterized in Proposition 6 and compare these contracts to the baseline PCF contract. Figure 3(a) shows the range of values of the FVF parameter (x-axis) that *can* achieve the socially optimal outcomes. Namely, for any FVF within the range shown in the figure, there exists a penalty (*i.e.*, a performance-based adjustment) that gives the provider incentives to choose care thresholds yielding the optimal social welfare. We find that the currently implemented value of FVF = \$40.82 (empty circle point) is within the range of admissible values that can achieve socially optimal care thresholds *for half of the considered states*. This observation implies that the level of FVF currently used by CMS is too low to align decisions with the first-best for all states. Indeed, the sub-optimality of the baseline visit fee translates into up to 1.8% lower social welfare. States for which the current value of \$40.82 is too low are those with relatively better health status and thus, from Figure 1, those with too much referral care at the current PCF contract compared to the social optimum. To align with the first-best, providers in these states require incentives to move care away from specialists and redirect it to the primary care setting. When the per-visit fee is too low, the provider defers more care of PCF beneficiaries to specialists in order to admit new patients. Hence, in order to achieve coordination in these states, a higher visit fee would be needed.

Figure 3(a) highlights that CMS may want to consider a non-homogeneous per-visit fee across states, recognizing that achieving first-best outcomes in all states may require adjusting the PCF contract to each state's specific characteristics. Indeed, the filled point in Figure 3(a) represents the per-visit fee that achieves coordination under the baseline performance-based adjustment. We notice that states with better health status would require a higher FVF, consistent with our previous observations. Moreover, for states with worse health status, the current (baseline) value of FVF at \$40.82 appears relatively close to the coordinating value, indicating that the baseline FVF is close to being coordinating for states with lower health status. In summary, it may be

Figure 3 Coordinating contracts



(a) Coordinating visit fee. The circle point corresponds to the baseline fee and the dark circle to the coordinating fee under the baseline penalty with PBP= \$28.

(b) Coordinating penalty. The points L, M, and H correspond to the baseline penalty with a PBP of \$28, 2×\$28, and 3×\$28, respectively.

beneficial to have a fee per visit that is higher in states where reliance on referral care is more excessive. Figure 3(b) shows the range of values of the health outcome penalty (x-axis) that *can* achieve socially optimal outcomes. Namely, for any health outcome penalty within the range shown in the figure, there exists an FVF parameter that gives the provider incentives to choose care thresholds yielding optimal social welfare. We find that the current (baseline) PCF performance-based adjustment is within the range of admissible values that can achieve socially optimal care thresholds for all states as long as the PBP is not too high. However, for higher values of the PBP, coordination would require a health outcome penalty higher than what is within the eligible range. Complementing Figure 1, this figure suggests that under low levels of capitation, the current PCF performance-based adjustment provides incentives to choose close to socially optimal modes of care delivery. However, under a high level of capitated payment (H point in the figure), which may be the case for higher-risk practices, the current performance-based adjustment should be lowered in order to yield socially optimal outcomes.

7. Concluding Remarks

While there is a consensus among experts that primary care needs reform, what such a reform should consist of remains controversial. Fee-for-service clearly does not provide the right incentives to reduce the volume of unnecessary care, reduce costs, and improve patient health outcomes and the patient experience. The COVID-19 pandemic has undoubtedly shown that both patients

and practitioners are open to remote care delivery; removing reimbursement barriers will be key in a post-pandemic environment to sustain this momentum. The newly proposed PCF initiative represents a step in the direction of achieving these goals. Through a blending of capitation and fee-for-service, PCF offers more flexibility to healthcare providers to offer care across different modes. By analyzing a payment model motivated by PCF, we find that if carefully calibrated, PCF can indeed drastically improve upon fee-for-service and can even achieve the first-best. One of the main advantages of PCF over fee-for-service is that it can incentivize remote care delivery which not only reduces cost but can also benefit many patients. Indeed, we show that remote care is always present at the social optimum.

We find that PCF improves incentives to deliver remote care, but the payment terms need to be carefully calibrated to avoid excessive fragmentation of the care delivered (*e.g.*, too many referrals). Importantly, in general, PCF can yield care delivery modes that align with the social optimum, for appropriate values of the performance incentives and visit reimbursement. However, to perform well, the PCF system must have contract terms that are tailored to the average health status of the local population. Overall, PCF appears a promising improvement over fee-for-service and a step toward better delivery of primary care. It will be interesting to validate empirically the performance of the PCF implementation when the initiative is concluded and assess how the characteristics of the population affect its performance.

CMS appears to be taking an incremental approach of testing a payment model on a relatively small scale, learning lessons from this implementation, then revising the payment model and testing it on a slightly larger scale, and iterating. A major aim of our analysis is precisely to identify how the current PCF model could be modified to improve outcomes, in order to help inform how the next generation of primary care payment systems should be designed. Our analysis suggests that the visit fee is generally too low to induce optimal outcomes, but the overall incentive structure is appropriate. It is possible that CMS will make changes to PCF in the next few years and roll out a modified program for primary care. If CMS down the road chooses to maintain the structure of the payment model and modify the strength of the incentives, our model could help to better understand the implications of those changes.

This paper is focused on using analytical modeling to shed light on the incentives driving decisions under a payment system motivated by PCF to derive managerial insights. To this end, we use a stylized model that abstracts away from some features that exist in reality, to capture the main effects, with the goal of gaining tractability in our analysis. We acknowledge that these simplifications represent a limitation of our work. Firstly, we consider waiting time as the primary metric driving patient service experience. In practice, the assessment of patient experience is more complex

and comprehensive and includes other metrics from the patient experience of care survey (which includes wait time), as well as monitoring high blood pressure and hemoglobin A1c for diabetes patients, colorectal cancer screening, and advance care planning. Secondly, we assume the PCF performance-based adjustments are linear in the number of patients. In reality, the adjustments are more nuanced and take into account how the practice stands relative to its peers. Thirdly, failure costs and patient convenience utility from remote care are modeled as independent of patient complexity. Similarly, considering strategic patients who could have a say in the mode of care delivery could represent an interesting direction for future research. Fourth, we do not consider any prevention activities aiming at improving the patient's health status, any adjustment in the Medicare patient panel size, nor the continuous improvement bonus present in reality in PCF. Finally, some of the insights we derive for the design of PCF were obtained through a numerical study, and therefore, one must be cautious in extrapolating these findings to different settings. However, the numerical study is calibrated using real data for 14 states in the U.S., and the results in these 14 states were remarkably similar, which brings robustness to the insights we derive.

References

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

Agency for Healthcare Research and Quality (2015) CAHPS Clinician & Group Survey. https://www.ahrq.gov/cahps/surveys-guidance/cg/index.html. Accessed July 2021.

Agency for Healthcare Research and Quality (2020) National inpatient hospital costs: The most expensive conditions by payer, 2017. https://www.hcup-us.ahrq.gov/reports/statbriefs/sb261-Most-Expensive-Hospital-Conditions-2017.jsp. Accessed January 2022.

ASPE (2020) Medicare Beneficiary Use of Telehealth Visits: Early Data from the Start of the Covid-19 Pandemic. Issue Brief, July 2020, https://aspe.hhs.gov/system/files/pdf/263866/hp-issue-brief-medicare-telehealth.pdf. Accessed July 2021.

Association of American Medical Colleges (2019) New findings confirm predictions on physician shortage. https://www.aamc.org/news-insights/press-releases/new-findings-confirm-predictions-physician-shortage. Accessed April 2020.

Barnett ML, Bitton A, Souza J, Landon BE (2021) Trends in outpatient care for medicare beneficiaries and implications for primary care, 2000 to 2019. *Annals of Internal Medicine* 174(12):1658–1665.

Basu S, Phillips RS, Song Z, Bitton A, Landon BE (2017) High levels of capitation payments needed to shift primary care toward proactive team and nonvisit care. *Health Affairs* 36(9):1599–1605.

Bavafa H, Savin S, Terwiesch C (2019) Managing patient panels with non-physician providers. *Production and Operations Management* 28(6):1577–1593.

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

Bravo F, Levi R, Perakis G, Romero G (2022) Care coordination for healthcare referrals under a shared-savings program. *Production and Operations Management* .

Burton R, Berenson RA, Zuckerman S (2017) Medicare's evolving approach to paying for primary care. Technical report, The Urban Institute, The Robert Wood Johnson Foundation, https://www.urban.org/sites/default/files/publication/95196/2001631_medicares_evolving_approach_to_paying_for_primary_care_0.pdf. Accessed November 2019.

Cachon GP, Lariviere MA (2005) Supply chain coordination with revenue-sharing contracts: Strengths and limitations. *Management Science* 51(1):30–44.

Çakıcı ÖE, Mills AF (2021) On the role of teletriage in healthcare demand management. *Manufacturing & Service Operations Management* 23(6):1483–1504.

Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M (2009) Effects of pay for performance on the quality of primary care in England. *New England Journal of Medicine* 361(4):368–378.

Center for Connected Health Policy (2020) COVID-19 telehealth coverage policies. https://www.cchpca.org/resources/covid-19-telehealth-coverage-policies. Accessed July 2020.

CMS (2011) Solicitation for the Comprehensive Primary Care Initiative. Centers for Medicare & Medicaid Services, Center for Medicare and Medicaid Innovation, https://innovation.cms.gov/Files/x/Comprehensive-Primary-Care-Initiative-Solicitation.pdf. Accessed November 2019.

CMS (2016) CMS launches largest-ever multi-payer initiative to improve primary care in America. Centers for Medicare & Medicaid Services Press release, https://www.cms.gov/newsroom/press-releases/cms-launches-largest-ever-multi-payer-initiative-improve-primary-care-america. Accessed November 2019.

CMS (2020) Medicare telemedicine health care provider fact sheet. URL https://www.cms.gov/newsroom/fact-sheets/medicare-telemedicine-health-care-provider-fact-sheet, accessed August 2022.

CMS (2021a) Primary Care First model options. Centers for Medicare & Medicaid Services, Center for Medicare and Medicaid Innovation, https://innovation.cms.gov/innovation-models/primary-care-first-model-options Accessed July 2021.

CMS (2021b) Primary Care First: Payment and attribution methodologies PY 2022. Technical report, U.S. Department of Health & Human Services, Centers for Medicare & Medicaid Services, = https://innovation.cms.gov/media/document/pcf-py22-payment-meth-vol1. Accessed January 2023.

Douthit N, Kiv S, Dwolatzky T, Biswas S (2015) Exposing some important barriers to health care access in the rural USA. *Public health* 129(6):611–620.

Dranove D (1996) Measuring costs. Sloan FA, ed., *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies*, chapter 4 (Cambridge University Press).

Fainman EZ, Kucukyazici B (2020) Design of financial incentives and payment schemes in healthcare systems: A review. *Socio-Economic Planning Sciences* 100901.

Fay B (2019) Doctor Visit Costs. https://www.debt.org/medical/doctor-visit-costs/. Accessed February 2020.

Fuloria PC, Zenios SA (2001) Outcomes-adjusted reimbursement in a health-care delivery system. *Management Science* 47(6):735–751.

Ginsburg PG, Darling M, Patel K (2016) CMMI's new Comprehensive Primary Care Plus: Its promise and missed opportunities. Health Affairs Blog, https://www.healthaffairs.org/doi/10.1377/hblog20160531.055050.

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

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

Jiang H, Pang Z, Savin S (2020) Performance incentives and competition in healthcare markets. *Production and Operations Management* 29(5):1145–1164.

LaPointe J (2018) Medicare reimbursement rules limit telehealth adoption. https://revcycleintelligence.com/news/medicare-reimbursement-rules-limit-telehealth-adoption. Accessed July 2020.

Maciag M (2017) Median Wages by State. https://www.governing.com/gov-data/wage-average-median-pay-data-for-states.html. Accessed February 2020.

Mahjoub R, Ødegaard F, Zaric GS (2018) Evaluation of a pharmaceutical risk-sharing agreement when patients are screened for the probability of success. *Health Economics* 27(1):e15–e25.

McDermott M, Roth J (2019) A closer look at Primary Care First. *National Law Review* 9(330), https://www.natlawreview.com/article/closer-look-primary-care-first.

Peikes D, Dale S, Ghosh A, Taylor EF, Swankoski K, OMalley AS, Day TJ, Duda N, Singh P, Anglin G, et al. (2018) The comprehensive primary care initiative: Effects on spending, quality, patients, and physicians. *Health Affairs* 37(6):890–899.

Peikes D, Swankoski K, Timmins L, Petersen D, Geonnotti K, Tu H, Singh P, Ghosh A, Dale S, Keith R, et al. (2021) Independent evaluation of comprehensive primary care plus (CPC+): third annual report. Technical report, Mathematica Policy Research.

Pew Research Center (2018) How far Americans live from the closest hospital differs by community type. https://www.pewresearch.org/fact-tank/2018/12/12/how-far-americans-live-from-the-closest-hospital-differs-by-community-type/. Accessed February 2020.

Rajan B, Tezcan T, Seidmann A (2018) Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care. *Management Science* 65(3):1236–1267.

Robinson JC (2001) Theory and practice in the design of physician payment incentives. *The Milbank Quarterly* 79(2):149–177.

Rohrer JE, Angstman KB, Adamson SC, Bernard ME, Bachman JW, Morgan ME (2010) Impact of online primary care visits on standard costs: A pilot study. *Population Health Management* 13(2):59–63.

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

Sessums LL, Basu S, Landon BE (2019) Primary Care First – Is it a step back? *The New England Journal of Medicine* 381(10):898–901.

Sessums LL, McHugh SJ, Rajkumar R (2016) Medicare's vision for advanced primary care: New directions for care delivery and payment. *Journal of the American Medical Association* 315(24):2665–2666.

Shachar C, Engel J, Elwyn G (2020) Implications for telehealth in a postpandemic future: Regulatory and privacy issues. *Journal of the American Medical Association* 323(23):2375–2376.

Shigekawa E, Fix M, Corbett G, Roby DH, Coffman J (2018) The current state of telehealth evidence: a rapid review. *Health Affairs* 37(12):1975–1982.

Sokol E (2020) Healthcare reimbursement still largely fee-for-service driven. Recycle Intelligence, https://revcycleintelligence.com/news/healthcare-reimbursement-still-largely-fee-for-service-driven. Accessed January 2023.

Thacker R (2021) CPC+ and Primary Care First: The new CMS payment model explained. MingleHealth, https://minglehealth.com/blog/cpc-plus-and-primary-care-first-cms-payment-model-explained. Accessed July 2021.

US Census Bureau (2019) Census: Population Density Data. https://www.census.gov/library/visualizations/time-series/demo/nia_county_maps.html. Accessed July 2021.

Zhong X, Hoonakker P, Bain PA, Musa AJ, Li J (2018) The impact of e-visits on patient access to primary care. *Health Care Management Science* 21(4):475–491.

Zhong X, Li J, Bain PA, Musa AJ (2016) Electronic visits in primary care: Modeling, analysis, and scheduling policies. *IEEE Transactions on Automation Science and Engineering* 14(3):1451–1466.

Online Supplement

Appendix A: Notation

Table A1 Notation

$x \in [0,1]$	patient complexity
e_0, e_1	patient complexity thresholds: maximum complexity to receive remote care and face-to-face care
λ	arrival rate of Medicare patients
$\lambda_N(e_0, e_1)$	arrival rate of new patients
$\mu, \bar{\mu}$	service rate for face-to-face care and for remote care
μ_{net}	average service rate
$c(x), \bar{c}(x), \tilde{c}(x)$	delivery cost for face-to-face care, remote care, and referral care
$p, \bar{p}, \tilde{p}$	patient co-payment for face-to-face care, remote care, and referral care
$q(x), \bar{q}(x), \tilde{q}(x)$	chance of treatment failure for face-to-face care, remote care, and referral care
$Q(x)$	$= \int_0^x q(y)dy$
α, β	decrease in failure chance due to referral and increase in failure chance due to remote care w.r.t. face-to-face care
u	patient utility from receiving remote care
t	provider coordination cost due to referral
v_H, z, w	disutility due to treatment failure for patient, provider and payer
$\tilde{f}(x)$	reimbursement to the specialist for referral visit
$f(x)$	reimbursement for face-to-face care under FFS
m^F	profit margin for face-to-face care under FFS
R	fixed payment from payer under PCF
r	reimbursement for face-to-face care and for remote care under PCF
p_H	PCF contract penalty per patient failure
ρ	target utilization of provider practice
$W(\cdot)$	expected wait time
$\bar{W}$	maximum average wait time to meet Quality Gateway qualification criterion
$\Pi_{patient}, \Pi_{payer}, \Pi_{provider}, \Pi_{spc}$	patient utility, payer profit, provider profit and specialist profit
Π_{social}	social welfare

Appendix B: Proofs

Proof of Proposition 1

The provider profit is decreasing in e_0, hence $e_0^F = 0$. The provider profit is concave in e_1. The solution follows from solving the first-order condition over $[0,1]$.

Proof of Proposition 2

Since $c(x) - \bar{c}(x)$ and $\tilde{c}(x) - c(x)$ are non-increasing in x, and other terms of the provider's objective are linear or proportional to $-Q(\cdot)$, which is concave, it follows that the provider's objective is jointly concave. Then, it suffices to solve the first-order conditions subject to the border conditions $e_0^{min} \le e_0 \le e_1 \le 1$ in the case when the provider meets the Quality Gateway qualification criterion and $0 \le e_0 < e_0^{min}, \quad e_0 \le e_1 \le 1$ in the case when the provider does not meet the Quality Gateway qualification criterion. We then compare the provider objective in the case with and without qualification to determine which of these two cases is optimal for the provider, and thus what the optimal thresholds are.

Consider contract parameters $\zeta := (R, r, p_H)$, we add the superscript q or nq for *qualified* and *non-qualified*, respectively, only when strictly needed. For a general contract ζ the FOC are:

$$\varphi_0(e_0, \zeta) \equiv \frac{1}{\lambda} \frac{\partial \Pi_{provider}^{PCF}}{\partial e_0} = \bar{p} - p + (m^F - zQ(1)) \left(1 - \frac{\mu}{\bar{\mu}}\right) + c(e_0) - \bar{c}(e_0) - \beta(z + p_H) q(e_0) = 0 \tag{B.10}$$

$$\varphi_1(e_1, \zeta) \equiv \frac{1}{\lambda} \frac{\partial \Pi_{provider}^{PCF}}{\partial e_1} = p - m^F + zQ(1) + r - c(e_1) + t - (z + \alpha p_H) q(e_1) = 0. \tag{B.11}$$

The first-order conditions Eqs. (B.10) and (B.11) are separable in e_0 and e_1.

We note that for any contract ζ, the optimal $e_1 < 1$. By assumption, we have $t + zQ(1) < zq(1)$. Since $r + p - c(x) < m^F \ \forall x$, it follows that $p - m^F + r - c(1) + t + zQ(1) < zq(1) < (z + \alpha p_H) q(1)$.

With qualification: Consider the inequality $\varphi_1(0,\zeta^q) \le 0$, *i.e.*,

$$p - m^F + zQ(1) + r^q - c(0) + t \le 0. \tag{B.12}$$

Solving Eq. (B.11) leads to

$$e_1 = \begin{cases} 0 & \text{if (B.12) holds} \\ \bar{e}_1^q & \text{else,} \end{cases}$$

where $\bar{e}_1^q \in (0,1)$ is the unique solution of the equation $\varphi_1(e_1,\zeta^q) = 0$.

By concavity of the objective with respect to e_0, $\varphi_0(e_0,\zeta^q)$ is decreasing in e_0. If

$$\varphi_0(e_0^{min},\zeta^q) = \bar{p} - p + (m^F - zQ(1))\left(1 - \frac{\mu}{\bar{\mu}}\right) + c(e_0^{min}) - \bar{c}(e_0^{min}) - \beta(z + p_H^q)q(e_0^{min}) < 0,$$

then solving the FOC on $[e_0^{min},1]$ leads to $e_0 = e_0^{min}$. If

$$\varphi_0(1,\zeta^q) = \bar{p} - p + (m^F - zQ(1))\left(1 - \frac{\mu}{\bar{\mu}}\right) + c(1) - \bar{c}(1) - \beta(z + p_H^q)q(1) \ge 0,$$

then solving the FOC on $[e_0^{min},1]$ leads to $e_0 = 1$. Else, the solution of Eq. (B.10) lies within $(e_0^{min},1)$. We denote this solution as $\bar{e}_0^q$, defined as the unique solution to $\varphi_0(e_0,\zeta^q) = 0$.

If the obtained thresholds (e_0,e_1) satisfy the inequalities $e_0 \le e_1$, they are the optimal solutions. We observe that this may only happen when $e_1 = \bar{e}_1^q$, $e_0 = \max\{\bar{e}_0^q, e_0^{min}\}$, and $\max\{\bar{e}_0^q, e_0^{min}\} \le \bar{e}_1^q$. The condition $\bar{e}_0^q \le \bar{e}_1^q$ is equivalent to $\varphi_1(\bar{e}_0^q,\zeta^q) \ge 0$, which can be written as

$$p - m^F + zQ(1) + r^q - c(\bar{e}_0^q) + t - (z + \alpha p_H^q)q(\bar{e}_0^q) \ge 0.$$

The condition $e_0^{min} \le \bar{e}_1^q$ is equivalent to $\varphi_1(e_0^{min},\zeta^q) \ge 0$, which can be written as

$$p - m^F + zQ(1) + r^q - c(e_0^{min}) + t - (z + \alpha p_H^q)q(e_0^{min}) \ge 0.$$

Otherwise, we have $e_0 = e_1 \equiv e$ at the optimal solution. The objective becomes

$$\begin{aligned} \Pi_{provider}^{PCF} =& R^q + (\bar{p} + r^q)\lambda e - \lambda \int_0^e \bar{c}(x)dx + (m^F - zQ(1))\lambda_N(e,e) - \lambda t(1-e) - \lambda(1+\beta)zQ(e) \\ & - p_H^q \lambda\left[(1-\alpha)Q(1) + (\alpha+\beta)Q(e)\right], \end{aligned}$$

which is a concave function of e. The first-order condition is

$$\psi(e,\zeta^q) \equiv \ \bar{p} + r^q - \bar{c}(e) - (m^F - zQ(1))\frac{\mu}{\bar{\mu}} + t - [(\alpha+\beta)p_H^q + (1+\beta)z]q(e) = 0.$$

Note $\psi(e,\zeta^q)$ is decreasing. Hence, the solution of the FOC lies within $(e_0^{min},1)$ iff $\psi(e_0^{min},\zeta^q) > 0$ and $\psi(1,\zeta^q) < 0$. If

$$\psi(e_0^{min},\zeta^q) = \bar{p} + r^q - \bar{c}(e_0^{min}) - (m^F - zQ(1))\frac{\mu}{\bar{\mu}} + t - [(\alpha+\beta)p_H^q + (1+\beta)z]q(e_0^{min}) \le 0,$$

then solving the FOC on $[e_0^{min},1]$ leads to $e = e_0^{min}$. If

$$\psi(1,\zeta^q) = \bar{p} + r^q - \bar{c}(1) - (m^F - zQ(1))\frac{\mu}{\bar{\mu}} + t - [(\alpha+\beta)p_H^q + (1+\beta)z]q(1) \ge 0,$$

then solving the FOC on $[e_0^{min},1]$ leads to $e = 1$. Else, the FOC has a solution on $(e_0^{min},1)$; we denote this solution as $\bar{e}^q$, defined as the unique solution to $\psi(e,\zeta^q) = 0$.

Without qualification: Consider the inequality $\hat{\varphi}_1(0,\zeta^{nq}) \le 0$, *i.e.*,

$$p - m^F + zQ(1) + r^{nq} - c(0) + t \le 0. \tag{B.13}$$

Similar to the case with qualification, solving Eq. (B.11) leads to

$$e_1 = \begin{cases} 0 & \text{if (B.13) holds} \\ \bar{e}_1^{nq} & \text{else,} \end{cases}$$

where $\bar{e}_1^{nq} \in (0,1)$ is the unique solution of the equation $\varphi_1(e_1, \zeta^{nq}) = 0$.

By concavity of the objective with respect to e_0, $\varphi_0(e_0, \zeta^{nq})$ is decreasing in e_0. If

$$\varphi_0(0, \zeta^{nq}) = \bar{p} - p + (m^F - zQ(1)) \left(1 - \frac{\mu}{\bar{\mu}}\right) + c(0) - \bar{c}(0) < 0,$$

then solving the FOC on $[0, e_0^{min})$ leads to $e_0 = 0$. If

$$\varphi_0(e_0^{min}, \zeta^{nq}) = \bar{p} - p + (m^F - zQ(1)) \left(1 - \frac{\mu}{\bar{\mu}}\right) + c(e_0^{min}) - \bar{c}(e_0^{min}) - \beta z q(e_0^{min}) > 0,$$

then solving the FOC on $[0, e_0^{min})$ leads to $e_0 = e_0^{min} - \epsilon$. Else, the solution of Eq. (B.10) lies within $(0, e_0^{min})$. We denote this solution as $\bar{e}_0^{nq}$, defined as the unique solution to $\varphi_0(e_0, \zeta^{nq}) = 0$. If the obtained thresholds (e_0, e_1) satisfy the inequalities $e_0 \le e_1$, they are the optimal solutions. We observe that this requires $e_1 = \bar{e}_1^{nq}$. The condition $e_0^{min} \le \bar{e}_1^{nq}$ is equivalent to $\varphi_1(e_0^{min}, \zeta^{nq}) \ge 0$, which can be written

$$p - m^F + zQ(1) + r^{nq} - c(e_0^{min}) + t - zq(e_0^{min}) \ge 0.$$

The condition $\bar{e}_0^{nq} \le \bar{e}_1^{nq}$ is equivalent to $\varphi_1(\bar{e}_0^{nq}, \zeta^{nq}) \ge 0$, which can be written

$$p - m^F + zQ(1) + r^{nq} - c(\bar{e}_0^{nq}) + t - zq(\hat{e}_0) \ge 0.$$

Otherwise, we have $e_0 = e_1 \equiv e$ at the optimal solution. The objective becomes

$$\Pi^{PCF}_{provider} = R^{nq} + (\bar{p} + r^{nq})\lambda e - \lambda \int_0^e \bar{c}(x)dx + (m^F - zQ(1))\lambda_N(e,e) - \lambda t(1-e) - \lambda(1+\beta)zQ(e),$$

which is a concave function of e. The first-order condition is

$$\psi(e, \zeta^{nq}) \equiv \ \bar{p} + r^{nq} - \bar{c}(e) - (m^F - zQ(1))\frac{\mu}{\bar{\mu}} + t - (1+\beta)zq(e) = 0.$$

Note $\psi(e, \zeta^{nq})$ is decreasing. Hence, the solution of the FOC lies within $(0, e_0^{min})$ iff $\psi(0, \zeta^{nq}) > 0$ and $\psi(e_0^{min}, \zeta^{nq}) < 0$. If

$$\psi(0, \zeta^{nq}) = \bar{p} + r^{nq} - \bar{c}(0) - (m^F - zQ(1))\frac{\mu}{\bar{\mu}} + t \le 0,$$

then solving the FOC on $[0, e_0^{min})$ leads to $e = 0$. If

$$\psi(e_0^{min}, \zeta^{nq}) = \bar{p} + r^{nq} - \bar{c}(e_0^{min}) - (m^F - zQ(1))\frac{\mu}{\bar{\mu}} + t - (1+\beta)zq(e_0^{min}) \ge 0,$$

then solving the FOC on $[0, e_0^{min})$ leads to $e = e_0^{min} - \epsilon$. Else, the FOC has a solution on $(0, e_0^{min})$; we denote this solution as $\bar{e}^{nq}$, defined as the unique solution to $\psi(e, \zeta^{nq}) = 0$.

Finally, the provider decides between the scenario with qualification and the scenario without qualification by comparing her objective at the optimal solutions, *i.e.*, the provider selects to qualify iff

$$\Pi^{PCF*}_{provider}(\zeta^q) \ge \Pi^{PCF*}_{provider}(\zeta^{nq}),$$

where $\Pi^{PCF*}_{provider}(\zeta^q)$ is the provider profit under PCF in the scenario with qualification when the provider selects the thresholds optimally as detailed above, and $\Pi^{PCF*}_{provider}(\zeta^{nq})$ is the provider profit under PCF in the scenario without qualification when the provider selects the thresholds optimally as detailed above.

Region $\mathcal{R}$ is the set of problem parameters such that the optimal thresholds are $0 < \max\{e_0^{min}, \bar{e}_0^q\} < \bar{e}_1^q < 1$, so that the three modes of care exist and the Quality Gateway qualification criterion is met. It is defined by the system of inequalities

$$\varphi_1(0, \zeta^q) > 0$$

$$\varphi_0(1, \zeta^q) < 0$$

$$\varphi_1(\bar{e}_0^q, \zeta^q) > 0$$

$$\varphi_1(e_0^{min}, \zeta^q) > 0$$

$$\Pi^{PCF*}_{provider}(\zeta^q) \geq \Pi^{PCF*}_{provider}.(\zeta^{nq})$$

Note that in the above system, the first inequality is unnecessary as the third inequality implies the first.

To obtain the optimal provider decisions, we propose the following algorithm.

Step 1: Find the optimal thresholds with qualification:

(i) Determine e_1: If $\varphi_1(0, \zeta^q) \leq 0$, set $e_1 = 0$. Else, solve $\varphi_1(\bar{e}_1^q, \zeta^q) = 0$ and set $e_1 = \bar{e}_1^q$.

(ii) Determine e_0: If $\varphi_0(e_0^{min}, \zeta^q) \leq 0$, set $e_0 = e_0^{min}$. Else, if $\varphi_0(1, \zeta^q) \geq 0$, set $e_0 = 1$. Else, solve $\varphi_0(\bar{e}_0^q, \zeta^q) = 0$ and set $e_0 = \bar{e}_0^q$.

(iii) If $e_0 \leq e_1$, go to Step (iv). Else, determine e: If $\psi(e_0^{min}, \zeta^q) \leq 0$, set $e = e_0^{min}$. Else, if $\psi(1, \zeta^q) \geq 0$, set $e = 1$. Else, solve $\psi(\bar{e}^q, \zeta^q) = 0$ and set $e = \bar{e}^q$. Set $e_0 = e_1 = e$.

(iv) Evaluate $\Pi^{PCF*}_{provider}(\zeta^q)$ as the provider profit at (e_0, e_1).

Step 2: Find the optimal thresholds without qualification:

(i) Determine e_1: If $\varphi_1(0, \zeta^{nq}) \leq 0$, set $e_1 = 0$. Else, solve $\varphi_1(\bar{e}_1^{nq}, \zeta^{nq}) = 0$ and set $e_1 = \bar{e}_1^{nq}$.

(ii) Determine e_0: If $\varphi_0(0, \zeta^{nq}) \leq 0$, set $e_0 = 0$. Else, if $\varphi_0(e_0^{min}, \zeta^{nq}) \geq 0$, set $e_0 = e_0^{min} - \epsilon$. Else, solve $\varphi_0(\hat{e}_0, \zeta^{nq}) = 0$ and set $e_0 = \bar{e}_0^{nq}$.

(iii) If $e_0 \leq e_1$, go to Step (iv). Else, determine e: If $\psi(0, \zeta^{nq}) \leq 0$, set $e = 0$. Else, if $\psi(e_0^{min}, \zeta^{nq}) \geq 0$, set $e = e_0^{min} - \epsilon$. Else, solve $\psi(\bar{e}^{nq}, \zeta^{nq}) = 0$ and set $e = \bar{e}^{nq}$. Set $e_0 = e_1 = e$.

(iv) Evaluate $\Pi^{PCF*}_{provider}(\zeta^{nq})$ as the provider profit at (e_0, e_1).

Step 3: The optimal solution is the one found in Step 1 iff $\Pi^{PCF*}_{provider}(\zeta^q) \geq \Pi^{PCF*}_{provider}(\zeta^{nq})$, otherwise it is the solution found in Step 2.

Proof of Lemma 1

We have $e_1^F = q^{-1}(t/z + Q(1))$. Moreover, we know $r + p - c(e_1^F) < f(e_1^F) + p - c(e_1^F) = m^F$. Based on the proof of Proposition 2, e_1^P satisfies $\varphi_1(e_1^P, \zeta) = 0$ where $\varphi_1(\cdot, \zeta)$ is decreasing. We thus obtain

$$\begin{aligned}\varphi_1(e_1^F, \zeta) =& r - c(e_1^F) + p - m^F + zQ(1) + t - (z + \alpha p_H)(t/z + Q(1)) = r - c(e_1^F) + p - m^F - \alpha p_H(t/z + Q(1)) \\ &< -\alpha p_H(t/z + Q(1)) < 0.\end{aligned}$$

It follows that $e_1^F > e_1^P$.

Proof of Proposition 3

Utilization: Since $e_0^F = 0 \leq e_0^P$, utilization of in-person services (face-to-face and referral) is higher under FFS than under PCF.

Failure rate: The failure rate under PCF is given by $\beta Q(e_0^P) + \alpha Q(e_1^P) + (1 - \alpha)Q(1)$. The failure rate under FFS is given by $\beta Q(e_0^F) + \alpha Q(e_1^F) + (1 - \alpha)Q(1)$. After simplifications, we obtain that the failure rate is lower under PCF iff $\Delta Q > 0$.

Wait time: Follows from noticing that $W(e_0)$ is decreasing in e_0 and that $e_0^F = 0 \leq e_0^P$.

Panel size: The panel size is given by $\lambda+\lambda_N(e_0,e_1)$ divided by the average number of visits per patient per year. Hence, the panel size is larger iff $\lambda_N(e_0,e_1)$ is larger. Using Eq. (1) and $e_0^F=0$, the panel size is larger under PCF iff $\rho-e_0^P\frac{\lambda}{\bar{\mu}}-(e_1^P-e_0^P)\frac{\lambda}{\mu}>\rho-e_1^F\frac{\lambda}{\mu}$, *i.e.*,

$$e_0^P\left(1-\frac{\mu}{\bar{\mu}}\right)>e_1^P-e_1^F.$$

A sufficient condition for the above inequality to hold is that $e_1^P<e_1^F$, which holds in the region $\mathcal{R}$.

Patient benefit: Patients are better off under PCF than FFS iff $\Pi_{patient}(e_0^F,e_1^F)\leq\Pi_{patient}(e_0^P,e_1^P)$ (Eq. (6)), namely

$$\begin{aligned}&-v_H[\alpha Q(e_1^F)+(1-\alpha)Q(1)]-v_W W(0)-pe_1^F-\tilde{p}(1-e_1^F)\leq\\&ue_0^P-v_H[\beta Q(e_0^P)+\alpha Q(e_1^P)+(1-\alpha)Q(1)]-v_W W(e_0^P)-\bar{p}e_0^P-p(e_1^P-e_0^P)-\tilde{p}(1-e_1^P),\end{aligned}$$

that is, $-v_H\Delta Q+(\tilde{p}-p)(e_1^F-e_1^P)\leq(u+p-\bar{p})e_0^P+v_W(W(0)-W(e_0^P))$.

Proof of Proposition 4

Payer reimbursement expenditures: The payer reimbursement expenditures are lower under PCF than FFS iff

$$\lambda\left(-\int_0^{e_1^F}f(x)dx-\int_{e_1^F}^1\tilde{f}(x)dx\right)<-R-\lambda\left(re_1^P+\int_{e_1^P}^1\tilde{f}(x)dx\right))+\lambda p_H[\beta Q(e_0^P)+\alpha Q(e_1^P)+(1-\alpha)Q(1)],$$

that is,

$$\frac{R}{\lambda}-\int_0^{e_1^F}f(x)dx+re_1^P-p_H[\beta Q(e_0^P)+\alpha Q(e_1^P)+(1-\alpha)Q(1)]<\int_{e_1^F}^{e_1^P}\tilde{f}(x)dx. \tag{B.14}$$

Since we set R to make the provider indifferent between PCF and FFS for the Medicare patient population, we have

$$\begin{aligned}\frac{R}{\lambda}=&-re_1^P+\int_0^{e_0^P}\bar{c}(x)dx+\int_{e_0^P}^{e_1^P}c(x)dx+m^Fe_1^F+t(e_1^F-e_1^P)-z[Q(e_1^F)-\beta Q(e_0^P)-Q(e_1^P)]\\&+p_H[\beta Q(e_0^P)+\alpha Q(e_1^P)+(1-\alpha)Q(1)]-\bar{p}e_0^P-p(e_1^P-e_0^P).\end{aligned} \tag{B.15}$$

Hence, inequality (B.14) can be re-written equivalently as

$$-\int_0^{e_1^F}c(x)dx+\int_0^{e_0^P}\bar{c}(x)dx+\int_{e_0^P}^{e_1^P}c(x)dx-t(e_1^P-e_1^F))-z[Q(e_1^F)-\beta Q(e_0^P)-Q(e_1^P)]+p(e_1^F-e_1^P+e_0^P)-\bar{p}e_0^P<\int_{e_1^F}^{e_1^P}\tilde{f}(x)dx,$$

that is, $\Delta X>0$.

Payer benefit: The payer is better off under PCF than FFS iff

$$\begin{aligned}&\lambda\left(-\int_0^{e_1^F}f(x)dx-\int_{e_1^F}^1\tilde{f}(x)dx-w\left[\alpha Q(e_1^F)+(1-\alpha)Q(1)\right]\right)<\\&-R-\lambda\left(re_1^P+\int_{e_1^P}^1\tilde{f}(x)dx\right)+\lambda(p_H-w)[\beta Q(e_0^P)+\alpha Q(e_1^P)+(1-\alpha)Q(1)].\end{aligned}$$

Using Eq. (B.15) and the definition of ΔX, we can rewrite the above inequality as $-(\Delta Q)w\leq\Delta X$. Since the thresholds e_0^P,e_1^P,e_1^F and function $Q(\cdot)$ are independent of w, so are ΔQ and ΔX, and thus the inequality $-(\Delta Q)w\leq\Delta X$ implies that, as long as $\Delta X>0$, the payer is better off under PCF whenever either $\Delta Q>0$, or $\Delta Q<0$ and the penalty w is less than $\Delta X/(-\Delta Q)$. Conversely, for $\Delta X<0$, the payer is better off under PCF whenever $\Delta Q>0$ and the penalty w is more than $(-\Delta X)/\Delta Q$. The table below summarizes which system the payer prefers:

	$\Delta X < 0$	$\Delta X > 0$
$\Delta Q < 0$	FFS	PCF iff w low enough
$\Delta Q > 0$	PCF iff w high enough	PCF

Proof of Proposition 5

We first show that Π_{social} is concave in e_0 and e_1. We compute the first and second-order partial derivatives of Π_{social} with respect to e_0 and e_1. Namely,

$$\varphi_0^S(e_0) \equiv \frac{\partial \Pi_{social}}{\partial e_0} = \lambda\left((m^F - zQ(1))(1-\mu/\bar{\mu}) + u - v_W W'(e_0) - \bar{c}(e_0) + c(e_0) - (v_H + w + z)\beta q(e_0)\right), \text{ and}$$
$$\varphi_0^{S\prime}(e_0) = \lambda(-v_W W''(e_0) - \bar{c}'(e_0) + c'(e_0) - (v_H + w + z)\beta q'(e_0)) \leq 0,$$

where the last inequality follows the assumption that $c(\cdot) - \bar{c}(\cdot)$ is non-increasing, $q(\cdot)$ is increasing, and the result that $W(\cdot)$ is convex-decreasing. Similarly, we compute

$$\varphi_1^S(e_1) \equiv \frac{\partial \Pi_{social}}{\partial e_1} = \lambda\left(-(m^F - zQ(1)) + t + \tilde{c}(e_1) - c(e_1) - (\alpha(v_H + w) + z)q(e_1)\right), \text{ and}$$
$$\varphi_1^{S\prime}(e_1) = -\lambda(c'(e_1) - \tilde{c}'(e_1) + (\alpha(v_H + w) + z)q'(e_1)) < 0.$$

The last inequality follows from $\tilde{c}(\cdot) - c(\cdot)$ being non-increasing and $q(\cdot)$ increasing.

Since there is no cross-term in the partial derivatives, we conclude that Π_{social} is jointly concave in e_0 and e_1. To characterize the optimal solution let us first ignore the constraint $e_0 \leq e_1$. We know $\varphi_0^S(e_0)$ is decreasing in e_0 and independent of e_1, and $\varphi_0^S(0) = \lambda\left((m^F - zQ(1))\left(1 - \frac{\mu}{\bar{\mu}}\right) + u + v_W \frac{\lambda(\bar{\mu}-\mu)}{(1-\rho)\rho\bar{\mu}\mu^2} - \bar{c}(0) + c(0)\right) > 0$, then it must be that $e_0^S = 0$. Then, if $\varphi_0^S(1) = \lambda\left((m^F - zQ(1))\left(1 - \frac{\mu}{\bar{\mu}}\right) + u + v_W \frac{\lambda\bar{\mu}(\bar{\mu}-\mu)\bar{\mu}\rho}{(1-\rho)(\lambda(\bar{\mu}-\mu)+\mu\bar{\mu}\rho)^2} - \bar{c}(1) + c(1) - (v_H + w + z)\beta q(1)\right) \geq 0$, then $e_0^S = 1$. If $\varphi_0^S(1) < 0$, then the optimal solution is $\bar{e}_0^S \in (0,1)$, where $\bar{e}_0^S$ is the unique solution to the first-order condition $\varphi_0^S(e_0) = 0$.

To determine the optimal e_1^S, we note that $\varphi_1^S(e_1)$ is decreasing in e_1 and independent of e_0. Thus, $\varphi_1^S(0) = \lambda(-(m^F - zQ(1)) + t + \tilde{c}(0) - c(0)) > 0$, then $e_1^S > 0$. If $\varphi_1^S(1) = \lambda(-(m^F - zQ(1)) + t + \tilde{c}(1) - c(1) - (\alpha(v + w) + z)q(1)) \geq 0$, then $e_1^S = 1$. If $\varphi_1^S(1) < 0$, then the optimal solution is $\bar{e}_1^S \in (0,1)$, where $\bar{e}_1^S$ is the unique solution to the first-order condition $\varphi_1^S(e_1) = 0$.

If $\varphi_0^S(1) < 0$ and $\varphi_1^S(1) \geq 0$, then the solution $(\bar{e}_0^S, 1)$ satisfies the constraint $e_0 \leq e_1$. If $\varphi_0^S(1) < 0$ and $\varphi_1^S(1) < 0$, and $\bar{e}_0^S < \bar{e}_1^S$, then the solution $(\bar{e}_0^S, \bar{e}_1^S)$ satisfies the constraint $e_0 \leq e_1$. Note that if $\bar{e}_0^S \geq \bar{e}_1^S$, then the constraint $e_0 \leq e_1$ is binding. We rewrite the social welfare assuming only two regions of care exist (*i.e.*, assuming $e_0 = e_1$). The resulting welfare function is concave and the first-order condition is

$$\psi^S(e) \equiv \frac{\partial \Pi_{social}(e,e)}{\partial e} = \lambda\left(-(m^F - zQ(1))\frac{\mu}{\bar{\mu}} + u + t + \tilde{c}(e) - \bar{c}(e) - v_W W'(e) - ((\alpha+\beta)(v+w) + (1+\beta)z)q(e)\right).$$

Because of concavity, ψ^S is decreasing and $\psi^S(0) = \lambda\left(-(m^F - zQ(1))\frac{\mu}{\bar{\mu}} + u + t + \tilde{c}(0) - \bar{c}(0) - v_W W'(0)\right) \geq 0$, then $e^S > 0$. If $\psi^S(1) = \lambda\left(-(m^F - zQ(1))\frac{\mu}{\bar{\mu}} + u + t + \tilde{c}(1) - \bar{c}(1) - v_W W'(1) - ((\alpha+\beta)(v+w) + (1+\beta)z)q(1)\right) < 0$ the optimal solution is $\bar{e}^S \in (0,1)$, which is the unique solution to the first-order condition $\psi^S(e) = 0$. In the case $\psi^S(1) > 0$, the optimal solution is $e_0^S = e_1^S = 1$.

We thus obtain the results in Table B2.

Proof of Proposition 6

Table B2 Social optimum

	Care modes	Optimal solution	Conditions
$\mathcal{R}^S$	Remote, face-to-face & referral	$(\bar{e}_0^S, \bar{e}_1^S)$	$\varphi_0^S(1)<0$ and $\varphi_1^S(1)<0$, and $\bar{e}_0^S < \bar{e}_1^S$
	Remote & referral	$(\bar{e}^S, \bar{e}^S)$	$\varphi_0^S(1)<0$ and $\varphi_1^S(1)<0$, and $\bar{e}_0^S \geq \bar{e}_1^S$ and $\psi^S(1)<0$
	Remote	$(1,1)$	$\varphi_0^S(1)\geq 0$ and $\varphi_1^S(1)\geq 0$, or
			$\varphi_0^S(1)\geq 0$ and $\varphi_1^S(1)<0$, and $\psi^S(1)\geq 0$
	Remote & face-to-face	$(\bar{e}_0^S, 1)$	$\varphi_0^S(1)<0$ and $\varphi_1^S(1)\geq 0$

To show the existence of a coordinating contract in region $\mathcal{R}^S$ (when the three modes of care co-exist) of the social optimum, we show that it is possible to find contract parameters (R, r, p_H) such that the provider solution described in Proposition 2 coincides with the first-best.

Consider the region $\mathcal{R}^S$ of the social optimum $(e_0^S, e_1^S) = (\bar{e}_0^S, \bar{e}_1^S)$, such that

$$q(e_0^S) = \frac{u + v_W|W'(e_0^S)| + c(e_0^S) - \bar{c}(e_0^S) + (m^F - zQ(1))(1-\mu/\bar{\mu})}{\beta(v_H + w + z)},$$
$$q(e_1^S) = \frac{t + \tilde{c}(e_1^S) - c(e_1^S) - (m^F - zQ(1))}{\alpha(v_H + w) + z}.$$

We characterize the coordinating contract by imposing that the social solution $(\bar{e}_0^S, \bar{e}_1^S)$ satisfies the provider's problem FOC (B.10) and (B.11). The contract has three levers to achieve coordination, namely, the wait time Quality Gateway criterion $\bar{W}$, the visit fee r^q, and the penalty for failed treatment p_H^q. Note that the capitation payment of R does not affect coordination. We also focus on the contract parameters with qualification only.

Case 1: $\bar{W} = W(e_0^S)$. In order to coordinate the care threshold $e_1 = e_1^S$, we must have $\varphi_1(e_1^S, \zeta^q) = 0$ (Eq. B.11), *i.e.*, the penalty must satisfy

$$p_H^q(r^q) := \frac{r^q + p + t - c(e_1^S) - (m^F - zQ(1)) - zq(e_1^S)}{\alpha q(e_1^S)}. \tag{B.16}$$

Note that for p_H^q to be non-negative, the visit fee r^q must satisfy

$$r^q \geq c(e_1^S) + m^F - zQ(1) + zq(e_1^S) - p - t \equiv \underline{r}_1. \tag{B.17}$$

To coordinate the lower threshold $e_0 = e_0^S$, we impose $\varphi_1(e_0^S, \zeta^q) \leq 0$ (Eq. B.10), so it is optimal for the provider to adopt the least amount of remote care while still satisfying the Quality Gateway. Using Eq. (B.16), we obtain the following lower bound in r^q,

$$r^q \geq (\bar{p} + c(e_0^S) - \bar{c}(e_0^S))\frac{\alpha q(e_1^S)}{\beta q(e_0^S)} + (m^F - zQ(1))\Big((1-\mu/\bar{\mu})\frac{\alpha q(e_1^S)}{\beta q(e_0^S)} + 1\Big) - p\Big(1 + \frac{\alpha q(e_1^S)}{\beta q(e_0^S)}\Big) - t + c(e_1^S) + z(1-\alpha)q(e_1^S) \equiv \underline{r}_2. \tag{B.18}$$

Thus, define $\underline{r} \equiv \max\{0, \underline{r}_1, \underline{r}_2\}$, there exists a family of coordinating contracts with $\bar{W} = W(e_0^S)$, $r^q \in [\bar{r}, f(0)]$, and $p_H^q(r^q)$ as defined in Eq. (B.16).

Case 2: $\bar{W} > W(e_0^S)$. In this case coordination requires to set both r^q and p_H^q to specific values to ensure the FOCs $\varphi_0(e_0^S, \zeta^q) = 0$ (Eq. B.10) and $\varphi_1(e_1^S, \zeta^q) = 0$ (Eq. B.11) are satisfied. Simple algebra leads to,

$$p_H^q := v_H + w - \frac{(p - \bar{p} + u + v_W|W'(e_0^S)|)}{\beta q(e_0^S)}, \tag{B.19}$$

$$r^q := \tilde{c}(e_1^S) - p - \frac{\alpha q(e_1^S)}{\beta q(e_0^S)}(p - \bar{p} + u + v_W|W'(e_0^S)|). \tag{B.20}$$

A coordinating contract exists whenever $0 \leq r^q \leq f(0)$, which corresponds to

$$p \leq \tilde{c}(e_1^S) - \frac{\alpha q(e_1^S)}{\beta q(e_0^S)}(p - \bar{p} + u + v_W|W'(e_0^S)|) \leq m^F + c(0),$$

and the penalty $p_H^q \geq 0$, that is, whenever

$$(v_H + w)\left(c(e_0^S) - p - (\bar{c}(e_0^S) - \bar{p}) + (m^F - zQ(1))(1 - \mu/\bar{\mu})\right) - z \geq 0.$$

There exists a family of coordinating contract with $\bar{W} > W(e_0^S)$, r^q as in Eq. (B.20), and p_H^q as in Eq. (B.19).

Appendix C: Parameter estimation

The summary of the parameters for all states is included in Table C3.

Table C3 Summary of parameters baseline values

Parameter	AR	CO	HI	KC	MI	MT	NJ	NY	OH	OK	OR	PA	RI	TN
Visits rate per year (λ)	1,638	1,139	634	1,979	1,055	1,980	1,096	1,126	1,091	1,477	1,276	1,011	1,238	1,515
FFS margin (m^F)	74	76	80	71	74	77	82	80	72	71	74	73	78	70
`|FFS_rate|`	71	73	76	69	72	74	78	76	70	69	71	71	75	68
Failure rate parameter (δ)	0.078	0.071	0.059	0.088	0.096	0.069	0.096	0.094	0.097	0.093	0.068	0.094	0.092	0.094
Patient utility (u)	3.658	4.754	4.095	10.263	7.980	27.977	4.147	12.289	4.904	9.310	20.632	15.263	3.866	5.110
Patient failure disutility (v_H)	2221	2257	2,490	2,168	2,297	2,157	2,579	2,625	2,254	2,224	2,299	2,327	2,374	2,197
Patient wait time disutility (v_W) per day	450	288	146	601	372	463	430	443	380	487	316	353	428	497
Payer failure cost (w)	11,700	11,700	11,700	11,700	11,700	11,700	11,700	11,700	11,700	11,700	11,700	11,700	11,700	11,700
Provider coordination cost (t)	50	50	50	50	50	50	50	50	50	50	50	50	50	50
Provider reputation cost (z)	30	30	30	30	30	30	30	30	30	30	30	30	30	30
Referral care quality gain (α)	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
Remote care quality loss (β)	30%	30%	30%	30%	30%	30%	30%	30%	30%	30%	30%	30%	30%	30%
FFS specialist care ($\tilde{f}(x) := \tilde{f}$)	224	237	222	215	245	223	245	267	272	224	242	245	230	225
Cost specialist care ($\tilde{c}(x) := \tilde{c}$)	182	186	191	178	184	189	196	193	180	178	182	181	190	177
Co-payment remote visit ($\bar{p}$)	18	18	18	18	18	18	18	18	18	18	18	18	18	18
Co-payment in-office visit (p)	20	20	20	20	20	20	20	20	20	20	20	20	20	20
Co-payment specialist visit ($\tilde{p}$)	40	40	40	40	40	40	40	40	40	40	40	40	40	40
Utilization (ρ)	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
Service rate in-office visit (μ) visits/year	2,048	1,424	792	2,473	1,319	2,475	1,370	1,408	1,364	1,846	1,595	1,264	1,547	1,894
Service rate remote visit ($\bar{\mu}$) visits/year	3,072	2,135	1,188	3,710	1,978	3,712	2,055	2,112	2,045	2,769	2392	1,896	2,321	2,840
Baseline PCF payment														
PBP per beneficiary per month	$\{28, 2\times28, 3\times28\}$													
FVF per visit	40.82													
PBA health (% of revenue)	$[-10\%, 34\%]$													

Note: PBP: Population-Based Payment, PBA: Performance-Based Adjustment, and FVF: Flat Fee per Visit. Sources: see Appendix C for details.

We next describe the details of how we calibrated the values in Table C3.

FFS rate: We estimate the state-specific average FFS rates according to the CMS fee schedule for specific CPT codes associated with in-office visits. We consider the average non-facility rates (services provided in the office) of the CPT codes related to office/outpatient visits for existing patients $\{99213 - 99215\}$. Table C4 shows the specific payment for each CPT code in primary care. The rate heterogeneity is a reflection of patient complexity as captured by the duration of the appointment. We use these rates to estimate the model FFS payment $f(x) = f_0 + \text{FFS\_rate}\ x$, where $x \in [0,1]$, for each state. We estimate $f_0 = \min_{i \in CPT} r_i$, where r_i is the average payment rate for CPT i in Table C4. The slope of the FFS rate is estimated as $\text{FFS\_rate} = \max_{i \in CPT} r_i - \min_{i \in CPT} r_i$.

For the specialist rate, we assume a constant rate independently of the patient complexity, *i.e.*, $\tilde{f}(x) := \tilde{f}$. To determine the specialist rate $\tilde{f}$, we use the Medicare Provider Utilization and Payment Data, Physician and Other Supplier PUF CY2015. We consider the same set of CPT codes as above but for all specialist visits, not just primary care. We then conservatively estimate the state-specific $\tilde{f}$ as the maximum payment rate among all specialists and CPT codes. The specialist rate is $\tilde{f}_{NJ} = \$245$ for NJ and $\tilde{f}_{OR} = \$242$ for OR. The list of considered specialties is available in Table C5.

Table C4 CPT codes and physician rates used for computing the in-office FFS rate.

State	99213 Low complexity	99214 Medium complexity	99215 High complexity
AR	73.8	108.2	144.9
CO	76.1	111.4	149.2
HI	79.9	116.6	155.4
KS	71.0	104.2	139.8
MI	74.5	109.1	146.6
MT	76.9	112.6	151.2
NJ	82.0	119.8	159.9
NY	79.7	116.5	156.1
OH	72.2	105.9	142.3
OK	71.1	104.4	140.3
OR	73.5	107.8	144.4
PA	73.0	107.0	143.7
RI	78.2	114.4	153.1
TN	70.4	103.4	138.6

Sources: CMS fee schedule FY2019 https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PFSlookup (Accessed 1-22-2022).

Table C5 List of medical specialties considered for estimating $\tilde{f}$.

Addiction Medicine	Hand Surgery	Osteopathic Manipulative Medicine
Allergy/Immunology	Hematology	Otolaryngology
Cardiac Electrophysiology	Hematology/Oncology	Pain Management
Cardiac Surgery	Hospice and Palliative Care	Pathology
Cardiology	Infectious Disease	Peripheral Vascular Disease
Chiropractic	Interventional Cardiology	Physical Medicine and Rehabilitation
Clinical Psychologist	Interventional Pain Management	Physical Therapist
Colorectal Surgery (formerly proctology)	Interventional Radiology	Plastic and Reconstructive Surgery
Critical Care (Intensivists)	Maxillofacial Surgery	Podiatry
CRNA	Medical Oncology	Psychiatry
Dermatology	Nephrology	Pulmonary Disease
Diagnostic Radiology	Neurology	Radiation Oncology
Emergency Medicine	Neuropsychiatry	Rheumatology
Endocrinology	Neurosurgery	Sleep Medicine
Gastroenterology	Nuclear Medicine	Sports Medicine
General Practice	Obstetrics/Gynecology	Surgical Oncology
General Surgery	Ophthalmology	Thoracic Surgery
Geriatric Medicine	Optometry	Urology
Geriatric Psychiatry	Oral Surgery (dentists only)	Vascular Surgery
Gynecological/Oncology	Orthopedic Surgery	

Source: Medicare Provider Utilization and Payment Data: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier (Accessed 1-20-2022)

Failure rate: We estimate the failure probability using the utilization of inpatient services as a proxy. We consider the following functional form for the treatment failure probability: $q(x) = \exp(\delta x) - 1$. Note that the lowest-complexity patients do not experience treatment failure, *i.e.*, $q(0) = 0$. To estimate the parameter δ for each state, we compute the proportion of non-elective inpatient encounters (either admitted from the ED or other hospital units) out of all the inpatient encounters (Medicare patients only). Thus, we compute the state-specific parameter δ by equating $\int_0^1 q(x)dx$ to the ratio of the proportion of Medicare non-elective inpatient services in a year by the average number of visits per year. The data used (and sources) is available in Table C6 and we estimate the average number of primary care visits in a year at 3 (Barnett et al. 2021).

We note that the maximum probability of treatment failure $q(1)$ is larger for states that have worse patient health status as captured by the proportion of the population with chronic conditions (second column in Table C6). We note that the aforementioned proportion data was not included in the calibration of the parameter δ, hence the correlation mentioned above serves as a simple validation of our calibration approach.

Table C6 **Medicare data for computing failure probability $q(x)$.**

State	% Pop. with chronic conditions	No. Non-elective IP from ED (i)	No. Non-elective IP from other units (ii)	No. All IP encounters (iii)	% Pop. had IP visit (iv)	Avg. Yearly Failure probability $((i)+(ii))/(iii)\times(iv)$	δ	$q(1)$
AR	55	196,458	35,911	279,392	0.145	0.120	0.078	0.081
CO	52	113,920	16,686	174,858	0.146	0.109	0.071	0.074
HI	52	33,907	2,234	41,266	0.103	0.090	0.059	0.061
KS	70	77,933	25,775	138,201	0.182	0.136	0.088	0.092
MI	74	236,210	52,229	354,653	0.184	0.149	0.096	0.101
MT	52	23,574	7,202	42,526	0.146	0.105	0.069	0.071
NJ	77	299,882	23,163	366,803	0.168	0.148	0.096	0.100
NY	74	707,056	87,534	923,466	0.169	0.146	0.094	0.099
OH	72	452,458	84,721	651,831	0.182	0.150	0.097	0.102
OK	75	34,908	108,079	182,279	0.184	0.144	0.093	0.098
OR	50	99,094	24,696	158,466	0.134	0.105	0.068	0.071
PA	70	530,260	86,366	738,630	0.175	0.146	0.094	0.099
RI	68	34,024	9,607	49,428	0.160	0.142	0.092	0.096
TN	76	249,971	48,671	360,309	0.176	0.145	0.094	0.099

Sources: The first column corresponds to the maximum prevalence across 21 chronic conditions listed in CMS.gov, 'Prevalence State Level: All Beneficiaries by Medicare-Medicaid Enrollment and Age', FY 2018 (https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Chronic-Conditions/CC_Main). Columns (i)-(iii) are obtained from HCUP, SID data FY2018: (i) from Table 8a, (ii) from Table 9a, and (iii) from Table 2a (https://www.hcup-us.ahrq.gov/reports/trendtables/summarytrendtables.jsp#export). Column (iv) = 'Total Original Medicare Part A Enrollees/Total Persons With Utilization' obtained from CMS.gov, 'Medicare Utilization and Payment Section', Table MDCR INPT HOSP 3. (https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier).

Patient utility: We estimate the convenience of remote care u as two times the average travel time to the nearest hospital multiplied by the state-specific cost of time. Using the data in Table C7, we estimate

$$u = 2 \times \text{Census region avg. travel time to nearest hospital} \times \frac{\text{Census region density}}{\text{State density}} \times \text{Median salary}.$$

Note that the distance to the nearest hospital used is at the census region level, thus to obtain a state-level estimate we adjust by the ratio of the population density of the census region and the state. We acknowledge that the convenience utility gained from a remote visit may encompass other hard-to-quantify benefits (*e.g.*, not leaving the house) in addition to the direct value of travel time saved. However, estimating such utility is challenging. To account for this additional convenience, we obtained results with an additional gain within $\{10, 20, 30, 40\}$ on top of the direct value of travel time saved per remote visit. Since the insights we gained did not significantly change with the value chosen, we present the results in the absence of this additional utility.

The disutility of treatment failure is estimated by assuming the patient loses several days of work for having to spend time at the hospital plus the inconvenience of having to go to the health care facility, which is assumed to be equal to u. Namely,

$$v_H = u + \text{Avg. days inpatient} \times \text{Median salary (8 hour workday)} + \text{Deductible},$$

where the deductible for inpatient care is \$1,484 (see, *e.g.*, https://www.medicare.gov/your-medicare-costs/medicare-costs-at-a-glance).

PCF contract and mapping to our contract parameters:

The baseline PCF payment has the following components (CMS 2021b):

1. Professional PBP: Practices receive a prospective, monthly PBP (paid quarterly) for each beneficiary attributed to their practice. The payment is risk-adjusted based on the practice risk score. There are four risk groups, and the payment per beneficiary per month is respectively \$28, \$45, \$100, \$175. We consider the value of \$28 per beneficiary per month as the baseline.

Table C7 Data for computing convenience utility u and disutility v_H and v_W in \$.

Census region	State	Avg. Travel time nearest hospital (min) (i)	Density census region (ppl. per sq mi) (ii)	Density state (ppl. per sq mi) (iii)	Median salary ($/hr) (iv)	Avg. Days IP (days) (v)	u ($/visit)	v_H ($/days)	v_W ($/day)
MOUNTAIN	AR	13.7	26.5	56.3	17.05	5.38	3.66	2,221	450
MOUNTAIN	CO	13.7	26.5	48.5	19.09	5.03	4.75	2,257	288
PACIFIC	HI	11.4	118.6	211.8	19.24	6.51	4.10	2,490	146
WEST NORTH CENTRAL	KS	15.8	41.0	34.9	16.57	5.08	10.26	2,168	601
EAST NORTH CENTRAL	MI	12.4	194.8	174.8	17.32	5.81	7.98	2,297	372
MOUNTAIN	MT	13.7	26.5	6.8	15.75	5.12	27.98	2,157	463
MIDDLE ATLANTIC	NJ	11.7	630.2	1195.5	20.17	6.76	4.15	2,579	430
MIDDLE ATLANTIC	NY	11.7	630.2	411.2	20.56	6.86	12.29	2,625	443
EAST NORTH CENTRAL	OH	12.4	194.8	282.3	17.19	5.56	4.90	2,254	380
WEST SOUTH CENTRAL	OK	12.3	78.0	54.7	15.93	5.73	9.31	2,224	487
PACIFIC	OR	11.4	118.6	39.9	18.26	5.44	20.63	2,299	316
MIDDLE ATLANTIC	PA	11.7	630.2	283.9	17.63	5.87	15.26	2,327	353
NEW ENGLAND	RI	13	475.6	1018.1	19.1	5.8	3.87	2,374	428
EAST SOUTH CENTRAL	TN	14.2	105.4	153.9	15.77	5.61	5.11	2,197	497

Sources: (i) and (ii) from Pew Research Center (https://www.pewresearch.org), (iii) from U.S. Census Bureau (density by state)(https://www.census.gov), and (iv) from Governing (https://www.governing.com/gov-data). (v) = 'Total Days of Care Per Person With Utilization' is obtained from CMS.gov, 'Medicare Utilization and Payment Section', Table MDCR INPT HOSP 3. (https://www.cms.gov/research-statistics-data-systems/cms-program-statistics/2018-medicare-utilization-and-payment).

2. FVF payments: Practices receive a flat Medicare payment for all face-to-face and remote primary care visits with their attributed beneficiaries. The flat fee amounts to \$40.82 per visit.

3. PBA: The performance-based adjustment (PBA) incentivizes practices to improve the quality of care while working to reduce the utilization of expensive resources (*e.g.*, hospitalizations). The adjustments are based on health outcomes and patient experience (or service quality). Depending on performance, the sum of the adjustments can vary from -10% to $+34\%$. We note that PCF has another bonus incentive for continuous improvement (up to 16% of revenue), which is not included in our paper.

4. Quality Gateway criterion ($\bar{W}$): This is a minimum service quality requirement in order to obtain the PBA. In practice, this is a composite metric including several aspects of care delivery and patient experience. We use time from scheduling to appointment, one of the access metrics used in practice (Thacker 2021), as a proxy.

Define $h \in [-10\%, 34\%]$ the bonus for health outcomes that is applied as a percentage of the total revenue. The PCF expected payment is the sum of its revenue and a performance-based adjustment (PBA), where:

$$\text{Revenue} = \text{CAP} + \text{FVF} \times \text{Avg. No. Visits, where CAP} = \text{PBP} \times 12 \times \text{Avg. No. Beneficiaries}$$

$$\text{PBA} = h \times \text{Revenue.}$$

Let us assume the performance adjustments are linear in the net failure rate, namely,

$$h \equiv a - b\left(\beta Q(e_0) + \alpha Q(e_1) + (1-\alpha)Q(1)\right),$$

for some constants a, and b. The total expected payment under PCF can be written as

$$\text{PCF Payment} = R + r \times \text{Avg. No. Visits}$$
$$- b \times (\text{CAP} + \text{FVF} \times \text{Avg. No. Visits}) \times (\beta Q(e_0) + \alpha Q(e_1) + (1-\alpha)Q(1))$$

where $R = \text{CAP} \times (1+a)$ and $r = \text{FVF} \times (1+a)$. Based on this reformulation, we estimate the baseline health penalty in our model as

$$p_H = b \times (\text{CAP} + \text{FVF} \times \text{Avg. No. Visits})/\text{Avg. No. Visits.} \tag{C.21}$$

We note that in the baseline implementation of PCF, the higher the capitation, the higher the revenue, and since the penalty payment is proportional to revenue, the higher the penalty payment.

To estimate the constants a and b, we assume that the largest bonus $h = 34\%$ is achieved when the net failure rate is the lowest, *i.e.*, when everyone is seen by the specialist ($e_0 = 0,\ e_1 = 0$), and the lowest bonus $h = -10\%$ is achieved when the net failure rate is highest, *i.e.*, when everyone is seen remotely $e_0 = 1,\ e_1 = 1$. Simple algebra leads to $a = 0.34 + 0.44(1-\alpha)/(\alpha+\beta),\ b = 0.44/((\alpha+\beta)Q(1))$. Table C8 reports all the values of the parameters introduced above. For the case without qualification $a \equiv -10\%$ and $b \equiv 0$. With this on hand and the PCF values of PBP and FVF, we estimate the baseline values for our contract parameters $(R^q,\ r^q,\ p_H^q)$ and $(R^{nq},\ r^{nq},\ p_H^{nq})$ using the equations above. For the Quality Gateway criterion, we set $\bar{W} := 1.05 \times W(e_0^S)$ to ensure the social welfare outcome leads to qualification.

Table C8 Parameters of PCF performance-based adjustment component under qualification.

State	a	b
AR	1.33	27.43
CO	1.33	30.26
HI	1.33	36.65
KS	1.33	24.19
MI	1.33	22.09
MT	1.33	31.28
NJ	1.33	22.29
NY	1.33	22.64
OH	1.33	21.96
OK	1.33	22.85
OR	1.33	31.44
PA	1.33	22.62
RI	1.33	23.29
TN	1.33	22.68

Table C9 Data for computing the average practice size and arrival rate.

State	No. of cumulative attributed beneficiaries (i)	No. CPC+ practices included in report (ii)	λ (visits/year) (iii)
AR	169,444	179	1,638
CO	170,246	202	1,139
HI	44,207	93	634
KC	118,892	99	1,979
MI	300,382	420	1,055
MT	68,936	49	1,979
NJ	317,201	431	1,096
NY	106,418	152	1,126
OH	410,433	557	1,090
OK	132,378	169	1,477
OR	132,807	154	1,276
PA	161,070	215	1,011
RI	30,314	31	1,238
TN	46,577	55	1,515

Sources (i-ii): CPC+ 2017 Quality and Utilization Performance Results. https://innovation.cms.gov/initiatives/comprehensive-primary-care-plus (Accessed 02-20-2020). $(iii) = (i)/(ii) \times$ Avg. number of visits per Medicare patient in a year, where the avg. number of visits per Medicare patient in a year is calculated from the 2016 Physician Compare utilization data: https://data.medicare.gov/data/physician-compare (Last visited: 06-04-2019) –it corresponds to the ratio of the state yearly total number of Medicare primary care visits over the state total Medicare population.

Appendix D: Impact of remote visits reflow

In this extension we consider the scenario where some of the patients who were initially seen remotely end up requiring a second, face-to-face, visit with the provider. We refer to these visits as "reflow" visits. Let us denote $\epsilon \in (0,1)$ the fraction of remote visits that will need a face-to-face follow-up appointment. We assume reflow visits have the same characteristics as other face-to-face visits, namely, their compensation, cost, co-payment, average duration, and the chance of failure are the same as regular face-to-face visits. Finally, we introduce a patient disutility d that captures the inconvenience of having to visit the provider's office for an issue that was already evaluated in a prior remote visit.

Under the above modeling assumptions, some of the expressions and results in our analysis must be modified to incorporate the impact of the fraction ϵ of reflow. We next present *only* the expressions that are affected by this addition, namely,

Rate of new non-Medicare patients:

$$\lambda_N^{\epsilon}(e_0,e_1)=\mu\Big(\rho-e_0\lambda\Big(\frac{1}{\bar{\mu}}-\frac{(1-\epsilon)}{\mu}\Big)-\frac{e_1\lambda}{\mu}\Big).$$

Note that the rate of visits from new non-Medicare patients decreases in the amount of reflow ϵ, as the reflow patients utilize some capacity which forces the provider to bring in fewer new patients.

Provider's profit under FFS and PCF:

$$\begin{aligned}\Pi_{provider}^{\text{FFS},\epsilon}=&m^F\lambda(e_1-(1-\epsilon)e_0)-\lambda\int_0^{e_0}\bar{c}(x)dx-t\lambda(1-e_1)-z\lambda((\beta+\epsilon)Q(e_0)+Q(e_1))\\&+(m^F-zQ(1))\lambda_N^{\epsilon}(e_0,e_1),\\\Pi_{provider}^{\text{PCF},\epsilon}=&R+r\lambda(e_1+\epsilon e_0)+\bar{p}\lambda e_0+p\lambda(e_1-(1-\epsilon)e_0)-\lambda\int_0^{e_0}(\bar{c}(x)+\epsilon c(x))dx-\lambda\int_{e_0}^{e_1}c(x)dx-t\lambda(1-e_1)\\&-z\lambda\left[(\beta+\epsilon)Q(e_0)+Q(e_1)\right]-p_H\lambda\left[(\beta+\epsilon)Q(e_0)+\alpha Q(e_1)+(1-\alpha)Q(1)\right]+(m^F-zQ(1))\lambda_N^{\epsilon}(e_0,e_1).\end{aligned}$$

We note that the provider's profit under PCF is still concave in e_0 and e_1. Moreover, the FOC (B.11) with respect to e_1 remains unchanged. That is, in the region $\mathcal{R}$, the reflow of remote patients only affects the adoption of remote and face-to-face care, but it does not affect the amount of referral care employed.

We next examine how the amount of patient reflow affects the utilization of face-to-face vs. remote care. We denote $e_0^{P,\epsilon}$ the optimal threshold under PCF with reflow. The following result shows that $e_0^{P,\epsilon}$ increases in ϵ when r^q is sufficiently high or p_H^q is sufficiently low and decreases in ϵ when r^q is sufficiently low or p_H^q is sufficiently high.

LEMMA 2. *If $r^q+p\geq c(1)+m^F-zQ(1)+q(1)(z+p_H^q)$, the optimal $e_0^{P,\epsilon}$ non-decreasing in ϵ. Alternatively, if $r^q+p\leq c(0)+m^F-zQ(1)+q(0)(z+p_H^q)$, the optimal $e_0^{P,\epsilon}$ non-increasing in ϵ.*

Proof: From the FOC of $\Pi_{provider}^{\text{PCF},\epsilon}$ with qualification, we obtain that $e_0^{P,\epsilon}$ satisfies the following equation

$$q(e_0^{P,\epsilon})=\frac{r^q\epsilon+p-(1-\epsilon)p-\bar{c}(e_0^{P,\epsilon})+(1-\epsilon)c(e_0^{P,\epsilon})+(m^F-zQ(1))(1-\epsilon-\mu/\bar{\mu})}{(z+p_H^q)(\beta+\epsilon)}.$$

Using the Implicit Function Theorem, we obtain that

$$\frac{\partial e_0^{p,\epsilon}}{\partial\epsilon}=\frac{r^q+p-c(e_0^{p,\epsilon})-m^F+zQ(1)-q(e_0^{p,\epsilon})(z+p_H^q)}{q'(e_0^{p,\epsilon})(z+p_H^q)(\beta+\epsilon)+\bar{c}'(e_0^{p,\epsilon})-(1-\epsilon)c'(e_0^{p,\epsilon})}.$$

The denominator of the above expression is always positive since $c(x)-\bar{c}(x)$ is non-increasing. Thus, if $r^q+p\geq c(1)+m^F-zQ(1)+q(1)(z+p_H^q)$ ($\geq c(e_0^{p,\epsilon})+m^F-zQ(1)+q(e_0^{p,\epsilon})(z+p_H^q)$), then the threshold $e_0^{p,\epsilon}$ is non-decreasing in ϵ. Further, if $r^q+p\leq c(0)+m^F-zQ(1)+q(0)(z+p_H^q)$ ($\leq c(e_0^{p,\epsilon})+m^F-zQ(1)+q(e_0^{p,\epsilon})(z+p_H^q)$), then the threshold $e_0^{p,\epsilon}$ is non-increasing in ϵ.

Patient's utility:

$$\begin{aligned}\Pi_{patient}^{\epsilon}(e_0,e_1)=&\lambda\left((u-\epsilon d)e_0-v_H\left[(\beta+\epsilon)Q(e_0)+\alpha Q(e_1)+(1-\alpha)Q(1)\right]-v_WW(e_0)-\right.\\&\left.\bar{p}e_0-p(e_1-(1-\epsilon)e_0)-\tilde{p}(1-e_1)\right).\end{aligned}$$

Payer's profit:

$$\Pi_{payer}^{\text{FFS},\epsilon}(e_0,e_1)=\lambda\left(-\int_{e_0}^{e_1}f(x)dx-\epsilon\int_0^{e_0}f(x)dx-w\left[(\beta+\epsilon)Q(e_0)+\alpha Q(e_1)+(1-\alpha)Q(1)\right]-\int_{e_1}^{1}\tilde{f}(x)dx\right).$$

$$\Pi_{payer}^{\text{PCF},\epsilon}(e_0,e_1)=-R-\lambda\left(r(e_1+\epsilon e_0)+\int_{e_1}^{1}\tilde{f}(x)dx\right)+\lambda(p_H-w)\left[(\beta+\epsilon)Q(e_0)+\alpha Q(e_1)+(1-\alpha)Q(1)\right].$$

Health outcome effect:

$$\Delta Q^{\epsilon} = \alpha Q(e_1^F) - (\beta+\epsilon)Q(e_0^{P,\epsilon}) - \alpha Q(e_1^P).$$

Expenditure effect:

$$\begin{aligned}\Delta X^{\epsilon} \equiv & \int_0^{e_1^F} c(x)dx - \int_0^{e_0^{P,\epsilon}} (\bar{c}(x)+\epsilon c(x))dx - \int_{e_0^{P,\epsilon}}^{e_1^P} c(x)dx + t(e_1^P - e_1^F) + \int_{e_1^F}^{e_1^P} \tilde{f}(x)dx \\ & - z[(\beta+\epsilon)Q(e_0^{P,\epsilon}) + Q(e_1^P) - Q(e_1^F)] - p(e_1^F - e_1^P + (1-\epsilon)e_0^{P,\epsilon}) + \bar{p}e_0^{P,\epsilon}.\end{aligned}$$

Social welfare:

$$\begin{aligned}\Pi^{\epsilon}_{social}(e_0,e_1) = \lambda^{\epsilon}_N(e_0,e_1)(m^F - zQ(1)) + \lambda\Big((u-\epsilon d)e_0 - v_W W(e_0) - \int_0^{e_0} \bar{c}(x)+\epsilon c(x)dx - \int_{e_0}^{e_1} c(x)dx - \int_{e_1}^{1} \tilde{c}(x)dx \\ - t(1-e_1) - (v_H+w+z)\left((\beta+\epsilon)Q(e_0) + Q(e_1)\right) - (v_H+w)(1-\alpha)(Q(1)-Q(e_1))\Big).\end{aligned}$$

We note that the social welfare is still concave in e_0 and e_1. Moreover, the FOC of $\Pi^{\epsilon}_{social}(e_0,e_1)$ with respect to e_0 depends on ϵ, but the one with respect to e_1 does not. That is, the reflow of remote patients only affects the adoption of remote and face-to-face care at the first-best, but it does not affect the amount of referral care employed.

We next examine how the amount of patient reflow affects the utilization of face-to-face vs. remote care at the first-best. We denote $e_0^{S,\epsilon}$ the optimal threshold at the first-best with reflow. The following result shows that reflow leads to less adoption of remote care at the social optimum solution.

LEMMA 3. *The social threshold $\bar{e}_0^{S,\epsilon}$ is non-increasing in ϵ.*

Proof: The FOC of $\Pi^{\epsilon}_{social}(e_0,e_1)$ with respect to e_0 is

$$\frac{\partial \Pi^{\epsilon}_{social}}{\partial e_0} = \lambda\left((m^F - zQ(1))(1-\epsilon-\mu/\bar{\mu}) + u - \epsilon d - v_W W'(e_0) - \bar{c}(e_0) + (1-\epsilon)c(e_0) - (v_H+w+z)(\beta+\epsilon)q(e_0)\right).$$

Thus, the optimal threshold $e_0^{S,\epsilon}$ satisfies

$$q(e_0^{S,\epsilon}) = \frac{u - \epsilon d + v_W|W'(e_0^{S,\epsilon})| + (1-\epsilon)c(e_0^{S,\epsilon}) - \bar{c}(e_0^{S,\epsilon}) + (m^F - zQ(1))(1-\epsilon-\mu/\bar{\mu})}{(\beta+\epsilon)(v_H+w+z)}.$$

Using the Implicit Function Theorem,

$$\frac{\partial e_0^{S,\epsilon}}{\partial \epsilon} = -\frac{q(e_0^{S,\epsilon})(v_H+w+z) + d + c(e_0^{S,\epsilon}) + m^F - zQ(1)}{q'(e_0^{S,\epsilon})(\beta+\epsilon)(v_H+w+z) + v_W W''(e_0^{S,\epsilon}) - (1-\epsilon)c(e_0^{S,\epsilon}) + \bar{c}(e_0^{S,\epsilon})} \leq 0,$$

where the last inequality follows from $c(e_0) - \bar{c}(e_0)$ non-increasing.

We have the following observations:

- The wait time Eq. (2) is independent of ϵ. This is because the new face-to-face visits caused by patient reflow are exactly compensated by a decrease in new non-Medicare patients, leaving the average wait time unchanged. Because of this, the Quality Gateway condition remains unchanged under ϵ reflow.
- The results in Lemma 1 and Propositions 1 and 2 remain valid for the model with ϵ reflow. Moreover, the optimal threshold $e_0^{P,\epsilon}$ can increase or decrease in ϵ depending on the value of r^q and p_H^q (see Lemma 2 above).
- Proposition 3 continue to hold with small modifications. Specifically, the panel size of non-Medicare patients increases under PCF as long as $\epsilon < 1 - \mu/\bar{\mu}$. That is, if there is a large number of patients requiring a second appointment in person after a remote visit, then the capacity implications are such that the provider

will admit fewer new non-Medicare patients under PCF than under FFS. Further, patients are better off under PCF iff the following updated condition holds:

$$-v_H \Delta Q^\epsilon + (\tilde{p} - p)(e_1^F - e_1^P) \leq (u - \epsilon d + (1-\epsilon)p - \bar{p})e_0^{P,\epsilon} + v_W(W(0) - W(e_0^{P,\epsilon})).$$

- Proposition 4 remain valid.
- Proposition 5 remains valid. The threshold $e_0^{S,\epsilon}$ decreases in ϵ (see Lemma 3 above).
- Proposition 6. The same logic used to derive the family of coordinating contracts can be used for the reflow model. The coordinating penalty in Eq. (B.16) is still valid, and the bound on the coordinating r^q, $\underline{r}_1$, remains the same. However, we have an updated Eq. (B.18), namely, a coordinating r^q must satisfy

$$r^q\Big(\frac{(\beta+\epsilon)q(e_0^{S,\epsilon})}{\alpha q(e_1^S)} - \epsilon\Big) \geq \bar{p} + (1-\epsilon)c(e_0^{S,\epsilon}) - \bar{c}(e_0^{S,\epsilon}) + (m^F - zQ(1))\Big(1 - \mu/\bar{\mu} + \frac{(\beta+\epsilon)q(e_0^{S,\epsilon})}{\alpha q(e_1^S)} - \epsilon\Big) - p\Big(1 + \frac{(\beta+\epsilon)q(e_0^{S,\epsilon})}{\alpha q(e_1^S)} - \epsilon\Big) - (t - c(e_1^S))\frac{(\beta+\epsilon)q(e_0^{S,\epsilon})}{\alpha q(e_1^S)} + z(\beta+\epsilon)q(e_0^{S,\epsilon})\Big(\frac{1}{\alpha} - 1\Big).$$

We explore the impact of patient reflow on the coordinating contract parameters numerically. Figure D1 shows, for an illustrative state (Colorado), the range of coordinating flat visit fee values (x-axis) for several values of the reflow ϵ shown on the y-axis (from 0 to 0.4 in increments of 0.05). We find that as more patients require a duplicate face-to-face visit, it is possible to achieve coordination using lower flat visit fees. This is because, with patient reflow, it is socially optimal to offer less remote care to avoid the costs and patient disutility associated with it (Lemma 3). Moreover, per Lemma 2, a lower fee per visit helps ensure that the amount of remote care selected by the provider decreases with the amount of patient reflow, consistent with the social optimum.

We also note that the baseline value of $40.82 could be a feasible coordinating flat fee value in this state, provided that the system has 20% or more reflow. Given that Colorado was among the states for which this value fell outside the range of coordinating flat fees (see Figure 3), this indicates that with a high enough level of patient reflow, the flat fee value currently in use would no longer need to be adjusted upwards to generate a coordinating contract.

Figure D1 Impact of patient reflow on coordinating flat visit fee