MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 25, No. 4, July–August 2023, pp. 1471–1488
ISSN 1523-4614 (print), ISSN 1526-5498 (online)
https://pubsonline.informs.org/journal/msom

Primary Care First Initiative: Impact on Care Delivery and Outcomes

Elodie Adida,[a] Fernanda Bravo[b,*]

[a] School of Business, University of California Riverside, Riverside, California 92521; [b] Anderson School of Management, University of California Los Angeles, Los Angeles, California 90095

*Corresponding author

Contact: elodie.goodman@ucr.edu, https://orcid.org/0000-0002-3638-1584 (EA); fernanda.bravo@anderson.ucla.edu, https://orcid.org/0000-0002-4625-7894 (FB)

Received: March 30, 2022
Revised: September 14, 2022; January 13, 2023
Accepted: February 20, 2023
Published Online in Articles in Advance: April 3, 2023

https://doi.org/10.1287/msom.2023.1207

Abstract. ***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 United States. 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.

Supplemental Material: The online appendix is available at https://doi.org/10.1287/msom.2023.1207.

Keywords: health care management • incentives and contracting • public policy

1. Introduction

Primary care occupies a key function in the U.S. healthcare 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 United States, in part because of the relatively lower compensation compared with 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 reinstated 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 United States, 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, aimed 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 five-year demonstration building upon CPCI: Comprehensive Primary Care Plus (CPC+). CPC+ had a broader range, involving 14 payers and 2,876 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, p. 2665). 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 that 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, p. 2665). 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 multipayer 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 overutilization 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 and 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, whereas Bravo et al. (2023) 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 nonphysician providers (e.g., nurse practitioners) 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, whereas 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 Online Appendix A.

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 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 Operations Management literature to model patient heterogeneity (e.g., Mahjoub et al. 2018, Adida 2021, and 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 $e_0 \le e_1 \in [0,1]$, 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 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 mode 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 that $\bar{c}(\cdot)$, $c(\cdot)$ 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 nonincreasing in x. These assumptions state that more complex patients incur more costs and that, 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 nonincreasing with patient complexity to capture the fact that there is less cost-savings potential for high-complexity patients. These conditions are sufficient to ensure that 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 copayment p for face-to-face care and $\bar{p}$ $(\le p)$ for remote care and pays a copayment $\tilde{p}$ $(> p)$ 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. Although 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 $\bar{q}(x) > q(x) > \tilde{q}(x) \quad \forall x \in [0,1]$. Moreover, $\bar{q}(\cdot) = (1+\beta)q(\cdot)$ and $\tilde{q}(\cdot) = (1-\alpha)q(\cdot)$, where $0 < \alpha < 1$, $0 < \beta \le 1/q(1) - 1$. 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 (respectively (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. Although 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 care was delivered 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

Table 1. Summary of Model Parameters for the Three Types of Care-Delivery Modes

Parameter	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 copayment	$\bar{p}$	p	$\tilde{p}$

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 the unresolved medical condition).

We consider two types of payment systems, described next.

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 $f(x) > c(x) \quad \forall x \in [0,1]$. 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). Although the delivery cost and the reimbursement depend on patient complexity, we assume that the provider's profit margin m^F does not, where $m^F \equiv f(x) + p - c(x) \quad \forall x \in [0,1]$. 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 $\tilde{f}(x) > f(x) \quad \forall x \in [0,1]$, 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—for example, 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 $r < f(x) \quad \forall x \in [0,1]$. (Currently, the flat fee is set at $40.82; see CMS 2021b.) Note that whereas remote care does not incur reimbursement under FFS, it does under PCF; see, for example, 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 Consumer Assessment of Healthcare Providers & Systems 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., numbers 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, p. 61).

Although 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 rescaling upward 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 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 to both 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. We denote the remote and face-to-face service rates as $\bar{\mu}$ and μ ($< \bar{\mu}$) (whether for new or existing patients), respectively. 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 that 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 upward or downward 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 because 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 that the utilization, after including new patients, remains at ρ—that is, $\lambda_N(e_0, e_1)$ satisfies the following equation:

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

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

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

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

We note that the expected wait time does not depend on the referral threshold e_1. This is because we assume that the service rate of face-to-face care is the same for a new patient as for an existing patient; that is, a patient 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 $1/\mu_{net}(e_0, e_1) = \rho/(\lambda e_0(1 - \mu/\bar{\mu}) + \mu\rho)$ decreases in e_0). It follows that the overall average wait time decreases. We note that if $\bar{W} \geq \rho/((1-\rho)\mu)$, 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

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

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 because $Q(1)$ 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 with 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, although 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, although 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 with the cost incurred by a patient failure). Note that profit margin m^F does not drive this decision because 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

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

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; and 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 Online 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 coexist.* (*Quantities $e_0^{min},\bar{e}_0,\bar{e}_1$ are defined in the proof.*)

The proof of the proposition provides the mathematical conditions when each of the cases described above arises. The 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 coexist, 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 also focus on the realistic scenario where it is socially optimal to pass the Quality Gateway and have the three modes of care coexist 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 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—that is, $e_0^F = 0$—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 PCF outcome 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, because 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 copayments p, $\bar{p}$ and $\tilde{p}$ for a face-to-face, remote, and specialist visit, respectively. Thus, the aggregate patient utility is given by

$$\begin{aligned}\Pi_{patient}(e_0, e_1) = \lambda(ue_0 - v_H[\beta Q(e_0) + \alpha Q(e_1)\\ + (1-\alpha)Q(1)] - v_W W(e_0)\\ - \bar{p}e_0 - p(e_1 - e_0) - \tilde{p}(1 - e_1)).\end{aligned} \tag{6}$$

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

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

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

$$\begin{aligned}\Pi_{payer}^{\mathrm{FFS}}(e_0, e_1) = \lambda\Bigg(-\int_{e_0}^{e_1} f(x)dx - w[\beta Q(e_0) + \alpha Q(e_1)\\ + (1-\alpha)Q(1)] - \int_{e_1}^{1} \tilde{f}(x)dx\Bigg).\end{aligned} \tag{8}$$

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 with 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 if and only if (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 Equation (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 with 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. Because 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 copayments, 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 copayment than face-to-face) than FFS hurts the patients financially. However, the fact that PCF has more remote care (with lower copayment than face-to-face) than FFS benefits the patients financially, due to both copayments 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

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

as the "expenditure effect." Intuitively, ΔX represents the *short-term* benefits of PCF, as measured by the payer reimbursement expenditures reduction under PCF compared with 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 with FFS depends on R—for example, if R is sufficiently high, the physician prefers PCF, whereas 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* $\Delta Q \geq 0$; *(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 characterizes when the payer is better off overall under PCF, depending on the expenditure and the health-outcome effects. 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 PCF can be designed to yield decisions that coincide with the social optimum.

Consistent with the literature (e.g., Dranove 1996, p. 62), we define social welfare as comprising the utility of all the stakeholders involved in the care-delivery process—that is, provider, payer, patients, and specialists. 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 that 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 copayment 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 nonavoidable—that is, 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):

$$\begin{aligned}\Pi_{social}(e_0, e_1) = {} & \lambda_N(e_0, e_1)(m^F - zQ(1)) \\ & + \lambda \Bigg(ue_0 - v_W W(e_0) - \int_0^{e_0} \bar{c}(x)dx \\ & - \int_{e_0}^{e_1} c(x)dx - \int_{e_1}^{1} \tilde{c}(x)dx \; - t(1 - e_1) \\ & - (v_H + w + z)(\beta Q(e_0) + Q(e_1)) \\ & - (v_H + w)(1 - \alpha)(Q(1) - Q(e_1)) \Bigg). \qquad (9)\end{aligned}$$

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 coexist.*

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, whereas all socially optimal outcomes lead to a nonzero 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 $e_1^P = e_1^S$. Because 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 that $\bar{W}$ 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 Online 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—that is, 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 five-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 1.8 million beneficiaries. See Table C9 in Online Appendix C for the size (practices and beneficiaries) of 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 people per square mile (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, whereas the rate is 77% in NJ (Table C7 in Online 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), whereas MT has the largest (due to the small number of practices in the state) (Table C9 in Online Appendix C). These observations highlight the heterogeneity among states, which we exploit to obtain insights into the optimal design of PCF.

We next summarize the calibration approach. The details and corresponding data are available in Online Appendix C. A summary of the values for each state is available in Table C3 in Online Appendix C.

The yearly average arrival rate λ is obtained from the 2016 "Physician Compare" utilization data (Table C9, column (iii), in the online appendix). 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 Current Procedural Terminology codes-based reimbursement rates for primary care in-office visits (see codes in Table C4 in the online appendix). It follows that $c(x)$ is linear in x, and we denote $c(x) := c_0 + \texttt{FFS\_rate} \times x$. We assume that the cost for the lowest-complexity patient is equal to the copayment—that is, $c_0 = p$. The cost of remote care is also modeled as linear in the care complexity x, and, for simplicity, we assume that 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—that is, $\tilde{c}(x) := 2 \times c(1)\ \forall x \in [0,1]$. 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 in Online Appendix C).

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 *nonelective* 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 in the online appendix). 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 (Lam et al. 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 healthcare 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 that the failure cost and the wait cost are of similar magnitude under the scenario where all care is delivered in person—that is, $v_W W(0) = v_H Q(1)$. The cost of access delay v_W thus depends on the population's health status and the state median hourly wage. See Table C7 in Online Appendix C for details.

Medicare's primary care and specialist visit copayments range from \$15 to \$25 and \$30 to \$50, respectively (Fay 2019). We set the primary care face-to-face copayment at $p = \$20$, the remote visit copayment at $\bar{p} = 90\% \times p = \$18$, and the specialist visit copayment 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 coexist for a wide range of contract 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 the 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, and 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 in Online Appendix C. We explain the mapping from PCF parameters to our equivalent contract parameters (R, r, p_H) in Online Appendix C.

6.2. Results

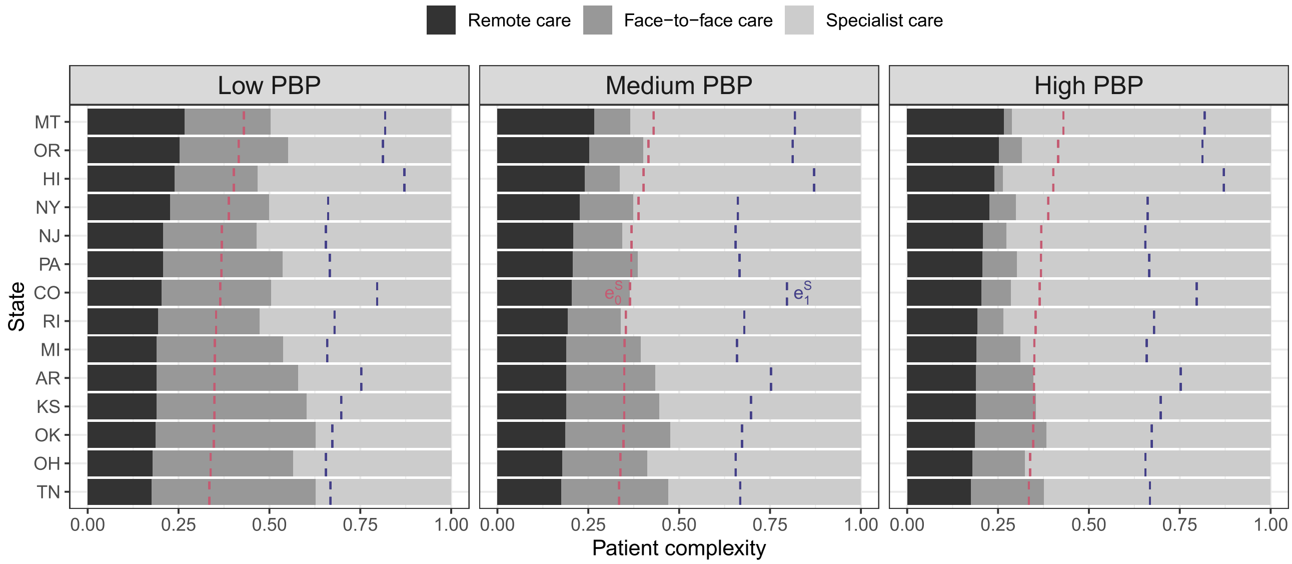
In the first part of this section, we analyze the adoption of the three modes of care under the baseline PCF and compare it to the socially optimal solution. In the second part, we investigate the impact of the Quality Gateway criterion on the care-delivery modes. Finally, in the third part, we study the family of coordinating contracts and discuss what population characteristics affect the coordinating PCF contract. Note that, although the insights provided in this section are based on numerical analysis only, we carefully calibrated the parameters based on realistic values across many states 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) (vertical dashed thresholds). The three panels correspond to the three considered possible values of the capitation payment (PBP).

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

Figure 1. (Color online) Care Thresholds under the Baseline PCF



Notes. States are ranked in decreasing value of e_0^S. 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 that e_0^S leads to qualification. Low PBP = \$28, Medium PBP = 2 × \$28, and High PBP = 3 × \$28.

higher penalties (see Online 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 to 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.

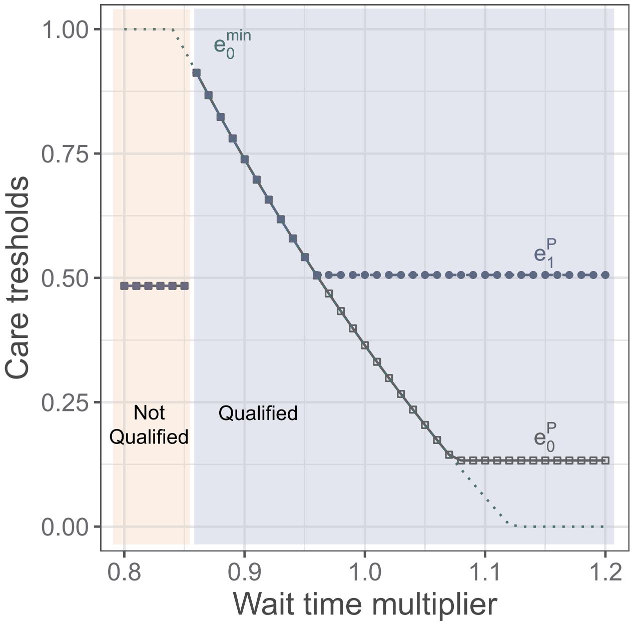
Secondly, we observe that the suboptimality of the baseline PCF manifests differently for the different states. Although all states exhibit less remote care and more referrals than the first-best, at all three PBP levels, the states' "health status" plays a noteworthy role. A state's health status, as measured by the utilization of inpatient services, is inversely related to the failure-rate parameter δ. In states with lower health status (e.g., Tennessee, Ohio, Oklahoma, and Kansas), 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, and Colorado) tends to overuse referrals to a larger extent than states with lower health status. This is because new patients are more profitable in states with high health status since the expected reputation costs of accepting new patients, which is proportional to the failure probability, is lower. 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 ±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, whereas 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

Figure 2. (Color online) Impact of Quality Gateway on Modes of Care Delivery (Colorado Depicted as an Illustrative State)



amount of remote care (e_0^{min}). We observe that when the maximum wait time is low—that is, 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.

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. Because 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.

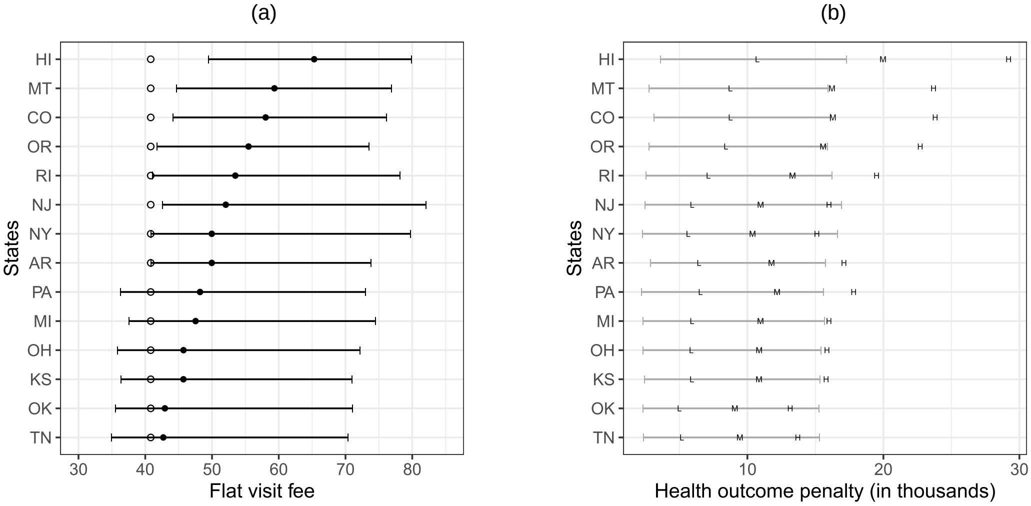
To summarize, we see that the maximum wait time in the Quality Gateway qualification condition can impact the provider's choice of care-delivery modes and, therefore, should be carefully selected, in conjunction with the other contract parameters. A too stringent requirement may induce the provider to 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 suboptimality 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 with 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 nonhomogeneous 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

Figure 3. Coordinating Contracts



Notes. (a) Coordinating visit fee. The circle point represents the baseline fee. The dark circle is the coordinating fee under the baseline penalty (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.

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 the socially optimal one. In summary, it may be 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

Although 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 postpandemic 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 that 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, or 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 United States, and the results in these 14 states were remarkably similar, which brings robustness to the insights we derive.

Acknowledgments

The authors thank the department editor, associate editor, and reviewers for their comments, which helped substantially improve the content and exposition of this work.

References

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

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

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

Agency for Healthcare Research and Quality (2015) CAHPS Clinician & Group Survey. Accessed March 13, 2023, https://www.ahrq.gov/cahps/surveys-guidance/cg/index.html.

Agency for Healthcare Research and Quality (2020) National inpatient hospital costs: The most expensive conditions by payer, 2017. Accessed March 13, 2023, https://www.hcup-us.ahrq.gov/reports/statbriefs/sb261-Most-Expensive-Hospital-Conditions-2017.jsp.

ASPE (2020) Medicare beneficiary use of telehealth visits: Early data from the start of the COVID-19 pandemic. *Issue Brief* (July). Accessed March 13, 2023, https://aspe.hhs.gov/system/files/pdf/263866/hp-issue-brief-medicare-telehealth.pdf.

Association of American Medical Colleges (2019) New findings confirm predictions on physician shortage. Accessed March 13, 2023, https://www.aamc.org/news-insights/press-releases/new-findings-confirm-predictions-physician-shortage.

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 Oper. Management* 28(6):1577–1593.

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

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

Burton R, Berenson RA, Zuckerman S (2017) Medicare's evolving approach to paying for primary care. Technical report, The Urban Institute, Washington, DC, and The Robert Wood Johnson Foundation, Princeton, NJ. Accessed March 13, 2023, https://www.urban.org/sites/default/files/publication/95196/2001631_medicares_evolving_approach_to_paying_for_primary_care_0.pdf.

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

Çakıcı ÖE, Mills AF (2021) On the role of teletriage in healthcare demand management. *Manufacturing Service Oper. 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. *N. Engl. J. Med.* 361(4):368–378.

Center for Connected Health Policy (2020) COVID-19 telehealth coverage policies. Accessed March 13, 2023, https://www.cchpca.org/resources/covid-19-telehealth-coverage-policies.

CMS (2011) Solicitation for the Comprehensive Primary Care Initiative. Center for Medicare and Medicaid Innovation, Centers for Medicare & Medicaid Services, Baltimore. Accessed March 14, 2023, https://innovation.cms.gov/Files/x/Comprehensive-Primary-Care-Initiative-Solicitation.pdf.

CMS (2016) CMS launches largest-ever multi-payer initiative to improve primary care in America. Press release, Centers for Medicare & Medicaid Services, Baltimore. Accessed March 14, 2023, https://www.cms.gov/newsroom/press-releases/cms-launches-largest-ever-multi-payer-initiative-improve-primary-care-america.

CMS (2020) Medicare telemedicine healthcare provider fact sheet. Accessed March 14, 2023, https://www.cms.gov/newsroom/fact-sheets/medicare-telemedicine-health-care-provider-fact-sheet.

CMS (2021a) Primary Care First model options. Center for Medicare and Medicaid Innovation, Centers for Medicare & Medicaid Services, Baltimore. Accessed March 14, 2023, https://innovation.cms.gov/innovation-models/primary-care-first-model-options.

CMS (2021b) Primary Care First: Payment and attribution methodologies PY 2022. Technical report, Centers for Medicare & Medicaid Services, Baltimore. Accessed March 14, 2023, https://innovation.cms.gov/media/document/pcf-py22-payment-meth-vol1.

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

Dranove D (1996) Measuring costs. Sloan FA, ed. *Valuing Healthcare: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies* (Cambridge University Press, Cambridge, UK), 61–76.

Fainman EZ, Kucukyazici B (2020) Design of financial incentives and payment schemes in healthcare systems: A review. *Socio-Econom. Planning Sci.* 72:100901.

Fay B (2019) Doctor visit costs. Accessed March 14, 2023, https://www.debt.org/medical/doctor-visit-costs/.

Fuloria PC, Zenios SA (2001) Outcomes-adjusted reimbursement in a health-care delivery system. *Management Sci.* 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* (May 31), https://www.healthaffairs.org/do/10.1377/forefront.20160531.055050.

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

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

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

Lam O, Broderick B, Toor S (2018) How far Americans live from the closest hospital differs by community type. Accessed March 14, 2023, https://www.pewresearch.org/fact-tank/2018/12/12/how-far-americans-live-from-the-closest-hospital-differs-by-community-type/.

LaPointe J (2018) Medicare reimbursement rules limit telehealth adoption. Accessed March 14, 2023, https://revcycleintelligence.com/news/medicare-reimbursement-rules-limit-telehealth-adoption.

Maciag M (2017) Median wages by state. Accessed March 14, 2023, https://www.governing.com/gov-data/wage-average-median-pay-data-for-states.html.

Mahjoub R, Ødegaard F, Zaric GS (2018) Evaluation of a pharmaceutical risk-sharing agreement when patients are screened for the probability of success. *Health Econom.* 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, O'Malley AS, Day TJ, 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, et al (2021) Independent evaluation of Comprehensive Primary Care Plus (CPC+): Third annual report. Technical report, Mathematica Policy Research, Princeton, NJ.

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

Robinson JC (2001) Theory and practice in the design of physician payment incentives. *Milbank Quart.* 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 Sci.* 65(7):3196–3215.

Sessums LL, Basu S, Landon BE (2019) Primary Care First—Is it a step back? *N. Engl. J. Med.* 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. *JAMA* 315(24):2665–2666.

Shachar C, Engel J, Elwyn G (2020) Implications for telehealth in a postpandemic future: Regulatory and privacy issues. *JAMA* 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. Accessed March 14, 2023, https://revcycleintelligence.com/news/healthcare-reimbursement-still-largely-fee-for-service-driven.

Thacker R (2021) CPC+ and Primary Care First: The new CMS payment model explained. Accessed March 14, 2023, https://minglehealth.com/blog/cpc-plus-and-primary-care-first-cms-payment-model-explained.

U.S. Census Bureau (2019) Select maps on the population 65 and older in the United States by County: 2013–2017. Accessed March 14, 2023, https://www.census.gov/library/visualizations/time-series/demo/nia_county_maps.html.

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 Sci.* 21(4):475–491.

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