Early Reservation for Follow-up Appointments: Enhancing Patient Care Continuity

Yichuan Ding
Desautels Faculty of Management, McGill University, Montreal, Quebec H3A 1G5, Canada, daniel.ding@mcgill.ca

Diwakar Gupta
McCombs School of Business, University of Texas, Austin, Texas 78712, Diwakar.Gupta@mccombs.utexas.edu

Shenghai Zhou
School of Business, Central South University, Changsha, China 410083, shzhou@csu.edu.cn

Problem Definition: In the context of outpatient care, physicians often decide at the end of a consultation session whether to schedule follow-up appointments for patients with potential needs. Those appointments are referred to as prioritized follow-up appointments (PFU) (Ding et al. 2023). We study mechanisms that encourage physicians to schedule the optimal quantity of PFUs, aiming to enhance continuity of care (COC) while minimizing no-shows and late-cancellations.

Methodology/Results: Utilizing both empirical analysis and modeling, our research examines strategies for enhancing COC within an appointment scheduling framework. Empirical evidence indicates that a greater frequency of PFUs is associated with improved COC. Subsequently, we introduce a queueing model that delineates the impact of PFU appointments on revenue generation and COC levels. The model reveals a discrepancy between physician and health system preferences regarding the number of PFUs, with doctors inclined to schedule fewer and health systems favoring more. To reconcile these differing objectives, we apply a principal-agent model. For situations involving symmetric information, we suggest a particular performance-based payment structure that effectively aligns the incentives of both stakeholders. When information is asymmetric, we theoretically evaluate and compare four distinct contract types.

Managerial Implications: Our findings suggest that health systems ought to implement incentive schemes that reward physicians for a higher proportion of PFUs in their scheduling. Such rewards are more cost-effective when based on the ratio of PFUs rather than the aggregate number.

Key words: Appointment Scheduling, Queueing Models, Continuity of Care, Principal-Agent Models

1. Introduction

Value based payment models, such as accountable care organizations and bundled payments, have proliferated in the US under the broader umbrella of healthcare reform. Health systems are increasingly paid a base amount either on an episode basis or on a population basis with additional payments associated with producing good health outcomes, rather than on the basis of the volume of services provided. However, many physicians are still paid on a fee for service (FFS) basis. This creates challenges for health systems who need

to align their priorities with those of doctors and other health professionals that provide care. In this paper, we focus on one such challenge in the context of outpatient clinics – namely, how to manage capacity when *continuity of care* (COC) affects a health system's profit, but doctors are paid on a fee-for-service (FFS) basis.

The medical literature has documented numerous benefits of higher COC, including increased patient satisfaction (Sans-Corrales et al. 2006, Saultz and Albedaiwi 2004), improved recognition of existing or previously identified health problems, fewer episodes of sickness and laboratory tests (Dreiher et al. 2012, Rogers and Curtis 1980), decreased hospitalizations and emergency department visits (Haggerty et al. 2003), higher acceptance of suggested preventive services, and completion of recommended care (Atlas et al. 2009). A systematic review of clinical trials on COC and its impact on quality-of-care has been provided by Van Servellen et al. (2006). The lack of continuity, in contrast, is found to be associated with lower patient satisfaction, higher morbidity, difficult consultations, non-attendance, and an increase in utilization of open-access clinics (Kikano et al. 2000). De Maeseneer et al. (2003) show that COC is one of the most important explanatory variable related to the total healthcare cost via a multivariate linear regression analysis.

We consider an outpatient care system that is motivated to achieve a high COC level, because COC is positively associated with health outcomes. Moreover, it may be either a performance metric that affects its incentives payments, or a determinant of its costs of providing care, or both. However, the doctors are paid on the FFS basis and not motivated to the same extent. Health systems could choose from a variety of strategies to realize higher COC. In this paper, we focus on one such strategy called "prioritized follow-up" (PFU), where doctors would be incentivized to schedule a follow-up appointment (FUA) at the conclusion of each visit when clinically appropriate. This strategy has been discussed in recent literature as an effective means of enhancing COC (RCGP 2019, Ding et al. 2023). The doctors choose which patients to advise to book PFU appointments, depending on the estimated likelihood that the patient will need an FUA. We assume that patients follow their doctors' advice, although they may cancel the booked appointments later.

If the doctor does not book a PFU for the patient, and the patient later needs an FUA, then they may book an appointment, which is referred to as "regular follow-up" (RFU). Some returning patients, however, may not be able to book an RFU with the same doctor on their desired dates because of appointments backlog. These patients may end up not

booking an FUA, or booking an FUA with another doctor. We refer to such behaviour as patient balking, which results in disruption of provider continuity and a lower COC level.

To quantify the effectiveness of the PFU strategy in enhancing COC, we conducted an empirical analysis using data comprising more than 530,000 appointments spanning a 12-month period. We employed propensity score matching to control for various clinic-related variables. The regression analysis revealed a positive association between a higher ratio of PFU/RFU and an increase in COC levels. This finding not only supports the efficacy of the PFU strategy in elevating COC, but also suggests that PFU can serve as a reliable indicator of a provider's COC performance. It holds particular significance because health systems typically lack information about patients' visit histories with unaffiliated providers.

Whereas PFU appointments help maintain high COC, they are associated with a higher spoilage rate compared to RFUs; see the empirical evidences provided by Ding et al. (2023) and Green and Savin (2008). Some patients with PFU appointments may not need to follow up with their doctors, and a fraction of them may fail to cancel in a timely fashion. This increases the likelihood of PFUs being cancelled in the last minute or patient no-shows, which elevates the risk of wasted appointment slots (Gallucci et al. 2005). Consequently, doctors may be hesitant to schedule as many PFUs as a health system may deem optimal. Note that higher spoilage rate can potentially reduce doctors' revenue, particularly under the FFS payment scheme. Given the aforementioned incentive misalignment, it is natural to ask "What incentives should health systems offer to doctors to realize the desired level of COC?" The primary objective of this paper is to develop a mathematical framework capable of addressing this question. Through models representing interactions between doctors, health systems, and patients, we explore strategies that align incentives and strike a balance between maximizing throughput and enhancing COC.

This study introduces several principal-agent (P-A) models in which doctors are incentivized to schedule an appropriate number of PFU appointments, thereby resulting in a high level of COC. The health system serves as the principal and the doctor serves as the agent. The P-A models rest on a chassis comprised of a queueing model, which calculates the steady-state service rates for different types of appointments. In our models, the principal's goals are to maximize the steady state throughput rate as well as the COC level. The principal's payment to the agent has two components: a volume based (per visit) payment

to incentivize the doctor to serve more patients, and a performance-based payment to incentivize the doctor to book more PFUs when clinically appropriate. However, in reality the principal may not observe the agent's effort level in booking PFUs because the number or the proportion of clinically-appropriate PFU appointments depends on the doctor's discretion as well as the distribution of the patients' revisit probability. If this distribution is known by both the principal and the agent, then that is referred to as the symmetric information case. If the distribution is privately known to the agent but unobservable by the principal, then that is referred to as the asymmetric information case. The key difference between the two cases is whether the principal can infer the agent's effort level, that is, the PFU-booking criterion, upon observing the number or proportion of booked PFU appointments.

In the case of symmetric information, we propose a typical bonus contract in which the principal rewards the agent if the latter's effort has reached a predefined target. This approach effectively aligns the objectives of both parties and incurs a cost that can be tightly bounded. In the case of asymmetric information, we theoretically compare four easy-to-implement contracts in which payments are proportional to different metrics that can be derived from appointment data. These metrics are: (1) the percentage of served FUAs among all served appointments, (2) the count of served FUAs, (3) the percentage of served PFUs among all served appointments, and (4) the count of served PFUs. We show that PFU based payments (3) and (4) outperform FUA based payments (1) and (2). We also show that ratio-based payments (1) and (3) outperform count-based payments (2) and (4). Moreover, we show that there is a gap between the first-best and the second-best solution, but that gap can be explicitly bounded when payments are based on ratio metrics (1) or (3). Notably, the PFU-ratio based payment (3) emerges as the most effective option among the four, positioning it as a front runner on the throughput-COC spectrum.

A summary of the contributions of this paper is as follows.

- We empirically show that a higher PFU/RFU rate contributes to a higher COC level, indicating that booking more PFU appointments is an effective means of enhancing COC. We also explain this observation via an analytical model. Our study offers a comprehensive view of how PFU appointments serve as a lever to improve COC and complements the recent work by Ding et al. (2023) that examines the impact of PFUs on the throughput rate.

- By conceptualizing the PFU booking process as a queueing control problem, we show that the doctor's revenue is a quasi-concave function of the number of PFU appointments, whereas the COC level strictly increases in that. Importantly, this may result in misalignment of the doctor and the health system's interests and different PFU booking strategies.
- Our exploration of a principal-agent model reveals how financial incentives can effectively encourage physicians to schedule a target number or percentage of PFU appointments, thus elevating COC levels. We delve into contract designs under both symmetric and asymmetric information scenarios, offering actionable strategies for enhancing the principal-agent relationship in healthcare settings. This approach aligns with and expands upon existing literature in personnel economics (Lazear and Shaw 2007), specifically applying it within the healthcare context.

2. Literature Review

In this section, we review several literature streams relevant to our research topic, including COC measurements, appointment scheduling with patient no-shows, queueing models with return customers, and principal-agent models in health care operations.

Previous studies have proposed several different measures of COC — see (Sualtz and John 2003) for a comprehensive overview. Examples include Continuity of Care Index, which examines the concentration of patient visits among different providers, and the Sequential Continuity *SECON* Index (Steinwachs 1979), which captures the sequential continuity of care for individual patients over a specific measurement period. We focus on COC in the appointment scheduling environment, which emphasizes a patient's ability to book a follow-up appointment with the same doctor (Nutting et al. 2003). This is aligned with the concept of sequential continuity captured by the SECON Index. We investigate the idea of booking PFU appointments to promote COC in outpatient care. The downside of this approach is the higher no-show rate of PFU appointments due to its long lead time (Ding et al. 2023, Green and Savin 2008). Literature has shown that no-shows lead to disruptions and inefficiencies in the delivery of outpatient care (Cayirli et al. 2006, Luo et al. 2012). Specifically, several papers have studied how to manage patient access in presence of no-shows through both inter-day (Liu et al. 2010, Feldman et al. 2014) and intra-day(Robinson and Chen 2010, Kong et al. 2013, 2020, Jiang et al. 2017)

appointment scheduling. Some other papers study open-access policies to reduce no-show rates (Steinbauer et al. 2006, Robinson and Chen 2010). In contrast, we show that by implementing an appropriate PFU-booking strategy, which may result in higher no-show rates and induce costs of incentives paid to the doctors, the health system can improve its COC level and overall profit.

We model the appointment system as a system of queues with re-entrant customers. Such queueing models have been utilized in a variety of service operations' settings (e.g., Armony and Maglaras 2004, and Kostami and Ward 2009). For such models, the exact characterization of the steady state probabilities of metrics such as the number in the system and the customer waiting time can be obtained only in some special cases, e.g., M/M/1 queue with exponentially distributed in-orbit time (Guo et al. 2019, Yom-Tov and Mandelbaum 2014, and Campello et al. 2017). In a majority of the models involving returning customers, it is usually not possible to derive a closed-form characterization of steady state probabilities, and researchers have used fluid and diffusion approximations to obtain asymptotic characteristics of the system-level metrics (Huang et al. 2015, Chan et al. 2014, Dobson et al. 2013). For queues arising in appointment booking systems, common assumptions of asymptotic analysis such as heavy traffic are difficult to justify. Therefore, instead of using asymptotic analysis, we study the system using a the approximation which is referred to as the *returning customers see time averages (RTA)* in the literature. This concept was first introduced by Greenberg and Wolff (1987) for M/M/c queues with orbits, where a customer is deemed to be in orbit in between two consecutive visits. This idea was further discussed in Yang and Templeton (1987) and Wolff (1989). More recently, Ding et al. (2023) used this approximation to analyze an appointment booking system for throughput maximization. We use a similar approach for the purpose of designing mechanisms that promote COC.

Our work uses a principal-agent model to align the incentives of the doctors and the health systems. The principal-agent model has been widely used in health care management and more generally in deciding how wages should be tied to performance. The latter has spawned a subfield of economics called personnel economics (Lazear and Shaw 2007). In the interest of brevity, we discuss only a few selected papers with healthcare focus. Gupta and Mehrotra (2015) model the early implementation of the Bundled Payments for Care Improvement by the CMS. In their paper, the principal (payer) selects the best agent

(proposer) among multiple agents and the proposers select bundles and target quality scores. They investigate whether the principal should reveal quality attribute weights it would use to evaluate proposals and show that the principal may prefer not to do so. Aswani and Shen (2019) studied the CMS' Medicare Shared Savings Program (MSSP), which aims to control escalating medicare spending and deliver healthcare more efficiently, as a principal-agent model. They propose a subsidy-based contract to partially reimburse the provider's investment and prove it to be more efficient than the current MSSP contract, leading to higher expected payoffs for both CMS and the provider. In a different example, Adida et al. (2019) proposed an outcome-based penalty contract to handle the moral hazard incurred by the provider that exerts effort.

A few papers have analyzed principal-agent models within a queueing framework. This requires the authors to characterize the steady state of the queueing system. Such analyses may not yield closed-form expressions, adding to their mathematical complexity. For example, Jiang et al. (2012) applied a principal-agent framework to analyze different contracts between a service purchaser (the principal) and the healthcare provider (the agent). The agent maximizes his payoff by allocating his service capacity among urgent patients, dedicated advance patients, and flexible advance patients. The principal aims to minimize the cost of purchasing services while ensuring a target waiting-time by designing appropriate contracts. They formulate the appointment dynamics as an $M/D/1$ queue and consider both adverse selection and moral hazard that are common consequences of informational asymmetry. Jiang et al. (2020) exploited a $G/G/m$ queue to describe the patient care process and use a principal-agent framework to study the performance-based contracting problem. They show that patient benefits can be enhanced through increased competition among hospitals and the introduction of incentives. Arifolu et al. (2020) used a principal-agent model to demonstrate that the Hospital Readmissions Reduction Program does not provide the right incentives for hospitals to reduce readmissions. In their model, patients are assumed to return to the hospital with a fixed probability. In contrast, in our model, the return probability depends endogenously on the PFU booking threshold decided by the doctor. Our work uses the principal-agent framework to compare mechanisms based on different functional forms of linear performance based payments, which has not been studied in the literature.

3. Empirical Evidence

In this section, we present empirical evidence to substantiate the assertion that a higher frequency of PFUs correlates with higher levels of COC. Our focus is to show that clinics scheduling a larger proportion of their appointments as PFUs achieve higher COC levels, even after controlling for factors such as patient demographics, insurance status, clinic location, and proportion of FUAs among all appointments. We utilize data on all patient visits over a 12-month period to 37 clinics in a particular geographical area. All 37 clinics were part of the same health system. We excluded data pertaining to doctors who had fewer than 400 appointments in the year and those who worked in multiple clinics because those doctors do not represent a regular practice pattern. After these exclusions, our data includes 534,220 visit records for 315 doctors.

To measure provider-level care continuity, we utilize the average SECON score, as introduced by Steinwachs (1979). For calculating the SECON score, we define the episodes and FUAs following the method introduced by Ding et al. (2023) – an appointment is regarded as an FUA if it occurs within 45 days of a preceding visit by the same patient. An episode of visits consists of one initiating appointment, and multiple FUAs. For episodes with at least one FUA, we calculate its *SECON* index using the following expression specified in (Eriksson and Mattsson 1983),

$$\text{SECON} = \frac{\sum_{i=2}^{I} s_i}{I-1}, \tag{1}$$

where I represents the total number of visits within an episode (including the initiating visit) and and s_i is an indicator for the sequential continuity which takes value 1 if the i^{th} visit is conducted by the same doctor as the preceding one, and 0 otherwise. We exclude episodes that do not include any FUA from the study because their SECON indices are undefined. For each doctor, we calculate their average SECON index across all episodes they were involved in during the study period. This aggregation results in a total of 315 doctor-level SECON indices, serving as the response variables in our regression analysis.

The treatment variable of our primary interest is the proportion of PFUs relative to the total number of appointments scheduled by each doctor. Our data includes detailed appointment booking times, which allow us to easily identify the PFUs as the appointments that were booked during the preceding consultation session of the same patient. We use the percentage instead of the absolute counts of PFUs to mitigate the influence of external

variables such as the variation in patient age demographics and the size of the doctor's patient panel. Moreover, to isolate the specific effect of PFUs on COC, we also incorporate the percentage of all follow-up appointments (FUAs) amidst all scheduled appointments as control variables within our analysis framework.

Given this setup, we proceed to apply a reduced-form model to our dataset, structured to precisely evaluate how the proportion of PFUs booked by each doctor influences the level of care continuity, while duly accounting for other pertinent variables that could affect this relationship. This modeling approach allows us to quantify the distinct contribution of PFUs to enhancing COC, offering a clear understanding of the strategic value of prioritizing follow-up appointments in outpatient care settings. Our regression model is as follows,

$$\text{SECON}_j = \alpha_0 + \alpha_1 PFURatio_j + \alpha_2 FUARatio_j + \varepsilon_j, \tag{2}$$

where SECON_j denotes the average SECON index for doctor j, while $PFURatio_j$ and $FUARatio_j$ denote the ratios of PFUs and FUAs among all appointments served by doctor j, respectively. Note that $FUARatio_j$ is the sum of $PFURatio_j$ and $RFURatio_j$, where $RFURatio_j$ is the percentage of RFUs among all appointments involving doctor j. However, directly fitting the above regression model to our data could lead to biased estimation due to potential confounders such as the patient mix, the clinic's location, and patients' insurance status. To minimize the effect if confounding factors, we employ the propensity score matching method (PSM) to create comparable samples of doctors for an unbiased estimation the treatment effect.

For the PSM process, we divide the providers into two groups: treatment group and control group, based on their PFU ratio. Doctors with PFU ratios below the median value are classified into the control group, while the rest form the treatment group. To ensure robustness of the results, we test different cutoff values other than the median to classify the treatment groups and control groups and the results are presented in Appendix B.

For the propensity score matching, we include covariates of each doctor to ensure a robust matching process. These covariates include the total number of visits (service volume) of the doctor's clinic, which would be related to the clinic's location, the doctor's service volume, the distribution of patients across five age groups (< 15, $15 \sim 30$, $30 \sim 50$, $50 \sim 65$, and > 65) treated by the doctor, and the distribution of patients across different insurance types (commercial, government, and self-pay) treated by the doctor. By incorporating these

variables into our matching algorithm, we aim to pair doctors in the control group with counterparts in the treatment group who have similar profiles.

To calculate the propensity scores, we employ logistic regression with the aforementioned covariates. We then perform nearest-neighbor matching within a specified caliper. Providers whose propensity score differences exceed this caliper threshold are dropped from the analysis as they cannot be adequately matched. This matching process significantly reduces the disparities between the control and treatment groups, making them nearly equivalent in terms of their covariates. The summary statistics of the covaraites for the two groups are presented in Table 1.

The comparison of the matched and unmatched samples is detailed in Table 1. The %bias column calculates the standardized difference of covariate values between treatment group and control group (Harder et al. 2010). Notably, the %bias for all covariates have absolute value of less than 10%, suggesting no systematic bias in the covariate values between the control and treatment groups (Li et al. 2022). In addition, the p-values (> 0.1) of the matched group also indicate no significant difference for covariate values between control group and treatment group. This result supports the effectiveness of our matching process and the validity of the subsequent empirical findings.

Table 1 Comparison Between the Treatment Group and Control Group

	Unmatched (U) Matched (M)	Mean Treatment	Control	%bias	p-value
NumApp_Cli	U	19933	25541	-59.1	0
	M	22749	22101	6.8	0.595
NumApp_Prov	U	1712.2	1679.8	4.3	0.706
	M	1579.7	1536.5	5.7	0.665
AgeCat1_Num*	U	238.73	271.06	-15.2	0.179
	M	240.94	241.57	-0.3	0.982
AgeCat2_Num	U	263.13	355.21	-50.1	0
	M	284.6	271.24	7.3	0.584
AgeCat3_Num	U	491.52	563.96	-23.5	0.038
	M	483.44	464.21	6.2	0.612
AgeCat4_Num	U	381.17	345.97	16.2	0.153
	M	346.02	341.65	2	0.878
Fin_cla1_Num**	U	915.98	1069.3	-34.2	0.003
	M	911.85	889.69	4.9	0.698
Fin_cla2_Num	U	615.23	403.51	60.3	0
	M	484.31	476.91	2.1	0.871

$*$ We define five age categories: < 15, $15 \sim 30$, $30 \sim 50$, and $50 \sim 65$, and > 65 (default).

$**$ There are three different financial methods: commercial insurance, government insurance, and self-pay (default).

The PSM process excludes about 20% of the observations, resulting in 142 and 110 observations in the control and treatment groups, respectively. These 252 observations are referred to as the PSM sample. We then apply the regression model (2) to both the full sample and the PSM sample. The coefficient estimates are summarized in Table 2. The estimate of α_1 (i.e., the coefficient of $PFURatio$ in Equation 2) captures the effect of PFUs on the COC level. As evident in Table 2, with the full sample, we observe that a ten percent increase in the PFU ratio correlates with an average rise in the SECON index by 0.1189. Moreover, the same relationship holds in the PSM sample, where a similar magnitude of effect, a 0.1072 increase in SECON index per ten-percent increment in the PFU ratio, is noted. Significantly, both estimates are statistically significant, underscoring that a higher proportion of PFUs is positively associated with an escalated SECON index. This finding substantiates the claim that booking more PFUs instead of RFUs can lead to measurable improvements in COC, as indicated by the SECON index.

Table 2 The PFUs effect on COC

	Full sample		PSM sample	
	Coeffecient	p-value	Coeffecient	p-value
PFURatio	1.189	0.001	1.072	0.008
FUARatio	0.338	0.105	0.190	0.394
Constant	0.343	0.000	0.365	0.000
Obs.	315		252	

4. Problem Description and Formulation

In this section, we describe a model created to investigate the impact of booking a potential follow-up appointment as a PFU or not. Adopting a methodology similar to that of Ding et al. (2023), we conceptualize the appointment scheduling system as a single-server priority queue with returning customers. That allows us to obtain the steady-state consequences of any PFU booking threshold in terms of the doctor's and the health system's revenue rate, and to subsequently endogenize the choice of the threshold in a P-A model with incentives. The latter is designed to align the doctor's decision with that of the health system. For clarity, we summarize the the notation used in model formulation in Table 3.

We assume that new (episode-initiating) appointment requests arrive according to a Poisson process with rate λ_n. All appointments have i.i.d. service times following an exponential distribution with rate $\mu = 1$. Similar assumptions have been made in earlier works

Table 3 Table of notation

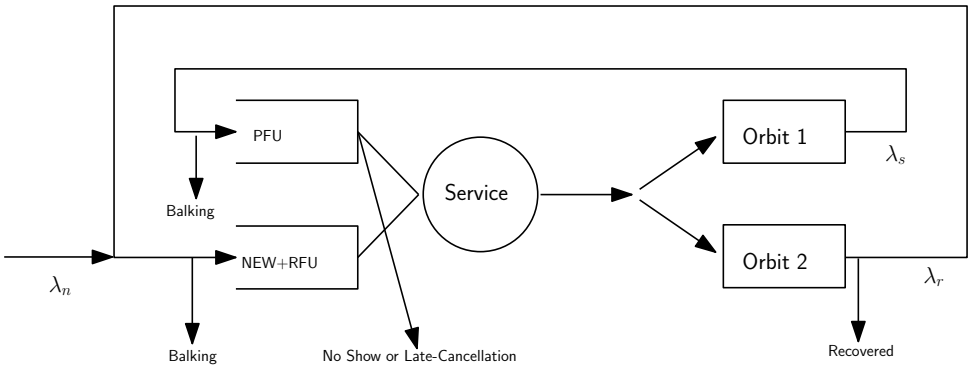
Notation	Explanation
μ	Service rate, assumed to be 1
λ_n	Poisson arrival rate for new arrivals
λ_p	Poisson arrival rate for PFUs
λ_r	Poisson arrival rate for RFUs
λ_d	Effective departure (service) rate
λ_d^P	The effective service rate of PFUs
λ_d^R	The effective service rate of RFUs
λ_v	The virtual arrival rate
p	Probability of requiring a follow-up appointment
$f(\cdot)$	Probability density function for p
$F(\cdot)$	Cumulative distribution function for the probability p
$G(x)$	Average revisit probability for the subpopulation with $p \in [0,x]$: $G(x) = \int_0^x pf(p)d(p)$
$\bar{p}$	$\bar{p} = G(1)$, the average revisit probability for the entire population
w	The PFU booking threshold
w_{FB}	The booking threshold that solves the first-best model
w_{SB}	The booking threshold that solves the second-best model
$\hat{w}$	The doctor's optimal booking threshold
w_O	The maximizer of function $\lambda_d(w)$
γ	The probability that a PFU appointment is canceled early enough to not be wasted
η	The no-show probability
b	The average balking rate of new and RFU appointments
θ	The slope parameter for balking rate
$P(w)$	The FUA balking rate given the threshold w
r	The average revenue per served appointment
h	The payment to the doctor per served appointment
ρ	The financial loss resulted from per-unit increment in the FUA balking rate
Π_{FB}	The optimal objective value of the first-best model
Π_{SB}	The optimal objective value of the second-best model under symmetric information
$R(w)$	The performance-based payment to the doctor

Green and Savin (2008), Guo et al. (2019) to ensure analytical tractability. Upon completion of an appointment, the patient may require a follow-up appointment (FUA) with the same doctor at a later time with probability p, which is a random variable drawn from a predefined distribution, characterized by a cumulative distribution function $F(\cdot)$ and a continuous density function $f(\cdot)$. We refer to this random variable p as the revisit probability, which follows an iid distribution $F(\cdot)$ regardless whether the preceding appointment is an episode-initiating appointment or an FUA. After the value of the revisit probability p is realized, the nature flips a biased coin with probability p of success to determine whether an FUA is generated or not. In case no FUA is generated, an episode of visits concludes.

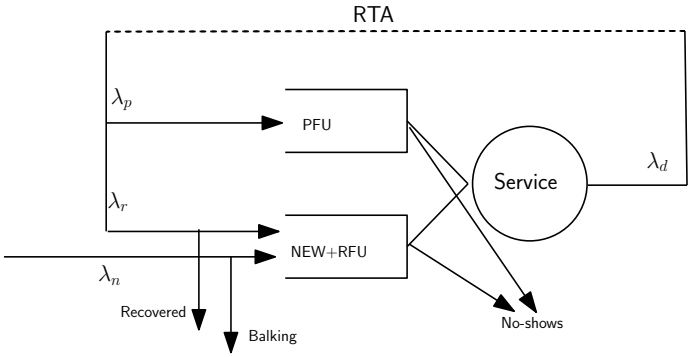
In our model, it is assumed that doctors can accurately estimate a patient's revisit probability at the end of each consultation. In other words, the doctor observes the realization of the random variable p before she decides on whether to book the potential FUA as a PFU or not. As argued in (Ding et al. 2023), it suffices to focus on *threshold-based PFU booking policies*, in which the doctor books a PFU for a patient if and only if p is greater

than a predetermined threshold $w \in [0, 1]$. Patients who do not have a PFU scheduled will book a RFU on their own if it becomes necessary later on.

Figure 1(a) depicts the typical process of scheduling FUAs, which we explain first. Figure 1(b) depicts an approximate model that we will discuss later.



(a) Priority Queue with Returning Customers



(b) An Approximate Model Using RTA

Figure 1 The RTA Approximation for a Priority Queue with Orbits (Ding et al. 2023)

Patients who complete their appointments are treated as "jobs" transitioning to either the PFU or the RFU orbit, based on the doctor's decision. These patients, or jobs, remain in their respective orbits for a stochastic observation period. Upon completion of this period, they may either recover, eliminating the need for further follow-up, or still require an FUA. If a patient in the RFU orbit recovers and no longer needs an FUA, then they exit the system. Such patients remain dormant until the next episode-initiating appointment. However, if they do not recover and an FUA is necessary, then they re-enter the queueing

system via the RFU queue. The dynamics for patients in the PFU orbit are different. If they recover, then they either proactively cancel the PFU appointment with a probability γ, or inadvertently fail to cancel with a probability of $1-\gamma$, resulting in spoilage. If they do not recover and need a FUA, then they return to the queueing system and join the PFU queue. The distinction between RFU and PFU recovered and returning patients helps model different likelihood of late cancellations or no-shows for these patients as well as different levels of priority when they return to the queueing system. Another difference is that PFU returning patients do not balk because they have high priority. Finally, for all non-canceled appointments, we assume that there is a probability of $\eta \in (0,1)$ that the patient may not show up for unanticipated reasons, such as emergencies, last-minute schedule changes, and traffic jams.

We assume that PFU requests have head-of-line priority over new (episode-initiating) visits and RFU requests in the queue. This is based on the premise that when a PFU request re-enters the queue from its orbit, it represents an appointment that was already booked during the previous visit, typically a few weeks earlier. Moreover, it is unlikely that a PFU appointment would be bumped in favor of a non-PFU appointment, justifying our treatment of PFU requests as having head-of-line priority in the queueing system.

In the literature, such a system is often referred to as a queue with returning customers. Analyzing these systems presents significant challenges, particularly in deriving closed-form expressions for the steady-state distribution of queue length. To circumvent these difficulties, we utilize the the RTA (Returning Customers See Time Averages) approximation Greenberg and Wolff (1987). Under the RTA assumption, PFUs and RFUs re-enter the queue according to time-homogeneous Poisson processes with mean arrival rates equal to the inverse of the average time spent in the orbit; see Figure 1(b). Consequently, the doctor's service queue has three independent Poisson arrival streams: new arrivals, RFUs, and PFUs, with rates denoted by λ_n, λ_r, and λ_p, respectively.

In addition, we introduce the notion of *average effective departure rate or service rate* and denote it by λ_d. The effective service rate λ_d only counts those patients who have actually been served, deliberately excluding individuals who scheduled appointments but subsequently canceled or were no-shows.

The conservation law applies to all patient flows. In particular, this means that the rate at which patients enter the orbits must equal their effective departure rate, resulting in

rate-balance equations that can be used to solve for parameters λ_d, λ_p and λ_r, for a given λ_d. Specifically, these rate balance equations are as follows:

$$\begin{aligned}\lambda_p &:= \lambda_d \int_w^1 (p + (1-p)(1-\gamma)) f(p) dp = \lambda_d \left[(1-\gamma)(1-F(w)) + \gamma(\bar{p} - G(w))\right] \\ \lambda_r &:= \lambda_d \int_0^w p f(p) dp = \lambda_d G(w),\end{aligned} \tag{3}$$

where $G(w) := \int_0^w f(p)\, dp$ denotes the average revisit probability for the subpopulation with $p \leq w$, and $\bar{p} := G(1)$ denotes the average revisit probability for all patients.

Let λ_v represent the *virtual arrival rate*, encompassing all appointment types: new appointments, RFUs, and PFUs. λ_v and λ_d differs in the amount of appointments that were late cancelled or no shows.

To account for the demand loss due to congestion within the system, both new and RFU patients may balk with probability b. In reality, patients may observe or be informed about the waiting time at the time of booking appointments and thus b may depend on the real-time queue length. However, with state-dependent balking, the steady-state waiting time has no closed-form expression, which makes the subsequent analysis of the principal-agent model very difficult. Therefore, for analytical tractability, we make an approximation by assuming that the balking rate b only depends on the steady-state expected sojourn time, W. This assumption is reasonable for health systems that only disclose average waiting time over a past period instead of the real-time waiting times.

To construct a relatively simple model that captures the influence of congestion on patient balking, we assume that the average balking rate b is a linear function of W with slope $\theta > 0$ until it reaches the upper limit 1. This assumption leads to the following expression,

$$b = \min\{\theta W, 1\} = \min\left\{\frac{\theta}{1-\lambda_v}, 1\right\}, \tag{4}$$

where $W = (1-\lambda_v)^{-1}$ is the steady-state average sojourn time in an M/M/1 queue. Knowing the expression for b, we next derive the following the rate-balance equation:

$$\begin{aligned}\lambda_d &:= (1-\eta)(\lambda_n + \lambda_r)(1-b) + (1-\eta)\lambda_d \int_w^1 p f(p) dp \\ &= (1-\eta)(\lambda_n + \lambda_d G(w))(1-b) + (1-\eta)(\bar{p} - G(w))\lambda_d.\end{aligned} \tag{5}$$

The above expression can be simplified as

$$\lambda_d = \frac{\lambda_n (1-\eta)(1-b)}{1 - (1-\eta)(\bar{p} - G(w)b)}, \tag{6}$$

where λ_n is a fixed system parameter that denotes the arrival rate of new patients, and b is a function of λ_v as given in (4).

Furthermore, we can express the virtual arrival rate λ_v as

$$\begin{aligned}\lambda_v(w,\lambda_d) &:= (\lambda_n+\lambda_r)(1-b)+\lambda_p \\ &= (\lambda_n+\lambda_d G(w))(1-b)+\lambda_d\left[(1-\gamma)(1-F(w))+\gamma(\bar{p}-G(w))\right] \\ &= \frac{\lambda_d}{1-\eta}+(1-\gamma)[1-F(w)-\bar{p}+G(w)]\lambda_d\end{aligned} \tag{7}$$

where the last equation follows from Equation (5).

Next, Lemma 1 states that the Equations (4)-(7) have a unique solution $\lambda_d \in (0,1)$. All proofs in this paper are provided in the Appendices.

LEMMA 1. *Given any fixed $w \in [0,1]$, there exists a unique vector $(\lambda_d, \lambda_v, b)$ that solves Equations (4)-(7) and has $\lambda_d, \lambda_v \in (0,1)$ and*

$$b = \frac{\theta}{1-\lambda_v(w)}. \tag{8}$$

The significance of Lemma 1 is that for a given $w \in [0,1]$, the values of λ_d, λ_v, and b are unique, and therefore can be expressed as functions of w. In fact, we can solve the expressions $\lambda_v(w)$, $\lambda_d(w)$, and d from equations (6), (7), and (8). Finally, we can first solve $\lambda_p(w)$ and $\lambda_r(w)$ from Equation (3), using the expressions of $\lambda_d(w)$.

We next investigate how $\lambda_d(w)$ changes with w, which provides key insights into the selection of the PFU booking threshold w.

LEMMA 2. *$\lambda_d(w)$ is quasi-concave in w. In addition, its maximizer $w_O := \arg\max_{w\in[0,1]} \lambda_d(w)$, is unique and characterized by the following equation,*

$$[1-(1-\eta)(\bar{p}-G(w_O))](1-\gamma)(1-w_O)\lambda_d = (1-\eta)(1-\theta-\lambda_v)w_O.$$

Unlike $\lambda_d(w)$, both $\lambda_v(w)$ and $b(w)$ monotonically decrease in w, as shown in the next Lemma.

LEMMA 3. *Both $\lambda_v(w)$ and $b(w)$ are non-increasing in $w \in [0,1]$. Moreover, if $\gamma < 1$ or $w > 0$, then both are strictly decreasing in w.*

The intuition behind Lemma 3 is that a larger PFU booking threshold w implies fewer PFUs will be booked and overall spoilage rate will decrease, which results in less congestion in the queue. Consequently, there are fewer patients balking.

The properties of the functions $\lambda_d(w)$, $\lambda_v(w)$, and $b(w)$ allow us to investigate the doctor's optimal choice of w under different incentive schemes, which is discussed in the next section.

5. Mechanism Design

We begin with calculating the health system's optimal PFU booking threshold in the first best model in which the health system fully observes and controls the doctor's effort level, which is captured by the PFU booking threshold w. We then formulate a second-best model in symmetric information case, in which the health system can no longer control the doctor's decision, but can infer the doctor's effort level by owning the same information as the doctor has. Finally, we formulate a second-best model in asymmetric information case in which the health system neither controls nor could infer the doctor's effort level.

5.1. First-Best Model

We consider a scenario where a health system's profit depends on the effective service rate λ_d and the COC level. In our model, COC is a key metric of care quality and may influence the health system's profit. For analytical tractability, we use the FUA balking rate as a measure for disruptions of COC, which has the following expression:

$$P(w) := \frac{b(w)G(w)(1-\eta)\lambda_d}{(1-\eta)\cdot\lambda_d\cdot\bar{p}} = \frac{b(w)\cdot G(w)}{\bar{p}}. \tag{9}$$

In the presented expression, the numerator $b(w)G(w)(1-\eta)\lambda_d$ quantifies the rate at which patients discontinue their follow-up consultations with the same doctor. The denominator $(1-\eta)\cdot\lambda_d\cdot\bar{p}$, on the other hand, calculates the arrival rate of effective FUAs. Thus, the ratio captures the percentage of balkings out of the total FUAs. This computation is conducted on a per-appointment basis, distinguishing it from the SECON index that calculates the balking rate on an episodic basis. Despite the technical difference, both metrics aim to assess the frequency at which the continuity of provider care is interrupted. For our modeling analysis, we use the FUA balking rate instead of the SECON index due to its simplicity in facilitating analysis. The next lemma shows that the FUA balking rate $P(w)$ is monotonically increasing in w.

Lemma 4. *$P(w)$ is increasing in w.*

Lemma 4 states that the FUA balking rate $P(w)$ increases as the threshold w increases. This implies that setting a higher threshold for scheduling PFU appointments leads to a reduction in the number of PFUs, thereby diminishing the Continuity of Care (COC).

We can then formulate the health system's decision problem as

$$\Pi_{FB} := \max_{w\in[0,1]} \ (r-h)\cdot\lambda_d(w) - \rho\cdot b(w)\cdot G(w)/\bar{p}, \tag{10}$$

where r represents the average revenue per served appointment, h represents the payment to the doctor per served appointment, and ρ signifies the monetary cost from per-unit increment in the FUA balking rate. This cost, ρ, may be determined by the health system based on its evaluation of the potential financial loss related to disruptions in provider care continuity and the resulting inferior care quality. Alternatively, in certain scenarios, ρ might reflect direct financial penalties as specified by pay-for-performance policies enacted by healthcare payers. This formulation allows the health system to consider the optimal level of effort that should be exerted (represented by w) to balance revenue maximization against the costs incurred from decreased COC due to patient balking at follow-up appointments.

We refer to the optimal solution to (10) as the health system's *first-best solution*, and denote it by w_{FB}. We call it first-best because it is the ideal PFU booking threshold under centralized decision making and no informational asymmetry. In particular, when $\rho=0$, the health system is solely maximizing the throughput rate $\lambda_d(w)$ and we let $w_O := \arg\max_{w\in[0,1]}\lambda_d(w)$ denote the throughput-maximizing PFU booking threshold. Whereas, as ρ increases, the health system is more and more concerned about minimizing $P(w)$, which requires to lower w due to Lemma 4. Therefore, the health system face conflicting objectives when choosing w over the interval $[0, w_O]$. The next proposition proves monotonic properties of w_{FB} with respect to ρ.

Proposition 1. *The optimal solution $w_{FB}(\rho)$ is decreasing with respect to the parameter ρ. Consequently, $w_{FB}(\rho) < w_O$ for all $\rho > 0$. Furthermore, the set of values of ρ for which $w_{FB}(\rho)$ has more than one maximizers is countable and has a measure of zero.*

Proposition 1 elucidates that if a health system places greater emphasis on improving COC, it should choose a lower PFU booking threshold w_{FB}, thereby scheduling more PFUs. This conclusion aligns with our empirical finding that higher COC levels are associated with increased ratio of PFUs. Such a scheduling approach is critical for fostering improved care continuity. Further insights along these lines will be explored in the analysis of the second-best model, which more closely mirrors real-world conditions and practices.

5.2. Second-Best Model under Symmetric Information

In practice, the health system cannot estimate a patient's revisit probability, and thus cannot decide whether to book a PFU or not. Instead, the PFUs are recommended by the doctor whose objective may differ from the clinic's, making it difficult to implement the first-best solution directly.

Because doctors are paid on a FFS basis, they would choose the throughput maximizing PFU booking threshold w_O to maximize their revenue. Therefore, the doctor's optimal decision is to book a PFU if and only if the patient's estimated revisit probability $p \geq w_O$. However, if the health system is interested in promoting provider continuity and reducing FUA balking rate, i.e., whenever $\rho > 0$, Proposition 1 implies that the health system's desired PFU booking threshold is $w_{FB}(\rho) < w_O$. That is, the health system would choose a lower PFU booking threshold than the doctor. The doctor's objective is thus not aligned with that of the health system.

Suppose the health system knows the distribution of patient revisit probability $G(\cdot)$. In this case, the health system can infer the threshold w used by the doctor by observing appointment data such as λ_d, λ_v, λ_r, and λ_p. From this reasoning, in our principal-agent model, the agent's effort level w is observable by the principal. This allows the principal to implement an incentive contract with two components. The doctor receives a volume-based payment $h\lambda_d$ as well as a performance-based payment $R(w)$, where $R(w)$ is decreasing in w so as to incentivize the agent to choose a smaller w and thereby to book more PFUs.

We assume that the health system selects the performance-based payment $R(\cdot) : [0,1] \to \mathbb{R}$ to maximize its objective; while the volume-based payment has a fixed rate h. The latter assumption is without loss of generality, because if the health system pays a different rate h', then the difference $(h' - h)\lambda_d(w)$ is a function of w can always be incorporated into $R(w)$. The second-best problem can be formulated as follows.

$$\Pi_{SB} := \max_{R(\cdot)\in[0,1]^{\mathbb{R}}} \quad (r-h)\lambda_d(\hat{w}) - R(\hat{w}) - \rho \cdot b(\hat{w}) \cdot G(\hat{w})/\overline{p} \tag{11}$$

$$s.t. \qquad R(w) \geq 0, \quad \forall w \in [0,1] \tag{12}$$

$$\hat{w} \in \arg\max_{w\in[0,1]} h\lambda_d(w) + R(w), \tag{13}$$

where (12) is the individual rationality (IR) constraint, which ensures the agent's participation by promising to pay them no less than their regular salary, which could be regarded

as the agent's outside option; (13) is the incentive compatibility (IC) constraint, which ensures optimality of the agent's action $\hat{w}$. We use Π_{SB} to denote the optimal objective values for the second-best formulation.

To solve the second best problem (11)-(13), we consider the following payment scheme. For a given $z \in [0,1]$, we define

$$\phi^z(w) := \begin{cases} 0 & \text{if } \ w \geq z \\ h\lambda_d(w_O) - h\lambda_d(z) & \text{if } \ w \leq z, \end{cases} \tag{14}$$

where $w_O := \arg\max \lambda_d(w)$ is the throughput maximizing threshold. Since health system prefers the doctor to adopt an even smaller w than w_O to further improves COC, it introduces a performance-based payment scheme, $\phi^z(\cdot)$, whereby a doctor receives a bonus payment $h\lambda_d(w_O) - h\lambda_d(z)$ by choosing a PFU booking threshold $w \leq z$; and receives no payment otherwise.

The incentive scheme characterized by $\phi^z(\cdot)$ is commonly implemented to calculate year-end bonuses for salespeople and managers, rewarding employees only when they surpass a specific performance benchmark, like a sales quota or a target level of operational efficiency, within an observation cycle. This approach is analogous to incentive structures studied in the literature, including works by Oyer (1998) and Lazear and Shaw (2007), highlighting the effectiveness of performance-based rewards in aligning individual actions with organizational goals.

The next proposition shows that the second best can be achieved through a payment scheme in the form of $\phi^z(\cdot)$, which is the minimum compensation that the principal must offer the agent to motivate the adoption of a PFU z that is lower than w_O. This strategic payment formula is designed to bridge the gap between the optimal and actual practices by financially encouraging physicians to prioritize COC in their scheduling decisions.

Proposition 2. *The second-best optimization problem* (11)-(13) *is solved by a payment scheme* $R(\cdot) = \phi^{w_{SB}}(\cdot)$*, where*

$$w_{SB} := \arg\max_{w \in [0,1]} \quad r\lambda_d(w) - \rho \cdot b(w) \cdot G(w)/\overline{p}. \tag{15}$$

According to Proposition 2, to solve the second-best problem (11)-(13), it suffices to perform a one-dimensional search for the maximizer of (15). The resulting w_{SB} is the PFU

booking threshold that the agent chooses under the payment scheme $\phi^{w_{SB}}(\cdot)$, at which the second best is achieved.

We next derive a lower bound for the second-best optimal value Π_{SB}, which further yields a closed-form upper bound for the gap $\Pi_{FB} - \Pi_{SB}$. To that end, we notice that if a performance-based payment scheme $\phi^{w_{FB}}(\cdot)$ is implemented in addition to a volume based incentive of $h\lambda_d(w)$, then the doctor will will voluntarily choose the first-best PFU booking threshold w_{FB}. In this case, the health system has to pay an extra $\phi^{w_{FB}}(w_{FB})$ compared to the amount that she pays in the first best case. As a result, $\phi^{w_{FB}}(w_{FB})$ provides an upper bound for the gap $\Pi_{FB} - \Pi_{SB}$ as the second-best model could potentially do better by adopting a payment scheme $\phi^{w_{SB}}(\cdot)$ instead of $\phi^{w_{FB}}(\cdot)$. A formal statement is presented in the following proposition.

PROPOSITION 3. *If the health system sets $R(w) = \phi^{w_{FB}}(\cdot)$, then the doctor will choose $\hat{w} = w_{FB}$. As a result,*

$$\Pi_{FB} - \Pi_{SB} \leq \phi^{w_{FB}}(w_{FB}) = h\lambda_d(w_O) - h\lambda_d(w_{FB}).$$

Furthermore, w_{SB} decreases in ρ.

Proposition 3 shows that when ρ increases, the PFU booking threshold w_{SB} decreases, resulting in a larger proportion of PFUs. Therefore, in both the first-best and the second-best scenario, our analytical results are consistent with the earlier empirical result revealing a positive correlation between elevated levels of COC and a higher proportion of PFUs.

When ρ approaches zero, we have $w_{FB} \to w_O$ and consequently $\phi^{w_{FB}}(w_{FB}) \to 0$ by its definition. Therefore, when ρ approaches zero, the relative gap $(\Pi_{FB} - \Pi_{SB})/\Pi_{FB}$ approaches zero. This is because when the health system is less concerned with COC, the health system's and the doctor's interests are closely aligned.

Finally, we want to compare the values of w_O, w_{FB}, and w_{SB}. Since the effective throughput rate $\lambda_d(w)$ is quasi-concave whereas the FUA balking rate $P(w)$ is strictly decreasing in w, the throughput-maximizing threshold w_O, is identified as having the largest value. w_{FB}, as the solution to the first best model, has the lowest value. This is because, in the first best scenario, the reward from minimizing $P(w)$ has been fully integrated into the decision-making process. Whereas in the second-best model, the reward from minimizing $P(w)$ is only partially incorporated into the doctor's objective, resulting in a w_{SB} positioned between w_{FB} and w_O. The comparison is summarized in the next Corollary. Its proof immediate follows the above logic and is omitted.

COROLLARY 1. If $\rho > 0$, then $w_{FB} < w_{SB} < w_O$.

The primary takeaway of Corollary 1 is that in a practical health system, which closely aligns with the second-best scenario, the proportion of PFUs should be higher than in a scenario without intervention (w_O), yet lower than in a fully coordinated system (w_{FB}).

5.3. Second-Best Model under Asymmetric Information

This section deals with the scenario in which the doctors can infer the distribution of their patients' revisit probability ($G(\cdot)$) based on knowledge of her patients' clinical histories. In contrast, while the health system can observe the appointment booking rates and service rates, i.e., λ_d and λ_v, it neither observes the PFU booking threshold w, nor the revisit probability distribution $G(\cdot)$. In practice, the health system might eventually approximate the distribution $G(\cdot)$, using extensive historical data, provided that the health system operates under steady conditions. However, in scenarios where the health system is newly established or does not have access to comprehensive data within a health network, accurately estimating $G(\cdot)$, or even other environmental parameters such as θ, becomes impractical.

This information asymmetry presents a significant hurdle in executing the contract proposed in Section 5.2. The inability to observe w directly, compounded by the lack of knowledge regarding $G(\cdot)$, prevents the health system from quantifying the performance based payment $R(\cdot)$. This information asymmetry thus necessitates alternative approaches for the health system to compensate the doctor. One alternative is to compensate the agent based on observable proxy measures instead of the unobservable effort level w. The set of possible observable measures includes the following candidates: λ_d, λ_v, $\lambda_d^R := (1-\eta) \cdot \lambda_d \cdot G(w) \cdot [1 - b(w)]$, and $\lambda_d^P := (1-\eta) \cdot \lambda_d \cdot (\bar{p} - G(w))$. Whereas λ_d and λ_v have been defined earlier, the two additional rates can be explained as follows.

- λ_d^R is the effective service rate of RFUs. The health system can identify RFUs using patients IDs and visit histories. The health system can further identify which RFUs were effectively served by looking into the appointment records.
- λ_d^P is the effective service rate of PFUs. The PFUs can be identified by matching the appointment booking time and patients previous visit time by patient ID. A PFU is effectively served if it was neither a cancellation nor a no show. Such information is available in the appointments data.

The functional form of the incentive payment R will then depend on the rates identified above. We analyze four functional forms as shown below:

$$\begin{aligned}
&\textbf{[FUA-Ratio]} \quad R_{FR}(w) := \frac{\lambda_d^R(w) + \lambda_d^P(w)}{\lambda_d(w)} \\
&\textbf{[FUA-Count]} \quad R_{FC}(w) := \lambda_d^R(w) + \lambda_d^P(w) \\
&\textbf{[PFU-Ratio]} \quad R_{PR}(w) := \frac{\lambda_d^P}{\lambda_d} \\
&\textbf{[PFU-Count]} \quad R_{PC}(w) := \lambda_d^P
\end{aligned} \tag{16}$$

Our comparison of the four types of contracts is first carried out under a bi-objective optimization framework in which the health system is assumed to maximize two objectives: (1) profit, which is the difference between the revenue and the financial compensation paid to the doctor; (2) COC, measured by the negative of FUA balking rate $-P(W)$. This leads to the following formulation for the four contracts $R_t(w)$ with $t \in \{\text{FR}, \text{FC}, \text{PR}, \text{PC}\}$.

$$\max_{c \geq 0} \quad (r-h)\lambda_d(\hat{w}(c)) - cR_t(\hat{w}(c)) \tag{17}$$

$$\min_{c \geq 0} \quad b(\hat{w}(c))G(\hat{w}(c))/\overline{p} \tag{18}$$

$$s.t. \quad R(w) \geq 0, \ \forall w \in [0,1] \tag{19}$$

$$\hat{w}(c) \in \arg\max_{w \in [0,1]} h\lambda_d(w) + cR_t(w), \tag{20}$$

where (19) denotes the individual rationality (IR) constraint and (20) denotes the incentive-compatibility (IC) constraint, which are the same as the constraints in the symmetric information case.

Our approach involves outlining the Pareto-frontier for each type of contracts on the profit-COC spectrum and then compare these frontiers through Pareto dominance to identify the most effective contract type. For a given contract type t, we calculate a pair of objective values $((r-h)\lambda_d(\hat{w}(c)) - cR_t, P(\hat{w}(c)))$ for each coefficient $c \geq 0$. By varying c, we generate a set of points $\{((r-h)\lambda_d(\hat{w}(c)) - cR_t, P(\hat{w}(c)))|c \geq 0\}$ on the profit-COC spectrum, which constitutes the Pareto-frontier of contract type t. To assess and compare the efficacy of different contract types, we introduce a partial order $\succeq$, where for any two contract types $t_1, t_2 \in \{\text{FR}, \text{FC}, \text{PR}, \text{PC}\}$, we say $t_1 \succ t_2$ if and only if the Pareteo frontier

of contract R_{t_1} dominates that of contract R_{t_2}. Mathematically, it means that there exists $c_1, c_2 \geq 0$, such that

$$P(\hat{w}(c_1)) = P(\hat{w}(c_2)) \text{ and } (r-h)\lambda_d(\hat{w}(c_1)) - c_1 R_{t_1}(\hat{w}(c_1)) \geq (r-h)\lambda_d(\hat{w}(c_2)) - c_2 R_{t_2}(\hat{w}(c_2)), \tag{21}$$

where $\hat{w}(c)$ is the solution to the IC condition (20). The following theorem compares the four types of contracts with respect to the above partial order.

THEOREM 1. *(1)* $\boldsymbol{PR \succeq FR \succeq FC}$*; (2)* $\boldsymbol{PR \succeq PC \succeq FC}$*.*

Theorem 1 suggests that the Pareto-frontier of the PFU ratio-based contract (**PR**) dominates all other contracts, while the FUA count-based contract (**FC**) is dominated by all other contracts. The performances of contracts **PC** and **FR** lie in the middle, whereas there is no dominance relationship between these two contracts as shown in Figure 2.

Figure 2 also illustrates that all four contracts are Pareto-dominated by the frontier of the second best model with symmetric information (11) and the first best model (10), and the gap could be significant when emphasis is placed on improving COC. Notably, the gap between the first best model and the second best model with symmetric information is much smaller than the gap between the second best model and the best performed contract (**PR**), suggesting information asymmetry as a major challenge in incentivizing doctors towards the optimal PFU booking practices.

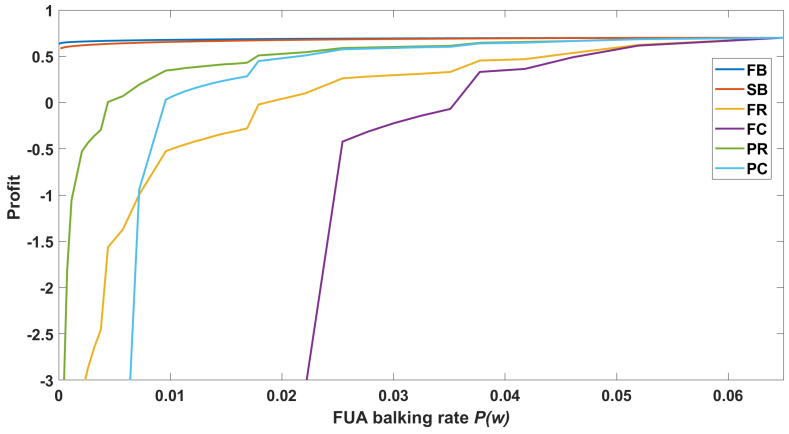


Figure 2 Pareto Frontiers for the first best, second best, and the four types of contracts ($f(\cdot) = Beta(0.5, 0.5)$, $r = 2$, $h = 1$, $\eta = 0.1$, $\gamma = 0.5$,$\lambda_n = 0.6$, $\theta = 0.05$)

In general, it is widely recognized that booking more FUA can enhance Continuity of Care. For instance, the Royal College of General Practitioners (RCGP) Guidelines for

Continuity of Care suggest that clinicians can negotiate with patients and proactively schedule FUAs to improve COC[1]. Similar arguments have been made in the context of resident training clinics, where researchers have found that enhancing scheduling support significantly increases rates of resident continuity(LaVine et al. 2020). However, how to effectively incentivize doctors to improve provider continuity of care and to maximize the clinic's profit remains unclear. Theorem 1 elucidates that incentivizing doctors through PFUs rather than general FUAs, and focusing on the relative ratio rather than the total counts, emerges as a more effective strategy.

We will then extend the result from Theorem 1 and compare the four contract types under a single-objective optimization framework. The formulation of the single objective follows (10) and (11), in which the health system is assumed to pay a penalty $\rho P(w)$ proportional to the FUA balking rate for a given parameter $\rho \geq 0$. Mathematically, for each contract type $t \in \{\text{FR}, \text{FC}, \text{PR}, \text{PC}\}$, the health system solves the following single-objective optimization problem:

$$\Pi_t := \max_{c \geq 0} \quad (r-h)\lambda_d(\hat{w}(c)) - cR_t(\hat{w}(c)) - \rho P(\hat{w}(c)), \tag{22}$$

$$s.t. \quad (19),\ (20). \tag{23}$$

For each fixed $\rho \geq 0$, We can compare the optimal objective values Π_t of the four contract types; see the next Corollary. The comparison result aligns with Theorem 1.

COROLLARY 2. (1) $\Pi_{PR} \geq \Pi_{FR} \geq \Pi_{FC}$; (2) $\Pi_{PR} \geq \Pi_{PC} \geq \Pi_{FC}$.

Furthermore, we can derive a constant upper bound for the gap between the first-best objective value Π_{FB} and that under a FUA-ratio based contract. Since $\Pi_{PR} \geq \Pi_{FB}$, this constant upper bound trivially applies to the gap $\Pi_{FB} - \Pi_{PR}$.

THEOREM 2. $\Pi_{FB} - \Pi_{FR} \leq h\rho/(r-h)$.

The main idea of the proof of Theorem 2 is to select $c = \frac{h\rho}{(r-h)(1-\eta)\bar{p}}$ in a FUA-ratio based contract, which yields a threshold $\hat{w}(c) = w_{FB}$ according to the IC condition. In other words, with the above c, the FUA-ratio based contract aligns the agent's effort level with that in the first best case. The resulting right-hand-side of (22) thus differs from Π_{FB} by

[1] https://www.rcgp.org.uk/getmedia/d77f39b7-3745-4942-acef-20f4a3118c31/RCGP-continuity-of-care-guide-141119.pdf

$\frac{h\rho}{r-h}$, which gives an upper bound for $\Pi_{FB} - \Pi_{FR}$ as the latter may do better by selecting a different c.

The result of Theorem 2 implies that when the penalty for care discontinuity, ρ, is small, achieving the first best outcome through a FUA-ratio based contract is feasible. This aligns with our earlier observation that aligns the interests of the health system and the doctor becomes more costly as the penalty for care discontinuity rises. Remarkably, in the hypothetical scenario where $\rho = 0$, the interests of the two parties can be perfectly aligned. Under such circumstances, the FUA-ratio based contract is capable of realizing the first best outcome. Moreover, when the ratio $h/(r-h)$ is low, it signifies a larger profit margin for the health system. This economic leverage makes it more cost-effective to offer financial incentives to physicians for scheduling more PFUs, thereby narrowing the gap between the current operational model and the ideal first-best scenario. This result underscores the nuanced interplay between financial structures, incentive strategies, and healthcare outcomes, emphasizing the critical role of contract design in enhancing the alignment of objectives within healthcare systems and improving overall care continuity.

6. Sensitivity Analyses

In this section, we conduct a series of numerical experiments to validate the effectiveness of the proposed contracts. Following Ding et al. (2023), in the base case setting, we assume $\lambda_n = 0.6$ and $f(\cdot) = Beta(0.5, 0.5)$, thereby resulting in $\bar{p} = 0.5$. Other parameters are set as follows: $r = 2$, $h = 1$, $\eta = 0.1$, $\gamma = 0.5$, and $\theta = 0.05$.

We first compare the optimal values of the first-best model Π_{FB}, the optimal value of the second best model Π_{SB}, and a lower bound for the second best model $\underline{\Pi}_{SB}$, which is obtained by using a payment contract $\phi^{w_{FB}}(w_{FB})$, across various values of ρ. The results are plotted in Figure 3.

Figure 3 reveals a consistent downward trend in the objective values for all three models as the parameter ρ increases. This behavior stems from the heightened penalty imposed on the FUA balking rate, resulting more PFUs being booked and a lower effective throughput rate as outlined in Proposition 1 for all three models. Additionally, the narrow gap between Π_{SB} and $\underline{\Pi}_{SB}$ suggests that the contract $\phi^{w_{FB}}(w_{FB})$ serves as a reliable approximation for the second-best model, showcasing its effectiveness.

Next, we study how the relative gap between the first-best model and the second-best model, $(\Pi_{FB} - \Pi_{SB})/\Pi_{FB}$, changes with key parameters such as ρ, λ_n and θ. The parameter

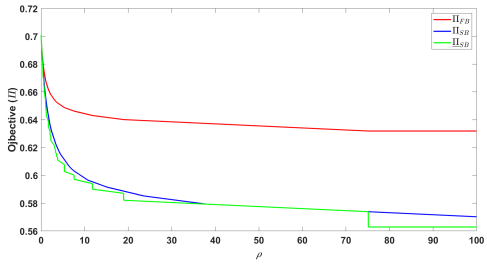
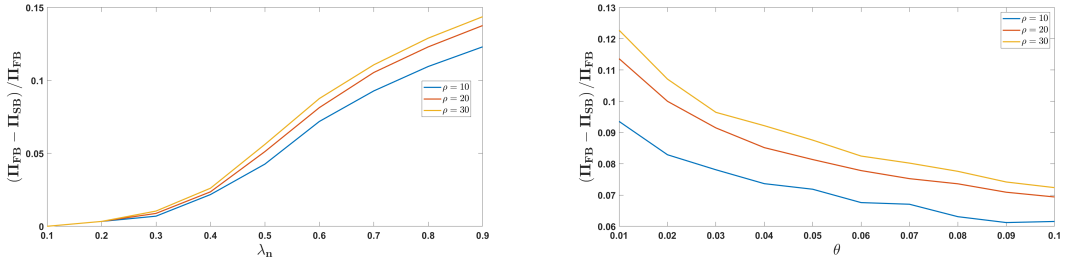
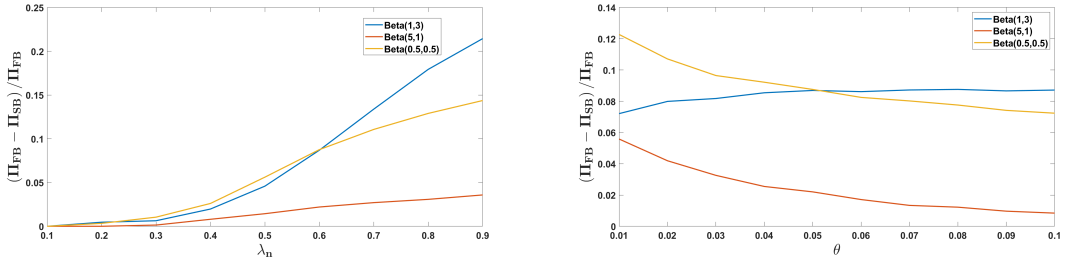


Figure 3 Π_{FB}, Π_{SB}, and $\underline{\Pi}_{SB}$, as a function of ρ ($f(\cdot) = Beta(0.5, 0.5)$, $r = 2$, $h = 1$, $\eta = 0.1$, $\gamma = 0.5$, $\lambda_n = 0.6$, $\theta = 0.05$)



(a) variation to λ_n ($\theta = 0.05$) (b) variation to θ ($\lambda_n = 0.6$)

Figure 4 Relative gap $(\Pi_{FB} - \Pi_{SB})/\Pi_{FB}$ for parameters ($f(\cdot) = Beta(0.5, 0.5)$, $r = 2$, $h = 1$, $\eta = 0.1$, $\gamma = 0.5$)

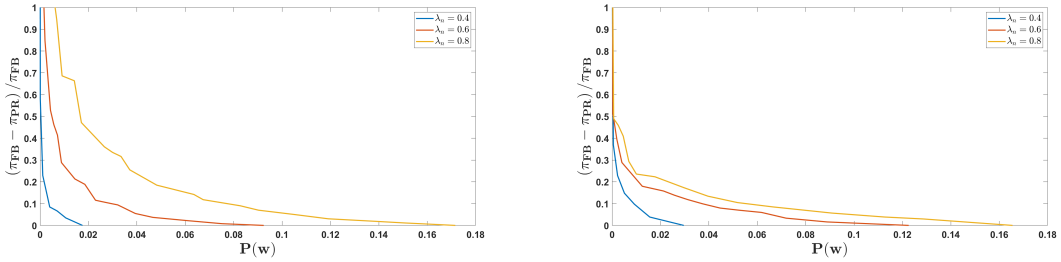


(a) variation to λ_n ($\theta = 0.05$) (b) variation to θ ($\lambda_n = 0.6$)

Figure 5 Relative gap $(\Pi_{FB} - \Pi_{SB})/\Pi_{FB}$ for different Beta distribution ($r = 2$, $h = 1$, $\eta = 0.1$, $\gamma = 0.5$, $\rho = 30$)

configuration remains consistent with that in Figure 3, except for the varying parameter. The comparative results are presented in Figure 4.

According to Figure 4, when the health system is more concerned with promoting COC, it is more costly to align the health system's and the doctor's interests. Similarly, when the system is more congested (larger λ_n and less θ), the alignment is also more costly. The explanation is that in a more congested system, it is more costly to incentivize the doctor to book more PFUs due to the increased opportunity value of an appointment slot.



(a) $Beta(1,3)$ (b) $Beta(5,1)$

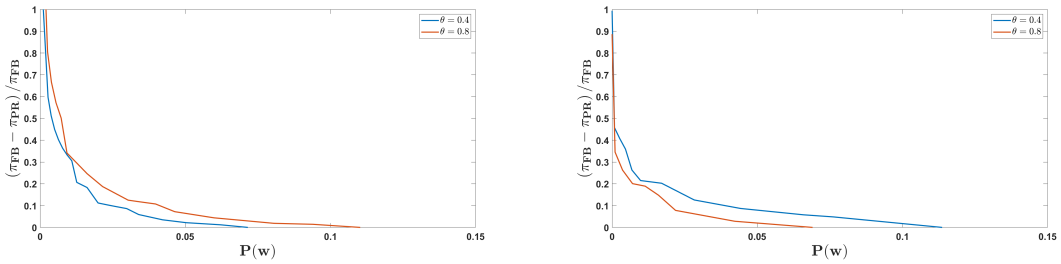
Figure 6 Relative gap $(\pi_{FB} - \pi_{PR})/\pi_{FB}$ as a function of FUA balking rate $P(w)$ for different λ_n ($r = 2$, $h = 1$, $\eta = 0.1$, $\gamma = 0.5$, $\theta = 0.05$)

We then investigate whether similar trend of the relative gap concerning parameters λ_n and θ holds across different distributions of revisit probability $F(\cdot)$. We fixed $\rho = 30$ and test two different distributions, i.e., $Beta(5,1)$ and $Beta(1,3)$. The results are depicted in Figure 5. It is evident from subfigure (a) of Figure 5 that the gap related to parameter λ_n is robust across different distributions. However, as demonstrated in subfigure (b) of Figure 5, the monotonic property does not exhibit robustness concerning the parameter θ.

Next, we compare the four contract types developed under the asymmetric information case. To begin, we compare the profit-FUA balking rate frontier for these contracts. Using the same parameter settings as in Figure 3, the frontiers of the first-best model, second-best model, and the four contracts in the asymmetric information case are presented in Figure 2 in Section 5. Figure 2 indicates that among the four proposed contracts, the contract **PR** stands out as the most favorable, while the contract **FC** is comparatively less effective, as outlined in Theorem 1.

We then proceed to compare the relative profit gap between the first-best model and the contract **PR** with a given FUA balking rate $P(w)$. Let π_{FB} and π_{PR} denote the profits for the first-best model and the **PR** contract, respectively. The relative profit gap is defined as $\frac{\pi_{FB} - \pi_{PR}}{\pi_{FB}}$. The results are illustrated in Figure 6 and Figure 7.

Figure 6 demonstrates that the smaller λ_n (larger $P(w)$), the smaller relative gap between the first-best model and the contract **PR**. These results hold consistently across different distributions $f(\cdot)$. However, Figure 7 reveals varying patterns concerning the parameter θ under different distribution functions $f(\cdot)$. In summary, both Figure 6 and Figure 7 indicate that the relative gap remains acceptable when the FUA balking rate is not extremely small.



(a) $Beta(1,3)$ (b) $Beta(5,1)$

Figure 7 Relative gap $(\pi_{FB} - \pi_{PR})/\pi_{FB}$ as a function of FUA balking rate $P(w)$ for different θ ($r = 2$, $h = 1$, $\eta = 0.1$, $\gamma = 0.5$,$\lambda_n = 0.6$)

7. Conclusion

In this study, we explored the integration of Continuity of Care (COC) into the early reservation of follow-up appointments, underscoring its significance in outpatient care management. Through empirical analysis, we established that the rate of prioritized follow-up (PFU) appointments serves as a reliable indicator of COC. Addressing the challenge of COC observability in decentralized systems, we developed various principal-agent models to encourage physicians to schedule an optimal number of PFUs, thereby enhancing COC. In these models, the health system (principal) offers a fee-for-service payment as well as a performance-based payment, while the physician (agent) determines the quantity of PFU appointments.

Our findings reveal that, in scenarios with symmetric information where both parties are aware of the revisit probability distribution, we can identify an optimal contract that approximates the second-best solution and suggest another that aligns the agent's actions with the first-best outcome. This suggests that carefully designed incentives can guide physicians towards decisions preferred by the health system. In cases of asymmetric information, where the distribution is known only to the agent, we examined four types of performance-based contracts. The analysis showed that contracts focused on PFUs outshine those based on general follow-up appointments (FUAs), and contracts emphasizing the ratio of PFUs are more effective than those based on absolute counts, highlighting the superiority of PFU ratio-based contracts.

This research, however, is not without limitations. We initially assume exponential service times, a simplification that future work could expand upon by considering more general service durations. Additionally, our current model focuses on a single-server setting for

contract design to improve COC, which could be extended to examine competitive servers within the principal-agent framework. Lastly, we model the balking rate based on the expected steady-state queue length rather than real-time queue lengths, an assumption that could be revisited for more dynamic modeling.

References

Adida, E., Bravo, F., and Science, M. Contracts for healthcare referral services: Coordination via outcome-based penalty contracts. *Management Science*, 65(3):1322–1241, 2019.

Amir, R. Supermodularity and complementarity in economics: An elementary survey. *Southern Economic Journal*, 71(3):636–660, 2005.

Arifolu, K., Ren, H., and Tezcan, T. Hospital readmissions reduction program does not provide the right incentives: Issues and remedies. *Management Science*, 67(4):2191–2210, 2020.

Armony, M. and Maglaras, C. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research*, 52(2):271–292, 2004.

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

Atlas, S. J., Grant, R. W., Ferris, T. G., Chang, Y., and Barry, M. J. Patient–physician connectedness and quality of primary care. *Annals of internal medicine*, 150(5):325–335, 2009.

Campello, F., Ingolfsson, A., and Shumsky, R. A. Queueing models of case managers. *Management Science*, 63(3):882–900, 2017.

Cayirli, T., Veral, E., and Rosen, H. Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, 9(1):47–58, 2006.

Chan, C. W., Yom-Tov, G., and Escobar, G. When to use speedup: An examination of service systems with returns. *Operations Research*, 62(2):462–482, 2014.

De Maeseneer, J. M., De Prins, L., Gosset, C., and Heyerick, J. Provider continuity in family medicine: does it make a difference for total health care costs? *The Annals of Family Medicine*, 1(3):144–148, 2003.

Ding, Y., Gupta, D., and Tang, X. Early reservation for follow-up appointments in a slotted-service queue. *Operations Research*, 71(3):917–938, 2023.

Dobson, G., Tezcan, T., and Tilson, V. Optimal workflow decisions for investigators in systems with interruptions. *Management Science*, 59(5):1125–1141, 2013.

Dreiher, J., Comaneshter, D. S., Rosenbluth, Y., Battat, E., Bitterman, H., and Cohen, A. D. The association between continuity of care in the community and health outcomes: a population-based study. *Israel journal of health policy research*, 1(1):1–12, 2012.

Eriksson, E. A. and Mattsson, L.-G. Quantitative measurement of continuity of care: measures in use and an alternative approach. *Medical care*, pages 858–875, 1983.

Feldman, J., Liu, N., Topaloglu, H., and Ziya, S. Appointment scheduling under patient preference and no-show behavior. *Operations Research*, 62(4):794–811, 2014.

Gallucci, G., Swartz, W., and Hackerman, F. Brief reports: Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services*, 56(3):344–346, 2005.

Green, L. V. and Savin, S. Reducing delays for medical appointments: A queueing approach. *Operations Research*, 56(6):1526–1538, 2008.

Greenberg, B. S. and Wolff, R. W. An upper bound on the performance of queues with returning customers. *Journal of Applied Probability*, pages 466–475, 1987.

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

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

Haggerty, J. L., Reid, R. J., Freeman, G. K., Starfield, B. H., Adair, C. E., and McKendry, R. Continuity of care: a multidisciplinary review. *BMJ: British Medical Journal*, 327(7425):1219, 2003.

Harder, V. S., Stuart, E. A., and Anthony, J. C. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, 15(3): 234–249, 2010.

Huang, J., Carmeli, B., and Mandelbaum, A. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, 63(4):892–908, 2015.

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

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

Jiang, R., Shen, S., and Zhang, Y. Integer programming approaches for appointment scheduling with random no-shows and service durations. *Operations Research*, 65(6):1638–1656, 2017.

Kikano, G. E., Flocke, S. A., Gotler, R. S., and Stange, K. C. 'my insurance changed': the negative effects of forced discontinuity of care. *Family practice management*, 7(10):44, 2000.

Kong, Q., Lee, C.-Y., Teo, C.-P., and Zheng, Z. Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research*, 61(3):711–726, 2013.

Kong, Q., Li, S., Liu, N., Teo, C.-P., and Yan, Z. Appointment scheduling under time-dependent patient no-show behavior. *Management Science*, 66(8):3480–3500, 2020.

Kostami, V. and Ward, A. R. Managing service systems with an offline waiting option and customer abandonment. *Manufacturing & Service Operations Management*, 11(4):644–656, 2009.

LaVine, N. A., Coletti, D. J., Verbsky, J., and Block, L. Enhanced scheduling support to improve continuity of care in a resident training clinic. *Journal of Graduate Medical Education*, 12(2):208–211, 2020.

Lazear, E. P. and Shaw, K. L. Personnel economics: The economist's view of human resources. *Journal of Economic Perspectives*, 21(4):91–114, December 2007. doi: 10.1257/jep.21.4.91.

Li, X., Wang, J., Luo, K., Liang, Y., and Wang, S. Exploring the spillover effects of urban renewal on local house prices using multi-source data and machine learning: The case of shenzhen, china. *Land*, 11(9): 1439, 2022.

Liu, N., Ziya, S., and Kulkarni, V. G. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management*, 12(2):347–364, 2010.

Luo, J., Kulkarni, V. G., and Ziya, S. Appointment scheduling under patient no-shows and service interruptions. *Manufacturing & Service Operations Management*, 14(4):670–684, 2012.

Nutting, P. A., Goodwin, M. A., Flocke, S. A., Zyzanski, S. J., and Stange, K. C. Continuity of primary care: to whom does it matter and when? *The Annals of Family Medicine*, 1(3):149–155, 2003.

Oyer, P. Fiscal year ends and nonlinear incentive contracts: The effect on business seasonality. *The Quarterly journal of economics*, 113(1):149–185, 1998. ISSN 0033-5533.

RCGP. Rcgp guidelines for continuity of care. 2019.

Robinson, L. W. and Chen, R. R. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing & Service Operations Management*, 12(2):330–346, 2010.

Rogers, J. and Curtis, P. The concept and measurement of continuity in primary care. *American journal of public health*, 70(2):122–127, 1980.

Sans-Corrales, M., Pujol-Ribera, E., Gene-Badia, J., Pasarín-Rua, M. I., Iglesias-Pérez, B., and Casajuana-Brunet, J. Family medicine attributes related to satisfaction, health and costs. *Family practice*, 23(3): 308–316, 2006.

Saultz, J. W. and Albedaiwi, W. Interpersonal continuity of care and patient satisfaction: a critical review. *The Annals of Family Medicine*, 2(5):445–451, 2004.

Steinbauer, J. R., Korell, K., Erdin, J., and Spann, S. J. Implementing open-access scheduling in an academic practice. *Family practice management*, 13(3):59–64, 2006.

Steinwachs, D. M. Measuring provider continuity in ambulatory care: an assessment of alternative approaches. *Medical care*, pages 551–565, 1979.

Sualtz and John, W. Defining and measuring interpersonal continuity of care. *Annals of Family Medicine*, 1(3):134–143, 2003.

Van Servellen, G., Fongwa, M., and D'Errico, E. M. Continuity of care and quality care outcomes for people experiencing chronic conditions: A literature review. *Nursing & health sciences*, 8(3):185–195, 2006.

Wolff, R. W. Stochastic modelling and the theory of queues. *Englewood Cliffs, NJ*, 96, 1989.

Yang, T. and Templeton, J. G. C. A survey on retrial queues. *Queueing systems*, 2(3):201–233, 1987.

Yom-Tov, G. B. and Mandelbaum, A. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299, 2014.

Appendices (Online Supplements)

A. Proofs

A.1. Proof of Lemma 1

Proof: To establish the uniqueness of the solution, we need to show that there exists a unique value of b that satisfies equation (4). Once we determine b, the remaining variables in the feasible performance vector, namely λ_d and λ_v, can be computed using equations (6) and (7), respectively. Thus, the uniqueness of b guarantees the uniqueness of the entire feasible performance vector.

$$(V(b,\lambda_v(b),\lambda_d(b)) :=) b - \min\left\{\theta\frac{1}{1-\lambda_v(b)}, 1\right\} = 0. \tag{A.1}$$

Note that if $\theta\frac{1}{1-\lambda_v(b)} \geq 1$, we have $b = 1$, resulting $\lambda_d = 0$. As a result, we must have $\theta\frac{1}{1-\lambda_v(b)} < 1$. Therefore, it is sufficient to show that the equation $V(b,\lambda_v(b),\lambda_d(b)) = 0$ has a unique solution $b \in (0,1)$. We will establish this by demonstrating the following two points: (1) the function $V(b,\lambda_v(b),\lambda_d(b))$ is negative when $b = 0$ and positive when $b = 1$. By continuity, this implies that there exists at least one solution to equation (A.1) in the interval $(0,1)$; (2) the function $V(b,\lambda_v(b),\lambda_d(b))$ strictly decreases with respect to b in the interval $(0,1)$. This guarantees that equation (A.1) has at most one solution in the interval $(0,1)$.

We first show (1). It is straightforward to see that

$$V(b,\lambda_v(b),\lambda_d(b))|_{b=0} = -\min\left\{\theta\frac{1}{1-\lambda_v(b)}, 1\right\} < 0$$

In addition, we have

$$V(b,\lambda_v(b),\lambda_d(b))|_{b=1} = 1 - \min\left\{\theta\frac{1}{1-\lambda_v(b)}, 1\right\} > 0$$

We next show (2). The derivative of $V(b,\lambda_v(b),\lambda_d(b))$ with respect to b can be calculated as

$$\frac{dV}{db} = 1 - \frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{db}\mathbb{I}\left(\frac{1}{1-\lambda_v(b)} \leq \frac{1}{\theta}\right)$$

Since

$$\begin{aligned}\frac{d\lambda_v}{db} &= \frac{d\lambda_v}{d\lambda_d}\frac{d\lambda_d}{db} \\ &= \left[\frac{1}{1-\eta} + (1-\gamma)(1-F(w)-\bar{p}+G(w))\right]\frac{-\lambda_n(1-\eta)\left[[1-(1-\eta)(\bar{p}-G(w)b)] + (1-b)(1-\eta)G(w)\right]}{[1-(1-\eta)(\bar{p}-G(w)b)]^2} \\ &\leq 0\end{aligned}$$

We have $\frac{dV}{db} \geq 1 > 0$. In other words, $V(b,\lambda_v(b),\lambda_d(b))$ is strictly increasing in b and the derivative is lower bounded by a positive constant 1. This completes the proof. ■

A.2. Proof of Lemma 2

Proof: We first build the relationship between $\frac{d\lambda_v}{dw}$ and $\frac{d\lambda_d}{dw}$ as follows:

$$\begin{aligned}\frac{d\lambda_v}{dw} &= \frac{1}{1-\eta}\frac{d\lambda_d}{dw} + (1-\gamma)(wf(w)-f(w))\lambda_d + (1-\gamma)[1-F(w)-\bar{p}+G(w)]\frac{d\lambda_d}{dw} \\ &= \left[\frac{1}{1-\eta} + (1-\gamma)[1-F(w)-\bar{p}+G(w)]\right]\frac{d\lambda_d}{dw} - (1-\gamma)f(w)(1-w)\lambda_d \\ &:= A\frac{d\lambda_d}{dw} - B\end{aligned}$$

where

$$\begin{aligned}A :=& \frac{1}{1-\eta} + (1-\gamma)[1-F(w)-\bar{p}+G(w)] \\ B :=& (1-\gamma)f(w)(1-w)\lambda_d\end{aligned}$$

Note that

$$\frac{db}{dw} = \frac{db}{d\lambda_v}\frac{d\lambda_v}{dw} = \frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw} \tag{A.2}$$

Then, we have

$$\begin{aligned}\frac{d\lambda_d}{dw} =& \frac{-\lambda_n(1-\eta)\frac{db}{dw}[1-(1-\eta)(\bar{p}-bG(w))] - \lambda_n(1-b)(1-\eta)^2(bwf(w)+G(w)\frac{db}{dw})}{[1-(1-\eta)(\bar{p}-bG(w))]^2} \\ =& \frac{-\lambda_n(1-\eta)\frac{db}{dw}[1-(1-\eta)(\bar{p}-G(w))] - \lambda_n(1-b)(1-\eta)^2bwf(w)}{[1-(1-\eta)(\bar{p}-bG(w))]^2} \\ =& \frac{-\lambda_n(1-\eta)[1-(1-\eta)(\bar{p}-G(w))]\frac{db}{d\lambda_v}\frac{d\lambda_v}{dw} - \lambda_n(1-b)(1-\eta)^2bwf(w)}{[1-(1-\eta)(\bar{p}-bG(w))]^2} \\ =& \frac{-\lambda_n(1-\eta)[1-(1-\eta)(\bar{p}-G(w))]\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw} - \lambda_n(1-b)(1-\eta)^2bwf(w)}{[1-(1-\eta)(\bar{p}-bG(w))]^2}\end{aligned}$$

Define

$$\begin{aligned}C :=& [1-(1-\eta)(\bar{p}-bG(w))]^2 \\ D :=& \lambda_n(1-\eta)[1-(1-\eta)(\bar{p}-G(w))]\frac{\theta}{(1-\lambda_v)^2} \\ E :=& \lambda_n(1-b)(1-\eta)^2bwf(w)\end{aligned}$$

Then, we have

$$C\frac{d\lambda_d}{dw} = -D\frac{d\lambda_v}{dw} - E = -D\left(A\frac{d\lambda_d}{dw} - B\right) - E$$

As a result, we have

$$\begin{aligned}&\frac{d\lambda_d}{dw} = \frac{DB-E}{C+DA} \\ &= \frac{\lambda_n(1-\eta)[1-(1-\eta)(\bar{p}-G(w))]\frac{\theta}{(1-\lambda_v)^2}(1-\gamma)f(w)(1-w)\lambda_d - \lambda_n(1-\eta)^2b(1-b)wf(w)}{[1-(1-\eta)(\bar{p}-bG(w))]^2 + \lambda_n(1-\eta)[1-(1-\eta)(\bar{p}-G(w))]\frac{\theta}{(1-\lambda_v)^2}\left[\frac{1}{1-\eta}+(1-\gamma)[1-F(w)-\bar{p}+G(w)]\right]}\end{aligned}$$

It is easy to show that $\frac{d\lambda_d}{dw}|w=0\geq 0$ and $\frac{d\lambda_d}{dw}|w=1\leq 0$. Let $\frac{d\lambda_d}{dw}=0$, we have

$$[1-(1-\eta)(\bar{p}-G(w_O))]\frac{\theta}{(1-\lambda_v)^2}(1-\gamma)(1-w_O)\lambda_d=(1-\eta)b(1-b)w_O$$

which is equivalent to

$$[1-(1-\eta)(\bar{p}-G(w_O))](1-\gamma)(1-w_O)\lambda_d=(1-\eta)(1-\theta-\lambda_v)w_O$$

Next, we prove quasi-concavity of λ_d. Since $C+DA>0$, we just need to identify the sign of the derivation of $\frac{d\lambda_d}{dw}\frac{C+DA}{\lambda_n(1-\eta)f(w)}\frac{(1-\lambda_v)^2}{\theta}$ at the point w_O. Note that

$$\frac{d\lambda_v}{dw}|_{w_O}=A\frac{d\lambda_d}{dw}-B|_{w_O}=-B=-(1-\gamma)(1-w)f(w)\lambda_d|_{w_O}$$

and

$$b=\frac{\theta}{1-\lambda_v}<1\Rightarrow 1-\lambda_v-\theta>0$$

Thus, we have

$$\begin{aligned}
&\left(\frac{d\lambda_d}{dw}\frac{C+DA}{\lambda_n(1-\eta)f(w)}\frac{(1-\lambda_v)^2}{\theta}\right)'|_{w_O}\\
=&[[1-(1-\eta)(\bar{p}-G(w))](1-\gamma)(1-w)\lambda_d-(1-\eta)(1-\theta-\lambda_v)w]'|_{w_O}\\
=&(1-\eta)wf(w)(1-\gamma)(1-w)\lambda_d\\
&+[1-(1-\eta)(\bar{p}-G(w))](1-\gamma)\left[-\lambda_d+(1-w)\frac{d\lambda_d}{dw}\right]\\
&+(1-\eta)\frac{d\lambda_v}{dw}w-(1-\eta)(1-\theta-\lambda_v)|_{w_O}\\
=&(1-\eta)wf(w)(1-\gamma)(1-w)\lambda_d-[1-(1-\eta)(\bar{p}-G(w))](1-\gamma)\lambda_d\\
&-(1-\eta)(1-\gamma)(1-w)f(w)\lambda_d w-(1-\eta)(1-\theta-\lambda_v)|_{w_O}\\
=&-[1-(1-\eta)(\bar{p}-G(w))](1-\gamma)\lambda_d-(1-\eta)(1-\theta-\lambda_v)|_{w_O}\\
\leq&0
\end{aligned}$$

The above inequality become strictly less than if $0<\eta<1$. This completes the proof. ■

A.3. Proof of Lemma 3

Proof: It suffice to prove $\frac{d\lambda_v}{dw}\leq 0$. From the proof of Lemma 2, we have

$$\begin{aligned}
\frac{d\lambda_d}{dw}=&\frac{-\lambda_n(1-\eta)[1-(1-\eta)(\bar{p}-G(w))]\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw}-\lambda_n(1-b)(1-\eta)^2bwf(w)}{[1-(1-\eta)(\bar{p}-bG(w))]^2}\\
=&\frac{-\lambda_n(1-\eta)[1-(1-\eta)(\bar{p}-G(w))]\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw}-E}{C}\\
=:&\frac{-D\frac{d\lambda_v}{dw}-E}{C}
\end{aligned}$$

Then, we have

$$\begin{aligned}\frac{d\lambda_v}{dw} &= \left[\frac{1}{1-\eta} + (1-\gamma)[1-F(w)-\bar{p}+G(w)]\right]\frac{d\lambda_d}{dw} + (1-\gamma)(wf(w)-f(w))\lambda_d \\ &= A\frac{d\lambda_d}{dw} - B \\ &= A\cdot\frac{-D\frac{d\lambda_v}{dw}-E}{C} - B\end{aligned}$$

Through algebraic manipulation, if $w>0$ or $\gamma<1$, we can deduce

$$\frac{d\lambda_v}{dw} = -\frac{AE+BC}{C+AD} < 0$$

Based on equation (A.2), we have

$$\frac{db}{dw} = \frac{db}{d\lambda_v}\frac{d\lambda_v}{dw} = \frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw} < 0$$

■

A.4. Proof of Lemma 4

Proof: Define $w_O := \arg\max \lambda_d(w)$. We prove Lemma 4 for two successive interval: $[0, w_O]$ and $[w_O, 1]$. Note that

$$\frac{dP(w)}{dw} = \frac{1}{\bar{p}}\left[\frac{db}{d\lambda_v}\frac{d\lambda_v}{dw}G(w) + bwf(w)\right] \tag{A.3}$$

For $w\in[0,w_O]$. We know that λ_d is increasing in $w\in[0,w_O]$. And by the definition we have

$$\begin{aligned}\frac{d\lambda_v}{dw} &= \left[\frac{1}{1-\eta} + (1-\gamma)[1-F(w)-\bar{p}+G(w)]\right]\frac{d\lambda_d}{dw} + (1-\gamma)(wf(w)-f(w))\lambda_d \\ &\geq (1-\gamma)(wf(w)-f(w))\lambda_d\end{aligned} \tag{A.4}$$

Then the derivative of $\bar{p}P(w)$ for $w\in[0,w_O]$ is

$$\begin{aligned}\frac{d(\bar{p}P(w))}{dw} &= \frac{db}{d\lambda_v}\frac{d\lambda_v}{dw}G(w) + bwf(w) \\ &= \frac{\theta}{(1-\lambda_v)^2}\left[G(w)\frac{d\lambda_v}{dw} + (1-\lambda_v)wf(w)\right] \\ &\geq \frac{\theta}{(1-\lambda_v)^2}\left[G(w)(1-\gamma)(wf(w)-f(w))\lambda_d + (1-\lambda_v)wf(w)\right] \\ &= \frac{\theta f(w)}{(1-\lambda_v)^2}\left[G(w)(1-\gamma)(w-1)\lambda_d + (1-\lambda_v)w\right]\end{aligned} \tag{A.5}$$

Note that for $w\in[0,w_O]$, each term in $[G(w)(1-\gamma)(w-1)\lambda_d + (1-\lambda_v)w]$ is increasing in w. As a result, we have

$$\begin{aligned}
&\frac{\theta f(w)}{(1-\lambda_v)^2}\left[G(w)(1-\gamma)(w-1)\lambda_d+(1-\lambda_v)w\right]\\
\geq&\frac{\theta f(w)}{(1-\lambda_v)^2}\left[G(w)(1-\gamma)(w-1)\lambda_d+(1-\lambda_v)w\right]|_{w=0}\\
=&0
\end{aligned} \tag{A.6}$$

indicating $P(w)$ is increasing in $[0, w_O]$.

For $w \in [w_O, 1]$**.** We know that λ_d is decreasing in $w \in [w_O, 1]$. And by the definition we have

$$-[1-(1-\eta)(\bar{p}-G(w))]\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw}-(1-b)(1-\eta)bwf(w)\leq 0 \tag{A.7}$$

Then the derivative of $\bar{p}P(w)$ for $w \in [w_O, 1]$ is

$$\begin{aligned}
\frac{d\left(\bar{p}P(w)\right)}{dw}=&\frac{db}{d\lambda_v}\frac{d\lambda_v}{dw}G(w)+bwf(w)\\
=&\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw}G(w)+bwf(w)\\
\geq&\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw}G(w)+\frac{-[1-(1-\eta)(\bar{p}-G(w))]}{(1-b)(1-\eta)}\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw}\\
=&-\frac{1-(1-\eta)(\bar{p}-b(w)G(w))}{(1-b)(1-\eta)}\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw}\\
\geq&0
\end{aligned} \tag{A.8}$$

indicating $P(w)$ is increasing in $[w_O, 1]$. ■

A.5. Proof of Proposition 1

Proof: Since $b(w)\cdot G(w)$ is increasing in w, we have $\frac{\partial^2 \Pi_{FB}}{\partial \rho \partial w} < 0$, which means Π_{FB} is strictly submodular in w and ρ. Based on the result of Theorem 2 in (Amir 2005), we can conclude that the set of values $w_{FB}(\rho)$ is strictly decreasing in ρ. As a result, $w_{FB}(\rho) < w_{FB}(0) = w_O$.

Since each element in the set $w_{FB}(\rho)$ is strictly decreasing with respect to ρ, the sets corresponding to different values of ρ must not overlap. Therefore, for each value of ρ that results in multiple maximizers, we can assign a distinct rational number to that ρ. Therefore, the number of ρ values that lead to multiple maximizers is upper bounded by the cardinality of rational numbers, which is countably infinite and has a measure of zero. ■

A.6. Proof of Proposition 2

Proof: Suppose that $R^*(\cdot)$ is the optimal payment scheme that solves the second-best optimization problem (11)-(13). Let $\hat{w}$ denote the doctor's PFU booking threshold under R^*. The incentive compataibility constraint (13) then implies that $R^*(\hat{w}) \geq h\lambda_d(w_O) - h\lambda_d(\hat{w})$. Thus, we have

$$\Pi_{SB} = (r-h)\lambda_d(\hat{w}) - \rho b(\hat{w})G(w)/\overline{p} - R^*(\hat{w}) \tag{A.9}$$

$$\leq r\lambda_d(\hat{w}) - \rho b(\hat{w})G(w)/\overline{p} - h\lambda_d(w_O) \tag{A.10}$$

$$\leq r\lambda_d(w_{SB}) - \rho b(w_{SB})G(w_{SB})/\overline{p} - h\lambda_d(w_O). \tag{A.11}$$

The last inequality follows (15). The right-hand-side of the above equation is the objective value under payment scheme $\phi^{w_{SB}}(\cdot)$. This proves $\phi^{w_{SB}}(\cdot)$ is the optimal payment scheme that solves (11)-(13). Moreover, $\frac{\partial^2 -\rho b(w)G(w)}{\partial w \partial \rho} < 0$, so the objective function is submodular in w and ρ. That implies that the maximizer of (15), w_{SB}, decreases in ρ following Theorem 2 in (Amir 2005). ■

A.7. Proof of Proposition 3

Proof: If $R(w) = \phi^{w_{FB}}(\cdot)$, We first examine the IC constraint (13):

$$\begin{aligned} &h\lambda_d(w_{FB}) + R(w_{FB}) = h\lambda_d(w_O) \\ \geq &h\lambda_d(w) + h\lambda_d(w_O) - h\lambda_d(w) = h\lambda_d(w) + R(w), \quad \forall w \end{aligned}$$

indicating the IC constraint (13) holds.

Since the first-best solution is the optimal solution for the doctor, we have $\hat{w} = w_{FB}$. Note that the IC constraint (13) indicates that $h\lambda_d(w_{FB}) + R(w_{SB}) \geq h\lambda_d(w_O) + R(w_O) = h\lambda_d(w_O) \geq 0$, indicating the IR constraint (12) holds. Therefore, solution $R(w) = \phi^{w_{FB}}(\cdot)$ is a feasible solution. Then, the gap can be bounded by:

$$\begin{aligned} \Pi_{FB} - \Pi_{SB} \leq &(r-h)\cdot\lambda_d(w_{FB}) - \rho\cdot b(w_{FB})/\bar{p}\cdot G(w_{FB}) \\ &- [(r-h)\lambda_d(w_{FB}) - \rho\cdot b(w_{FB})/\bar{p}\cdot G(w_{FB}) - \phi^{w_{FB}}(w_{FB})] \\ = &\phi^{w_{FB}}(w_{FB}) \\ = &h\lambda_d(w_O) - h\lambda_d(w_{FB}) \end{aligned}$$

■

A.8. Proof of Theorem 1

Proof: We prove the Theorem 1 by comparing the profits under the identical FUA balking rate $P(w*)$, thus the identical PFU booking threshold w^*. Let $c, \hat{c}, \tilde{c}, \bar{c}$ denote parameter chosen by the health system such that $\hat{w}(\cdot) = w^*$ for the proposed four contracts, respectively.

We first prove the statement (1). For **FR** and **FC**, we first reformulate two contracts as the form which involves effective service rate and FUA ratio. Note that the contract (16) (**FR**) is equivalent to

$$\begin{aligned}
&\max_{w} \quad h\lambda_d + c\frac{\lambda_d^R + \lambda_d^P}{\lambda_d} \\
=&\max_{w} \quad h\lambda_d + c(1-\eta)(\bar{p} - b(w)\cdot G(w)) \\
=&\max_{w} \quad h\lambda_d - c(1-\eta)b(w)\cdot G(w) + c(1-\eta)\bar{p}
\end{aligned} \tag{A.12}$$

And the contract **FC** is equivalent to

$$\begin{aligned}
&\max_{w} \quad h\lambda_d + \hat{c}(\lambda_d^R + \lambda_d^P) \\
=&\max_{w} \quad h\lambda_d + \hat{c}(1-\eta)(\bar{p}\lambda_d - b(w)\cdot G(w)\lambda_d) \\
=&\max_{w} \quad (h + \hat{c}(1-\eta)\bar{p})\lambda_d - \hat{c}(1-\eta)b(w)\cdot G(w)\lambda_d
\end{aligned} \tag{A.13}$$

For any given specific level of FUA balking rate $P(w^*)$ for two contracts, the first order conditions of (A.12) and (A.13) must satisfy

$$\begin{aligned}
&h\frac{d\lambda_d}{dw} - c(1-\eta)wf(w)b(w) - c(1-\eta)G(w)\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw} \mid_{w^*} = 0 \\
&[h + \hat{c}(1-\eta)\bar{p}]\frac{d\lambda_d}{dw} - \hat{c}(1-\eta)\left[\frac{d\lambda_d}{dw}G(w)b(w) + \lambda_d wf(w)b(w) + \lambda_d G(w)\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw}\right] \mid_{w^*} = 0
\end{aligned} \tag{A.14}$$

The relationship in (A.14) indicates that

$$[h + \hat{c}(1-\eta)\bar{p}] \mid_{w^*} = \hat{c}(1-\eta)\left[G(w^*)b(w^*) + \frac{h\lambda_d(w^*)}{c(1-\eta)}\right] \tag{A.15}$$

which is equivalent to

$$\hat{c}(1-\eta)[\bar{p} - G(w^*)b(w^*)] = h\frac{\hat{c}\lambda_d(w^*) - c}{c} \tag{A.16}$$

Note that (A.16) indicating $\hat{c}\lambda_d(w^*) - c \geq 0$. Then, under the specific level of FUA balking rate with respect to $(\lambda_d(w^*), b(w^*)G(w^*))$, the profit difference between two contracts is:

$$\begin{aligned}
&[(r-h)\lambda_d(w^*) + c(1-\eta)b(w^*)G(w^*) - c(1-\eta)\bar{p}] - [(r-h)\lambda_d(w^*) - \hat{c}(1-\eta)\bar{p}\lambda_d(w^*) + \hat{c}(1-\eta)b(w^*)G(w^*)\lambda_d] \\
=&(1-\eta)b(w^*)G(w^*)(c - \hat{c}\lambda_d(w^*)) + (1-\eta)\bar{p}(\hat{c}\lambda_d(w^*) - c) \\
=&(1-\eta)[\bar{p} - b(w^*)G(w^*)](\hat{c}\lambda_d(w^*) - c) \\
\geq&0
\end{aligned} \tag{A.17}$$

This leads to $\mathbf{FR} \succeq \mathbf{FC}$.

Next we compare **PR** and **FR**. Note that the contract **PR** is equivalent to

$$\mathbf{PR} \quad \max_{w} \quad h\lambda_d + \tilde{c}(1-\eta)(\bar{p} - G(w)) \tag{A.18}$$

For any given level of FUA balking rate $P(w^*)$ for two contracts, the first order conditions of contract **FR** and **PR** must satisfy

$$
\begin{aligned}
&h\frac{d\lambda_d}{dw} - c(1-\eta)wf(w)b(w) - c(1-\eta)G(w)\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw} \mid_{w^*}=0\\
&h\frac{d\lambda_d}{dw} - \tilde{c}(1-\eta)wf(w) \mid_{w^*}=0
\end{aligned}
\tag{A.19}
$$

The relationship in (A.19) indicates that

$$
h\frac{d\lambda_d}{dw}\frac{\tilde{c}-cb(w)}{\tilde{c}} \mid_{w^*}= c(1-\eta)G(w)\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw} \mid_{w^*}\leq 0 \tag{A.20}
$$

indicating $\tilde{c}-cb(w^*)\leq 0$.

Then, for the clinic, under a certain level of FUA balking rate with respect to $(\lambda_d(w^*), b(w^*)G(w^*))$, the profit difference between contract **PR** and **FR** is:

$$
\begin{aligned}
&[(r-h)\lambda_d(w^*) - \tilde{c}(1-\eta)\bar{p} + \tilde{c}(1-\eta)G(w^*)] - [(r-h)\lambda_d(w^*) + c(1-\eta)b(w^*)G(w^*) - c(1-\eta)\bar{p}]\\
=&(1-\eta)\bar{p}(c-\tilde{c}) + (1-\eta)G(w^*)(\tilde{c}-cb(w^*))\\
\geq&(1-\eta)\bar{p}(cb(w^*)-\tilde{c}) + (1-\eta)G(w^*)(\tilde{c}-cb(w^*))\\
=&-(1-\eta)(\tilde{c}-cb(w^*))(\bar{p}-G(w^*))\geq 0
\end{aligned}
\tag{A.21}
$$

indicating $\mathbf{PR}\succeq\mathbf{FR}$.

Next, we prove the statement (2). We first prove $\mathbf{PR}\succeq\mathbf{PC}$. Note that the contract **PC** is equivalent to

$$
\mathbf{PC} \quad \max_{w} \quad h\lambda_d + \bar{c}(1-\eta)(\bar{p}-G(w))\lambda_d \tag{A.22}
$$

Then, for any given level of FUA balking rate $P(w^*)$ for two contracts, the first order conditions of contract **PR** and **PC** must satisfy

$$
\begin{aligned}
&h\frac{d\lambda_d}{dw} - \tilde{c}(1-\eta)wf(w) \mid_{w^*}=0\\
&h\frac{d\lambda_d}{dw} - \bar{c}(1-\eta)wf(w)\lambda_d + \bar{c}(1-\eta)\frac{d\lambda_d}{dw}(\bar{p}-G(w)) \mid_{w^*}=0
\end{aligned}
\tag{A.23}
$$

The relationship in A.23 indicates that

$$
\frac{d\lambda_d}{dw}\left[h - \frac{h\bar{c}}{\tilde{c}}\lambda_d + \bar{c}(1-\eta)(\bar{p}-G(w))\right] \mid_{w^*}=0 \tag{A.24}
$$

indicating

$$
\frac{h}{\tilde{c}}(\tilde{c}-\bar{c}\lambda_d(w^*)) \mid_{w^*}= -\bar{c}(1-\eta)(\bar{p}-G(w)) \mid_{w^*}\leq 0 \tag{A.25}
$$

Then, under a certain level of FUA balking rate with respect to $(\lambda_d(w^*), b(w^*)G(w^*))$, the profit difference between contract **PR** and **PC** is:

$$
\begin{aligned}
&[(r-h)\lambda_d(w^*) - \tilde{c}(1-\eta)(\bar{p}-G(w^*))] - [(r-h)\lambda_d(w^*) - \bar{c}(1-\eta)(\bar{p}-G(w^*))\lambda_d(w^*)]\\
=&(1-\eta)(\bar{p}-G(w^*))(\bar{c}\lambda_d(w^*)-\tilde{c})\\
\geq&0
\end{aligned}
\tag{A.26}
$$

indicating $\mathbf{PR} \succeq \mathbf{PC}$.

Similarly, the first order conditions of contract **PC** and **FC** must satisfy

$$\begin{aligned} & h\frac{d\lambda_d}{dw} - \bar{c}(1-\eta)wf(w)\lambda_d + \bar{c}(1-\eta)\frac{d\lambda_d}{dw}(\bar{p} - G(w)) \mid_{w^*} = 0 \\ & [h + \hat{c}(1-\eta)\bar{p}]\frac{d\lambda_d}{dw} - \hat{c}(1-\eta)\left[\frac{d\lambda_d}{dw}G(w)b(w) + \lambda_d wf(w)b(w) + \lambda_d G(w)\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw}\right] \mid_{w^*} = 0 \end{aligned} \tag{A.27}$$

The relationship in A.27 indicates that

$$\frac{d\lambda_d}{dw}\frac{h}{\bar{c}}(\bar{c} - \hat{c}b(w)) \mid_{w^*} = \hat{c}(1-\eta)\lambda_d G(w)\frac{\theta}{(1-\lambda_v)^2}\frac{d\lambda_v}{dw} \mid_{w^*} \leq 0 \tag{A.28}$$

indicating $\bar{c} - \hat{c}b(w) \leq 0$

Therefore, under a certain level of FUA balking rate $P(w^*)$, the profit difference between contract **PC** and **FC** is:

$$\begin{aligned} & [(r-h)\lambda_d(w^*) - \bar{c}(1-\eta)(\bar{p} - G(w^*))\lambda_d(w^*)] - [(r-h)\lambda_d(w^*) - \hat{c}(1-\eta)(\bar{p} - b(w^*)G(w^*))\lambda_d(w^*)] \\ =& (1-\eta)\bar{p}\lambda_d(w^*)(\hat{c} - \bar{c}) + (1-\eta)G(w^*)\lambda_d(w^*)(\bar{c} - \hat{c}b(w^*)) \\ \geq& (1-\eta)\bar{p}\lambda_d(w^*)[\hat{c}b(w^*) - \bar{c}] + (1-\eta)G(w^*)\lambda_d(w^*)[\bar{c} - \hat{c}b(w^*)] \\ =& (1-\eta)\lambda_d(w^*)[\bar{p} - G(w^*)][\hat{c}b(w^*) - \bar{c}] \\ \geq& 0 \end{aligned} \tag{A.29}$$

indicating $\mathbf{PC} \succeq \mathbf{FC}$. This completes the proof. ■

A.9. Proof of Corollary 2

Proof: It suffices to prove $\Pi_{PR} \geq \Pi_{FR}$ as the proofs of the rest inequalities follow the same logic. Let c_{FR} denote the maximizer to the single-objective optimization (22) for contract **FR**.

By Theorem 1, we know $\mathbf{PR} \succeq \mathbf{FR}$. Then by the definition of Pareto dominance (21), there must exist $\hat{c}$ such that the $P(\hat{w}(\hat{c})) = P(\hat{w}(c_{FR}))$ and

$$(r-h)\lambda_d(\hat{w}(\hat{c})) - \hat{c}R_{PR}(\hat{w}(\hat{c})) \geq (r-h)\lambda_d(\hat{w}(c_{FR})) - c_{FR}R_{FR}(\hat{w}(c_{FR})).$$

Therefore,

$$\Pi_{PR} \geq (r-h)\lambda_d(\hat{w}(\hat{c})) - \hat{c}R_{PR}(\hat{w}(\hat{c})) - \rho P(\hat{w}(\hat{c})) \tag{A.30}$$

$$\geq (r-h)\lambda_d(\hat{w}(c_{FR})) - c_{FR}R_{FR}(\hat{w}(c_{FR})) - \rho P(\hat{w}(c_{FR})) \tag{A.31}$$

$$= \Pi_{FR}, \tag{A.32}$$

where the first inequality follows that the optimal objective value of the single-objective optimization problem for contract type PR should be at least what is achieved by using the contract $\hat{c}R_{PR}(\cdot)$, and the second inequality follows from $\mathbf{PR} \succeq \mathbf{FR}$, and the equality follows that c_{FR} is the maximizer. ■

A.10. Proof of Theorem 2

Proof: Let $c = \frac{h\rho}{(r-h)(1-\eta)\bar{p}}$. Then, for Contract **FR**, the doctor's objective becomes

$$\begin{aligned}
&\max_{w} \quad h\lambda_d(w) + c\frac{\lambda_d^R(w) + \lambda_d^P(w)}{\lambda_d} \\
=&\max_{w} \quad h\lambda_d(w) + \frac{h\rho}{(r-h)(1-\eta)\bar{p}}\frac{\lambda_d^R(w) + \lambda_d^P(w)}{\lambda_d} \\
=&\max_{w} \quad h\lambda_d(w) + \frac{h\rho}{(r-h)\bar{p}}(\bar{p} - b(w)\cdot G(w)) \\
=&\max_{w} \quad h\left[\lambda_d(w) - \frac{\rho}{(r-h)\bar{p}}b(w)\cdot G(w)\right] + \frac{h\rho}{r-h}
\end{aligned} \tag{A.33}$$

Note that the first-best model equals to

$$\begin{aligned}
\max_{w\in[0,1]} \Pi_{FB} &= (r-h)\lambda_d(w) - \rho b(w)G(w)/\bar{p} \\
&= (r-h)\left[\lambda_d(w) - \frac{\rho}{(r-h)\bar{p}}b(w)G(w)\right]
\end{aligned}$$

which would result in the same solution with the problem A.33. Therefore, we can conclude that the threshold $\hat{w}(c) = w_{FB}$ aligns the agent's effort level with that in the first best case. In addition, the above c resulting a lower bound for Π_{FR}. Then, we have

$$\begin{aligned}
\Pi_{FB} - \Pi_{FR} \leq& [(r-h)\lambda_d(w_{FB}) - \rho b(w_{FB})G(w_{FB})/\bar{p}] - \left[(r-h)\lambda_d(w_{FB}) - c\frac{\lambda_d^R(w_{FB}) + \lambda_d^P(w_{FB})}{\lambda_d(w_{FB})} - \rho b(w_{FB})G(w_{FB})/\bar{p}\right] \\
=& c\frac{\lambda_d^R(w_{FB}) + \lambda_d^P(w_{FB})}{\lambda_d(w_{FB})} \\
=& \frac{h\rho}{(r-h)(1-\eta)\bar{p}}\frac{\lambda_d^R(w_{FB}) + \lambda_d^P(w_{FB})}{\lambda_d(w_{FB})} \\
=& \frac{h\rho}{(r-h)(1-\eta)\bar{p}}\left[(1-\eta)(\bar{p} - b(w_{FB})G(w_{FB}))\right] \\
=& \frac{h\rho}{(r-h)\bar{p}}\left[\bar{p} - b(w_{FB})G(w_{FB})\right] \\
\leq& \frac{h\rho}{r-h}
\end{aligned}$$

■

B. Robustness Test for Empirical Analyses

In Section 3, we employ the propensity score matching method (PSM) to create comparable samples of doctors for an unbiased estimation the treatment effect. During the PSM process, doctors with PFU ratios below the median value are classified into the control group, and those above the median value are classified into the treatment group. We now test the robutness of the results by trying other cutoff values that are used to classify the control and treatment groups. Specifically, we test two the 40-quantile and the 60-quantile as the cutoff values. Doctors with PFU ratios below the cutoff are classified into the control group while the other doctors are classified into the treatment group.

Using the same covariates and same PSM process as in Section 3, comparison of the matched and unmatched samples with the 40-quantile and 60-quantile cutoffs are presented in Table 4. We can see that for both cutoffs, the %bias for all covariates, except for covariate AgeCat1_ Num (AgeCat4_ Num), have absolute value of less than 10%, suggesting no significant systematic bias in the covariate values between the control and treatment groups. After the matching procedure, using the 40-quantile threshold yields a PSM sample comprising 107 and 133 observations in the control and treatment groups, respectively. Using the 60-quantile threshold results in 167 observations in the control group and 104 observations in the treatment group.

Table 4 Comparison Between the Treatment Group and Control Group for Robust Test

		40-quantile threshold				60-quantile threshold			
	Unmatched (U) Matched (M)	Mean Treated	Control	%bias	p-value	Mean Treated	Control	%bias	p-value
NumApp_Cli	U	20490	26131	-59.3	0	19431	24956	-57.9	0
	M	23990	23080	9.6	0.4	21204	20369	8.7	0.513
NumApp_Prov	U	1692.2	1701.6	-1.2	0.915	1709.3	1687	2.9	0.799
	M	1598.7	1626.7	-3.7	0.766	1681.9	1719.5	-4.9	0.727
AgeCat1_Num	U	242.29	273.92	-14.4	0.198	242.4	263.3	-9.9	0.395
	M	254.68	284.32	-13.5	0.325	254.3	248.99	2.5	0.847
AgeCat2_Num	U	275.78	359.62	-44.8	0	264.1	339.46	-40.6	0
	M	289.13	300.85	-6.3	0.581	278.28	284.82	-3.5	0.792
AgeCat3_Num	U	496.71	574.57	-25	0.029	481.63	558.67	-25.3	0.03
	M	495.34	496.72	-0.4	0.97	492.41	521.09	-9.4	0.474
AgeCat4_Num	U	369.41	354.67	6.7	0.558	378.31	353.65	11.4	0.327
	M	350.11	342.76	3.3	0.789	365.01	393.47	-13.2	0.351
Fin_cla1_Num	U	928.41	1089.6	-35.9	0.002	902.12	1053.4	-33.9	0.004
	M	949.82	957.41	-1.7	0.885	927.78	962.5	-7.8	0.557
Fin_cla2_Num	U	583.86	396.79	54.6	0	628.11	429.65	55	0
	M	461.6	483.32	-6.3	0.607	569.27	565.97	0.9	0.951

We then estimate the regression model in Section 3 for both classifications using the 40-quantile and 60-quantile cutoffs. The estimated results are presented in Table 5. For the 40-quantile threshold case, we observe that a ten percent increase in the PFU ratio correlates with an average rise

in the SECON index by 0.1115. While for the 60-quantile threshold case, a similar magnitude of effect, a 0.0996 increase in SECON index per ten-percent increment in the PFU ratio, is observed. In addition, both estimates are statistically significant, indicating our result, i.e., booking more PFUs instead of RFUs can lead to measurable improvements in the SECON index, is robust.

Table 5 The PFUs effect on COC (Robust test)

	40-quantile threshold		60-quantile threshold	
	Coefficient	p-value	Coefficient	p-value
PFURatio	1.115	0.006	0.996	0.013
FUARatio	0.124	0.600	0.248	0.270
Constant	0.378	0.000	0.362	0.000
Obs.	240		271	