Is Continuity All We Need? A Modeling Approach to Evaluating Relational Continuity in Primary Care

Naireet Ghosh
Management Science and Operations, London Business School, nghosh@london.edu

Nicos Savva
Management Science and Operations, London Business School, nsavva@london.edu

Yueyang Zhong
Management Science and Operations, London Business School, yzhong@london.edu

Relational continuity of care—the sustained relationship between a patient and their usual clinician—improves outcomes but limits resource pooling in congested primary care systems. We study this trade-off in a sequence of mean-field queueing models in which routing a patient away from their usual general practitioner generates additional downstream workload for future consultations, creating an endogenous feedback between continuity decisions and system load. In a homogeneous-patient model, neither full continuity nor full pooling is generally optimal: the planner trades off workload amplification from disrupted continuity against delay reduction from pooling. Decentralized patient behavior, by contrast, is bang-bang—patients either preserve full continuity or switch to the earliest available clinician—so equilibrium may deviate from the optimum, though inefficiency remains bounded and, for empirically relevant parameters, modest. With heterogeneous patients, type-dependent policies strictly dominate uniform ones: patients who benefit from continuity should receive it, while others should pool maximally. Threshold-based routing rules that condition switching on observed queue-length gaps perform at least as well as the probabilistic benchmark in our numerical experiments. Calibration to NHS practice data shows that, under the high utilization typical of English general practice, the optimal policy is full continuity for complex patients and full pooling for non-complex patients, and that under these calibrated conditions, decentralized patient choice replicates this optimum exactly. The results support a simple, implementable prescription: protect continuity for those who benefit, pool access for those who do not, and empower patients to choose.

Key words: Relational continuity, multi-server queues, equilibrium, mean-field approximation, join-the-shortest-queue.

1. Introduction

Relational continuity (RC)—a sustained therapeutic relationship between a patient and a specific healthcare provider—is widely regarded as a cornerstone of effective primary care (Haggerty et al. 2003, 2007). RC is especially valuable for patients with chronic or complex conditions, as it supports better diagnosis and management, strengthens trust, and improves clinical outcomes (Jeffers

and Baker 2016). Beyond these patient-level benefits, continuity can also affect system workload. Using English primary care data, Kajaria-Montag et al. (2024) find that when patients see their usual general practitioner (GP), the interval to the next consultation increases by 18.1%, without lengthening consultation time. Aggregated to the system level, this effect can meaningfully reduce demand: if all practices achieved the continuity levels of the top decile, total consultations in the English National Health Service (NHS)—which delivers over 300 million general practice appointments annually, often under utilization above 90%—could fall by up to 5.2%.

Despite these documented benefits, continuity in English general practice has fallen sharply in recent years—declining from 29.3% in 2018 to 19% in 2022—with an especially pronounced drop following the COVID-19 lockdown (University of Leicester 2024). Some researchers warn that continuity "could effectively disappear" unless urgent reforms are introduced (Wilkinson 2024). National media, such as the *Guardian*, have called for restoring continuity as both a driver of better health outcomes and a means to reduce clinician workload (Gregory 2024). At the same time, NHS policy has increasingly promoted *resource pooling* through centralized and digital access pathways (e.g., the *Digital First* toolkit and related booking initiatives), which are designed to route patients to the earliest available clinician rather than to their usual GP (NHS England 2019, 2023). These two imperatives—restoring continuity and expanding pooled access—pull in opposite directions. Continuity can improve long-run outcomes and reduce future demand, but it may also reduce operational flexibility by restricting pooling, potentially increasing waits and reducing access. Pooled access, in contrast, improves short-run access and efficiency but may sacrifice RC, potentially degrading care and increasing future utilization. Understanding how to manage this fundamental continuity-access tension is the central goal of this paper. Specifically, we focus on three main questions.

Our first research question is: **When is high continuity operationally optimal, and when is partial pooling desirable?** While pooling is well understood to reduce delay in systems with exogenous demand, our setting features an endogenous demand, driven by a feedback loop in which routing decisions today affect future system load. In particular, disrupting continuity increases patients' downstream consultation workload and thereby inflates long-run demand, so the system's effective load is endogenous to its routing policy. As a result, increasing pooling can improve short-run access yet worsen long-run congestion through higher effective load, so the net effect on total costs need not be monotone.

Our second research question is: **How should continuity targets depend on patient heterogeneity in continuity benefits?** Continuity benefits are highly uneven across patients. For example, patients with chronic or complex needs may derive substantial value from seeing their usual GP, whereas many acute or low-complexity visits may benefit little. Under limited capacity, a uniform continuity target, although operationally simpler to implement, can misallocate scarce

access, either sacrificing pooling for patients who do not benefit or sacrificing continuity for those who do. We therefore ask when type-dependent continuity targets dominate uniform policies.

Finally, we examine the third research question: **Should continuity decisions be left to patients—who trade off their own waiting time against continuity—or centrally coordinated by the system?** Patient choices can create externalities that operate not only through contemporaneous congestion but also through future demand. Switching away from one's usual GP may reduce current waiting (a short-run benefit to others) but can accelerate future revisits and increase long-run congestion (a long-run cost to others). It is therefore not clear a priori whether decentralized behavior yields too much pooling, too little pooling, or outcomes close to the social optimum.

Our Contributions

We develop and analyze a sequence of queueing models that link continuity, routing, and endogenous revisit demand. We model the demand amplification from disrupted continuity as additional follow-up workload, a reduced-form representation motivated by the hazard-rate evidence in Kajaria-Montag et al. (2024) (see Section 3 and Remark 1 for a formal discussion). We begin with the simplest possible model of a GP practice: an $M/M/n$ parallel-server queue with n GPs, where patients are homogeneous in the benefit they derive from continuity of care, yet heterogeneous in that each has their own usual GP. The key modeling feature is that disrupted continuity inflates future demand, creating an endogenous feedback loop between routing and load. In the baseline model, continuity is implemented via a simple *probabilistic* policy governed by a switching probability $p \in [0,1]$: when a patient's usual GP does not have the shortest queue, the patient is routed to the shortest queue with probability p (pooling) and stays with their usual GP with probability $1-p$ (continuity). The extremes recover two familiar benchmarks: $p=1$ yields a join-the-shortest-queue (JSQ) policy (Haight 1958, Kingman 1961), while $p=0$ yields n decoupled $M/M/1$ queues. Exact analysis is intractable even in this stylized setting, so we adopt a mean-field approximation in which the number of GPs grows large while system load remains fixed (Vvedenskaya et al. 1996). This approximation yields closed-form performance metrics and allows us to analyze both (i) centralized control, where p minimizes aggregate costs, and (ii) decentralized choice, where each patient selects p to minimize individual cost.

Three findings emerge from the homogeneous-patient model. First, in the centralized case, *full* continuity is generally *not* optimal. Preserving some pooling can reduce congestion enough to outweigh the continuity-induced reduction in future demand. Second, in the decentralized case, equilibrium behavior often bifurcates—patients either follow full continuity when revisit costs dominate waiting costs, or abandon it entirely when waiting costs dominate, with possible mixing only when the two

costs are closely balanced. Third, decentralized choices can diverge from the centralized optimum because patients do not internalize the externalities of their decisions. When system load is high, patients tend to preserve too much continuity because they undervalue the congestion relief their switching creates for others; when load is lower and visit costs loom larger relative to waiting, they may instead switch too much because they undervalue the downstream workload their discontinuity imposes on the system. Even so, the resulting inefficiency is bounded and, for empirically relevant parameters, modest.

We then extend the model to account for patient heterogeneity in continuity benefits, motivated by evidence that RC matters far more for complex patients than for patients with simple or acute needs. We first analyze uniform policies that apply the same switching probability p to all patients, irrespective of whether they benefit from continuity or not, and show that the optimal policy is typically interior ($0 < p < 1$). We then study type-dependent policies and find that they dominate uniform policies. For patients with no measurable continuity benefit, the optimal choice is full pooling ($p = 1$), whereas for those who benefit, the optimal continuity level depends on the cost structure and the severity of the demand penalty. Importantly, patients without measurable continuity benefits serve as a load-balancing mechanism: even a modest share of such patients helps equalize workloads across GPs, thereby reducing waiting times for complex patients and enabling them to sustain high levels of continuity. As in the homogeneous-patient setting, centralized and decentralized outcomes can differ, particularly when waiting and revisit costs are comparable for patients who benefit from continuity or when such patients comprise a large share of demand.

Next, we study a more implementable class of *threshold-based* continuity rules. Under these policies, continuity is preserved if the expected waiting time difference between the usual GP and the best alternative is below a threshold. Although threshold-based policies are analytically intractable, numerical analysis shows that they consistently yield lower costs than probabilistic policies, suggesting that the latter (our core focus and the object of theoretical analysis in this paper) provide a conservative lower bound on system performance. The optimal threshold being finite further confirms our theoretical result that, in general, full continuity of care is not optimal and that preserving some degree of resource pooling is beneficial.

Finally, we calibrate patient arrivals, patient composition, service times, practice size, utilization, and cost parameters using NHS practice-level data. In addition, we relax the assumptions of exponential inter-arrival and service times. The numerical results suggest that, under the high utilization typical of NHS general practice, the optimal policy assigns full continuity to patients with chronic or complex conditions and none to those for whom continuity yields no measurable benefit. Importantly, under our NHS calibration, the externalities from decentralized patient choice are modest enough that whether continuity decisions are made centrally or left to patients never

changes the outcome. These results provide a clear and simple operational prescription for resolving the continuity-access trade-off: safeguard continuity for patients who benefit from it, pool access for those who do not, and in virtually all practices, empower patients to choose.

While our motivation is rooted in primary care, with an emphasis on NHS, the mechanism we study—assignment decisions today shaping demand tomorrow—arises broadly in relational service systems. In specialty and mental health care, continuity with the same provider improves outcomes and reduces future utilization; in physical therapy, sustained contact with the same therapist lowers the likelihood of surgery and reduces costs; and in customer service and professional services, repeat interactions with the same agent can improve effectiveness and enhance satisfaction (Gans et al. 2003). Across these settings, the key insight of our analysis is that continuity is rarely "all we need": performance is maximized by balancing selective continuity with pooling, with appropriate central coordination guided by the strength of externalities.

2. Literature Review

This paper relates to two streams of research: (i) medical and healthcare operations work on continuity of care, and (ii) queueing literature on resource pooling, endogenous demand, and strategic customer behavior.

Medical and Healthcare Operations Literature The clinical benefits of continuity of care are well documented. Patients who regularly see the same doctor have better quality-of-life outcomes (Ye et al. 2016, Chen et al. 2017, Drury et al. 2020), enhanced control of chronic conditions such as hypertension and diabetes (Leniz and Gulliford 2019, Ahuja et al. 2022), lower mortality risk (Cho et al. 2015, Gray et al. 2018), and higher medication adherence (Dossa et al. 2017). These benefits are especially pronounced for older adults and patients with complex medical or mental health needs (Nyweide et al. 2013). Related medical literature also associates higher continuity with lower downstream utilization, including fewer emergency-department visits and fewer unplanned hospitalizations (Huntley et al. 2014, Pourat et al. 2015, Barker et al. 2017). While these system-level spillovers are important, we do not endogenize downstream acute-care use and instead study the operational consequences of continuity within primary care. Namely, disrupted continuity can inflate effective demand for primary care itself. Using over 10 million consultations, Kajaria-Montag et al. (2024) find that seeing one's regular doctor increases the interval to the next consultation by 18.1% without lengthening consultation time, with stronger effects for complex patients (e.g., 20.8% among those aged 70+ versus 13.5% among younger adults). This evidence motivates our modeling assumption that disrupted continuity generates additional downstream primary-care workload (see Section 3 for the formal model and Remark 1 for the relationship to the hazard-rate evidence above). The resulting question is not whether continuity matters clinically, but how it should be managed operationally when preserving it limits pooling and when breaking it may increase future workload.

Recent work in healthcare operations has begun to study mechanisms for preserving continuity. For example, Ding et al. (2026) analyze how prioritized follow-up appointments and associated physician incentives can reduce patient balking and thereby improve provider continuity. However, the existing literature has largely overlooked the operational trade-off between continuity and access that arises under capacity constraints and congestion. A notable exception is Liu et al. (2018), who use a discrete-choice experiment to estimate how patients trade off waiting time against seeing their regular doctor. We complement and extend this work by incorporating the endogenous demand effect of continuity—absent from the preference-estimation framework of Liu et al. (2018)—and show that full continuity is generally not optimal and that decentralized patient choices can produce equilibrium outcomes that diverge from the centralized optimum.

Similar trade-offs also arise in the recent expansion of telemedicine: virtual visits can improve access but may provide lower resolution, potentially inducing follow-up in-person utilization. Empirical studies find that telemedicine adoption in primary care can increase subsequent in-person demand (Bavafa et al. 2018, Lekwijit et al. 2023). In the same spirit as our paper, a recent growing analytical literature integrates queueing models with patient behavior to optimize operational decisions in telemedicine settings, including workload control (Saghafian et al. 2018), visit-interval management (Bavafa et al. 2021), capacity allocation (Liu and Armony 2024, Liu et al. 2025), and scheduling design (Zang et al. 2024). Our setting fundamentally differs from the telemedicine context, in which telemedicine and in-person capacity can often be treated as separable resources. In primary care continuity, however, each GP simultaneously serves as the usual provider for some patients and as an alternative provider for others, so continuity and pooling are intrinsically intertwined within a single capacity pool.

Queueing Literature Our paper sits at the intersection of three themes in the queueing literature: resource pooling in multiserver systems, endogenous (congestion- or routing-dependent) demand, and strategic customer behavior in queues. We briefly discuss each.

The literature on load balancing and routing in parallel-server systems provides our technical foundation. A classical benchmark is the Join-the-Shortest-Queue (JSQ) policy, which routes each arrival to the shortest queue and is known to perform well in reducing congestion (Haight 1958, Kingman 1961). We extend this line of work by introducing a probabilistic JSQ policy, in which joining the shortest queue is randomized rather than deterministic. This routing rule interpolates between JSQ (full pooling) and dedicated routing (full continuity) through a switching probability p, enabling a direct operational study of the continuity-pooling trade-off. Sunar et al. (2021) show that pooling can actually harm welfare when customers are delay-sensitive, because strategic joining behavior offsets the congestion benefits; our model identifies a complementary channel in which pooling hurts by disrupting continuity and amplifying future demand. In the appointment-scheduling

literature, Kuiper and Lee (2022) study multiserver appointment systems and explicitly note that continuity of care acts as a restriction on pooling; our queueing framework complements their scheduling perspective by endogenizing the demand consequences of that restriction. Nevertheless, exact analysis of state-dependent routing in parallel-server systems is difficult, even in a two-queue system under the classical JSQ policy. To address this analytical intractability, we adopt a mean-field approximation, in which the number of servers and the arrival rate grow large together, yielding tractable steady-state characterizations. This method has been successfully employed to analyze JSQ-type policies and their variants in previous work (e.g., Mitzenmacher (1996), Braverman (2020), Mukherjee et al. (2020)).

A distinguishing feature of our model is that demand is endogenous: routing decisions that disrupt continuity generate additional workload, so the system's effective arrival rate depends on its own routing policy. This contrasts with most of the routing and scheduling literature, which treats demand as fixed and exogenous. The closest queueing-theoretic antecedent is Whitt (2003), who studies how multiserver queues scale when demand grows with congestion; our mechanism is similar in spirit but the feedback channel is provider-assignment quality rather than congestion itself. Within healthcare operations, some models study centralized assignment rules (e.g., Gupta and Wang (2008)); others examine patient choice among service options (e.g., Green et al. (2013)). What is missing is a tractable framework that incorporates endogenous demand arising from provider-assignment quality and compares centralized versus patient-driven routing. Our paper fills this gap.

Additionally, our decentralized choice setting relates to the queueing game literature (see Hassin and Haviv (2003), Hassin (2016) for extensive surveys) and, more specifically, to models in which customer decisions interact with routing rules and system design (Armony and Maglaras 2004). A classic result in this literature, beginning with Naor (1969), is that customers may over-join because they do not internalize the waiting-time externality imposed on those behind them. In contrast, our setting reveals two countervailing externalities: pooling generates a *negative* externality by increasing downstream workload due to disrupted continuity and thus lengthening long-run waiting time, but also a *positive* externality by reducing the total number of customers in queue and thus lowering immediate waiting time. Such countervailing forces imply that decentralized behavior can lead to either excessive pooling or excessive continuity, depending on which force dominates—a phenomenon not captured by standard queueing game models with exogenous demand.

3. Model

The goal of this paper is to analyze the fundamental trade-off between relational continuity and operational efficiency in primary care delivery. High continuity improves outcomes and reduces future consultation demand, but constrains pooling of capacity and can lengthen current waits; conversely,

low continuity eases congestion today but may generate more visits tomorrow. We develop a modeling framework that quantifies this trade-off and clarifies when continuity should be prioritized, relaxed, or coordinated centrally.

To this end, we build a sequence of queueing models of increasing realism and complexity. Each step isolates a different dimension of the continuity-efficiency trade-off while remaining analytically tractable. Because exact analysis is intractable even in stylized settings, we rely on approximations (specifically, mean-field approximations) to obtain analytical insights into system behavior.

3.1. Baseline Model: Homogeneous Patients

We begin with the simplest possible setting. We model a primary care practice with n identical full-time equivalent (FTE) general practitioners (GPs) as an $M/M/n$ parallel-server queue.

Arrivals and service. Patients' requests for appointments arrive to the practice according to a homogeneous Poisson process with rate $n\lambda$. Equivalently, each individual GP would see appointment requests at a rate λ if demand were evenly split. Because continuity affects how quickly patients return, the steady-state effective arrival rate per GP—which we will define later—generally differs from λ. Patients are served on a first-come, first-served (FCFS) basis, and the system is work-conserving, namely, GPs do not idle when patients are waiting. Service times are independent and exponentially distributed with rate μ, which we normalize to $\mu = 1$ (unit mean). The exponential inter-arrival and service time assumption is made for analytical tractability. (In Section 8, where we calibrate the model to data, we will relax this assumption and adopt empirically calibrated distributions to assess robustness.) The unit mean assumption is without loss of generality, as it just scales the arrival rate λ in the units of the service rate μ.

Mapping model primitives to practice operations. Because our model is couched in queueing terminology, we briefly clarify the correspondence with appointment operations. A "service completion" maps to a concluded consultation; "queue length" corresponds to the number of outstanding appointment requests awaiting a given GP; and FCFS scheduling corresponds to patients being seen in the order their appointments were booked, which is the default in most GP booking systems. Key operational frictions that we abstract away include triage by nurses or receptionists and asynchronous contacts such as e-consultations, which do not change the fundamental continuity-pooling trade-off that is the focus of our analysis.

Continuity and demand feedback. Each patient in the system is associated with a usual GP who is familiar with their medical history and care needs. If a patient is seen by their usual GP, continuity is preserved and the visit resolves the patient's current episode of need, generating no additional follow-up demand. In contrast, if the patient's visit is with a different GP, continuity is broken and the encounter generates, in expectation, $\alpha - 1$ additional follow-up consultations ($\alpha \geq 1$), reflecting

the well-documented tendency for disrupted continuity to produce incomplete problem resolution, duplicated investigations, or deferred decisions that necessitate return visits (Kajaria-Montag et al. 2024). We model these follow-ups as terminal: they contribute to system workload but do not themselves propagate further demand.[1] Thus, each disrupted visit produces a total of α consultation-equivalents of workload (including the original visit), while each continuity-preserving visit produces exactly one. The parameter α quantifies the demand *penalty* from disrupted continuity. In this baseline model, patients are assumed to be homogeneous in the benefit they receive from continuity; that is, the parameter α is the same for all patients. We will relax this assumption in Section 6.

Routing policy and effective demand. To model the trade-off between preserving continuity of care and pooling capacity to improve access, we introduce a simple probabilistic routing policy governed by a routing parameter $p \in [0,1]$. The routing decision is made at the time of booking and is assumed to remain fixed thereafter, consistent with appointment scheduling practices in NHS primary care, notwithstanding occasional ex post adjustments due to operational disruptions. Upon arrival, the system manager compares the queue length of the patient's usual GP to the minimum queue length across all GPs. If the patient's usual GP is among the shortest queues (ties allowed), then the patient is routed to their usual GP (because this both minimizes their immediate wait and preserves continuity). Otherwise, if the usual GP's queue is *not* the shortest, then the patient is routed to a GP with the shortest queue with probability p (ties among shortest queues can be broken uniformly at random) and is routed to their usual GP with probability $1-p$. Thus, p is a control parameter, with $1-p$ representing the degree of continuity. In particular, $p=0$ corresponds to full continuity (dedicated routing), while $p=1$ corresponds to full pooling (JSQ routing). We note that even when $p=1$, continuity is not completely eliminated, because some patients who happen to arrive at a time when their usual GP has the shortest queue will still be seen by their usual GP. This "partial pooling" (probabilistic JSQ) class is practically relevant. In many practices and call-handling/triage workflows, continuity is treated as a soft priority—patients are steered toward their usual GP when feasible, but are redirected to another GP to protect access and balance workload.

The routing policy influences the long-run *effective* demand rate seen by the practice. When continuity is preserved, each visit generates exactly one consultation. When continuity is broken (which happens in our model on a fraction of p of those arrivals whose usual GP's queue is not the shortest among all GPs), the disrupted visit generates α consultation-equivalents of workload in total. This feedback loop will determine an *endogenous* effective arrival rate. Fix $p \in [0,1]$ and consider the probability that a (random) patient's usual GP is not among the shortest queues at arrival, with respect to the steady-state distribution induced by this policy; denote this probability by $\sigma(p)$. The probability that an arrival breaks continuity is then $p\sigma(p)$. Let $\hat{\lambda}(p)$ be the long-run effective arrival rate per GP, accounting for the continuity disruptions under policy p. Since a

fraction $1-p\sigma(p)$ of visits preserve continuity (each generating one consultation-equivalent) and a fraction $p\sigma(p)$ break continuity (each generating α consultation-equivalents), we have

$$\hat{\lambda}(p) = \big(1-\sigma(p)p\big)\cdot\lambda + \sigma(p)p\cdot\alpha\lambda = \big(1+(\alpha-1)p\sigma(p)\big)\lambda. \tag{1}$$

Here, the additional follow-up workload from disrupted visits is aggregated into the effective arrival rate $\hat{\lambda}(p)$ rather than modeled as separate endogenous arrivals with their own routing decisions; this reduced-form treatment is what makes the linear closure in (1) tractable. Importantly, (1) is a self-consistency relation, where $\sigma(p)$ depends on the steady-state distribution, which in turn depends on $\hat{\lambda}(p)$ and the routing policy. We use a mean-field approximation to characterize $\sigma(p)$ and $\hat{\lambda}(p)$ in closed form in the next section.

The boundary cases are intuitive. If $p=0$ (full continuity), then the system reduces to n independent $M/M/1$ queues with arrival rate λ. If $p=1$ (full pooling), then the system becomes a standard $M/M/n$ queue with inflated arrival rate, in consultation-equivalent terms, of $\hat{\lambda}(1) = \big(1+(\alpha-1)\sigma(1)\big)\lambda$.

Who makes the decision? In this baseline model, we assume that p is chosen centrally by the system. Delegating decision rights to a central planner closely reflects institutional practice in centralized healthcare systems such as the NHS, where routing policies are set at the practice or system level. In Section 5, we relax this assumption and analyze a decentralized setting where patients make their own decisions.

System costs. Routing policies directly affect how frequently patients generate consultations and how long patients wait for care. As such, we measure policy performance using a cost rate that captures both the burden of additional consultations and the disutility from delayed access. Each consultation generates a *visit cost* C_v, summarizing the patient burden associated with having a health problem that warrants a consultation (e.g., symptoms and disruptions, travel and time off work) and, if adopting a system perspective, the resource cost of providing the visit. Since patients routed to non-usual GPs generate more downstream workload (captured by $\hat{\lambda}(p)$ increasing in p), the total visit cost increases as continuity is relaxed. In addition, patients incur a *waiting cost* C_w per unit time from delay in receiving care (e.g., physical and psychological discomfort, avoidable deterioration while access is delayed). A system planner may care about waiting time not only because it affects patient welfare and experience, but also because it can influence practice reputation and demand, enter quality-of-care metrics under value-based reimbursement, and increase downstream resource needs through avoidable clinical deterioration. Let $W(p)$ denote the steady-state expected time spent in the system (including the time spent waiting in queue and the time in service) under policy p. Normalizing by the number of GPs, the steady-state cost rate (per unit capacity) under policy p is

$$\mathcal{C}(p) := \hat{\lambda}(p)\,\big(C_v + C_w\,W(p)\big), \tag{2}$$

where $\hat{\lambda}(p)$ is the effective appointment arrival rate, given by (1). (The system-wide cost rate is $n\mathcal{C}(p)$, so optimizing (2) is equivalent to optimizing total system costs.) The objective (2) formalizes the continuity-access trade-off. When $C_v = 0$, waiting is the only cost. Unless disrupted continuity amplifies future demand too strongly, this pushes the planner toward full pooling ($p = 1$), routing patients to the earliest available GP. When $C_w = 0$, waiting is costless and the optimal policy is full continuity ($p = 0$), regardless of queue lengths. More generally, the optimal degree of continuity depends on the relative magnitudes of C_v and C_w and on the strength of demand feedback from disrupted continuity (captured by α).

Evaluating (2) requires characterizing both $\hat{\lambda}(p)$ and $W(p)$, which in turn requires knowing the steady-state distribution of the system. Exact analysis is intractable because the routing policy couples the n queues in a state-dependent way. This challenge is already present even in the classical JSQ system (a special case of our model with full pooling $p = 1$ and no demand penalty from disrupted continuity $\alpha = 1$), for which closed-form steady-state expressions are not available for general $n \geq 2$. We therefore, in the next section, adopt a mean-field approximation that enables tractable analysis when n is large.

3.2. Mean-Field Approximation

We consider a sequence of queueing systems indexed by the number of GPs n, and let n grow large to infinity. For a system of size n, we append a superscript "(n)" to denote the processes or quantities associated with that system. Specifically, each GP operates at unit service rate $\mu^{(n)} = 1$, and patients arrive to each GP at rate $\lambda^{(n)} = \lambda$. This scaling ensures that the system load remains constant as n increases.

For a given system size n and a routing policy p, let $Q_j^{(n)}(t;p)$ be the number of patients (both in queue and in service) at GP j at time t (for $j \in \{1,\ldots,n\}$). We describe the evolution of the system (namely, the system state) using the *occupancy process*, which tracks the distribution of queue lengths across GPs. Specifically, for $i = 0,1,2,\ldots$, define

$$s_i^{(n)}(t;p) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\left\{Q_j^{(n)}(t;p) \geq i\right\},$$

as the fraction of GPs that have at least i patients at time t. By definition, $s_0^{(n)}(t;p) \equiv 1$ for all t (every GP has at least 0 patients), and the sequence is non-increasing in i (i.e., $s_i^{(n)}(t;p) \geq s_{i+1}^{(n)}(t;p)$ for all i). We assume a finite initial condition; that is, there exists some $J < \infty$ such that $s_J^{(n)}(0;p) = 0$ (initially, no GP has J or more patients). This state descriptor applies for any finite n, but analyzing the exact finite-n dynamics is intractable due to the complex interactions among queues under our routing policy. Fortunately, as $n \to \infty$, stochastic fluctuations average out and flows become deterministic, so there are no stochastic dependencies across queues. In particular, the stochastic

occupancy processes concentrate and converge to a deterministic mean-field limit. For any quantity $X^{(n)}$, we denote by $X^{(\infty)}$ the limiting counterpart of $X^{(n)}$ as $n \to \infty$. For example, $s_i^{(\infty)}(t;p)$ denotes the fraction of GPs with at least i patients at time t in the infinite-system limit.

Since disrupted continuity can inflate demand, not every value of p can yield a stable system. We henceforth restrict attention to parameters satisfying the stability condition $\hat{\lambda}^{(n)}(p) < 1$, i.e., the effective per-GP arrival rate remains below the service rate. From (1) and the fact that $\sigma(p) \leq 1$, a sufficient condition for stability, as $n \to \infty$, is

$$\big(1 + (\alpha - 1)p\big)\lambda < 1, \tag{3}$$

which ensures a strictly positive fraction of idle capacity. Condition (3) implies that, in the limit, a strictly positive fraction of GPs is idle, so the shortest queue is zero. This greatly simplifies the routing dynamics because in the mean-field limit, a patient's decision depends only on whether the usual GP is idle or busy, not on the full system state. The next lemma formalizes the evolution of the system in the mean-field limit (as $n \to \infty$) as an infinite-dimensional system of ordinary differential equations (ODEs). The lemma involves the mean-field effective arrival rate $\hat{\lambda}^{(\infty)}(p)$, which we treat as a parameter for now.

LEMMA 1 **(Mean-field system evolution)**. *Fix $p \in [0,1]$ satisfying (3), and assume that $s^{(\infty)}(0;p)$ is a finite initial condition. Assume that a mean-field effective arrival rate per GP $\hat{\lambda}^{(\infty)}(p) \in (0,1)$ exists. Then, the mean-field occupancy process $s^{(\infty)}(t;p)$ satisfies the following for all $t \geq 0$:*

$$\frac{d}{dt}s_1^{(\infty)}(t;p) = \Big(1 - (1-p)s_1^{(\infty)}(t;p)\Big)\hat{\lambda}^{(\infty)}(p) - \Big(s_1^{(\infty)}(t;p) - s_2^{(\infty)}(t;p)\Big), \tag{4}$$

$$\frac{d}{dt}s_i^{(\infty)}(t;p) = (1-p)\Big(s_{i-1}^{(\infty)}(t;p) - s_i^{(\infty)}(t;p)\Big)\hat{\lambda}^{(\infty)}(p) - \Big(s_i^{(\infty)}(t;p) - s_{i+1}^{(\infty)}(t;p)\Big), \quad \forall i \geq 2. \tag{5}$$

Each equation in Lemma 1 admits a flow-balance interpretation. The inflow term captures arrivals that increase a GP's queue from $i-1$ to i. For $i = 1$, idle GPs absorb arrivals both from their own patients and from patients pooled away from busy usual GPs, yielding the compound rate $\big(1-(1-p)s_1^{(\infty)}\big)\hat{\lambda}^{(\infty)}$. For $i \geq 2$, only continuity-insisting patients join a non-shortest queue, so the inflow rate is $(1-p)(s_{i-1}^{(\infty)} - s_i^{(\infty)})\hat{\lambda}^{(\infty)}$. The outflow term $s_i^{(\infty)} - s_{i+1}^{(\infty)}$ captures service completions that reduce a queue from i to $i-1$.

3.3. Steady-State Analysis

Lemma 1 governs the transient dynamics of the mean-field system. We now turn from transient dynamics to the steady state. Setting $\frac{d}{dt}s_i^{(\infty)}(t;p) = 0$ for all i in (4)–(5), we obtain a system of linear equations for the stationary occupancy distribution, denoted by $s_i^{(\infty)}(\infty;p) := \lim_{t\to\infty} s_i^{(\infty)}(t;p)$. As the next result shows, this fixed point has a simple closed-form solution.

PROPOSITION 1 **(Mean-field fixed point)**. *Fix* $p \in [0,1]$ *satisfying* (3)*, and assume that* $s^{(\infty)}(0;p)$ *is a finite initial condition. Assume that a mean-field effective arrival rate per GP* $\hat{\lambda}^{(\infty)}(p) \in (0,1)$ *exists. The ODE system* (4)–(5) *admits a unique fixed point* $s^{(\infty)}(\infty;p)$ *given by*

$$s_i^{(\infty)}(\infty;p) = (1-p)^{i-1}\big(\hat{\lambda}^{(\infty)}(p)\big)^i, \qquad \forall i \geq 1. \tag{6}$$

When $p=0$, the fixed point reduces to the geometric distribution of an $M/M/1$ queue. When $p=1$, $s_i^{(\infty)}(\infty;1)=0$ for all $i \geq 2$, implying that no GP ever has more than one patient and essentially no waiting occurs. For intermediate $0<p<1$, the tail decays geometrically at rate $(1-p)\hat{\lambda}^{(\infty)}(p)$, which is strictly faster than the $p=0$ rate, so even moderate pooling can substantially reduce long queues while preserving continuity for most patients. This fixed-point characterization provides the analytical foundation for evaluating how different values of p impact performance measures including the probability that a patient's usual GP is not among the shortest queues, the effective arrival rate per GP, and the expected time in system.

PROPOSITION 2 **(Steady-state performance)**. *Fix* $p \in [0,1]$ *satisfying* (3)*. In the mean-field steady state, the following hold:*

$$\sigma^{(\infty)}(p) = \hat{\lambda}^{(\infty)}(p) = \frac{\lambda}{1-p(\alpha-1)\lambda}, \tag{7}$$

and

$$W^{(\infty)}(p) = 1 + \frac{(1-p)\lambda}{1-\big(p(\alpha-1)+(1-p)\big)\lambda}. \tag{8}$$

Proposition 2 verifies the existence of $\hat{\lambda}^{(\infty)}(p)$, which we assumed in Lemma 1 and Proposition 1, and provides the closed-form expression for $\hat{\lambda}^{(\infty)}(p)$. The key step is the identity $\sigma^{(\infty)}(p) = \hat{\lambda}^{(\infty)}(p)$, which can be understood as follows. Under the stability condition (3), the minimum number of patients a GP has is 0 with probability one in the mean-field limit. Consequently, a patient's usual GP (which is, by symmetry, a uniformly random GP) is *not* among the shortest queues if and only if it is busy. The fraction of busy GPs is $s_1^{(\infty)}(\infty;p)$, which by the fixed-point characterization (Proposition 1) equals $\hat{\lambda}^{(\infty)}(p)$; hence $\sigma^{(\infty)}(p) = \hat{\lambda}^{(\infty)}(p)$. Substituting into the workload closure (1) yields the self-consistency equation $\hat{\lambda} = \lambda(1+(\alpha-1)p\hat{\lambda})$, whose unique solution is (7).

Equation (7) shows that $\sigma^{(\infty)}(p)$ is increasing in both the baseline load λ and the continuity penalty α. This means that higher underlying demand or stronger demand amplification from disrupted continuity raise the probability that a randomly selected usual GP is not among the shortest queues. Moreover, when $\alpha > 1$, increasing p raises $\hat{\lambda}^{(\infty)}(p)$ through the demand-feedback channel. Equation (8) decomposes the expected time in system into service time plus a queueing component that arises only when a patient stays with a non-idle usual GP. Both λ and α increase $W^{(\infty)}(p)$

by raising the total workload. The dependence on p is more nuanced: increasing p improves access by routing patients to idle GPs but also increases $\hat{\lambda}^{(\infty)}(p)$ through disrupted-continuity demand feedback. This tension between immediate load balancing and endogenous demand amplification drives the continuity-access trade-off.

Taken together, (7)–(8) provide a closed-form characterization of $\mathcal{C}^{(\infty)}(p)$, given by (2), and show that the effect of α and λ is monotone (higher demand or stronger continuity penalties worsen performance for any fixed policy), whereas the effect of p can be non-monotone.

Remark 1 (Hazard-rate alternative). Our baseline model treats a disrupted visit as generating α total consultation-equivalents of workload. An alternative micro-foundation instead models disrupted continuity as elevating the patient's next-visit *hazard rate* from λ to $\alpha\lambda$, so that a patient in the disrupted state returns sooner on average. Under that interpretation, the steady-state effective arrival rate $\hat{\lambda}_H^{(\infty)}(p)$ satisfies a quadratic self-consistency equation rather than the linear expression (7):

$$(\alpha-1)\,p\,(\hat{\lambda}_H^{(\infty)}(p))^2 - \alpha\,\hat{\lambda}_H^{(\infty)}(p) + \alpha\lambda = 0. \tag{9}$$

The distinction arises because, under the hazard-rate formulation, patients in the disrupted state cycle through visits faster and therefore spend *less* calendar time in that state—an effect that is absent when the disruption is modeled as additional workload. Formally, if a fraction $p\sigma$ of visits break continuity, the hazard-rate model requires averaging inter-visit *times* (harmonically) rather than visit *rates* (arithmetically), yielding $\hat{\lambda}_H(p) = \alpha\lambda/(\alpha-(\alpha-1)p\sigma)$ in place of $\hat{\lambda}(p) = \lambda(1+(\alpha-1)p\sigma)$.

The two formulations agree to first order in $(\alpha-1)$. Writing $\varepsilon := \alpha - 1$, a Taylor expansion of the stable root of (9) gives

$$\hat{\lambda}^{(\infty)}(p) - \hat{\lambda}_H^{(\infty)}(p) = O(\varepsilon^2), \tag{10}$$

with the leading correction term equal to $\varepsilon^2 p\lambda^2(1-p\lambda) + O(\varepsilon^3)$, which is strictly positive in the stability region. At the empirically calibrated value $\alpha = 1.169$ (Kajaria-Montag et al. 2024), the relative discrepancy $|\hat{\lambda}^{(\infty)}(p) - \hat{\lambda}_H^{(\infty)}(p)|/\hat{\lambda}^{(\infty)}(p)$ is below 0.7% across all (λ, p) in the stability region. In particular, numerical evaluation confirms that the regime classifications in Theorems 1 and 2, and the NHS case study results in Section 8, are unchanged when the hazard-rate formulation is used in place of the workload formulation. In the interest of analytical tractability and closed-form characterizations, we adopt the workload formulation.

3.4. Performance of the Approximation

We now show that the mean-field approximation is asymptotically accurate; that is, as $n \to \infty$, the stochastic finite-n system concentrates around the deterministic mean-field trajectory.

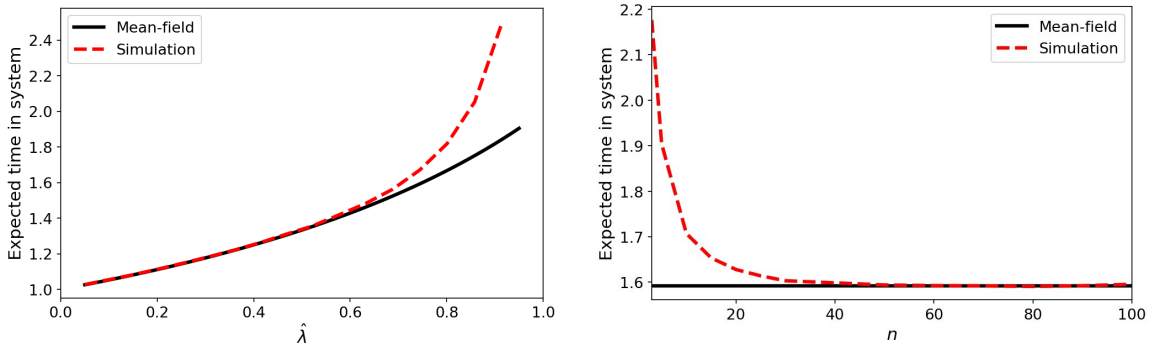
PROPOSITION 3 **(Convergence)**. *Fix* $p \in [0,1]$ *satisfying* (3)*, and assume that* $s^{(n)}(0;p) \Rightarrow s^{(\infty)}(0;p)$*, as* $n \to \infty$*, for some finite initial condition* $s^{(\infty)}(0;p)$*. Then, as* $n \to \infty$*,* $s^{(n)}(\,\cdot\,;p) \Rightarrow s^{(\infty)}(\,\cdot\,;p)$*, where* $s^{(\infty)}(\,\cdot\,;p)$ *is as defined in Lemma 1. Furthermore,* $\lim_{n\to\infty} W^{(n)}(p) = W^{(\infty)}(p)$*, where* $W^{(\infty)}(p)$ *is given by* (8)*.*

In addition to the above convergence result, we further establish a central-limit-type bound on the rate of convergence. Specifically, the finite-n system deviates from the mean-field limit by at most $\mathcal{O}(1/\sqrt{n})$ on any finite time horizon, consistent with the $1/\sqrt{n}$ scaling expected from a central-limit theorem. For an integer $m \geq 1$, define the truncated occupancy state for any finite-n system and for the infinite-server system (as $n \to \infty$) as $s^{(n)}_{\leq m}(t;p) := \left(s^{(n)}_1(t;p), \ldots, s^{(n)}_m(t;p)\right)$ and $s^{(\infty)}_{\leq m}(t;p) := \left(s^{(\infty)}_1(t;p), \ldots, s^{(\infty)}_m(t;p)\right)$.

PROPOSITION 4 **(Convergence rate)**. *Fix* $p \in [0,1]$ *satisfying* (3) *and a finite time horizon* $T < \infty$*. Assume the initial condition satisfies* $\mathbb{E}\|s^{(n)}_{\leq m}(0;p) - s^{(\infty)}_{\leq m}(0;p)\|_1 \leq C_0/\sqrt{n}$ *for some constant* C_0*, for all* n*. Then there exists a constant* $C_{T,m} < \infty$ *such that for all* n*:*

$$\mathbb{E}\left[\sup_{0\leq t\leq T}\left\|s^{(n)}_{\leq m}(t;p) - s^{(\infty)}_{\leq m}(t;p)\right\|_1\right] \leq \frac{C_{T,m}}{\sqrt{n}}. \tag{11}$$

The $1/\sqrt{n}$ bound holds for any fixed finite time horizon T. Extending it to steady-state metrics like $W^{(n)}(p)$ requires proving additional technical conditions (e.g., uniform mixing or an interchange-of-limits argument), which are beyond our scope. We instead validate the steady-state approximation via simulation, setting $\alpha = 1.169$[2] and averaging over 30 independent replications per parameter set.



(a) $n = 15$ (b) $\hat{\lambda} = 0.744$

Figure 1 Comparison of steady-state expected time in system from simulation $W^{(n)}$ (dashed) vs. mean-field predictions $W^{(\infty)}$ (solid). In this figure, $p = 0.5$, $\alpha = 1.169$.

Figure 1a compares the steady-state expected time in system from simulation (dashed) with the mean-field prediction (8) (solid), for a system with $n = 15$ GPs, routing parameter $p = 0.5$, and various per-GP effective arrival rates $\hat{\lambda}$. We observe that the mean-field approximation matches

simulation closely for light to moderate loads (e.g. $\hat{\lambda} = 0.4$ to 0.8). A noticeable gap emerges as $\hat{\lambda}$ approaches capacity (while preserving the correct qualitative trend), because at very high utilization, the finite system experiences strong queueing correlations and random fluctuations (which the deterministic mean-field model neglects), leading to larger discrepancies. Figure 1b examines how the approximation error diminishes as the number of GPs n grows, with $\hat{\lambda} = 0.744$ and $p = 0.5$. As n increases, the approximation error shrinks rapidly. Notably, increasing the system size from $n = 5$ to $n = 15$ reduces the error by well over half, consistent with the error bound in Proposition 4. Overall, these numerical experiments demonstrate that the mean-field model provides a good approximation even for moderately sized systems, and that its accuracy improves with system scale.

In the remainder of the paper, we leverage the tractability of the mean-field model to optimize the continuity-pooling trade-off. For ease of notation, we henceforth suppress the "(∞)" superscript when referring to mean-field quantities.

4. The Optimal Level of Continuity

We now use the closed-form mean-field expressions from Section 3 to study the *centralized* design problem: how should a system planner choose p to balance the short-run access benefit from pooling (higher p) against the long-run demand inflation from disrupted continuity (higher p when $\alpha > 1$)? This section proceeds in three steps. First, we express the planner's mean-field objective $\mathcal{C}(p)$ in a tractable form that isolates the aforementioned two channels. Second, we establish structural properties (convexity and boundary regimes) and characterize the optimal level of continuity and its comparative statics. Finally, we validate the analytical prescription against finite-n simulation to assess its accuracy for practice sizes observed in the NHS.

4.1. Mean-Field Formulation

In centralized systems where the provider controls assignment rules (e.g., the NHS), the planner selects a switching probability $p \in [0, 1]$ to minimize the steady-state cost rate. For any fixed p, the mean-field effective arrival rate per GP and the mean-field expected time in system are given by (7) and (8). Substituting these expressions into the cost rate per unit capacity (2) yields

$$\mathcal{C}(p) = \frac{\lambda C_v}{1 - p(\alpha - 1)\lambda} + \frac{\lambda C_w}{1 - \big(p(\alpha - 1) + (1 - p)\big)\lambda}. \tag{12}$$

Equation (12) isolates the two forces that determine the optimal continuity level. The first term captures *endogenous demand*: when $\alpha > 1$, increasing p breaks continuity more often, generates additional follow-up workload, and raises the long-run consultation-equivalent load and therefore the cumulative visit cost. The second term captures *access*: pooling reduces the fraction of arrivals that must join a busy usual GP and therefore can reduce delay costs. Writing the second denominator as $1 - \lambda - p(\alpha - 2)\lambda$ shows a useful dichotomy: the access term decreases in p when $\alpha < 2$ (pooling

improves access), but increases in p when $\alpha > 2$ (demand inflation is so strong that it dominates even the short-run congestion relief). This observation foreshadows why the optimal policy collapses to full continuity when α is sufficiently large.

Because disrupted continuity can inflate demand, not every $p \in [0,1]$ yields a stable system. The planner's choice is therefore restricted to the *stability region* where the stability condition (3) holds:

$$\mathcal{P} := \{p \in [0,1] : \big(1 + (\alpha - 1)p\big)\lambda < 1\}. \tag{13}$$

Let $\overline{p} := \sup \mathcal{P}$ denote the largest pooling level consistent with stability. If full pooling is stable, then $\overline{p} = 1$; otherwise, $\overline{p} < 1$ and the planner can interpret choosing $p = \overline{p}$ as choosing p arbitrarily close to $\overline{p}$ from below. The planner's mean-field problem is thus

$$p^\star \in \arg\min_{p \in [0,\overline{p}]} \mathcal{C}(p). \tag{14}$$

4.2. Structure of the Optimizer

We first establish convexity of the mean-field objective on the stability region. This implies the optimization problem is well behaved: if the optimum is interior, it is unique and characterized by a first-order condition.

PROPOSITION 5 **(Convexity)**. *Fix $\lambda \in (0,1)$ and $\alpha \geq 1$. The mean-field cost $\mathcal{C}(p)$ in (12) is convex on $[0,\overline{p}]$. Consequently, any interior minimizer is unique and characterized by the first-order condition $\mathcal{C}'(p) = 0$.*

Convexity in Proposition 5 implies that the optimal policy is either at a boundary of the feasible set (full continuity, $p = 0$, or maximal stable pooling, $p = \overline{p}$) or the unique interior point where the marginal access benefit of pooling is exactly offset by the marginal demand-inflation cost from disrupted continuity.

THEOREM 1 **(Optimal switching probability)**. *Fix $\lambda \in (0,1)$ and $\alpha \geq 1$. When $\alpha > 1$, define*

$$y_l := \frac{(2-\alpha)\left(1 - \overline{p}(\alpha-1)\lambda\right)^2}{(\alpha-1)\left(1 - \lambda - \overline{p}(\alpha-2)\lambda\right)^2}, \qquad y_u := \frac{2-\alpha}{(\alpha-1)(1-\lambda)^2}.$$

(i) If $\alpha = 1$, then $p^\star = 1$.

(ii) If $\alpha \in (1,2)$, then $0 < y_l < y_u$ and:

(a) if $C_v/C_w \leq y_l$, then $p^\star = \overline{p}$.

(b) if $C_v/C_w \in (y_l, y_u)$, then $p^\star \in (0,\overline{p})$ is the unique solution to $\mathcal{C}'(p) = 0$, given by

$$p^\star = \frac{\sqrt{2-\alpha} - (1-\lambda)\sqrt{(\alpha-1)C_v/C_w}}{\lambda\Big((2-\alpha)\sqrt{(\alpha-1)C_v/C_w} + (\alpha-1)\sqrt{2-\alpha}\Big)}.$$

Moreover, $p^\star$ is strictly decreasing in C_v/C_w and α, and strictly increasing in λ.

(c) if $C_v/C_w \geq y_u$, *then* $p^\star = 0$.

(iii) If $\alpha \geq 2$, *then* $p^\star = 0$.

Theorem 1 provides a simple regime-based prescription. The extreme cases are intuitive: when $\alpha = 1$, switching never increases workload, so full pooling is optimal; when $\alpha \geq 2$, demand amplification dominates access gains, so full continuity is optimal. In the empirically relevant region $\alpha \in (1, 2)$, the parameter space is partitioned into three regimes via cost-ratio cutoffs y_l and y_u: maximal stable pooling when visits are relatively cheap compared to delay ($C_v/C_w \leq y_l$), interior partial pooling when visit and delay costs are closely balanced ($C_v/C_w \in (y_l, y_u)$), and full continuity when visits are sufficiently costly ($C_v/C_w \geq y_u$). In the interior region, higher baseline load λ increases the marginal value of pooling for access, raising the interior optimum $p^\star$. By contrast, a larger demand penalty from disrupted continuity α or a larger visit-to-wait cost ratio C_v/C_w makes the system less tolerant of discontinuity-induced revisits, lowering $p^\star$.

Taken together, these results show that full continuity is not generically optimal even when continuity reduces downstream consultation workload. Instead, the optimal continuity target depends jointly on load (λ), the strength of demand feedback (α), and how the system values delay relative to visits (C_v/C_w). Managerially, this implies continuity targets should be practice-specific: tighter when continuity disruptions materially increase downstream workload (high α) or when additional consultations are costly (high C_v), and looser when heavy load makes timely access particularly valuable (high λ).

4.3. Finite-n Validation via Simulation

We next compare the mean-field prescription to finite-n simulation to assess its accuracy at realistic practice sizes and evaluate how deviations vary with practice size and load.

For each parameter setting, we simulate a finite-n system under the probabilistic routing policy over a fine grid of $p \in [0, 1]$ and compute the induced steady-state cost rate $\mathcal{C}^{(n)}(p)$ using the same cost structure as (2), over 30 independent simulation runs. We then define the simulation-based optimizer $p^{\star,(n)}$ as the minimizer of $\mathcal{C}^{(n)}(p)$, and compare it with the mean-field optimizer $p^\star$ from Theorem 1. We vary the cost ratio C_v/C_w (normalizing $C_w = 1$ without loss of generality), consider $n \in \{3, 7, 15\}$, and examine three baseline loads $\lambda \in \{0.65, 0.75, 0.85\}$,[3] which span a representative range of practice sizes and demand levels.

Our first observation from Figure 2 is that, across all system sizes (n) and demand levels (λ), the optimal switching probability $p^{\star,(n)}$ decreases in C_v/C_w, which matches the comparative statics in Theorem 1(ii)(b). Second, for any given visit-to-wait cost ratio (C_v/C_w), higher λ shifts the optimum $p^{\star,(n)}$ toward more pooling, again consistent with the comparative statics in Theorem 1(ii)(b). Third, the mean-field optimizer closely tracks finite-n optima for $\lambda \in \{0.65, 0.75\}$. Nevertheless, at

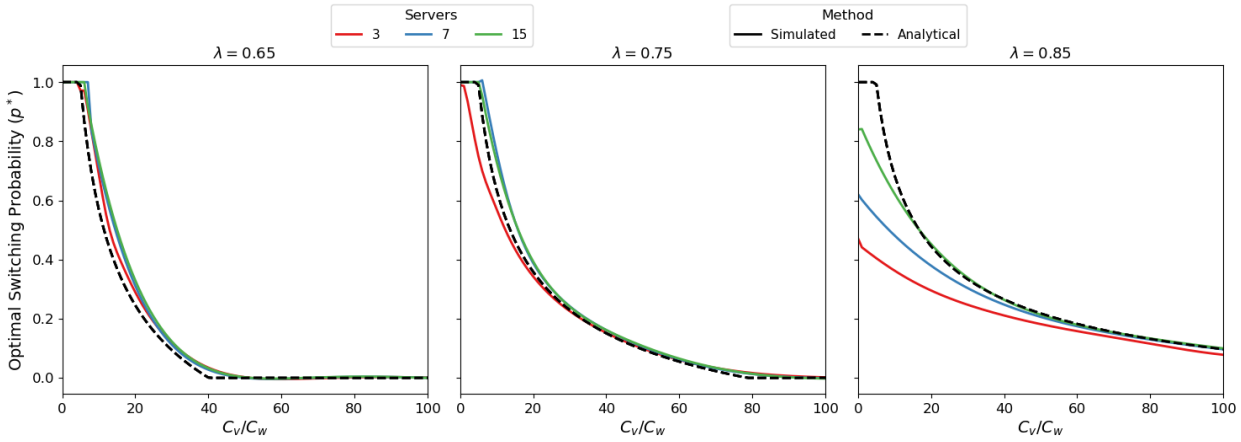


Figure 2 Optimal switching probability as a function of the cost ratio C_v/C_w. Solid colored curves show simulation-based optima $p^{\star,(n)}$ for $n \in \{3, 7, 15\}$; the dashed black curve show the mean-field optimizer $p^\star$ from Theorem 1. Panels correspond to $\lambda \in \{0.65, 0.75, 0.85\}$, with $\alpha = 1.169$ held fixed. In the figure, $\overline{p} = 1$.

$\lambda = 0.85$, deviations become larger, especially at low to moderate cost ratios, with the gap narrowing as n increases. Notably, the mean-field optimizer tends to recommend more pooling than is optimal in finite-n systems. This reflects a core finite-size effect that the mean-field model effectively treats switching as accessing an "instantaneously available" empty queue (an increasingly accurate approximation as n grows), whereas in small, highly utilized systems, the shortest queue is often nonempty. Consequently, the immediate access gains from switching are smaller in finite-n systems, leading to more continuity (lower $p^{\star,(n)}$) than the mean-field prescription.

Overall, the numerical results suggest that the mean-field optimizer provides a useful policy prescription for moderate to large practices and remains qualitatively informative even for smaller practices.

5. Patient Behavior and Equilibrium Analysis

Thus far, we study a *centralized* setting in which a healthcare system (e.g., the UK's NHS) assigns patients to GPs to minimize overall system cost. However, many real-world environments—such as private healthcare systems, decentralized online appointment platforms, or self-directed care models—grant patients significant autonomy over scheduling decisions. In these *decentralized* settings, each patient can choose whether to prioritize continuity by waiting for their usual GP or to switch to the shortest queue to reduce delay.

In this section, we examine such decentralized decision-making, where patients act independently and strategically to minimize their own expected cost. We define and characterize the equilibrium behavior of patients, compare the decentralized equilibrium to the system-optimal benchmark from Section 4, and assess the accuracy of a mean-field approximation in predicting equilibrium outcomes.

5.1. Modeling Patient Individual Optimization

We begin by modeling the behavior of individual patients acting in their own self-interest. When booking an appointment, a patient observes the queue length at the usual GP and the shortest queue among alternatives. If the usual GP is among the shortest (allowing ties), staying is strictly optimal: it preserves continuity without increasing waiting time. The only interesting case is when another GP has a strictly shorter queue. The patient then faces a genuine trade-off between continuity and access.

Consider a representative patient i who adopts a switching probability $p_i \in [0,1]$. If patient i's usual GP does not have the shortest queue at the time of an appointment request, she will switch to the shortest queue with probability p_i and stay with her usual GP with probability $1-p_i$. We model patient i's objective as minimizing her own long-run expected cost rate (cost per unit time), given by

$$\mathcal{C}_i(p_i, p_{-i}) = \hat{\lambda}_i(p_i, p_{-i}) \left(C_v + C_w\, W_i(p_i, p_{-i})\right), \tag{15}$$

where p_{-i} denotes the switching probabilities of all other patients. In this expression, $W_i(p_i, p_{-i})$ is the steady-state expected time in system experienced by patient i, and $\hat{\lambda}_i(p_i, p_{-i})$ is her effective appointment arrival rate accounting for continuity effects. Similar to (1) in Section 3, the effective arrival rate captures how disrupting continuity (by switching GP) can increase a patient's downstream consultation-equivalent workload. Specifically, we write:

$$\hat{\lambda}_i(p_i, p_{-i}) = \big(1 - \sigma_i(p_i, p_{-i})p_i\big)\lambda + \sigma_i(p_i, p_{-i})p_i \cdot \alpha\lambda = \big(1 + (\alpha - 1)p_i\sigma_i(p_i, p_{-i})\big)\lambda, \tag{16}$$

where λ is the baseline appointment arrival rate for patient i under full continuity, $\alpha > 1$ is a penalty factor capturing the increase in follow-up consultations when continuity is broken, and $\sigma_i(p_i, p_{-i})$ is the steady-state probability that patient i's usual GP does not have the shortest queue. In words, with probability $1 - \sigma_i(p_i, p_{-i})p_i$, the patient sticks with her usual GP, so her visit generates one consultation-equivalent of workload; with probability $\sigma_i(p_i, p_{-i})p_i$, she switches and breaks continuity, generating α consultation-equivalents of workload in total.

Each patient i independently chooses her switching probability p_i. However, her decision is complicated by the fact her waiting time W_i and the probability σ_i of facing a longer queue depend on the choices of all other patients. In other words, $\mathcal{C}_i(p_i, p_{-i})$ is influenced by the entire profile of others' switching probabilities. Patients thus engage in a non-cooperative queueing game, where each patient seeks to minimize her own cost (15) while anticipating the switching behavior of others. We focus on Nash equilibria, in which no patient can unilaterally reduce her expected cost by changing her strategy. Formally, $\tilde{p}$ is a Nash equilibrium if for every patient i, $\mathcal{C}_i(\tilde{p}_i, \tilde{p}_{-i}) \le \mathcal{C}_i(p_i, \tilde{p}_{-i})$ for all $p_i \in [0,1]$.

Given the ex ante homogeneity of patients and GPs, we restrict attention to symmetric Nash equilibria, in which all patients follow the same strategy. These strategies may be pure or mixed. In a pure-strategy equilibrium, $\tilde{p} \in [0,1]$ is a common fixed switching probability adopted by all patients. In a mixed-strategy equilibrium, $\tilde{p}$ is a common probability distribution over switching probabilities from which each patient independently draws a permanent type; that is, each patient commits to a single switching probability for all her visits rather than re-randomizing on each visit.

5.2. Mean-Field Equilibrium Analysis

Working in the mean-field regime (suppressing the "(∞)" superscript), a key property is that each patient's cost is *concave* in her own switching probability, which pins down the equilibrium structure.

PROPOSITION 6 **(Concavity)**. *Fix $\lambda \in (0,1)$ and $\alpha \geq 1$. For any given profile of other patients' strategies p_{-i}, the mean-field cost $\mathcal{C}_i(p_i, p_{-i})$ is a concave function of p_i under the stability condition* (3).

Because p_i is restricted to the feasible interval $[0, \overline{p}]$ (where $\overline{p}$ is the maximal switching probability consistent with stability), Proposition 6 implies a *bang-bang* best response: a cost-minimizing choice of p_i must occur at an endpoint, $p_i = 0$ (full continuity) or $p_i = \overline{p}$ (maximal stable pooling). Intuitively, increasing switching improves access by reallocating the patient toward shorter queues, but (when $\alpha > 1$) it also increases downstream workload by disrupting continuity. In the mean-field limit, this trade-off is concave in the switching probability, so there is no strict incentive to choose an interior switching probability.

Consequently, any *symmetric* Nash equilibrium must take one of three forms: (i) the pure strategy $\tilde{p} = 0$, (ii) the pure strategy $\tilde{p} = \overline{p}$ (and in particular $\overline{p} = 1$ when full pooling can stabilize the system), or (iii) a mixed strategy that randomizes between these two extremes.

THEOREM 2 **(Equilibrium switching probability)**. *Fix $\lambda \in (0,1)$ and $\alpha \geq 1$. When $\alpha > 1$, define*

$$z_l := \frac{(2-\alpha)(1-\overline{p}(\alpha-1)\lambda)}{(\alpha-1)(1-\lambda-\overline{p}(\alpha-2)\lambda)}, \qquad z_u := \frac{(2-\alpha)+(\alpha-1)\overline{p}\lambda}{(\alpha-1)(1-\lambda)}.$$

(i) If $\alpha = 1$, then $\tilde{p} = 1$.

(ii) If $\alpha \in (1, \frac{2-\overline{p}\lambda}{1-\overline{p}\lambda})$, then $z_l < z_u$ and $z_u > 0$:

- *(a) if $C_v/C_w \leq z_l$, then $\tilde{p} = \overline{p}$.*
- *(b) if $C_v/C_w \in (z_l, z_u)$, then $\tilde{p} = \overline{p}$ with probability $\tilde{q}$ and $\tilde{p} = 0$ with probability $1-\tilde{q}$, where $\tilde{q}$ is given by*

$$\tilde{q} = \frac{\hat{\lambda}^\dagger - \lambda}{(\alpha-1)\overline{p}\,\lambda\,\hat{\lambda}^\dagger}, \tag{17}$$

and $\hat{\lambda}^\dagger \in \big(\lambda,\, \lambda/\big[1-\overline{p}(\alpha-1)\lambda\big]\big)$ *is the unique root in that interval of*

$$(\alpha-1)\frac{C_v}{C_w}(1-\overline{p})\,\hat{\lambda}^2 - \left[(\alpha-1)\Big(\frac{C_v}{C_w}+1\Big)(1-\overline{p}\lambda) + \frac{C_v}{C_w}\right]\hat{\lambda} + \lambda\Big(\frac{C_v}{C_w}+1\Big) = 0. \tag{18}$$

(c) if $C_v/C_w \geq z_u$*, then* $\tilde{p}=0$*.*

(iii) If $\alpha \geq \frac{2-\overline{p}\lambda}{1-\overline{p}\lambda}$*, then* $\tilde{p}=0$*.*

Remark 2. Theorem 2(ii)(a) is non-empty only when $z_l > 0$, equivalently $\alpha < 2$. When $\alpha \geq 2$, $z_l \leq 0$ so the maximal pooling pure equilibrium vanishes and the equilibrium reduces to two regions: mixed or full continuity. When full pooling is feasible ($\overline{p}=1$), the quadratic (18) reduces to a linear equation, yielding the closed form $\hat{\lambda}^\dagger = \lambda(C_v/C_w+1)/\big[(\alpha-1)(C_v/C_w+1)(1-\lambda)+C_v/C_w\big]$.

Theorem 2 shows that equilibrium behavior is bang-bang: patients either never switch ($\tilde{p}=0$) or switch as much as stability permits ($\tilde{p}=\overline{p}$). Accordingly, the symmetric equilibrium is either one of these pure profiles or a mixed equilibrium in which the population splits between "switchers" ($\tilde{p}=\overline{p}$) and "stayers" ($\tilde{p}=0$) types, with $\tilde{q}$ denoting the population share. Specifically, a fraction $\tilde{q}$ of patients permanently adopt $\overline{p}$, while the remaining $1-\tilde{q}$ of patients permanently adopt 0. Which equilibrium outcome arises is determined jointly by the continuity-disruption penalty α and the visit-to-wait cost ratio C_v/C_w.

At one extreme, when $\alpha=1$, breaking continuity does not increase future consultation demand, so switching purely improves access, yielding $\tilde{p}=1$. At the other extreme, when $\alpha \geq \frac{2-\overline{p}\lambda}{1-\overline{p}\lambda}$, disruptions to continuity amplify demand so strongly that the induced additional workload outweighs any access gains from pooling. In this case, full continuity is individually optimal, i.e., $\tilde{p}=0$.

Between these extremes, the thresholds z_l and z_u partition the cost-ratio space into up to three regions: maximal stable pooling when waiting is sufficiently costly ($C_v/C_w \leq z_l$), full continuity when visit costs are sufficiently large ($C_v/C_w \geq z_u$), and a mixed-strategy equilibrium when the two costs are balanced ($C_v/C_w \in (z_l, z_u)$). In this intermediate region, neither pure strategy is self-sustaining: if all patients switch, queues equalize and the marginal access benefit of switching vanishes, making continuity strictly preferable; if all maintain continuity, queues become uneven and the incentive to defect to a shorter queue becomes strong. The unique resolution is a mixed-strategy equilibrium in which a fraction $\tilde{q}$ of patients choose $\overline{p}$ and the remaining $1-\tilde{q}$ choose 0, with $\tilde{q}$ exactly such that a representative patient is indifferent between the two extremes. Because stayers and switchers face different individual arrival rates (stayers at λ, switchers at $\lambda(1+(\alpha-1)\overline{p}\hat{\lambda})$), the indifference condition yields a quadratic in $\hat{\lambda}$ rather than a linear equation in $\tilde{q}$. The mixing probability $\tilde{q}$ bridges the two pure regimes: as $C_v/C_w \downarrow z_l$, $\tilde{q}\uparrow\overline{p}$; as $C_v/C_w \uparrow z_u$, $\tilde{q}\downarrow 0$.

5.3. Finite-n Validation via Simulation

We validate the mean-field equilibrium by simulating finite systems under the same setup as Section 4.3 (demand penalty $\alpha = 1.169$, number of servers $n \in \{3, 7, 15\}$, baseline load $\lambda \in \{0.65, 0.75, 0.85\}$). Figure 3 plots the equilibrium outcome $\tilde{q}$ against C_v/C_w, with solid curves representing simulation results and dashed curves the mean-field prediction (Theorem 2). Values $\tilde{q} = 0$ and $\tilde{q} = 1$ correspond to full continuity ($\tilde{p} = 0$) and full pooling ($\tilde{p} = \overline{p} = 1$), respectively, while interior values indicate mixed-strategy equilibria.

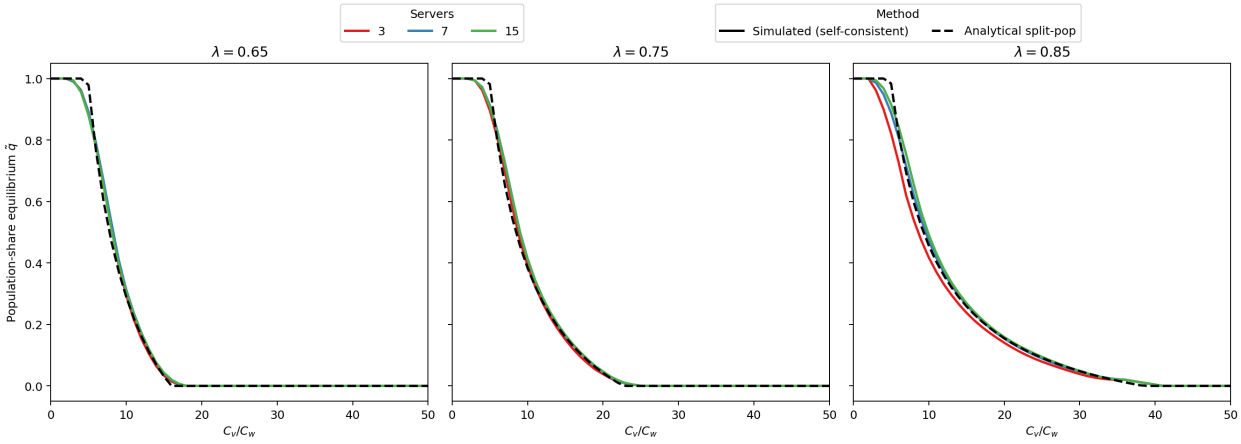


Figure 3 Population-share equilibrium $\tilde{q}$ as a function of the cost ratio C_v/C_w. Solid colored curves show simulation-based equilibria for $n \in \{3, 7, 15\}$; the dashed black curve shows the mean-field analytical prediction from Theorem 2. Panels correspond to $\lambda \in \{0.65, 0.75, 0.85\}$, with $\alpha = 1.169$ held fixed. In the figure, $\overline{p} = 1$.

The mean-field equilibrium closely tracks the finite-n equilibrium across nearly all parameter values. A slight deviation arises when the baseline arrival rate is high ($\lambda = 0.85$), as the effective load approaches capacity and the shortest queue in a finite system is often nonempty, consistent with Figure 2. Overall, the mean-field equilibrium provides a good approximation for moderately large systems.

5.4. Comparing Equilibrium to the Optimum

We next quantify how decentralized patient behavior compares to the optimal switching policy $p^\star$, characterized in Theorem 1. We summarize inefficiency using the *price of anarchy (PoA)*, defined as

$$\text{PoA} := \frac{\mathcal{C}_{\text{eq}}}{\mathcal{C}(p^\star)}, \tag{19}$$

where $\mathcal{C}$ is the cost rate defined in (2), $\mathcal{C}(p^\star)$ is the planner's minimum cost rate, and $\mathcal{C}_{\text{eq}}$ is the population-average cost rate at equilibrium, characterized in Theorem 2. In a pure-strategy equilibrium ($\tilde{p} \in \{0, \overline{p}\}$), $\mathcal{C}_{\text{eq}} = \mathcal{C}(\tilde{p})$. In a mixed-strategy equilibrium, where a fraction $\tilde{q}$ of patients

permanently adopt $\overline{p}$ and the rest adopt 0, $\mathcal{C}_{\text{eq}}$ is the population-weighted average of the two types' costs in the resulting population-split steady state. By construction, PoA ≥ 1; in particular, PoA $= 1$ means no inefficiency, and larger PoA indicates greater inefficiency due to uncoordinated patients' decisions.

PROPOSITION 7 **(Price of anarchy)**. *Fix $\lambda \in (0,1)$ and $\alpha \geq 1$. Let y_l and y_u be as defined in Theorem 1 and z_l and z_u be as defined in Theorem 2.*

(i) If $\alpha = 1$, then PoA $= 1$.

(ii) If $\alpha \in (1, \frac{2-\overline{p}\lambda}{1-\overline{p}\lambda})$, then PoA $= 1$ when $C_v/C_w \leq z_l$ or $C_v/C_w \geq \max\{y_u, z_u\}$, and $1 <$ PoA $\leq 1 + \max\left\{\frac{(2-\alpha)(\alpha-1)\,\lambda^2}{1-\lambda}, (\alpha-2)\lambda\right\}$ otherwise.

(iii) If $\alpha \geq \frac{2-\overline{p}\lambda}{1-\overline{p}\lambda}$, then PoA $= 1$.

Proposition 7 shows that PoA $= 1$ at the cost-ratio extremes and in the polar α-regimes, and is bounded in between. Inefficiency arises only when $\alpha \in (1, \frac{2-\overline{p}\lambda}{1-\overline{p}\lambda})$ and C_v/C_w is intermediate, where the continuity-access trade-off is nontrivial. In this case, the planner prefers an interior policy, but individual best responses are bang-bang, so decentralized behavior cannot replicate the planner's optimum. Fortunately, we can bound the maximum inefficiency by a closed-form upper bound (albeit not tight) that is strictly increasing in load λ.

The threshold ordering clarifies the *direction* of the distortion. One can show $z_l < y_l$, implying that at low-to-moderate cost ratios, the planner prefers maximal pooling, while patients already mix with full continuity, leading to excessive continuity. The mechanism is that patients do not internalize the congestion-relief benefit that their switching would provide to others—a positive externality of pooling. On the high-cost side, the ordering of z_u and y_u depends on load. When λ is high, $z_u < y_u$, so patients choose full continuity while the planner still prefers partial pooling, again implying excessive continuity. Under high load, therefore, excessive continuity is the predominant distortion across the cost-ratio spectrum, because the positive externality of pooling grows with load. Conversely, when λ is low, $z_u > y_u$, so patients continue mixing after the planner prefers full continuity, resulting in patients exercising excessive pooling. In this case, patients fail to internalize the demand externality from disrupted continuity.

Figure 4 plots PoA against C_v/C_w for a representative parameter set. Consistent with Proposition 7(ii), PoA equals 1 at very low and very high C_v/C_w, where equilibrium and optimum coincide. PoA rises above 1 in the intermediate range, peaking where the continuity-access trade-off is tightest. In this numerical example, however, the maximum PoA is approximately 1.05, indicating a relatively modest welfare loss. Moreover, PoA is larger under heavier load, as the congestion externality of each switching decision is amplified when queues are longer and more sensitive to incremental demand. Managerially, coordination is most valuable when α and C_v/C_w are intermediate and the

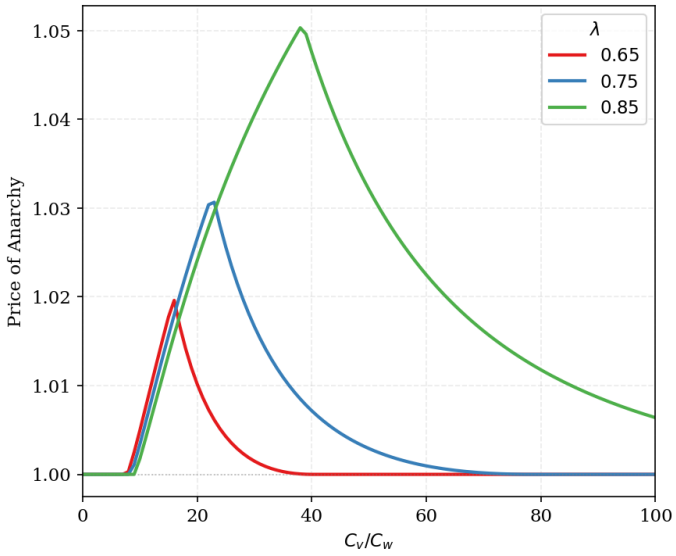


Figure 4 **PoA as a function of the cost ratio C_v/C_w. In the figure, $\alpha = 1.169$.**

system is highly utilized. In this region, interventions that partially internalize externalities—such as structured triage or targeted appointment guidance—can move individual behavior toward the optimum and reduce welfare loss.

6. Heterogeneous Patient Model

The homogeneous-patient model captures the fundamental continuity-pooling trade-off, but in practice, healthcare systems serve diverse populations. Patients with chronic or complex conditions benefit substantially from relational continuity, whereas those with acute, routine concerns may prioritize shorter waiting times and derive little value from seeing the same physician. To reflect this heterogeneity, we extend the model to two distinct patient types—those who benefit from continuity and those who do not.

This section proceeds as follows. Section 6.1 introduces the two-type model and characterizes its mean-field steady state. We then study the centralized problem under (i) a type-agnostic policy assigning a common switching probability regardless of type (Section 6.2) and (ii) a type-dependent policy (Section 6.3). Section 6.4 analyzes the decentralized equilibrium and compares it with the centralized optimum.

6.1. Problem Setting and Mean-Field Model

We consider a primary care practice serving two distinct types of patients: type 1 patients, who benefit from continuity of care, and type 2 patients, whose health outcomes are independent of relational continuity. The queueing structure (n parallel servers, Poisson arrivals, FCFS, exponential service at rate $\mu = 1$) is as in Section 3. Type $k \in \{1, 2\}$ patients arrive to the system at rate $n\lambda_k$

and switch to the shortest queue with probability $p_k \in [0,1]$ when their usual GP's queue is not the shortest.

Switching disrupts continuity for type 1 patients, resulting in a workload amplification factor $\alpha_1 \geq 1$: each disrupted visit generates α_1 consultation-equivalents of workload in total, with larger α_1 corresponding to greater penalties from disrupted care. For type 2 patients, continuity is not relevant, so $\alpha_2 = 1$. Let $C_{v,k}$ and $C_{w,k}$ denote the unit visit and waiting cost for type k patients, respectively. When type 1 and type 2 patients adopt switching probabilities p_1 and p_2, the expected cost per unit time for a type k patient is

$$\mathcal{C}_k(p_1,p_2) := \hat{\lambda}_k(p_1,p_2)\,\left(C_{v,k} + C_{w,k}\,W_k(p_1,p_2)\right), \tag{20}$$

where

$$\hat{\lambda}_k(p_1,p_2) = (1-\sigma_k(p_1,p_2)\cdot p_k)\,\lambda_k + \sigma_k(p_1,p_2)\cdot p_k \cdot \alpha_k\lambda_k = (1+(\alpha_k-1)p_k\sigma_k(p_1,p_2))\,\lambda_k \tag{21}$$

denotes the effective arrival rate for type k patients. Here, $\sigma(p_1,p_2) := \sigma_1(p_1,p_2) = \sigma_2(p_1,p_2)$ is the type-independent steady-state probability that a patient's usual GP is not among the shortest queues, and $W_k(p_1,p_2)$ is the steady-state expected waiting time for a type k patient. Let $\hat{\lambda}(p_1,p_2) = \hat{\lambda}_1(p_1,p_2) + \hat{\lambda}_2(p_1,p_2)$ denote the total effective arrival rate of patients to the system. For stability, we require that $\hat{\lambda}(p_1,p_2) < 1$. Since $\sigma(p_1,p_2) \leq 1$, a sufficient condition is

$$(1+(\alpha_1-1)p_1))\,\lambda_1 + (1+(\alpha_2-1)p_2))\,\lambda_2 < 1 \quad \Leftrightarrow \quad (1+(\alpha_1-1)p_1))\,\lambda_1 + \lambda_2 < 1. \tag{22}$$

As in Section 3.2, we analyze the mean-field limit as $n \to \infty$, using the occupancy process $s_i^{(n)}(t;p_1,p_2)$ (fraction of GPs with at least i patients). We use superscript "(∞)" to denote limiting quantities as $n \to \infty$.

Lemma 2 **(Mean-field system evolution (2-type))**. *Fix $p_1 \in [0,1]$ and $p_2 \in [0,1]$ satisfying (22), and assume that $s^{(\infty)}(0;p_1,p_2)$ is a finite initial condition. Assume that mean-field effective arrival rates $\hat{\lambda}_k^{(\infty)}(p_1,p_2) \in (0,1)$ for $k \in \{1,2\}$ exist. Then, the mean-field occupancy process $s^{(\infty)}(t;p_1,p_2)$ satisfies the following for all $t \geq 0$:*

$$\frac{d}{dt}s_1^{(\infty)}(t;p_1,p_2) = \sum_{k=1}^{2}\Big(1-(1-p_k)s_1^{(\infty)}(t;p_1,p_2)\Big)\hat{\lambda}_k^{(\infty)}(p_1,p_2) - \Big(s_1^{(\infty)}(t;p_1,p_2) - s_2^{(\infty)}(t;p_1,p_2)\Big), \tag{23}$$

$$\frac{d}{dt}s_i^{(\infty)}(t;p_1,p_2) = \sum_{k=1}^{2}(1-p_k)\Big(s_{i-1}^{(\infty)}(t;p_1,p_2) - s_i^{(\infty)}(t;p_1,p_2)\Big)\hat{\lambda}_k^{(\infty)}(p_1,p_2) - \Big(s_i^{(\infty)}(t;p_1,p_2) - s_{i+1}^{(\infty)}(t;p_1,p_2)\Big), \forall i \geq 2. \tag{24}$$

The ODE system generalizes Lemma 1 to two patient types. The arrival-rate terms now sum over both types, each weighted by its pooling probability $1-p_k$.

Setting the ODEs to zero yields the fixed point of the limiting system.

PROPOSITION 8 **(Mean-field fixed point (2-type))**. *Fix $p_1 \in [0,1]$ and $p_2 \in [0,1]$ satisfying (22), and assume that $s^{(\infty)}(0;p_1,p_2)$ is a finite initial condition. Assume that mean-field effective arrival rates $\hat{\lambda}_k^{(\infty)}(p_1,p_2) \in (0,1)$ for $k \in \{1,2\}$ exist, with $\hat{\lambda}^{(\infty)}(p_1,p_2) = \hat{\lambda}_1^{(\infty)} + \hat{\lambda}_2^{(\infty)}$ denoting the total. The ODE system (23)–(24) admits a unique fixed point $s^{(\infty)}(\infty;p_1,p_2)$ given by*

$$s_i^{(\infty)}(\infty;p_1,p_2) = \Big((1-p_1)\hat{\lambda}_1^{(\infty)}(p_1,p_2) + (1-p_2)\hat{\lambda}_2^{(\infty)}(p_1,p_2)\Big)^{i-1} \cdot \hat{\lambda}^{(\infty)}(p_1,p_2), \quad \forall i \geq 1.$$

As in Proposition 1, the steady-state occupancy decays geometrically, with tail rate $(1-p_1)\hat{\lambda}_1^{(\infty)} + (1-p_2)\hat{\lambda}_2^{(\infty)}$. From this fixed point, we derive the type-specific performance measures.

PROPOSITION 9 **(Steady-state performance (2-type))**. *Fix $p_1 \in [0,1]$ and $p_2 \in [0,1]$ satisfying (22). In the mean-field steady state, the following hold for each $k \in \{1,2\}$:*

$$\begin{aligned}
\sigma^{(\infty)}(p_1,p_2) &= \hat{\lambda}^{(\infty)}(p_1,p_2) = \frac{\lambda_1+\lambda_2}{1-p_1(\alpha_1-1)\lambda_1},\\
\hat{\lambda}_k^{(\infty)}(p_1,p_2) &= \big(1+(\alpha_k-1)\,p_k\,\hat{\lambda}^{(\infty)}(p_1,p_2)\big)\lambda_k, \quad k\in\{1,2\},\\
W_k^{(\infty)}(p_1,p_2) &= 1 + \frac{(1-p_k)(\lambda_1+\lambda_2)}{1-\big(p_1(\alpha_1-1)+(1-p_1)\big)\lambda_1-(1-p_2)\lambda_2+(\alpha_1-1)p_1\lambda_1\lambda_2(p_1-p_2)}.
\end{aligned}$$

In particular, since $\alpha_2 = 1$, we have $\hat{\lambda}_2^{(\infty)} = \lambda_2$ and $\hat{\lambda}_1^{(\infty)} = \lambda_1\big(1+(\alpha_1-1)p_1\lambda_2\big)/\big(1-p_1(\alpha_1-1)\lambda_1\big)$.

As in the case of the homogeneous-patient model, Proposition 9 verifies the existence of $\hat{\lambda}_k^{(\infty)}$ assumed in Lemma 2 and Proposition 8. The key structural difference from the homogeneous-patient model is that only type 1 switching inflates total effective demand ($\hat{\lambda}_2^{(\infty)} = \lambda_2$ regardless of p_2), while type 2 switching affects waiting times by reshaping the queue-length distribution. Additionally, the denominator of W_k contains a cross-type interaction term $(\alpha_1-1)p_1\lambda_1\lambda_2(p_1-p_2)$ that vanishes when $p_1 = p_2$ (as in an across-the-board policy that we study next) or when $\lambda_2 = 0$ (as in the homogeneous-patient setting). This cross-type externality will play a central role in the type-dependent analysis later.

For ease of notation, we suppress the "(∞)" superscript when referring to mean-field quantities in the remainder of the section.

6.2. Across-the-board Policy

We begin with an *across-the-board* policy, under which every patient regardless of type follows the same switching probability, denoted by $p_a \in [0,1]$. This rule requires no clinical information at booking and provides the benchmark against which type-dependent rules (Section 6.3) are measured.

For a given across-the-board switching probability $p_a \in [0,1]$, the total cost function is given by

$$\mathcal{C}_a(p_a) := \mathcal{C}(p_a,p_a) = \sum_{k=1}^{2} \hat{\lambda}_k(p_a,p_a)\,(C_{v,k} + C_{w,k}\,W_k(p_a,p_a)). \tag{25}$$

Let $\overline{p}_1 := \sup\{p_a \in [0,1] : (1+(\alpha_1-1)p_a)\lambda_1 + \lambda_2 < 1\}$ denote the maximal stable switching probability. Although both types adopt the same p_a, the stability constraint (22) is driven entirely by type 1 patients, because type 2 patients generate no demand amplification ($\alpha_2 = 1$). The planner solves

$$p_a^\star \in \arg\min_{p_a \in [0,\overline{p}_1]} \mathcal{C}_a(p_a). \tag{26}$$

The following result characterizes the cost rate $\mathcal{C}_a(p_a)$ and the unique across-the-board switching probability that minimizes it.

THEOREM 3 **(Optimal across-the-board switching probability (2-type))**. *Fix $\lambda_1 \in (0,1)$, $\lambda_2 \in (0,1)$ with $\lambda_1 + \lambda_2 < 1$, and $\alpha_1 \geq 1$. When $\alpha_1 > 1$, define*

$$y_l^a := \frac{(\alpha_1-1)\lambda_1\,\varphi}{(1-\lambda_1-\lambda_2)^2}, \qquad y_u^a := \frac{\varphi}{(\alpha_1-1)\lambda_1\,(1-\lambda_1-\lambda_2)^2},$$

where $\varphi := \lambda_1\big[(2-\alpha_1)+(\alpha_1-1)\lambda_2\big] + \frac{C_{w,2}}{C_{w,1}}\,\lambda_2\big[1-(\alpha_1-1)\lambda_1\big]$.

(i) *If $\alpha_1 = 1$, then $p_a^\star = 1$.*

(ii) *If $\alpha_1 \in (1,2)$, then $\mathcal{C}_a$ is convex on $[0,\overline{p}_1]$. Moreover, $0 < y_l^a < y_u^a$ and:*

(a) If $C_{v,1}/C_{w,1} \leq y_l^a$, then $p_a^\star = \overline{p}_1$.

(b) If $C_{v,1}/C_{w,1} \in (y_l^a, y_u^a)$, then $p_a^\star \in (0,\overline{p}_1)$ is the unique solution to $\mathcal{C}_a'(p_a) = 0$, given by

$$p_a^\star = \frac{\sqrt{\varphi} \;-\; (1-\lambda_1-\lambda_2)\,\sqrt{(\alpha_1-1)\lambda_1\,C_{v,1}/C_{w,1}}}{(\alpha_1-1)\lambda_1\,\sqrt{\varphi} \;+\; \big[(2-\alpha_1)\lambda_1+\lambda_2\big]\,\sqrt{(\alpha_1-1)\lambda_1\,C_{v,1}/C_{w,1}}}.$$

(c) If $C_{v,1}/C_{w,1} \geq y_u^a$, then $p_a^\star = 0$.

(iii) *If $\alpha_1 \geq 2$:*

(a) If $\varphi \leq 0$, then $\mathcal{C}_a'(p_a) > 0$ for all $p_a \in (0,\overline{p}_1)$ and $p_a^\star = 0$.

(b) If $\varphi > 0$ and $(2-\alpha_1)\lambda_1 + \lambda_2 \geq 0$, then $\mathcal{C}_a$ is convex on $[0,\overline{p}_1]$. Moreover, $0 < y_l^a < y_u^a$, and $p_a^\star = \overline{p}_1$ if $C_{v,1}/C_{w,1} \leq y_l^a$, $p_a^\star \in (0,\overline{p}_1)$ is the unique solution to $\mathcal{C}_a'(p_a) = 0$ if $C_{v,1}/C_{w,1} \in (y_l^a, y_u^a)$, and $p_a^\star = 0$ if $C_{v,1}/C_{w,1} \geq y_u^a$.

(c) If $\varphi > 0$ and $(2-\alpha_1)\lambda_1 + \lambda_2 < 0$, then $\mathcal{C}_a$ need not be convex. The optimum $p_a^\star$ is found by comparing $\mathcal{C}_a(0)$, $\mathcal{C}_a(\overline{p}_1)$, and $\mathcal{C}_a$ at an interior critical point.

When $\alpha_1 \in (1,2)$, $\mathcal{C}_a$ is convex and the three-regime characterization of Theorem 1(ii) carries over. The novel feature arises when $\alpha_1 \geq 2$: the homogeneous-patient model prescribes full continuity unconditionally (Theorem 1(iii)), but type 2 patients can restore the value of pooling. This happens when $\varphi > 0$, i.e., type 2's pure pooling benefit outweighs type 1's demand amplification. A second condition, $\lambda_2 \geq (\alpha_1 - 2)\lambda_1$, ensures type 2 patients are numerous enough for their pooling to keep the access channel improving with p_a, thus preserving convexity of $\mathcal{C}_a$. When both hold, the three-regime structure when $\alpha_1 \in (1,2)$ re-emerges intact. When $\varphi > 0$ but type 2 patients are too rare to

preserve convexity, the optimum is found by direct comparison of candidate points. The across-the-board result exposes the fundamental limitation of a one-size-fits-all policy. By constraining both types to a common switching probability, it necessarily under-pools for type 2 patients (who would optimally always switch) and over-pools for type 1 patients whenever $p_a^\star < 1$.

6.3. Type-dependent Optimal Policy

We now allow the planner to assign distinct switching probabilities to each patient type. For a given pair of type-dependent switching probabilities $(p_1, p_2) \in [0,1]^2$, the total cost function is given by

$$\mathcal{C}(p_1, p_2) = \sum_{k=1}^{2} \hat{\lambda}_k(p_1, p_2)\left(C_{v,k} + C_{w,k} W_k(p_1, p_2)\right). \tag{27}$$

As in the across-the-board case, the stability constraint binds only through type 1 ($\alpha_2 = 1$), so $\overline{p}_1$ is the same bound defined above. Type 2 faces no stability constraint and can choose any $p_2 \in [0,1]$. The planner solves

$$(p_1^\star, p_2^\star) \in \underset{p_1 \in [0,\overline{p}_1], p_2 \in [0,1]}{\arg\min} \mathcal{C}(p_1, p_2). \tag{28}$$

The following result identifies the optimal type-dependent switching probabilities.

THEOREM 4 **(Optimal type-dependent switching probability (2-type))**. *Fix $\lambda_1 \in (0,1)$, $\lambda_2 \in (0,1)$ with $\lambda_1 + \lambda_2 < 1$, and $\alpha_1 \geq 1$. The mean-field cost $\mathcal{C}(p_1, p_2)$ in (27) is strictly decreasing in $p_2 \in [0,1]$ for all p_1, so $p_2^\star = 1$. The mean-field cost $\mathcal{C}(p_1, 1)$ has at most one interior local minimum and one interior local maximum on $(0, \overline{p}_1)$. When $\alpha_1 > 1$, define*

$$y_l' := \frac{\left[(2-\alpha_1) + (\alpha_1 - 1)\lambda_2(2\overline{p}_1 - 1)\right]\left(1 - \overline{p}_1(\alpha_1 - 1)\lambda_1\right)^2}{(\alpha_1 - 1)\left[(1-\lambda_1) + \overline{p}_1\lambda_1\big((2-\alpha_1) - (\alpha_1-1)\lambda_2\big) + (\alpha_1 - 1)\overline{p}_1^2\lambda_1\lambda_2\right]^2}, \quad y_u' := \frac{(2-\alpha_1) - (\alpha_1 - 1)\lambda_2}{(\alpha_1 - 1)(1-\lambda_1)^2}.$$

Let $\overline{\alpha}_1 > 1$ be the unique value at which $y_l' = 0$.

(i) *If $\alpha_1 = 1$, then $p_1^\star = 1$.*

(ii) *If $\alpha_1 \in (1, \overline{\alpha}_1)$, then $y_l' > 0$ and:*
   - *(a) if $C_{v,1}/C_{w,1} \leq \min(y_l', y_u')$, then $p_1^\star = \overline{p}_1$.*
   - *(b) if $C_{v,1}/C_{w,1} \in (y_l', y_u')$, then $p_1^\star \in (0, \overline{p}_1)$ is the unique solution to $\mathcal{C}'(p_1, 1) = 0$.*
   - *(c) if $C_{v,1}/C_{w,1} \in [y_u', y_l']$, then $p_1^\star = \arg\min_{p_1 \in \{0, \overline{p}_1\}} \mathcal{C}(p_1, 1)$.*
   - *(d) if $C_{v,1}/C_{w,1} \geq \max(y_l', y_u')$, then $p_1^\star$ is found by comparing $\mathcal{C}(0,1)$ with $\mathcal{C}(p_1, 1)$ at any interior local minimum.*

(iii) *If $\alpha_1 \geq \overline{\alpha}_1$, then $y_l' \leq 0$ and $p_1^\star = 0$.*

REMARK 3. When $\lambda_2 = 0$, $\overline{\alpha}_1 = 2$ and the thresholds y_l', y_u' reduce to y_l, y_u in Theorem 1, recovering the homogeneous-patient case exactly. Let $\alpha_1^\dagger \in (1, \overline{\alpha}_1)$ be the unique value at which $y_l' = y_u'$. Then, Theorem 4(ii)(b) is non-empty only when $\alpha_1 < \alpha_1^\dagger$, in which case $y_l' < y_u'$, and Theorem 4(ii)(c) is non-empty only when $\alpha_1 \geq \alpha_1^\dagger$, in which case $y_u' \leq y_l'$.

Theorem 4 delivers three structural prescriptions. First, type 2 patients should always join the shortest queue ($p_2^\star = 1$), because switching carries no demand penalty ($\alpha_2 = 1$) and purely improves access. Second, when the penalty of disrupted continuity is sufficiently severe ($\alpha_1 \geq \overline{\alpha}_1$), type 1 patients should always stay with their usual GP ($p_1^\star = 0$) regardless of the cost ratio, which is analogous to the $\alpha_1 \geq 2$ case in the homogeneous-patient model (Theorem 1(iii)). Third, for moderate penalties ($\alpha_1 < \overline{\alpha}_1$), the optimal type 1 policy exhibits the same cost-ratio threshold structure as in Theorem 1(ii). Specifically, if visits are cheap relative to waiting, the planner prescribes maximal stable pooling ($p_1^\star = \overline{p}_1$); if visits are sufficiently costly, full continuity is optimal ($p_1^\star = 0$); for intermediate cost ratios, the planner implements partial pooling through a unique interior optimum when the cost function is convex.

The key departure from the homogeneous-patient setting is that convexity of the cost $\mathcal{C}(p_1, 1)$ in p_1 is not guaranteed when $\lambda_2 > 0$. Intuitively, type 2 pooling can either soften or reinforce the case for type 1 switching—softening it when type 2 already equalizes queues (substitution), reinforcing it when type 2 pooling reduces the congestion cost of type 1 switching (complementarity)—and this ambiguity is what breaks convexity. In the homogeneous-patient model ($\lambda_2 = 0$), the cost is convex and the interior optimum, when it exists, is unique (Proposition 5). With two types, the quadratic interaction between type 1 switching and type 2 pooling can create a local cost maximum, so the cost landscape over $[0, \overline{p}_1]$ may be non-monotone. When this occurs, we show that $\mathcal{C}(p_1, 1)$ has at most one interior local minimum and one interior local maximum, so the optimization reduces to evaluating at most three candidate points.

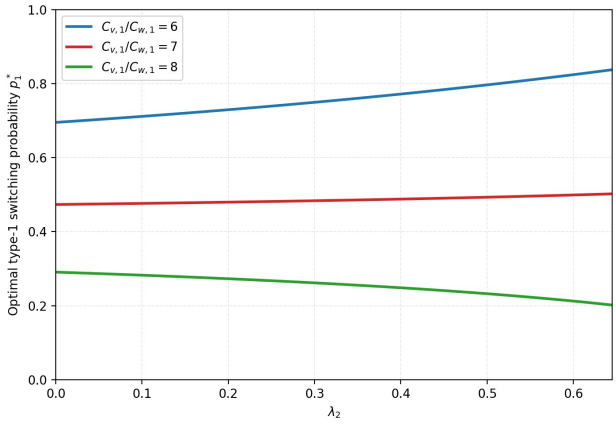


Figure 5 **Optimal type 1 switching probability $p_1^\star$ as a function of the type 2 arrival rate λ_2, for three values of the cost ratio $C_{v,1}/C_{w,1}$. In this figure, $\lambda_1 = 0.3$, $\alpha_1 = 1.169$.**

Figure 5 illustrates the substitution-complementarity pattern. When visit costs are high ($C_{v,1}/C_{w,1} = 8$), type 2 pooling substitutes for type 1 switching, enabling the planner to reduce

$p_1^\star$ and preserve more continuity. When visit costs are low ($C_{v,1}/C_{w,1}=6$), type 2 pooling complements type 1 switching by reducing its congestion cost, pushing $p_1^\star$ higher. At an intermediate ratio, the effects cancel and $p_1^\star$ is insensitive to λ_2. This highlights that a planner who can differentiate between patient types can leverage the load-balancing externality of type 2 pooling to fine-tune the continuity prescription for type 1 patients.

6.4. Equilibrium Analysis

We now analyze the decentralized equilibrium, in which each patient type independently chooses a switching probability to minimize her own expected cost. A Nash equilibrium is defined as a pair of switching probabilities $(\tilde{p}_1, \tilde{p}_2)$ that satisfy the following two best-response conditions simultaneously:

$$\tilde{p}_1 \in \arg\min_{p_1\in[0,\overline{p}_1]} \mathcal{C}_1(p_1,\tilde{p}_1,\tilde{p}_2), \qquad \tilde{p}_2 \in \arg\min_{p_2\in[0,1]} \mathcal{C}_2(p_2,\tilde{p}_1,\tilde{p}_2), \tag{29}$$

where $\mathcal{C}_k(p_k,\tilde{p}_1,\tilde{p}_2)$ denotes the cost rate for a representative type k patient switching with probability p_k, given the common strategies $(\tilde{p}_1,\tilde{p}_2)$ adopted by all others.

THEOREM 5 **(Equilibrium switching probability (2-type))**. *Fix $\lambda_1\in(0,1)$, $\lambda_2\in(0,1)$ with $\lambda_1+\lambda_2<1$, and $\alpha_1\geq 1$. When $\alpha_1>1$, define*

$$z_l' := \frac{\big[(2-\alpha_1)+(\alpha_1-1)\lambda_2(\overline{p}_1-1)\big]\,D(\overline{p}_1)}{(\alpha_1-1)\,E(\overline{p}_1)}, \qquad z_u' := \frac{(2-\alpha_1)+(\alpha_1-1)\big(\overline{p}_1(\lambda_1+\lambda_2)-\lambda_2\big)}{(\alpha_1-1)(1-\lambda_1)},$$

where $D(p_1) := 1-(\alpha_1-1)p_1\lambda_1$ and $E(p_1) := (1-\lambda_1)+p_1\lambda_1\big[(2-\alpha_1)-(\alpha_1-1)\lambda_2\big]+(\alpha_1-1)p_1^2\lambda_1\lambda_2$. Let $\alpha_1^{\ddagger\prime} := \frac{2-\overline{p}_1\lambda_1+(1-\overline{p}_1)\lambda_2}{1-\overline{p}_1\lambda_1+(1-\overline{p}_1)\lambda_2}>1$ at which $z_u'=0$. Then, type 2 patients play $\tilde{p}_2=1$ in every equilibrium. For type 1 patients:

(i) *If $\alpha_1=1$, then $\tilde{p}_1=1$.*

(ii) *If $\alpha_1\in(1,\alpha_1^{\ddagger\prime})$, then $z_l'<z_u'$ and $z_u'>0$:*

(a) if $C_{v,1}/C_{w,1}\leq z_l'$, then $\tilde{p}_1=\overline{p}_1$.

(b) if $C_{v,1}/C_{w,1}\in(z_l',z_u')$, then $\tilde{p}_1=\overline{p}_1$ with probability $\tilde{q}_1$ and $\tilde{p}_1=0$ with probability $1-\tilde{q}_1$, where

$$\tilde{q}_1 = \frac{\hat{\lambda}^\dagger-\Lambda}{(\alpha_1-1)\overline{p}_1\lambda_1\,\hat{\lambda}^\dagger}, \tag{30}$$

and $\hat{\lambda}^\dagger\in\big(\Lambda,\,\Lambda/D(\overline{p}_1)\big)$ is the unique root in that interval of

$$(\alpha_1-1)\frac{C_{v,1}}{C_{w,1}}(1-\overline{p}_1)\,\hat{\lambda}^2-\left[(\alpha_1-1)\Big(\frac{C_{v,1}}{C_{w,1}}+1\Big)(1+\lambda_2-\overline{p}_1\Lambda)+\frac{C_{v,1}}{C_{w,1}}\right]\hat{\lambda}+\Lambda\Big(\frac{C_{v,1}}{C_{w,1}}+1\Big)=0. \tag{31}$$

(c) if $C_{v,1}/C_{w,1}\geq z_u'$, then $\tilde{p}_1=0$.

(iii) *If* $\alpha_1 \geq \alpha_1^{\ddagger\prime}$, *then* $\tilde{p}_1 = 0$.

REMARK 4. Theorem 5(ii)(a) is non-empty only when $z_l' > 0$, equivalently $\alpha_1 < \alpha_1^{\dagger\prime} := \frac{2+(1-\overline{p}_1)\lambda_2}{1+(1-\overline{p}_1)\lambda_2}$. When $\alpha_1 \geq \alpha_1^{\dagger\prime}$, $z_l' \leq 0$ and the equilibrium reduces to two sub-cases: mixed or full continuity. When $\lambda_2 = 0$, $\alpha_1^{\dagger\prime} = 2$ and $\alpha_1^{\ddagger\prime} = (2-\overline{p}_1\lambda_1)/(1-\overline{p}_1\lambda_1)$, z_l', z_u' reduce to z_l, z_u, and (31) reduces to (18) in Theorem 2, recovering the single-type equilibrium characterization.

Theorem 5 shows that equilibrium behavior retains the bang-bang structure of the homogeneous-patient model (Theorem 2). Type 2 patients adopt $\tilde{p}_2 = 1$ in every equilibrium, since switching is purely access-improving. For type 1 patients, the equilibrium is governed by the thresholds z_l' and z_u', which partition the cost-ratio space into three regimes: maximal stable pooling when waiting costs dominate, full continuity when visit costs dominate, and a mixed-strategy equilibrium when waiting and visit costs are balanced.

Comparing Theorems 4 and 5, equilibrium and optimum coincide for type 2 patients ($\tilde{p}_2 = p_2^\star = 1$). For type 1 patients, alignment obtains at the type 1 cost-ratio extremes, but can diverge for intermediate $C_{v,1}/C_{w,1}$ because bang-bang equilibrium strategies cannot replicate the planner's interior optimum.

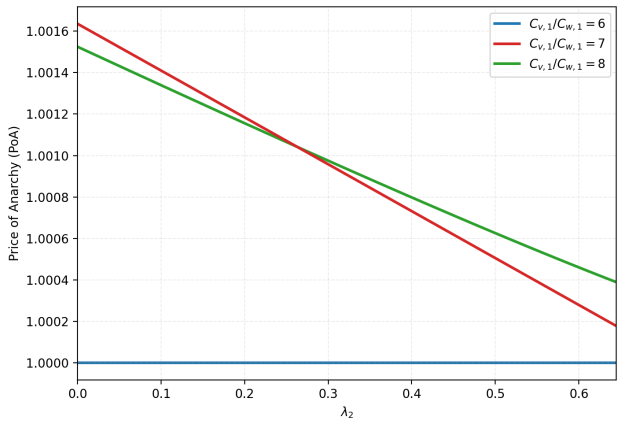


Figure 6 Price of anarchy as a function of the type 2 arrival rate λ_2, for three values of the cost ratio $C_{v,1}/C_{w,1}$. In this figure, $\lambda_1 = 0.3$, $\alpha_1 = 1.169$.

Figure 6 illustrates how population composition affects the efficiency of decentralized behavior. It shows that PoA is non-increasing in λ_2. A larger type 2 population weakly reduces the welfare loss from self-routing. This efficiency improvement does not arise from a change in type 1 patients' equilibrium behavior—the gap between $\tilde{p}_1$ and $p_1^\star$ persists across all values of λ_2—but rather from an externality absorption effect. Type 2 patients pool unconditionally under both equilibrium and optimum, providing queue equalization that is costless from a demand-feedback perspective ($\alpha_2 = 1$). As the type 2 share of total arrivals grows, this queue equalization absorbs a larger portion of

the congestion externality imposed by type 1 patients' suboptimal switching, diluting the welfare consequences of their self-interested behavior. Central coordination is therefore most valuable when type 1 patients comprise a large share of demand.

7. Practical Implementation

The probabilistic switching model analyzed in previous sections utilizes only binary queue-length information: a patient observes whether or not her usual GP has the shortest queue, and, if not, switches with a fixed probability that is independent of the actual queue lengths. A natural generalization is to exploit richer information—the queue-length gap—and let the switching decision depend on the size of the gap. In this section, we study a simple and practical instance of such a policy: the *absolute threshold policy*.

An absolute threshold policy is governed by a parameter $\tau \in \{0, 1, 2, \ldots\}$. A patient switches to the shortest queue if and only if

$$q_{\text{usual}} - q_{\text{shortest}} > \tau, \tag{32}$$

where q_{usual} and q_{shortest} denote the queue lengths at the patient's usual GP and at the shortest queue, respectively, at the moment of arrival. Setting $\tau = 0$ recovers full pooling, while $\tau \to \infty$ recovers full continuity, and intermediate values calibrate how large a gap patients tolerate before switching. More generally, one could allow the switching probability to vary with the gap. If the equilibrium exhibits a bang-bang structure at any given gap τ, as characterized in Section 5, we conjecture that the optimal gap-dependent probabilistic policy reduces to a deterministic threshold rule of the form (32). A formal proof is left for future work.

To compare the absolute threshold policy with the probabilistic switching policy that we analytically studied on a common scale, we map each τ to its implied switching probability, denoted by p_τ, which is defined as the steady-state fraction of non-shortest arrivals (whose usual GP does not have the shortest queue) that actually switch under the threshold rule. Computing p_τ requires a fixed-point iteration. Specifically, a candidate p determines the effective arrival rate $\hat{\lambda}(p)$ and hence the queue-length distribution, which in turn determines the fraction of arrivals whose gap exceeds τ. We iterate until the assumed p and the realized switching fraction agree. Using p_τ as the horizontal axis places both policies on the same metric and enables a fair cost comparison at matched switching frequency.

Figure 7 presents results across $\lambda \in \{0.65, 0.75, 0.85\}$ with $n = 15$, $\alpha = 1.169$, and $C_v/C_w = 20$. The blue curve shows the simulated cost of the probabilistic switching policy, and the red circles show the simulated cost under the absolute threshold policy for a range of τ values, each mapped to its implied p_τ.

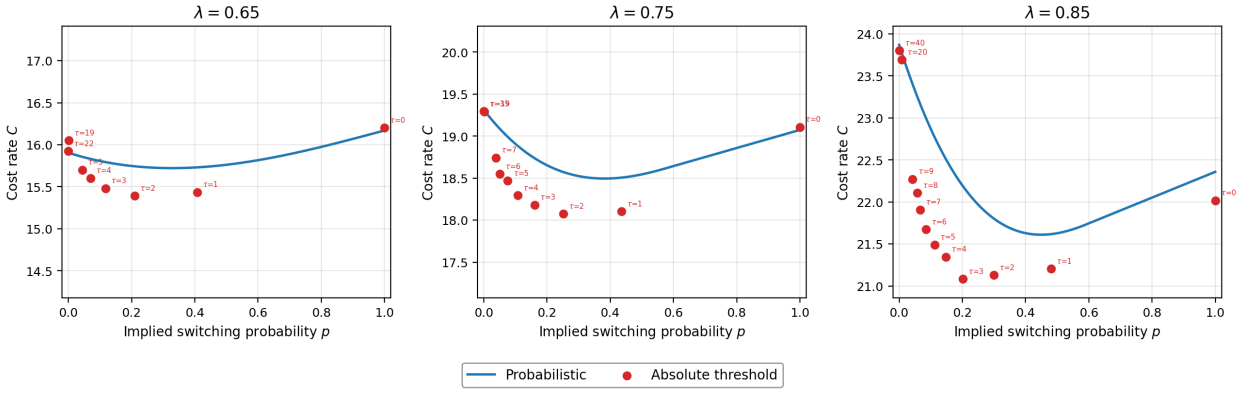


Figure 7 **Cost rate $\mathcal{C}$ as a function of the implied switching probability p under two policies: the probabilistic policy (blue line) and the absolute threshold policy (red circles, labeled by threshold τ). Each panel corresponds to a different arrival rate λ. In this figure, $n = 15$, $\alpha = 1.169$, and $C_v/C_w = 20$.**

Two observations emerge from the figure. First, the absolute threshold policy sits at or below the cost curve of the probabilistic switching policy at every matched implied p, confirming that conditioning the switching decision on the queue-length gap—rather than randomizing independently of it—improves system performance. The mechanism is intuitive: the threshold rule switches only when the gap is large enough to indicate a genuine wait-time advantage, whereas the probabilistic policy may switch when the queues are nearly equal or fail to switch when the gap is large. That said, the cost gap between the two policies is modest, so the analytical results of the probabilistic switching policy derived in Sections 3–6 serve as conservative performance benchmarks for the more informative but analytically intractable threshold rule. Second, by varying τ from ∞ to 0, the threshold policy spans essentially the full interval $p_\tau \in [0, 1]$, so any probabilistic switching rate p can be equivalently realized by a suitable threshold τ. Additionally, the threshold policy is easier to implement in practice, since it requires only the observable rule "switch if your usual GP has more than τ extra patients waiting." This framing maps naturally onto digital booking interfaces that display differences in next-available appointment dates, and nests the NHS *Digital First* pathway, where patients are routed to the earliest available clinician (i.e., $\tau = 0$) as a special case.

8. NHS Case Study

Relational continuity is widely regarded as a cornerstone of high-quality primary care, yet its optimal level remains actively debated within the United Kingdom's National Health Service (NHS). This section brings our analytical framework to bear on this debate by calibrating the model with practice-level data from the NHS and conducting a series of numerical experiments that address three policy questions:

(i) Does full continuity maximize system-wide welfare?

(ii) Should continuity rules be uniform across patients or differentiated by patient type?

(iii) Can practices delegate continuity decisions to patients without sacrificing efficiency?

8.1. Model Calibration

Patients with at least one chronic condition constitute type 1, and they generate additional downstream consultation workload when continuity is disrupted. All other patients form type 2, for whom we assume no continuity benefit. Based on Kajaria-Montag et al. (2024), we set the workload amplification factor to $\alpha_1 = 1.169$, meaning each disrupted type 1 visit generates about 1.169 consultation-equivalents of total workload (see Remark 1 for the relationship to the underlying hazard-rate evidence). We set $\alpha_2 = 1$. The proportion of each type in the patient cohort at every practice is estimated from the 2023–24 Quality and Outcomes Framework (QOF) prevalence statistics (NHS England 2024).

Beside patient mix, capacity and system utilization are two further practice-specific determinants of the optimal continuity policy. We measure capacity by the number of general practitioner full-time equivalents (GP FTEs), extracted from the General Practice Workforce census (NHS England 2025b). To gauge utilization, we obtain the appointment activity recorded in the Appointments in General Practice data for the same month (NHS England 2025a). For each practice, we compute utilization as the ratio of appointments attended to appointments offered. Table 1 summarizes the calibration inputs.

	Mean	Std. Deviation	10th Percentile	90th Percentile
Proportion of type 1 patients (%)	46.52	9.28	34.48	56.78
Total GP FTE at each practice	6.06	4.67	1.48	11.60
Utilization at each practice (%)	95.35	2.61	92.13	97.88

Table 1 Descriptive Statistics of GP Practices Data

A critical input to our simulations is the set of nominal arrival rates, λ_k, $k \in \{1, 2\}$, that would prevail if every patient always saw their usual GP. The rates observed in the NHS appointments data, by contrast, provide an estimate of effective arrival rates, $\hat{\lambda}_k$, already shaped by existing continuity policies. To recover the nominal values, we combine the average patient mix reported in Table 1 (46.5% type 1 patients), the estimate that type 1 patients account for 80.6% of all GP appointments from Kajaria-Montag et al. (2024), and the most recent estimate of NHS continuity level, which stands at 36% (Kajaria-Montag and Freeman 2020). These imply that complex patients consult roughly 4.4 times more frequently than non-complex patients under full continuity, consistent with QOF data. We apply this visit-frequency ratio to each practice's observed arrival rate and patient mix to deduce the corresponding nominal arrival rates.

We impose identical cost parameters on both patient types and on both the patient's and planner's objective. Any PoA exceeding one therefore arises purely from structural misalignment between self-routing incentives and system-level objectives, rather than from differential cost attributions. In practice, the NHS as planner may assign additional weight to costs that individual patients do not fully internalize at the point of booking—most notably, downstream expenditures triggered by continuity disruption. Consequently, equal costs provide a conservative lower bound on the true externality of decentralization.

We calibrate $C_v \approx$ £64.67 per consultation and $C_w \approx$ £0.53 per working hour from UK cost and stated-preference evidence, yielding a baseline ratio $C_v/C_w = 122$ working hours per visit—that is, a patient would need to wait approximately 122 working hours ($\approx$ 12 days) before the accumulated waiting cost equals the cost of one consultation. The visit cost combines an NHS resource component (£42 per consultation (Jones et al. 2024)) and a patient-borne component (travel and time costs (NIHR 2023)); the waiting cost derives from stated-preference willingness-to-pay estimates (Cheraghi-Sohi et al. 2008). All values are expressed in 2023/24 pounds sterling. Appendix EC.5.1 provides details on the data sources and cost calibration.

The simulations adopt log-normal inter-arrival and service times (with four and two times the exponential variance, respectively) to match the heavier tails observed in NHS appointment data.

8.2. Results and Implications for NHS Practices

The key structural predictions of our analytical model (which assumes exponential inter-arrival and service times) carry over to the case study, where the inter-arrival and service time distributions are no longer exponential. Specifically, we continue to observe an interior optimal across-the-board switching probability, $p_1^\star = 0$ and $p_2^\star = 1$ under the type-dependent policy, and bang-bang equilibrium strategies.

NHS practices differ markedly in patient case-mix, physician capacity, and system utilization. To translate findings into practice-specific guidance, we partition over 6,000 practices into eight groups using k-means clustering on the aforementioned three practice-specific features. Figure 8 displays the within-cluster sum of squares as a function of the number of clusters, and the curve exhibits a reasonable elbow at $k = 8$, supporting this as a parsimonious choice. We treat each cluster centroid as a representative practice (Table 2), spanning from small low-congestion practices (Cluster 8) to large highly congested ones (Cluster 4).

Table 3 reports results for each representative practice under three regimes: type-dependent policy $(p_1^\star, p_2^\star)$, across-the-board policy $(p_a^\star)$, and decentralized equilibrium $(\tilde{p}_1, \tilde{p}_2)$. Equilibria are computed via finite-n simulation.[4] The table reports the implied continuity levels and mean waiting times under each regime, the percentage welfare loss of the across-the-board policy relative to the type-dependent optimum, and the price of anarchy of the decentralized equilibrium.

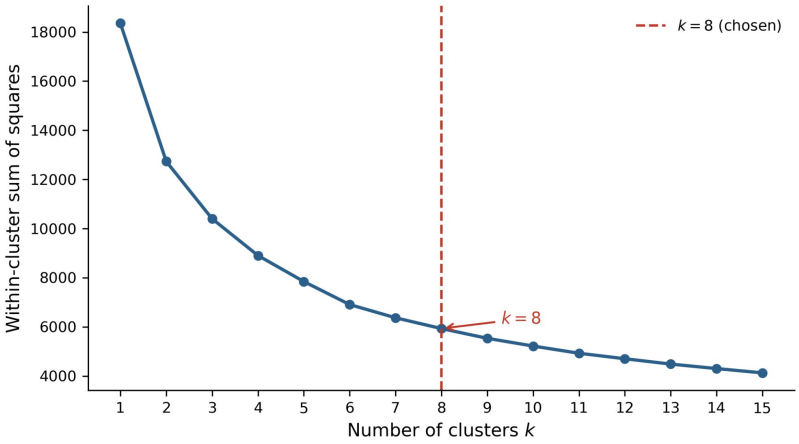


Figure 8 **The elbow curve of the k-means clustering.**

Cluster	Type 1 patients (%)	Total GP FTEs	Utilization (%)
1	45.92	7	97.42
2	57.45	5	97.11
3	43.72	2	96.93
4	52.07	14	96.49
5	31.67	9	94.02
6	48.96	7	93.92
7	50.73	2	93.84
8	36.57	2	91.82

Table 2 **Cluster Centroids of GP Practices Data**

Cluster	Type-dependent policy				Across-the-board policy				Decentralized equilibrium		
	$p_1^\star$	$p_2^\star$	Continuity (%)	Avg wait (days)	p_a^*	Continuity (%)	Avg wait (days)	Optimality gap (%)	$\tilde{p}_1$	$\tilde{p}_2$	PoA
1	0	1	83.98	0.92	0.13	89.76	1.13	3.00	0	1	1
2	0	1	89.55	1.09	0.09	93.37	1.20	1.85	0	1	1
3	0	1	90.12	1.56	0.09	95.80	1.76	1.99	0	1	1
4	0	1	85.54	0.79	0.15	87.26	0.88	2.55	0	1	1
5	0	1	75.06	0.65	0.15	87.92	0.88	3.14	0	1	1
6	0	1	85.66	0.75	0.14	89.25	0.87	2.41	0	1	1
7	0	1	92.13	1.23	0.05	97.67	1.39	1.46	0	1	1
8	0	1	88.14	1.12	0.07	96.75	1.23	1.25	0	1	1

Table 3 **Results for representative practices under the baseline cost calibration ($C_v/C_w = 122$ hours).**

Across-the-board policy. The optimal across-the-board policy seldom prescribes 100% continuity: $p_a^\star$ lies between 0.05 and 0.15, producing continuity levels of 87–98%. Switching is greatest in practices with more type 2 patients (for whom switching is purely access-improving) and more GPs

(where a wider pool of alternative queues amplifies the potential waiting-time reduction). Nevertheless, a uniform rule leaves value on the table (welfare loss of 1.25–3.14% across clusters) because it uses type 1 patients as partial load balancers when type 2 patients can provide nearly all of the pooling benefit without workload amplification.

Type-dependent policy. Across all eight representative practices, the type-dependent optimum prescribes full continuity for type 1 patients ($p_1^\star = 0$) and full pooling for type 2 patients ($p_2^\star = 1$). Under the high utilization typical of the NHS (91–98%), breaking continuity for type 1 patients offers only modest wait reductions while generating additional workload ($\alpha_1 = 1.169$), so full continuity is welfare-maximizing for complex patients. Type 2 patients serve as a natural load-balancing mechanism: by joining the shortest queue unconditionally, they redistribute demand across GPs, thereby reducing average waiting times for *all* patients without imposing any penalty of disrupted continuity on complex ones. The resulting continuity levels range from 75% to 92% across clusters, largely driven by each practice's type 2 share. Notably, the type-dependent policy also reduces average waiting times: mean waits range from 0.65 to 1.56 days under the type-dependent policy versus 0.87 to 1.76 days under the across-the-board policy.

Delegating continuity decisions to patients. Implementing type-dependent rules requires classifying patients by complexity and enforcing differentiated scheduling, which may not be feasible at the practice level. Equilibrium analysis reveals that delegation is a viable alternative: under the calibrated NHS conditions, individual choice coincides with the centralized optimum in all eight representative practices (PoA = 1). Delegation therefore relieves practices of the administrative burden while granting patients greater autonomy and perceived fairness in scheduling. This aligns with the broader NHS emphasis on patient empowerment and shared decision-making, as patients are given the information and autonomy to balance continuity against access according to their own clinical needs.

Under extreme cost parameterizations, multiple equilibria can arise, including an inefficient "all-switch" outcome ($\tilde{p}_1 = \tilde{p}_2 = 1$). Table 4 shows that four clusters—all operating at moderate utilization (92–94%)—admit such a second equilibrium, but only at C_v/C_w ratios far below the calibrated value of 122 working hours per visit. In these cases, the inefficient equilibrium reduces continuity drastically to 35–67% and raises the price of anarchy to 1.08–1.15. However, such scenarios require waiting costs several times larger than any empirical estimate and are unlikely to arise in practice (see Appendix EC.5.3 for details).

In summary, our case study under typical NHS utilization reveals that: (i) 100% continuity is not optimal—modest pooling reduces waits without materially increasing demand; (ii) type-dependent rules capture virtually all attainable surplus: full continuity for complex patients, full pooling for non-complex; and (iii) delegation to patients replicates the optimum (PoA = 1) across all clusters.

Cluster	C_v/C_w range (working hours per visit)	Type-dependent policy				Decentralized equilibrium				
		$p_1^\star$	$p_2^\star$	Continuity (%)	Avg wait (days)	$\tilde{p}_1$	$\tilde{p}_2$	Continuity (%)	Avg wait (days)	PoA
5	28.6–52.0	0	1	75.06	0.65	0	1	75.06	0.65	1
						1	1	35.30	0.87	1.12–1.15
6	35.3–61.7	0	1	85.66	0.75	0	1	85.66	0.75	1
						1	1	38.70	0.97	1.13–1.15
7	59.3–87.6	0	1	92.13	1.23	0	1	92.13	1.23	1
						1	1	67.10	1.57	1.09–1.11
8	58.3–79.5	0	1	88.14	1.12	0	1	88.14	1.12	1
						1	1	67.10	1.41	1.08–1.09

Table 4 Possible inefficient equilibria for four clusters under hypothetical values of C_v/C_w.

Threshold-based policies (analyzed in Section 7) deliver consistent performance and do not alter these qualitative conclusions. Appendix EC.5.3 provides sensitivity analysis to the full range of empirically plausible cost parameters.

9. Concluding Remarks

This paper develops a mean-field queueing framework to analyze the trade-off between relational continuity and operational efficiency in primary care. The key modeling innovation is an endogenous demand-feedback channel; that is, routing a patient away from her usual GP disrupts continuity and generates additional downstream workload, so that the effective arrival rate is itself a function of the routing policy. Within this framework, we study three progressively richer settings. First, we analyze a homogeneous-patient model in which a central planner selects the probability with which patients switch to a shorter queue, and characterize both the centralized optimal policy and the decentralized Nash equilibrium. Second, we extend the model to two patient types—those who benefit from continuity and those who do not—and compare across-the-board policies and type-dependent ones, with the decentralized Nash equilibrium. Third, we evaluate threshold-based routing rules numerically and calibrate the model to NHS practice-level data, relaxing the distributional assumption of exponential inter-arrival and service times.

Several results emerge. In the centralized homogeneous-patient model, neither full continuity nor full pooling is generally optimal. The optimal switching probability balances the long-run workload amplification caused by disrupted continuity against the short-run congestion relief from pooling. In the decentralized setting, each patient's cost is concave in her own switching probability, producing bang-bang best responses. Patients either maintain full continuity or switch maximally, with a mixed-strategy equilibrium emerging for intermediate cost ratios in which a fraction of the population permanently adopts maximal switching and the rest permanently maintain full continuity. Because the planner's interior optimum cannot be replicated by such extreme individual strategies,

the equilibrium and optimum can diverge for intermediate values of C_v/C_w, though the resulting price of anarchy remains bounded and, for empirically relevant parameters, modest. With heterogeneous patients, type-dependent policies strictly dominate uniform ones. Non-complex patients with no benefit from continuity should always join the shortest queue, while the optimal policy for complex patients follows the same structure as in the homogeneous-patient setting. Importantly, non-complex patients who pool unconditionally provide queue equalization that absorbs the congestion externality imposed by complex patients' suboptimal switching, diluting the welfare consequences of self-interested behavior and reducing the price of anarchy. Threshold-based routing, which conditions switching on the observed queue-length gap, consistently outperforms the probabilistic switching policy at matched implied switching rates.

These findings translate into actionable prescriptions for healthcare managers and policymakers. First, neither full continuity nor full pooling is generally best. The best policy trades off workload amplification from disrupted continuity against congestion relief from pooling. Second, differentiating by patient type captures virtually all attainable surplus: complex patients should maintain full continuity, while non-complex patients should be directed to the earliest available clinician. By unconditionally joining the shortest queue, non-complex patients serve as a natural load-balancing mechanism that redistributes demand across GPs, reducing waiting times for *all* patients without imposing any continuity penalty on complex ones. Third, under the high utilization characteristic of NHS primary care and empirically calibrated cost parameters, individual incentives and system-wide objectives are aligned, so practices can empower patients to choose without sacrificing system efficiency; such delegation eliminates the need for patient classification, removing the associated administrative burden and fairness concerns. Threshold-based switching rules, which condition switching on the observed queue-length gap, are both practical and superior to probabilistic assignment, and do not alter any of these conclusions.

Endnotes

1. An alternative micro-foundation models disrupted continuity as elevating the patient's next-visit *hazard rate* from λ to $\alpha\lambda$, rather than generating additional visits. We discuss this formulation and its relationship to our model in Remark 1 below.

2. The value $\alpha = 1.169$ is calibrated from the hazard-model estimates in Kajaria-Montag et al. (2024), who find that breaking relational continuity increases the rate of follow-up consultations by approximately 16.9%. Under our workload formulation, this translates to each disrupted visit generating, in expectation, 1.169 total consultation-equivalents. As shown in Remark 1, the workload and hazard-rate formulations agree to first order in $(\alpha - 1)$.

3. We select $\lambda = 0.85$ as the largest baseline arrival rate because stability of the queueing system requires the effective arrival rate to remain strictly below the service rate $\mu = 1$. When all patients switch to the

shortest queue ($p = 1$), the continuity-disruption arrival-amplification factor $\alpha = 1.169$ inflates the effective arrival rate to $\hat{\lambda} = \alpha\lambda$. At $\lambda = 0.85$, this yields $\hat{\lambda} = 0.85 \times 1.169 \approx 0.994$, which is close to but still below $\mu = 1$. Any baseline rate above $1/\alpha \approx 0.855$ would violate the stability condition for at least some values of p, making the full range $p \in [0, 1]$ infeasible. Thus, $\lambda = 0.85$ is approximately the largest baseline load that permits the entire switching-probability range to be explored while respecting stability.

4. We calculate equilibria directly from the finite-n simulation rather than from the mean-field analytical formulas. Specifically, for each candidate pure-strategy profile, we simulate the queueing system, collect the steady-state occupancy distribution, and compute the cost-ratio threshold at which a focal patient is indifferent between staying and switching. This finite-n procedure accounts for effects absent from the mean-field limit—most notably, the shortest queue being non-empty under high utilization—and can produce dual pure-strategy equilibria in parameter ranges where the mean-field model predicts a unique mixed equilibrium (see Appendix EC.5.2 for details).

References

Ahuja V, Alvarez CA, Staats BR (2022) An operations approach for reducing glycemic variability: evidence from a primary care setting. *Manufacturing & Service Operations Management* 24(3):1474–1493.

Armony M, Maglaras C (2004) Customer contact centers: Research problems and results. *Manufacturing & Service Operations Management* 6(4):282–306.

Barker I, Steventon A, Deeny SR (2017) Association between continuity of care in general practice and hospital admissions for ambulatory care sensitive conditions: cross sectional study of routinely collected, person level data. *BMJ* 356:j84.

Bavafa H, Hitt LM, Terwiesch C (2018) The impact of e-visits on visit frequencies and patient health: Evidence from primary care. *Management Science* 64(12):5461–5480.

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

Benaïm M, Le Boudec JY (2008) A class of mean field interaction models for computer and communication systems. *Performance Evaluation* 65(11-12):823–838.

Braverman A (2020) Steady-state analysis of the join-the-shortest-queue model in the Halfin–Whitt regime. *Mathematics of Operations Research* 45(3):1069–1103.

Chen HM, Tu YH, Chen CM (2017) Effect of continuity of care on quality of life in older adults with chronic diseases: a meta-analysis. *Clinical Nursing Research* 26(3):266–284.

Cheraghi-Sohi S, Hole AR, Mead N, McDonald R, Whalley D, Bower P, Roland M (2008) What patients want from primary care consultations: a discrete choice experiment to identify patients' priorities. *The Annals of Family Medicine* 6(2):107–115.

Cho KH, Kim YS, Nam CM, Kim TH, Kim SJ, Han KT, Park EC (2015) The association between continuity of care and all-cause mortality in patients with newly diagnosed obstructive pulmonary disease: a population-based retrospective cohort study, 2005–2012. *PloS one* 10(11):e0141465.

Ding Y, Gupta D, Zhou S (2026) Early reservation for follow-up appointments: Enhancing patient care continuity. *Manufacturing & Service Operations Management* .

Dossa AR, Moisan J, Guénette L, Lauzier S, Grégoire JP (2017) Association between interpersonal continuity of care and medication adherence in type 2 diabetes: an observational cohort study. *Canadian Medical Association Open Access Journal* 5(2):E359–E364.

Drury A, Payne S, Brady AM (2020) Identifying associations between quality of life outcomes and healthcare-related variables among colorectal cancer survivors: A cross-sectional survey study. *International Journal of Nursing Studies* 101:103434.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 5(2):79–141.

Gray DJP, Sidaway-Lee K, White E, Thorne A, Evans PH (2018) Continuity of care with doctors—a matter of life and death? a systematic review of continuity of care and mortality. *BMJ open* 8(6):e021161.

Green LV, Savin S, Savva N (2013) "nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science* 59(10):2237–2256.

Gregory A (2024) Seeing same GP 'improves patient health and cuts workload of doctors'. `https://www.theguardian.com/society/2024/feb/23/seeing-same-gp-improves-patient-health-and-cuts-workload-of-doctors`.

Gupta D, Wang L (2008) Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research* 56(3):576–592.

Haggerty J, Burge F, Lévesque JF, Gass D, Pineault R, Beaulieu MD, Santor D (2007) Operational definitions of attributes of primary health care: consensus among canadian experts. *The Annals of Family Medicine* 5(4):336–344.

Haggerty JL, Reid RJ, Freeman GK, Starfield BH, Adair CE, McKendry R (2003) Continuity of care: a multidisciplinary review. *BMJ* 327(7425):1219–1221.

Haight FA (1958) Two queues in parallel. *Biometrika* 45(3-4):401–410.

Hassin R (2016) *Rational queueing* (CRC press).

Hassin R, Haviv M (2003) *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59 (Springer Science & Business Media).

Huntley A, Lasserson D, Wye L, Morris R, Checkland K, England H, Salisbury C, Purdy S (2014) Which features of primary care affect unscheduled secondary care use? a systematic review. *BMJ open* 4(5):e004746.

Jeffers H, Baker M (2016) Continuity of care: still important in modern-day general practice. *The British Journal of General Practice* 66(649):396–397.

Jones KC, Weatherly H, Birch S, Castelli A, Chalkley M, Dargan A, Forder JE, Gao M, Hinde S, Markham S, Premji S, Findlay D, Teo H (2024) Unit costs of health and social care 2023 manual. Technical report, Personal Social Services Research Unit (University of Kent) & Centre for Health Economics (University of York), Kent, UK, URL `http://dx.doi.org/10.22024/UniKent/01.02.105685`, kAR id:105685.

Kajaria-Montag H, Freeman M (2020) Explaining the erosion of relational care continuity: an empirical analysis of primary care in england .

Kajaria-Montag H, Freeman M, Scholtes S (2024) Continuity of care increases physician productivity in primary care. *Management Science* 70(11):7943–7960.

Kingman JF (1961) Two similar queues in parallel. *The Annals of Mathematical Statistics* 32(4):1314–1323.

Kuiper A, Lee CH (2022) Appointment scheduling for multiple servers. *Operations Research* 70(5):2892–2908.

Kurtz TG (1970) Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability* 7(1):49–58.

Kurtz TG (1971) Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability* 8(2):344–356.

Lekwijit TS, Song H, Terwiesch C, Chaiyachati K (2023) Multi-channel healthcare operations: The impact of video visits on the usage of in-person care. *Available at SSRN 4397550* .

Leniz J, Gulliford MC (2019) Continuity of care and delivery of diabetes and hypertensive care among regular users of primary care services in Chile: a cross-sectional study. *BMJ Open* 9(10):e027830.

Liu N, Finkelstein SR, Kruk ME, Rosenthal D (2018) When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling. *Management Science* 64(5):1975–1996.

Liu N, Wang S, Zychlinski N (2025) Channel management in outpatient care: Implications of telemedicine and transportation support. *Available at SSRN 4383199* .

Liu XA, Armony M (2024) Telemedicine versus in-person outpatient care: Equilibrium, capacity, and quality differentiation. *Available at SSRN 5048340* .

Mitzenmacher M (1996) Load balancing and density dependent jump Markov processes. *Proceedings of 37th Conference on Foundations of Computer Science*, 213–222 (IEEE).

Mukherjee D, Borst SC, Van Leeuwaarden JS, Whiting PA (2020) Asymptotic optimality of power-of-d load balancing in large-scale systems. *Mathematics of Operations Research* 45(4):1535–1571.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

NHS England (2019) Digital-first primary care: Policy consultation. `https://www.england.nhs.uk/wp-content/uploads/2019/06/digital-first-primary-care-consultation.pdf`.

NHS England (2023) Digital requirements guidance for GP practices. https://www.england.nhs.uk/gp/investment/gp-contract/digital-requirements-guidance/.

NHS England (2024) Quality and outcomes framework, 2023-24. https://shorturl.at/sx7ns.

NHS England (2025a) Appointments in general practice, march 2025. https://shorturl.at/KWBwe.

NHS England (2025b) General practice workforce, 31 march 2025. https://shorturl.at/rFCxq.

NIHR (2023) Time and travel analysis.

Nyweide DJ, Anthony DL, Bynum JP, Strawderman RL, Weeks WB, Casalino LP, Fisher ES (2013) Continuity of care and the risk of preventable hospitalization in older adults. *JAMA Internal Medicine* 173(20):1879–1885.

Pourat N, Davis AC, Chen X, Vrungos S, Kominski GF (2015) In California, primary care continuity was associated with reduced emergency department use and fewer hospitalizations. *Health Affairs* 34(7):1113–1120.

Saghafian S, Hopp WJ, Iravani SM, Cheng Y, Diermeier D (2018) Workload management in telemedical physician triage and other knowledge-based service systems. *Management Science* 64(11):5180–5197.

Sunar N, Tu Y, Ziya S (2021) Pooled vs. dedicated queues when customers are delay-sensitive. *Management Science* 67(6):3785–3802.

University of Leicester (2024) Fewer GP appointments and COVID lockdown have exacerbated declining continuity of care in English general practices. https://medicalxpress.com/news/2024-06-gp-covid-lockdown-exacerbated-declining.html.

Vvedenskaya ND, Dobrushin RL, Karpelevich FI (1996) Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii* 32(1):20–34.

Whitt W (2003) How multiserver queues scale with growing congestion-dependent demand. *Operations Research* 51(4):531–542.

Wilkinson E (2024) GP continuity of care could 'effectively disappear' unless action taken. https://www.pulsetoday.co.uk/news/workforce/gp-continuity-of-care-could-effectively-disappear-unless-action-taken.

Ye T, Sun X, Tang W, Miao Y, Zhang Y, Zhang L (2016) Effect of continuity of care on health-related quality of life in adult patients with hypertension: a cohort study in China. *BMC Health Services Research* 16:1–8.

Zang L, Hu Y, Roet-Green R, Sun S (2024) The impact of information-granularity and prioritization on patients' care modality choice. *Available at SSRN 4733392* .

Electronic Companion

In this technical appendix, we provide supporting technical material for the manuscript titled: "Is Continuity All We Need? A Modeling Approach to Evaluating Relational Continuity in Primary Care". The proofs of these results are in the order in which they appear in the main body.

EC.1. Proofs from Section 3

EC.1.1. Proof of Lemma 1

We begin by noting that, under the stability condition (3), as $n \to \infty$, there is a positive fraction of idle GPs, so the system minimum queue length is 0 with probability approaching 1. For any fixed $i \geq 1$, $s_i^{(\infty)}(t)$ experiences transitions as patients arrive and depart. In particular, $s_i^{(\infty)}(t)$ increases when GP's queue length increases from $i-1$ to i, and $s_i^{(\infty)}(t)$ decreases when GP's queue length decreases from i to $i-1$. We examine these transitions and their rates below.

Arrivals that increase $s_i^{(\infty)}$: Such an event occurs when a new patient joins a GP who has exactly $i-1$ patients (both in queue and in service), causing that GP's queue length to increase from $i-1$ to i.

For $i=1$, an increase in $s_1^{(n)}$ occurs when an arrival is routed to an idle GP (queue length 0). An arriving patient's usual GP is idle with probability $1-s_1^{(\infty)}(t;p)$, in which case the patient is routed there and immediately makes that GP busy. If the usual GP is busy (probability $s_1^{(\infty)}(t;p)$), then the patient switches to an idle shortest-queue GP with probability p, and stays with the busy usual GP with probability $1-p$. Hence, conditional on the state at time t, the probability the arrival is routed to an idle GP is

$$\left(1-s_1^{(\infty)}(t;p)\right)\cdot 1 + s_1^{(\infty)}(t;p)\cdot p = 1-(1-p)s_1^{(\infty)}(t;p).$$

which yields the arrival term in (4) after multiplying by the effective arrival rate $\hat{\lambda}^{(\infty)}(p)$. Thus, the rate at which arrivals cause an $i-1 \to i$ queue-length transition, for $i=1$, is $\left(1-(1-p)s_1^{(\infty)}(t;p)\right)\hat{\lambda}^{(\infty)}(p)$.

For $i \geq 2$, an increase in $s_i^{(\infty)}$ occurs when an arrival joins a GP that currently has exactly $i-1$ patients. Recall that the minimum queue length is 0, any arrival that switches (probability p when the usual GP is not among the shortest) is routed to an idle GP, and therefore cannot increase $s_i^{(\infty)}$ for $i \geq 2$. The only way to join a queue of length $i-1 \geq 1$ is to stay with the usual GP (which occurs with probability $1-p$) when it has exactly $i-1$ patients. The probability that an arriving patient finds their usual GP with $i-1$ patients is $s_{i-1}^{(\infty)}(t;p) - s_i^{(\infty)}(t;p)$, because the fraction of GPs with exactly $i-1$ patients is $s_{i-1}^{(\infty)}(t;p) - s_i^{(\infty)}(t;p)$. Hence, the rate at which arrivals cause an $i-1 \to i$ queue-length transition, for $i \geq 2$, is $(1-p)\left(s_{i-1}^{(\infty)}(t;p) - s_i^{(\infty)}(t;p)\right)\hat{\lambda}^{(\infty)}(p)$, yielding the arrival term in (5).

Departures that decrease $s_i^{(\infty)}$: Such an event occurs when a service completion happens at a GP who has i patients (both in queue and in service), causing that queue length to drop from i to $i-1$. The fraction of GP with exactly i patients at time t is $s_i^{(\infty)}(t;p) - s_{i+1}^{(\infty)}(t;p)$. Each such GP, if it has i patients, completes a service at rate 1, and after service the GP will have $i-1$ patients (thus it will no longer count toward the $\geq i$ group). Therefore, the rate at which departures cause an $i \to i-1$ queue-length transition is $\Big(s_i^{(\infty)}(t;p) - s_{i+1}^{(\infty)}(t;p)\Big)$.

Combining the arrival and departure effects, we can write the drift (time derivative) of the occupancy state $s_i^{(\infty)}$, for $i \geq 1$, as:

$$\begin{aligned}
\frac{d}{dt}s_1^{(\infty)}(t;p) =& \Big(1-(1-p)s_1^{(\infty)}(t;p)\Big)\hat{\lambda}^{(\infty)}(p) - \Big(s_1^{(\infty)}(t;p) - s_2^{(\infty)}(t;p)\Big),\\
\frac{d}{dt}s_i^{(\infty)}(t;p) =& (1-p)\Big(s_{i-1}^{(\infty)}(t;p) - s_i^{(\infty)}(t;p)\Big)\hat{\lambda}^{(\infty)}(p) - \Big(s_i^{(\infty)}(t;p) - s_{i+1}^{(\infty)}(t;p)\Big), \quad \forall i \geq 2.
\end{aligned}$$

∎

EC.1.2. Proof of Proposition 1

Setting the right-hand sides of (4) to zero implies

$$s_i^{(\infty)}(\infty;p) - s_{i+1}^{(\infty)}(\infty;p) = (1-p)\left(s_{i-1}^{(\infty)}(\infty;p) - s_i^{(\infty)}(\infty;p)\right)\hat{\lambda}^{(\infty)}(p), \quad i \geq 2.$$

Let $\Delta_i := s_i^{(\infty)}(\infty;p) - s_{i+1}^{(\infty)}(\infty;p)$ for $i \geq 1$. Then, the above display can be equivalently written as

$$\Delta_i = (1-p)\hat{\lambda}^{(\infty)}(p)\Delta_{i-1}, \quad i \geq 2.$$

Hence, $\Delta_i = \Delta_1\big((1-p)\hat{\lambda}^{(\infty)}(p)\big)^{i-1}$ for $i \geq 1$. Summing the increments yields

$$s_i^{(\infty)}(\infty;p) = \sum_{k\geq i}\Delta_k = \Delta_1\sum_{k\geq i}\big((1-p)\hat{\lambda}^{(\infty)}(p)\big)^{k-1} = \Delta_1\frac{\big((1-p)\hat{\lambda}^{(\infty)}(p)\big)^{i-1}}{1-(1-p)\hat{\lambda}^{(\infty)}(p)}, \quad i \geq 1. \tag{EC.1}$$

Setting the right-hand side of (4) to zero implies

$$\Delta_1 = s_1^{(\infty)}(\infty;p) - s_2^{(\infty)}(\infty;p) = \Big(1-(1-p)s_1^{(\infty)}(\infty;p)\Big)\hat{\lambda}^{(\infty)}(p). \tag{EC.2}$$

Substituting for Δ_1 using (EC.2) into (EC.1) and letting $i=1$:

$$s_1^{(\infty)}(\infty;p) = \Big(1-(1-p)s_1^{(\infty)}(\infty;p)\Big)\hat{\lambda}^{(\infty)}(p)\frac{1}{1-(1-p)\hat{\lambda}^{(\infty)}(p)},$$

Solving for $s_1^{(\infty)}(\infty;p)$ using the above equation:

$$s_1^{(\infty)}(\infty;p) = \hat{\lambda}^{(\infty)}(p).$$

Using this as well as (EC.1) and (EC.2), we obtain

$$\begin{aligned}
s_i^{(\infty)}(\infty;p) =& \Big(1-(1-p)\hat{\lambda}^{(\infty)}(p)\Big)\hat{\lambda}^{(\infty)}(p)\frac{\big((1-p)\hat{\lambda}^{(\infty)}(p)\big)^{i-1}}{1-(1-p)\hat{\lambda}^{(\infty)}(p)}\\
=& (1-p)^{i-1}\left(\hat{\lambda}^{(\infty)}(p)\right)^{i}, \quad i \geq 1.
\end{aligned}$$

∎

EC.1.3. Proof of Proposition 2

Derivation of $\sigma^{(\infty)}(p)$ and $\hat{\lambda}^{(\infty)}(p)$. Under the stability condition (3), the shortest queue length is 0 in the mean-field limit, so the event "the usual GP is not among the shortest queues" is equivalent to "the usual GP is busy", which occurs with probability $s_1^{(\infty)}(\infty;p)$. By Proposition 1, $s_1^{(\infty)}(\infty;p) = \hat{\lambda}^{(\infty)}(p)$. Therefore,

$$\sigma^{(\infty)}(p) = s_1^{(\infty)}(\infty;p) = \hat{\lambda}^{(\infty)}(p).$$

Substituting $\sigma^{(\infty)}(p) = \hat{\lambda}^{(\infty)}(p)$ into the workload closure (1) yields the self-consistency equation

$$\hat{\lambda}^{(\infty)}(p) = \big(1 + (\alpha - 1)\, p\, \hat{\lambda}^{(\infty)}(p)\big)\lambda.$$

Since this equation is linear in $\hat{\lambda}^{(\infty)}(p)$, it has the unique solution

$$\hat{\lambda}^{(\infty)}(p) = \frac{\lambda}{1 - p(\alpha-1)\lambda},$$

establishing (7).

Derivation of $W^{(\infty)}(p)$. If the usual GP is idle, the patient joins it and immediately enters service; if the usual GP is busy and the patient switches, the patient is routed to an idle GP (recalling that the minimum queue length is 0 in the mean-field limit) and also enters service immediately. Hence the only source of delay (waiting in queue) beyond service is when the patient stays with a busy usual GP (probability $1-p$), in which case the expected number of customers in the chosen GP's system is $\sum_{i\geq 1} s_i^{(\infty)}(\infty;p)$ by the tail-sum formula. Since service times are exponential with unit mean, the expected time in system equals 1 (expected time in service) plus the expected number of customers in queue:

$$\begin{aligned}
W^{(\infty)}(p) =& 1 + (1-p)\sum_{i\geq 1} s_i^{(\infty)}(\infty;p) \\
\overset{(1)}{=}& 1 + (1-p)\sum_{i\geq 1} (1-p)^{i-1}\left(\hat{\lambda}^{(\infty)}(p)\right)^i \\
=& 1 + \sum_{i\geq 1}\left((1-p)\hat{\lambda}^{(\infty)}(p)\right)^i \\
=& 1 + \frac{(1-p)\hat{\lambda}^{(\infty)}(p)}{1-(1-p)\hat{\lambda}^{(\infty)}(p)} \\
\overset{(2)}{=}& 1 + \frac{(1-p)\frac{\lambda}{1-p(\alpha-1)\lambda}}{1-(1-p)\frac{\lambda}{1-p(\alpha-1)\lambda}} \\
=& 1 + \frac{(1-p)\lambda}{1-\big(p(\alpha-1)+(1-p)\big)\lambda},
\end{aligned}$$

where (1) follows from (6) in Proposition 1 and (2) follows from (7), proved in the above. ∎

EC.1.4. Proof of Proposition 3

Fix $p \in [0,1]$ satisfying (3). Throughout the proof, write $\hat{\lambda} := \hat{\lambda}^{(\infty)}(p) = \frac{\lambda}{1-p(\alpha-1)\lambda} \in (0,1)$, and consider the n-server system with total arrival rate $n\hat{\lambda}$ and i.i.d. Exp(1) services. We begin by formalizing the state space and its associated topology.

State space and topology. Define the state space as

$$\mathcal{S} := \Big\{ s = (s_i)_{i\geq 0} : s_0 = 1,\ 1 \geq s_1 \geq s_2 \geq \cdots \geq 0,\ \lim_{i\to\infty} s_i = 0 \Big\}.$$

For technical convenience, endow $\mathcal{S}$ with the *weighted* ℓ_1 norm

$$\|s\|_w := \sum_{i\geq 1} 2^{-i}|s_i|.$$

This induces a metric under which $\mathcal{S}$ is compact.

For each n, the occupancy process $s^{(n)}(t;p) \in \mathcal{S}$ is a càdlàg Markov process.

Step 1: Drift representation via time-changed Poisson processes. For every $i \geq 1$, the coordinate $s_i^{(n)}$ increases by $1/n$ exactly when an arrival joins a queue of length $i-1$, and decreases by $1/n$ exactly when a service completion occurs at a queue of length exactly i. Hence we can write, for every $i \geq 1$,

$$s_i^{(n)}(t) = s_i^{(n)}(0) + \frac{1}{n}A_i^{(n)}(t) - \frac{1}{n}D_i^{(n)}(t), \tag{EC.3}$$

where $A_i^{(n)}(t)$ counts arrivals up to time t that join a queue of length $i-1$, and $D_i^{(n)}(t)$ counts service completions up to time t from queues of length exactly i.

Conditional on the filtration $\mathcal{F}_t$, arrivals occur at rate $n\hat{\lambda}$. Let $a_i^{(n)}(s)$ be the conditional probability that an arriving customer joins a queue of length $i-1$ given the occupancy state $s \in \mathcal{S}$. Assume that we can establish the following auxiliary result, whose proof is delayed until the end.

Claim EC.1. *For any $s \in \mathcal{S}$, define the minimum queue length $m(s) := \min\{k \geq 0 : s_k > s_{k+1}\}$. Then:*

(i) $a_i^{(n)}(s) = 0$ *for all* $i \leq m(s)$;

(ii) $a_{m(s)+1}^{(n)}(s) = \big(s_{m(s)} - s_{m(s)+1}\big) + p\, s_{m(s)+1}$;

(iii) $a_i^{(n)}(s) = (1-p)\big(s_{i-1} - s_i\big)$ *for all* $i \geq m(s)+2$.

In particular, when $s_1 < 1$ (i.e., $m(s) = 0$):

$$a_1^{(n)}(s) = 1 - (1-p)\, s_1, \tag{EC.4}$$

$$a_i^{(n)}(s) = (1-p)(s_{i-1} - s_i), \quad i \geq 2. \tag{EC.5}$$

Note that the expressions (EC.4)–(EC.5) are identical to the mean-field limiting routing probabilities. That is, on the set $\{s \in \mathcal{S} : s_1 < 1\}$, the finite-$n$ drift coincides with the limiting drift *exactly*, without any approximation error.

By Claim EC.1, $A_i^{(n)}$ admits the random time-change representation

$$A_i^{(n)}(t) = N_i^A\left(\int_0^t n\hat{\lambda}\, a_i^{(n)}\big(s^{(n)}(u)\big)\, du\right),$$

where $N_i^A(\cdot)$ is a unit-rate Poisson process.

Similarly, the number of servers with exactly i jobs equals $n\left(s_i^{(n)} - s_{i+1}^{(n)}\right)$, and each such server completes service at rate 1, so

$$D_i^{(n)}(t) = N_i^D\left(\int_0^t n\left(s_i^{(n)}(u) - s_{i+1}^{(n)}(u)\right) du\right),$$

where $N_i^D(\cdot)$ is another unit-rate Poisson process (independent across i and from $\{N_i^A\}$).

Write $N(x) = x + M(x)$ where $M(\cdot)$ is the centered Poisson martingale. Substituting into (EC.3) yields the semimartingale decomposition

$$s_i^{(n)}(t) = s_i^{(n)}(0) + \int_0^t F_i^{(n)}\big(s^{(n)}(u)\big)\, du + M_i^{(n)}(t), \tag{EC.6}$$

where the drift $F^{(n)}(s)$ is determined by the routing probabilities in Claim EC.1. On $\{s \in \mathcal{S} : s_1 < 1\}$, Claim EC.1 gives

$$F_1^{(n)}(s) = \big(1 - (1-p)s_1\big)\hat{\lambda} - (s_1 - s_2), \tag{EC.7}$$

$$F_i^{(n)}(s) = (1-p)(s_{i-1} - s_i)\hat{\lambda} - (s_i - s_{i+1}), \quad i \geq 2, \tag{EC.8}$$

and $M_i^{(n)}(t)$ is the difference of two time-changed centered Poisson martingales scaled by $1/n$. When $s_1 = 1$ (all GPs are busy), the drift $F^{(n)}(s)$ differs from $F^{(\infty)}(s)$ because the minimum queue length $m(s) \geq 1$ alters the routing probabilities (Claim EC.1); however, this discrepancy occurs on a set that the mean-field trajectory avoids under stability (by Kurtz's theorem), and thus does not affect the convergence argument.

Step 2: Identification of the limiting ODE drift. By Claim EC.1, for any $s \in \mathcal{S}$ with $s_1 < 1$, the finite-n drift (EC.7)–(EC.8) coincides exactly with the limiting ODE drift:

$$F_1^{(\infty)}(s) = \big(1 - (1-p)s_1\big)\hat{\lambda} - (s_1 - s_2), \tag{EC.9}$$

$$F_i^{(\infty)}(s) = (1-p)(s_{i-1} - s_i)\hat{\lambda} - (s_i - s_{i+1}), \quad i \geq 2. \tag{EC.10}$$

This is exactly the ODE system in Lemma 1. In particular, on the set $\{s \in \mathcal{S} : s_1 \leq 1 - \varepsilon\}$ for any $\varepsilon > 0$, $F^{(n)}(s) = F^{(\infty)}(s)$ identically, so $\sup_{s:\, s_1 \leq 1-\varepsilon} \|F^{(n)}(s) - F^{(\infty)}(s)\|_w = 0$.

Step 3: Lipschitz property of the limiting drift. Define $F^{(\infty)}(s) = (F_i^{(\infty)}(s))_{i\geq 1}$ by (EC.9)–(EC.10). Each $F_i^{(\infty)}$ depends only on (s_{i-1}, s_i, s_{i+1}) and is affine, hence globally Lipschitz. A direct weighted-norm calculation gives that there exists $L < \infty$ such that for all $s, \tilde{s} \in \mathcal{S}$,

$$\|F^{(\infty)}(s) - F^{(\infty)}(\tilde{s})\|_w \leq L \|s - \tilde{s}\|_w. \tag{EC.11}$$

(For example, one may take $L = 4(1 + \hat{\lambda})$ using triangle inequalities and the relations between the weights 2^{-i}.) Therefore, the ODE $\frac{d}{dt}s(t) = F^{(\infty)}(s(t))$ admits a unique solution for each initial condition $s(0) \in \mathcal{S}$.

Step 4: Mean-field limit convergence. Equations (EC.6)–(EC.8) show that $s^{(n)}$ is a density-dependent Markov process with jump sizes $O(1/n)$ and total jump intensity $O(n)$. By (EC.11) and the fact that $F^{(n)}(s) = F^{(\infty)}(s)$ for all s with $s_1 < 1$ (Step 2), we may apply Kurtz's functional law of large numbers for density-dependent Markov processes (e.g., Kurtz (1970, 1971)): for each fixed $T < \infty$,

$$s^{(n)}(\,\cdot\,; p) \Rightarrow s^{(\infty)}(\,\cdot\,; p) \quad \text{in } D([0,T], \mathcal{S}),$$

where $s^{(\infty)}(\,\cdot\,; p)$ is as defined in Lemma 1 with initial condition $s^{(\infty)}(0; p)$. Because T is arbitrary, this proves the first statement of the proposition.

Step 5: Convergence of steady-state expected time in system. Let π_n be the unique stationary distribution of the n-server Markov chain (positive recurrence follows from $\hat{\lambda} < 1$). Let $s^{(\infty)}(\infty; p)$ denote the unique fixed point of the limiting ODE (given explicitly by Proposition 1). Standard mean-field stationary-concentration results for density-dependent Markov processes with a globally attracting equilibrium (e.g., Kurtz (1971) or Theorem 1 of Benaïm and Le Boudec (2008)) imply $\pi_n \Rightarrow s^{(\infty)}(\infty; p)$, as $n \to \infty$.

We now map stationary occupancy to stationary expected time in system. Consider an arrival in steady state. By PASTA, the pre-arrival state has distribution π_n. Conditional on occupancy state s, an arriving patient is routed to a non-shortest (busy) usual GP only if: (i) the usual GP is busy, and (ii) the patient does not switch, which occurs with probability $(1-p)$. In the mean-field limit, pooled customers join an idle server and incur no waiting. Thus, the conditional expected time in system given state s equals

$$G(s) := 1 + (1-p)\sum_{i\geq 1} s_i, \tag{EC.12}$$

because $\sum_{i\geq 1} s_i = \mathbb{E}[L]$ for the queue length L of a uniformly random GP, and the patient's usual GP is uniformly random. (The identity $\mathbb{E}[L] = \sum_{i\geq 1} \mathbb{P}(L \geq i) = \sum_{i\geq 1} s_i$ is standard.) Hence $W^{(n)}(p) =$

$\mathbb{E}_{\pi_n}[G(s)]$. To pass to the limit, we use uniform integrability. Under any routing policy with stability parameter $\hat{\lambda} < 1$, each individual GP's stationary queue length in the n-server system is stochastically dominated by an $M/M/1$ queue with load $\hat{\lambda}$ (since pooling can only reduce queue lengths relative to dedicated service). Therefore $\mathbb{E}_{\pi_n}[\sum_{i\geq 1} s_i] \leq \hat{\lambda}/(1-\hat{\lambda})$ uniformly in n, which implies that $\{G(s) : s \sim \pi_n\}_{n\geq 1}$ is uniformly integrable. Combined with the weak convergence $\pi_n \Rightarrow s^{(\infty)}(\infty; p)$, this gives

$$\lim_{n\to\infty} W^{(n)}(p) = G(s^{(\infty)}(\infty; p)).$$

Finally, plugging the explicit fixed point $s_i^{(\infty)}(\infty; p) = (1-p)^{i-1}\hat{\lambda}^i$ (from Proposition 1) into (EC.12) yields

$$G(s^{(\infty)}(\infty; p)) = 1 + (1-p)\sum_{i\geq 1}(1-p)^{i-1}\hat{\lambda}^i = 1 + \frac{(1-p)\hat{\lambda}}{1-(1-p)\hat{\lambda}} = 1 + \frac{(1-p)\lambda}{1-\big(p(\alpha-1)+(1-p)\big)\lambda} = W^{(\infty)}(p),$$

which is exactly (8). This proves $\lim_{n\to\infty} W^{(n)}(p) = W^{(\infty)}(p)$. ∎

To complete the proof, we verify Claim EC.1 below.

Proof of Claim EC.1. Consider an arriving patient and let $s \in \mathcal{S}$ be the system state upon arrival. The arrival's usual GP is selected uniformly at random from the n GPs. Let J denote the queue length of the usual GP, and let M denote the minimum queue length among the remaining $n-1$ (non-usual) GPs. Under the routing policy, if $J \leq M$ the patient stays with the usual GP; if $J > M$ the patient switches to a shortest non-usual GP with probability p and stays with probability $1-p$.

Define the minimum queue length in the system as $m := m(s) = \min\{k \geq 0 : s_k > s_{k+1}\}$. There are $n(s_m - s_{m+1}) \geq 1$ GPs at queue length m. One key observation is that if the usual GP has queue length $J > m$, then none of the GPs at the minimum queue length m is the usual GP, so all $n(s_m - s_{m+1})$ GPs at length m are among the $n-1$ non-usual GPs. Therefore, $M = m < J$, and the patient switches to a GP at length m with probability p or stays at the usual GP (length J) with probability $1-p$. Conversely, if the usual GP has queue length $J = m$, then the remaining $n-1$ GPs include $n(s_m - s_{m+1}) - 1$ GPs at length m and possibly GPs at longer lengths. In either case $M \geq m = J$, so the usual GP is among the shortest queues and the patient stays.

We now compute $a_i^{(n)}(s)$, the conditional probability that an arriving customer joins a queue of length $i-1$ given the occupancy state $s \in \mathcal{S}$.

(i): For $i \leq m$, the arriving customer joins a queue of length less than the minimum length m, which is impossible, so $a_i^{(n)}(s) = 0$.

(ii): For $i = m+1$, the arriving customer joins a queue at the minimum length m. This occurs if (i) the usual GP has length m (probability $s_m - s_{m+1}$), in which case the patient stays; or (ii) the

usual GP has length $> m$ (probability s_{m+1}) and the patient switches (probability p) to a GP at length m. Hence

$$a_{m+1}^{(n)}(s) = (s_m - s_{m+1}) + p\, s_{m+1}.$$

(iii): For $i \geq m+2$, the arriving customer joins a queue of length $i-1 > m$. The patient can only join such a queue by staying at a usual GP of length $i-1$. Since $i-1 > m$, the usual GP is not among the shortest, so the patient stays with probability $1-p$. Hence

$$a_i^{(n)}(s) = (1-p)(s_{i-1} - s_i).$$

Specializing to $s_1 < 1$, we have $m = 0$ (since $s_0 = 1 > s_1$), and the formulas above reduce to:

$$\begin{aligned} a_1^{(n)}(s) &= (s_0 - s_1) + p\, s_1 = (1 - s_1) + p\, s_1 = 1 - (1-p)\, s_1, \\ a_i^{(n)}(s) &= (1-p)(s_{i-1} - s_i), \quad i \geq 2, \end{aligned}$$

establishing (EC.4)–(EC.5). ∎

EC.1.5. Proof of Proposition 4

Let $\Delta^{(n)}(t) := s^{(n)}(t) - s^{(\infty)}(t)$. We work in the weighted ℓ_1 norm $\|\cdot\|_w$, namely, $\|s\|_w = \sum_{i\geq 1} 2^{-i}|s_i|$, and then translate to the first m coordinates.

Step 1: Semimartingale decomposition and error equation. From (EC.6) in the proof of Proposition 3, we have

$$s^{(n)}(t) = s^{(n)}(0) + \int_0^t F^{(n)}\big(s^{(n)}(u)\big)du + M^{(n)}(t),$$

where $F^{(n)}$ is defined in (EC.7)–(EC.8), and $M^{(n)}(t)$ is the difference of two time-changed centered Poisson martingales scaled by $1/n$, and

$$s^{(\infty)}(t) = s^{(\infty)}(0) + \int_0^t F^{(\infty)}\big(s(u)\big)du,$$

where $F^{(\infty)}$ is defined in (EC.9)–(EC.10). Subtracting the above two displays yields

$$\Delta^{(n)}(t) = \Delta^{(n)}(0) + \int_0^t \Big(F^{(n)}(s^{(n)}(u)) - F^{(\infty)}(s(u))\Big)du + M^{(n)}(t). \tag{EC.13}$$

We next split the drift difference into two parts:

$$F^{(n)}(s^{(n)}) - F^{(\infty)}(s) = \underbrace{\Big(F^{(n)}(s^{(n)}) - F^{(\infty)}(s^{(n)})\Big)}_{\text{finite-}n\text{ drift bias}} + \underbrace{\Big(F^{(\infty)}(s^{(n)}) - F^{(\infty)}(s)\Big)}_{\text{Lipschitz term}}.$$

On any set $\{s \in \mathcal{S} : s_1 < 1\}$, Claim EC.1 gives $F^{(n)}(s) = F^{(\infty)}(s)$ exactly (Step 2 of the proof of Proposition 3), so the finite-n drift bias vanishes identically. We therefore focus on the Lipschitz term; by (EC.11),

$$\|F^{(\infty)}(s^{(n)}(u)) - F^{(\infty)}(s(u))\|_w \le L\|\Delta^{(n)}(u)\|_w.$$

Taking weighted norms on both sides of (EC.13) and using triangle inequality implies

$$\|\Delta^{(n)}(t)\|_w \le \|\Delta^{(n)}(0)\|_w + L\int_0^t \|\Delta^{(n)}(u)\|_w du + \sup_{0\le u\le t}\|M^{(n)}(u)\|_w.$$

Taking supremum over $[0,T]$ and applying Grönwall's inequality yields

$$\sup_{0\le t\le T}\|\Delta^{(n)}(t)\|_w \le e^{LT}\Big(\|\Delta^{(n)}(0)\|_w + \sup_{0\le t\le T}\|M^{(n)}(t)\|_w\Big). \tag{EC.14}$$

Step 2: Bounding the martingale term in expectation. Each jump of $s^{(n)}$ changes the occupancy state by at most $O(1/n)$ in $\|\cdot\|_w$. Indeed, an arrival to a queue of length k increases s_i by $1/n$ for all $i \ge k+1$, hence the weighted jump size is

$$\frac{1}{n}\sum_{i\ge k+1} 2^{-i} = \frac{1}{n}2^{-k}.$$

A departure from a queue of length $k \ge 1$ decreases s_i by $1/n$ for all $i \ge k$, hence the weighted jump size is

$$\frac{1}{n}\sum_{i\ge k} 2^{-i} = \frac{1}{n}2^{-(k-1)} \le \frac{2}{n}.$$

Therefore $\|\Delta s^{(n)}\|_w \le 2/n$ for every single jump.

Meanwhile, the total jump rate is $O(n)$, because arrivals occur at rate $n\hat{\lambda}$ and services complete at total rate at most n. Hence the predictable quadratic variation of the $\|\cdot\|_w$-martingale satisfies

$$\mathbb{E}\big[\langle M^{(n)}\rangle_T\big] \;\le\; \Big(\frac{2}{n}\Big)^2 \cdot (n\hat{\lambda}+n)\,T \;\le\; \frac{C\,T}{n}$$

for some constant C depending only on $\hat{\lambda}$. By the Burkholder-Davis-Gundy inequality for càdlàg martingales,

$$\mathbb{E}\left[\sup_{0\le t\le T}\|M^{(n)}(t)\|_w\right] \le C'\sqrt{\mathbb{E}[\langle M^{(n)}\rangle_T]} \le \frac{C''\sqrt{T}}{\sqrt{n}}, \tag{EC.15}$$

for constants C', C'' independent of n.

Step 3: Conclude the $1/\sqrt{n}$ bound in $\|\cdot\|_w$, then translate to the first m coordinates. Take expectations in (EC.14) and use the martingale bound (EC.15)

$$\mathbb{E}\left[\sup_{0\le t\le T}\|\Delta^{(n)}(t)\|_w\right] \le e^{LT}\left(\mathbb{E}\|\Delta^{(n)}(0)\|_w + \frac{C''\sqrt{T}}{\sqrt{n}}\right).$$

Because the initial conditions have finite support and the assumption bounds the first m coordinates by $C_0/\sqrt{n}$, we have $\mathbb{E}\|\Delta^{(n)}(0)\|_w \le C_0/\sqrt{n}$ up to a constant factor. Thus,

$$\mathbb{E}\left[\sup_{0\le t\le T}\|\Delta^{(n)}(t)\|_w\right] \le \frac{C_T}{\sqrt{n}},$$

for some $C_T < \infty$ independent of n.

Finally, we relate $\|\cdot\|_w$ to the first m coordinates. For any vector $x = (x_i)_{i\ge 1}$,

$$\sum_{i=1}^{m}|x_i| \le \sum_{i=1}^{m} 2^i \sum_{j\ge 1} 2^{-j}|x_j| \le (2^{m+1})\|x\|_w.$$

Applying this to $x = \Delta^{(n)}(t)$ and taking supremum over $[0,T]$ and expectation gives

$$\mathbb{E}\left[\sup_{0\le t\le T}\|s^{(n)}_{\le m}(t;p) - s^{(\infty)}_{\le m}(t;p)\|_1\right] \le 2^{m+1}\mathbb{E}\left[\sup_{0\le t\le T}\|\Delta^{(n)}(t)\|_w\right] \le \frac{C_{T,m}}{\sqrt{n}},$$

with $C_{T,m} := 2^{m+1}C_T$. ∎

EC.2. Proofs from Section 4

EC.2.1. Proof of Proposition 5

Let $A(p) := 1 - p(\alpha-1)\lambda$ and $B(p) := 1 - \big(p(\alpha-1) + (1-p)\big)\lambda$. Then, the planner's mean-field objective function (12) can be written as

$$\mathcal{C}(p) = \frac{\lambda C_v}{A(p)} + \frac{\lambda C_w}{B(p)}.$$

By definition, every $p \in [0,\overline{p}]$ satisfies the stability condition (3):

$$(1 + (\alpha-1)p)\,\lambda < 1,$$

which implies that

$$A(p) = 1 - p(\alpha-1)\lambda > \lambda > 0,$$

and that

$$B(p) = p\lambda + [1 - (1 + (\alpha-1)p)\,\lambda] > 0.$$

Differentiating $\mathcal{C}(p)$ with respect to p yields

$$\begin{aligned}\mathcal{C}'(p) =& \lambda C_v \cdot \frac{(\alpha-1)\lambda}{A(p)^2} + \lambda C_w \cdot \frac{(\alpha-2)\lambda}{B(p)^2} \\ =& \lambda^2 \left(\frac{(\alpha-1)C_v}{A(p)^2} + \frac{(\alpha-2)C_w}{B(p)^2} \right), \\ \mathcal{C}''(p) =& 2\lambda^3 \left(\frac{(\alpha-1)^2 C_v}{A(p)^3} + \frac{(\alpha-2)^2 C_w}{B(p)^3} \right).\end{aligned}$$

Since $A(p) > 0$ and $B(p) > 0$ on the stability region $[0, \overline{p}]$, and $(\alpha-1)^2 \geq 0$ and $(\alpha-2)^2 \geq 0$, we have $\mathcal{C}''(p) \geq 0$ for all $p \in [0, \overline{p}]$. Hence $\mathcal{C}$ is convex. As a consequence, if the optimal solution is an interior point, then the solution is unique and can be characterized by the first-order condition $\mathcal{C}'(p) = 0$.

∎

EC.2.2. Proof of Theorem 1

Let $A(p) := 1 - p(\alpha-1)\lambda$ and $B(p) := 1 - \big(p(\alpha-1) + (1-p)\big)\lambda = 1 - \lambda - p(\alpha-2)\lambda$. Recall from the proof of Proposition 5 that the derivative of $\mathcal{C}(p)$ is given by

$$\mathcal{C}'(p) = \lambda^2 \left(\frac{(\alpha-1)C_v}{A(p)^2} + \frac{(\alpha-2)C_w}{B(p)^2} \right), \tag{EC.16}$$

(i): If $\alpha = 1$, then $\alpha - 1 = 0$ and $\alpha - 2 < 0$. Hence, the first term of $\mathcal{C}'(p)$ in (EC.16) is zero and the second term of $\mathcal{C}'(p)$ in (EC.16) is negative. Therefore, $\mathcal{C}'(p) < 0$ for all $p \in [0, \overline{p}]$, which implies that $\mathcal{C}(p)$ is decreasing in $p \in [0, \overline{p}]$. From (13), $\overline{p} = 1$ when $\alpha = 1$ (given that $\lambda \in (0,1)$). Thus, $p^\star = 1$.
(ii): If $\alpha \in (1,2)$, by the convexity result (Proposition 5), $\mathcal{C}'(p)$ is increasing in $p \in [0, \overline{p}]$. Therefore: (a) if $\mathcal{C}'(0) \geq 0$, then $\mathcal{C}'(p) \geq 0$ for all $p \in [0, \overline{p}]$ and $\mathcal{C}$ is increasing, so $p^\star = 0$; (b) if $\mathcal{C}'(\overline{p}) \leq 0$, then $\mathcal{C}'(p) \leq 0$ for all $p \in [0, \overline{p}]$ and $\mathcal{C}$ is decreasing, so $p^\star = \overline{p}$. Next, we compute the boundary derivatives from (EC.16).

From (EC.16), $\mathcal{C}'(0) \geq 0$ reads

$$\lambda^2 \left(\frac{(\alpha-1)C_v}{A(0)^2} + \frac{(\alpha-2)C_w}{B(0)^2} \right) \geq 0.$$

Since $A(0) = 1$, $B(0) = 1 - \lambda$, the above display is equivalent to

$$\frac{C_v}{C_w} \geq \frac{2-\alpha}{(\alpha-1)(1-\lambda)^2} =: y_u.$$

It is clear that $y_u > 0$, given that $\alpha \in (1,2)$.

Similarly, $\mathcal{C}'(\overline{p}) \leq 0$ reads

$$\lambda^2 \left(\frac{(\alpha-1)C_v}{A(\overline{p})^2} + \frac{(\alpha-2)C_w}{B(\overline{p})^2} \right) \leq 0,$$

which is equivalent to

$$\frac{C_v}{C_w} \leq \frac{(2-\alpha)A(\overline{p})^2}{(\alpha-1)B(\overline{p})^2} = \frac{(2-\alpha)\left(1 - \overline{p}(\alpha-1)\lambda\right)^2}{(\alpha-1)\left(1 - \lambda - \overline{p}(\alpha-2)\lambda\right)^2} =: y_l.$$

It is clear that $y_l > 0$, given that $\alpha \in (1,2)$.

Additionally, we show $y_l < y_u$. Since $y_l = (2-\alpha)A(\overline{p})^2/[(\alpha-1)B(\overline{p})^2]$ and $y_u = (2-\alpha)/[(\alpha-1)(1-\lambda)^2]$, the inequality $y_l < y_u$ is equivalent to $A(\overline{p})^2(1-\lambda)^2 < B(\overline{p})^2$, i.e., $A(\overline{p})(1-\lambda) < B(\overline{p})$ (both sides are positive). Expanding, this becomes

$$\big(1-\overline{p}(\alpha-1)\lambda\big)(1-\lambda) < 1-\lambda-\overline{p}(\alpha-2)\lambda,$$

which simplifies to $\overline{p}\lambda\big[1-(\alpha-1)\lambda\big] > 0$. This holds because $\overline{p} > 0$, $\lambda > 0$, and $(\alpha-1)\lambda < 1$ under the stability condition (3). This proves that $0 < y_l < y_u$.

The above establishes the boundary regimes, namely, statements **(ii)(a)** and **(ii)(c)**. The remainder of the proof is devoted to establishing statement **(ii)(b)**.

Given that $\alpha \in (1,2)$, let $a := \alpha - 1 \in (0,1)$ and $b := 2-\alpha \in (0,1)$ (and $a+b=1$). If $C_v/C_w \in (y_l, y_u)$, then $\mathcal{C}$ is strictly convex and $\mathcal{C}'(0) < 0 < \mathcal{C}'(\overline{p})$, so there is a unique interior minimizer $p^\star \in (0,\overline{p})$ characterized by $\mathcal{C}'(p^\star) = 0$. Setting (EC.16) to zero and using $a = \alpha - 1$ and $\alpha - 2 = -b$ gives

$$\frac{aC_v}{(1-a\lambda p^\star)^2} = \frac{bC_w}{(1-\lambda+b\lambda p^\star)^2},$$

where $1-a\lambda p^\star = A(p^\star) > 0$ and $1-\lambda+b\lambda p^\star = B(p^\star) > 0$ under the stability condition (3), recalling from the proof of Proposition 5 that $A(p) > 0$ and $B(p) > 0$ for all p on the stability region. Taking square roots on both sides of the above display,

$$\frac{\sqrt{aC_v}}{1-a\lambda p^\star} = \frac{\sqrt{bC_w}}{1-\lambda+b\lambda p^\star}. \tag{EC.17}$$

Rearranging (EC.17) yields

$$\sqrt{aC_v}\,\big(1-\lambda+b\lambda p^\star\big) = \sqrt{bC_w}\,\big(1-a\lambda p^\star\big).$$

Solving for $p^\star$ gives

$$p^\star = \frac{\sqrt{2-\alpha} - (1-\lambda)\sqrt{(\alpha-1)C_v/C_w}}{\lambda\Big((2-\alpha)\sqrt{(\alpha-1)C_v/C_w} + (\alpha-1)\sqrt{2-\alpha}\Big)}.$$

Next, we establish comparative statics of $p^\star$ in this interior regime.

Monotonicity in C_v/C_w and in α. Rewrite (EC.17) as

$$\sqrt{\frac{C_v}{C_w}}\sqrt{\frac{a}{b}} = \frac{1-a\lambda p^\star}{1-\lambda+b\lambda p^\star}.$$

For fixed (α,λ), the right-hand side is strictly decreasing in $p^\star$ because the numerator decreases in $p^\star$ while the denominator increases in $p^\star$. The left-hand side is strictly increasing in C_v/C_w and strictly increasing in α (equivalently, increasing in a/b). Therefore, increasing C_v/C_w or increasing α raises the left-hand side and, to restore equality, $p^\star$ must weakly decrease. Hence $p^\star$ is decreasing in C_v/C_w and decreasing in α.

Monotonicity in λ. For fixed $(\alpha, C_v/C_w)$ in the interior regime, the closed-form expression for $p^\star$ can be written as

$$p^\star(\lambda) = \frac{U-(1-\lambda)V}{\lambda K} = \frac{U-V}{\lambda K} + \frac{V}{K},$$

where $U := \sqrt{2-\alpha}$, $V := \sqrt{(\alpha-1)C_v/C_w}$, and $K := (2-\alpha)V + (\alpha-1)U$ are constants w.r.t. λ. In the interior regime, $C_v/C_w > y_l = b/a$ implies $V > U$, hence $U - V < 0$ and therefore

$$\frac{d}{d\lambda}p^\star(\lambda) = -\frac{U-V}{\lambda^2 K} > 0.$$

Thus $p^\star$ is increasing in λ. This completes the proof of statement **(ii)(c)**.

(iii): If $\alpha \geq 2$, then $\alpha - 1 \geq 0$ and $\alpha - 2 \geq 0$. Hence, both terms in $\mathcal{C}'(p)$ in (EC.16) are nonnegative for all p. Therefore, $\mathcal{C}(p)$ is nondecreasing in $p \in [0, \overline{p}]$ and thus $p^\star = 0$.

∎

EC.3. Proofs from Section 5

EC.3.1. Proof of Proposition 6

In the mean-field regime, a single patient's actions have negligible impact on the overall queue-length distribution. We may treat $\sigma_i(p_i, p_{-i})$ as effectively independent of p_i (it is determined by the other patients' strategies p_{-i}). Similarly, the patient's waiting time can be expressed as a convex combination of the two extreme cases $p_i = 0$ and $p_i = 1$. In particular, letting $W_i(0, p_{-i})$ denote patient i's expected steady-state wait if she never switches, and $W_i(1, p_{-i})$ the wait if she always switches, we can write

$$W_i(p_i, p_{-i}) = (1-p_i)\, W_i(0, p_{-i}) + p_i\, W_i(1, p_{-i}).$$

This follows because with probability $1 - p_i$ the patient stays with her own GP (incurring wait $W_i(0, p_{-i})$), and with probability p_i she switches to the shortest queue (incurring wait $W_i(1, p_{-i})$). Thus W_i is an affine function of p_i. Meanwhile, from (16), the effective arrival rate can be written $\hat{\lambda}_i(p_i, p_{-i}) = \big(1 + (\alpha-1)p_i\sigma_i(p_i, p_{-i})\big)\lambda$, which is also affine in p_i, recalling that $\sigma_i(p_i, p_{-i})$ is effectively independent of p_i, which we simply denote by σ_i. Therefore, the cost (15) can be expanded as

$$\begin{aligned}
\mathcal{C}_i(p_i, p_{-i}) =& \big(1 + (\alpha-1)p_i\sigma_i\big)\,\lambda\Big(C_v + C_w\,\left[(1-p_i)\, W_i(0, p_{-i}) + p_i\, W_i(1, p_{-i})\right]\Big) \\
=& \lambda\big(C_v + C_w \cdot W_i(0, p_{-i})\big) \\
&+ p_i \cdot \Big[\lambda(\alpha-1)\sigma_i\big(C_v + C_w \cdot W_i(0, p_{-i})\big) + \lambda C_w\big(W_i(1, p_{-i}) - W_i(0, p_{-i})\big)\Big] \\
&+ p_i^2 \cdot \lambda(\alpha-1)\sigma_i C_w\big(W_i(1, p_{-i}) - W_i(0, p_{-i})\big).
\end{aligned}$$

This is a quadratic function of p_i. To determine its concavity, consider the coefficient of the p_i^2 term:

$$\lambda(\alpha-1)\sigma_i C_w\big(W_i(1, p_{-i}) - W_i(0, p_{-i})\big).$$

Since $W_i(1,p_{-i}) < W_i(0,p_{-i})$ (always switching yields a shorter wait than never switching), the difference $W_i(1,p_{-i}) - W_i(0,p_{-i})$ is negative. Thus the p_i^2 coefficient is negative, implying $\mathcal{C}_i(p_i,p_{-i})$ is a concave (in fact, concave quadratic) function of p_i, for any given p_{-i}.

∎

EC.3.2. Proof of Theorem 2

Recall from the proof of Proposition 6 that the asymmetric cost function is given by

$$\begin{aligned}\mathcal{C}_i(p_i,p_{-i}) =& \big(1+(\alpha-1)p_i\sigma_i\big)\,\lambda\Big(C_v + C_w\,\big[(1-p_i)\,W_i(0,p_{-i}) + p_i\,W_i(1,p_{-i})\big]\Big)\\ =&\lambda\big(C_v + C_w\cdot W_i(0,p_{-i})\big)\\ &+p_i\cdot\Big[\lambda(\alpha-1)\sigma_i\big(C_v + C_w\cdot W_i(0,p_{-i})\big) + \lambda C_w\big(W_i(1,p_{-i}) - W_i(0,p_{-i})\big)\Big]\\ &+p_i^2\cdot\lambda(\alpha-1)\sigma_i C_w\big(W_i(1,p_{-i}) - W_i(0,p_{-i})\big). \end{aligned} \tag{EC.18}$$

Differentiating (EC.18) with respect to p_i yields

$$\begin{aligned}\frac{d}{dp_i}\mathcal{C}_i(p_i,p_{-i}) =&\lambda(\alpha-1)\sigma_i\big(C_v + C_w\cdot W_i(0,p_{-i})\big) + \lambda C_w\big(W_i(1,p_{-i}) - W_i(0,p_{-i})\big)\\ &+2p_i\cdot\lambda(\alpha-1)\sigma_i C_w\big(W_i(1,p_{-i}) - W_i(0,p_{-i})\big),\end{aligned} \tag{EC.19}$$

which is strictly decreasing in p_i, given that $W_i(1,p_{-i}) < W_i(0,p_{-i})$ because always switching yields a shorter wait than never switching.

(i): If $\alpha = 1$, from (EC.19), the partial derivative of $\mathcal{C}_i(p_i,p_{-i})$ with respect to p_i becomes

$$\frac{d}{dp_i}\mathcal{C}_i(p_i,p_{-i}) = \lambda C_w\big(W_i(1,p_{-i}) - W_i(0,p_{-i})\big) < 0,$$

where $W_i(1,p_{-i}) < W_i(0,p_{-i})$. Hence, $\mathcal{C}_i(p_i,p_{-i})$ is strictly decreasing in $p_i \in [0,\overline{p}]$. From (13), $\overline{p} = 1$ when $\alpha = 1$ (given that $\lambda \in (0,1)$). Thus, $\tilde{p} = 1$.

(ii): If $\alpha \in (1, \frac{2-\overline{p}\lambda}{1-\overline{p}\lambda})$, we study pure-strategy equilibria first and then followed by mixed-strategy equilibria.

Pure-Strategy Equilibria: By the concavity result (Proposition 6), any symmetric pure equilibrium must lie at $\tilde{p} = \overline{p}$ or $\tilde{p} = 0$. We examine each candidate in turn.

First, suppose all patients adopt $\tilde{p} = \overline{p}$ (maximal stable pooling). This is a Nash equilibrium if and only if no single patient can reduce her cost by deviating to an alternate strategy $p_i \neq \overline{p}$. Given that the only potentially better alternatives are at the extreme $p_i = 0$ (never switch) or some mixture thereof, the relevant condition is that a single patient would not benefit by unilaterally switching to $p_i = 0$. In formal terms, $\tilde{p} = \overline{p}$ is an equilibrium if $\mathcal{C}_i(\overline{p},\overline{p}) \le \mathcal{C}_i(0,\overline{p})$, i.e., the cost for a representative patient when everyone (including herself) always maximally switches is no greater than the cost if she alone were to always stay while others maximally switch. This inequality can be rearranged to a

condition on the cost ratio C_v/C_w. Specifically, the cost function of the focal patient i who chooses switching probability p_i while all other choose switching probability p, is given by

$$\begin{aligned}\mathcal{C}_i(p_i,p) =&\hat{\lambda}_i(p_i,p)\left(C_v + C_w \cdot W_i(p_i,p)\right)\\ =&\left((1-\hat{\lambda}(p)p_i)\lambda + \hat{\lambda}(p)p_i\alpha\lambda\right)\cdot\left(C_v + C_w\cdot\left(1+\frac{(1-p_i)\hat{\lambda}(p)}{1-(1-p)\hat{\lambda}(p)}\right)\right)\\ =&-p_i^2\cdot\frac{(\alpha-1)\hat{\lambda}(p)^2\lambda}{1-(1-p)\hat{\lambda}(p)}C_w + p_i\cdot\left((\alpha-1)(C_w+C_v)\hat{\lambda}(p)\lambda + \frac{C_w(\alpha-1)\hat{\lambda}(p)^2\lambda}{1-(1-p)\hat{\lambda}(p)} - \frac{C_w\hat{\lambda}(p)\lambda}{1-(1-p)\hat{\lambda}(p)}\right)\\ &+\frac{C_w\hat{\lambda}(p)\lambda}{1-(1-p)\hat{\lambda}(p)} + \lambda(C_w+C_v). \end{aligned} \tag{EC.20}$$

When $p=\overline{p}$, we have $\hat{\lambda}(\overline{p}) = \frac{\lambda}{1-\overline{p}(\alpha-1)\lambda}$, and then from (EC.20), one can show that

$$\mathcal{C}_i(0,\overline{p}) = \lambda(C_w+C_v) + \frac{\lambda^2}{1-\lambda-\overline{p}(\alpha-2)\lambda}C_w,$$

and

$$\mathcal{C}_i(\overline{p},\overline{p}) = \frac{\lambda C_v}{1-\overline{p}(\alpha-1)\lambda} + \frac{\lambda C_w}{1-\lambda-\overline{p}(\alpha-2)\lambda}.$$

Then, it follows that

$$\mathcal{C}_i(\overline{p},\overline{p}) \le \mathcal{C}_i(0,\overline{p}) \Leftrightarrow C_v/C_w \le \frac{(2-\alpha)(1-\overline{p}(\alpha-1)\lambda)}{(\alpha-1)(1-\lambda-\overline{p}(\alpha-2)\lambda)} =: z_l.$$

Therefore, if $C_v/C_w \le z_l$, then we have $\mathcal{C}_i(\overline{p},\overline{p}) \le \mathcal{C}_i(0,\overline{p})$, meaning a single patient would not benefit by deviating to $p_i=0$, implying that $\tilde{p}=\overline{p}$ is a Nash equilibrium. Moreover, $\tilde{p}=\overline{p}$ is the *unique* equilibrium if the inequality is strict, that is, $C_v/C_w < z_l$. Conversely, if $C_v/C_w > z_l$, then $\mathcal{C}_i(\overline{p},\overline{p}) > \mathcal{C}_i(0,\overline{p})$, so a patient could strictly reduce her cost by deviating to never switch, implying $\tilde{p}=\overline{p}$ is not an equilibrium in that case.

Next, suppose all patients adopt $\tilde{p}=0$ (full continuity). By a parallel argument, $\tilde{p}=0$ is a Nash equilibrium if and only if $\mathcal{C}_i(0,0) \le \mathcal{C}_i(\overline{p},0)$, i.e., no patient can reduce cost by deviating to maximally switch. When $p=0$, we have $\hat{\lambda}(0)=\lambda$ and then from (EC.20), one can show that

$$\mathcal{C}_i(0,0) = \lambda C_v + \frac{\lambda}{1-\lambda}C_w,$$

and

$$\begin{aligned}\mathcal{C}_i(\overline{p},0) =&-\overline{p}^2\cdot\frac{(\alpha-1)\lambda^3}{1-\lambda}C_w + \overline{p}\cdot\left((\alpha-1)(C_w+C_v)\lambda^2 + \frac{C_w(\alpha-1)\lambda^3}{1-\lambda} - \frac{C_w\lambda^2}{1-\lambda}\right) + \frac{C_w\lambda^2}{1-\lambda} + \lambda(C_w+C_v)\\ =&-\overline{p}^2\cdot\frac{(\alpha-1)\lambda^3}{1-\lambda}C_w + \overline{p}\cdot\left((\alpha-1)\lambda^2 C_v + \frac{(\alpha-2)\lambda^2}{1-\lambda}C_w\right) + \lambda C_v + \frac{\lambda}{1-\lambda}C_w\\ =&\lambda\left(1+\overline{p}(\alpha-1)\lambda\right)\left(C_v + \frac{1-\lambda\overline{p}}{1-\lambda}C_w\right).\end{aligned}$$

Then, it follows that

$$\mathcal{C}_i(0,0) \le \mathcal{C}_i(\overline{p},0) \Leftrightarrow C_v/C_w \ge \frac{(2-\alpha)+(\alpha-1)\overline{p}\lambda}{(\alpha-1)(1-\lambda)} =: z_u.$$

Therefore, if $C_v/C_w \ge z_u$, then we have $\mathcal{C}_i(0,0) \le \mathcal{C}_i(\overline{p},0)$, meaning a lone deviator would not save cost by maximally switching, so $\tilde{p}=0$ is an equilibrium. Moreover, $\tilde{p}=0$ is the *unique* equilibrium if the inequality is strict, that is, $C_v/C_w > z_u$. Conversely, if $C_v/C_w < z_u$, then $\mathcal{C}_i(0,0) > \mathcal{C}_i(\overline{p},0)$, so a patient could strictly lower her cost by maximally switching, implying $\tilde{p}=0$ is not an equilibrium for that range.

We establish the following ordering result, whose proof is delayed until the end.

Claim EC.2. *For any $\alpha \in (1, \frac{2-\overline{p}\lambda}{1-\overline{p}\lambda})$, we have $z_l < z_u$.*

When $\alpha \in (1,2)$, both z_l and z_u are strictly positive. When $\alpha \ge 2$, $z_l \le 0$ while $z_u > 0$, so the condition $C_v/C_w \le z_l$ in case (a) is vacuously infeasible and the equilibrium has only two regions.

The above establishes the boundary regimes, namely, statements **(ii)(a)** and **(ii)(c)**. The remainder of the proof is devoted to establishing statement **(ii)(b)**.

Mixed-Strategy Equilibria: In the mean-field limit with a continuum of patients, if each patient independently plays $\overline{p}$ with probability $\tilde{q}$ and 0 with probability $1-\tilde{q}$, the law of large numbers implies that a deterministic fraction $\tilde{q}$ of the population permanently adopts the "always-switch" strategy $\overline{p}$, while the remaining $1-\tilde{q}$ permanently adopts the "always-stay" strategy 0. We call these two subpopulations *switchers* and *stayers*.

Because stayers never switch, they never disrupt continuity and therefore face no demand inflation; each stayer generates appointments at the base rate λ. Switchers disrupt continuity with probability $\overline{p}\sigma$ on each visit (where $\sigma = \hat{\lambda}$ is the busy probability), so each switcher's effective arrival rate is $\lambda(1+(\alpha-1)\overline{p}\hat{\lambda})$. The aggregate effective arrival rate satisfies the fixed-point equation

$$\hat{\lambda}(\tilde{q}) = (1-\tilde{q})\lambda + \tilde{q}\lambda\big(1+(\alpha-1)\overline{p}\,\hat{\lambda}(\tilde{q})\big) = \frac{\lambda}{1-(\alpha-1)\overline{p}\lambda\,\tilde{q}}.$$

In the mean-field regime, the minimum queue length is zero (positive idle probability), so switchers are always routed to idle servers. Only patients who stay with a busy GP contribute to queues of length two or more. The "dedicated" traffic rate is therefore

$$r(\tilde{q}) = (1-\tilde{q})\lambda + \tilde{q}(1-\overline{p})\lambda\big(1+(\alpha-1)\overline{p}\,\hat{\lambda}(\tilde{q})\big),$$

and the expected queue backlog is $S(\tilde{q}) = \hat{\lambda}(\tilde{q})/\big(1-r(\tilde{q})\big)$.

A focal stayer's cost rate is $\mathcal{C}_0(\tilde{q}) = \lambda\big(C_v + C_w(1+S(\tilde{q}))\big)$, while a focal switcher's cost rate is $\mathcal{C}_{\overline{p}}(\tilde{q}) = \lambda(1+(\alpha-1)\overline{p}\,\hat{\lambda}(\tilde{q}))\big(C_v + C_w(1+(1-\overline{p})S(\tilde{q}))\big)$.

The indifference condition $\mathcal{C}_0(\tilde{q}) = \mathcal{C}_{\overline{p}}(\tilde{q})$ can be simplified. Writing $c := C_v/C_w$ and $a := \alpha - 1$, and substituting $\tilde{q} = (\hat{\lambda} - \lambda)/(a\overline{p}\lambda\hat{\lambda})$ to eliminate $\tilde{q}$, the indifference condition becomes a quadratic equation in $\hat{\lambda}$:

$$ac(1-\overline{p})\hat{\lambda}^2 - \Big[a(c+1)(1-\overline{p}\lambda) + c\Big]\hat{\lambda} + \lambda(c+1) = 0. \tag{EC.21}$$

Define $H(\hat{\lambda})$ to be the left-hand side of (EC.21). Then H is convex (the leading coefficient $ac(1-\overline{p}) \geq 0$), and one can verify that

$$H(\lambda) = a\lambda(1-\lambda)(z_u - c), \qquad H\left(\frac{\lambda}{1-a\overline{p}\lambda}\right) = \frac{a\lambda(1-\lambda-\overline{p}(\alpha-2)\lambda)}{(1-\overline{p}(\alpha-1)\lambda)^2}(z_l - c).$$

For $c \in (z_l, z_u)$, we have $H(\lambda) > 0$ and $H(\lambda/(1-a\overline{p}\lambda)) < 0$. By the intermediate value theorem and convexity, H has a unique root $\hat{\lambda}^\dagger$ in $(\lambda, \lambda/(1-a\overline{p}\lambda))$, which yields a unique $\tilde{q} \in (0,1)$ via $\tilde{q} = (\hat{\lambda}^\dagger - \lambda)/(a\overline{p}\lambda\hat{\lambda}^\dagger)$. This proves statement **(ii)(b)**.

Combining the above analysis, the equilibrium is $\tilde{p} = \overline{p}$ if $C_v/C_w \leq z_l$, $\tilde{p} = 0$ if $C_v/C_w \geq z_u$, and mixed with probability $\tilde{q}$ if $C_v/C_w \in (z_l, z_u)$. When $\alpha \geq 2$, $z_l \leq 0$ so the first case is vacuously empty and the equilibrium has only two regions.

(iii): If $\alpha \geq \frac{2-\overline{p}\lambda}{1-\overline{p}\lambda}$, again following the analysis in part **(ii)**, we note that $z_l \leq 0$ and $z_u \leq 0$. Hence, the condition $C_v/C_w \leq z_l$ becomes infeasible, while the condition $C_v/C_w \geq z_u$ is automatically satisfied. Therefore, the equilibrium is a pure-strategy equilibrium $\tilde{p} = 0$.

∎

To complete the proof, we verify Claim EC.2 below.

Proof of Claim EC.2. We consider the following two cases.

Case (I): When $\overline{p} = 1$ (i.e., $\lambda < 1/\alpha$), $z_l = (2-\alpha)/(\alpha-1)$ and $z_u = ((2-\alpha)+(\alpha-1)\lambda)/((\alpha-1)(1-\lambda))$. Direct computation gives

$$z_u - z_l = \frac{(2-\alpha)+(\alpha-1)\lambda}{(\alpha-1)(1-\lambda)} - \frac{2-\alpha}{\alpha-1} = \frac{(2-\alpha)+((\alpha-1)\lambda-(2-\alpha)(1-\lambda))}{(\alpha-1)(1-\lambda)} = \frac{\lambda}{(\alpha-1)(1-\lambda)} > 0.$$

Case (II): When $\overline{p} < 1$ (i.e., $\lambda \geq 1/\alpha$), $z_l = (2-\alpha)\lambda/(1-\lambda)$ and $z_u = (3-\alpha-\lambda)/((\alpha-1)(1-\lambda))$. Then $z_l < z_u$ is equivalent to

$$(2-\alpha)(\alpha-1)\lambda < 3-\alpha-\lambda. \tag{EC.22}$$

Let $f(\lambda) := 3-\alpha-\lambda-(2-\alpha)(\alpha-1)\lambda = 3-\alpha+(\alpha^2-3\alpha+1)\lambda$, which is linear in λ. Over $\lambda \in [1/\alpha, 1)$:

$$f\left(\tfrac{1}{\alpha}\right) = 3-\alpha+\frac{\alpha^2-3\alpha+1}{\alpha} = \frac{1}{\alpha} > 0, \qquad f(1) = (\alpha-2)^2 \geq 0.$$

Since f is linear, $f(\lambda) \geq \min\{f(1/\alpha), f(1^-)\} > 0$ for all $\lambda \in [1/\alpha, 1)$, establishing (EC.22).

∎

EC.3.3. Proof of Proposition 7

(i): When $\alpha = 1$, both Theorem 1(i) and Theorem 2(i) give $p^\star = \tilde{p} = 1$. Hence, PoA $= \mathcal{C}(1)/\mathcal{C}(1) = 1$.

(ii): When $\alpha \in (1, \frac{2-\overline{p}\lambda}{1-\overline{p}\lambda})$.

(ii-1): When $\alpha \in (1,2)$, Theorem 1(ii) gives $p^\star = \overline{p}$ when $C_v/C_w \le y_l$, interior $p^\star \in (0,\overline{p})$ when $C_v/C_w \in (y_l, y_u)$, and $p^\star = 0$ when $C_v/C_w \ge y_u$. From Theorem 2(ii), $\tilde{p} = \overline{p}$ when $C_v/C_w \le z_l$, mixed with probability $\tilde{q}$ when $C_v/C_w \in (z_l, z_u)$, and $\tilde{p} = 0$ when $C_v/C_w \ge z_u$.

To proceed, we need the following claim, whose proof is deferred to the end.

Claim EC.3. *For $\alpha \in (1,2)$ and $\lambda \in (0,1)$, $y_l \ge z_l$.*

Efficient regimes. When $C_v/C_w \le z_l$, the planner's optimal solution is $p^\star = \overline{p}$ (by Theorem 1(ii)(a) and Claim EC.3), and the equilibrium is $\tilde{p} = \overline{p}$ (by Theorem 2(ii)(a)). Hence, PoA $= 1$. When $C_v/C_w \ge \max\{y_u, z_u\}$, the planner's optimal solution is $p^\star = 0$ (by Theorem 1(ii)(c)) and the equilibrium is $\tilde{p} = 0$ (by Theorem 2(ii)(c)). Hence, PoA $= 1$.

Inefficient regime. When $C_v/C_w \in \big(z_l, \max\{y_u, z_u\}\big)$, the planner's solutions and equilibrium do not coincide, so PoA > 1. In the mixed-strategy equilibrium ($C_v/C_w \in (z_l, z_u)$), the equilibrium assigns patients to a bang-bang mixture of 0 and $\overline{p}$, whereas the planner prescribes a common interior $p^\star$ (or $\overline{p}$ if $C_v/C_w < y_l$). Since $\tilde{p} \ne p^\star$, the cost under $\tilde{p}$ strictly exceeds $\mathcal{C}(p^\star)$. When $C_v/C_w \in [z_u, \max\{y_u, z_u\})$, the equilibrium is $\tilde{p} = 0$ while $p^\star > 0$, so $\mathcal{C}(0) > \mathcal{C}(p^\star)$ and again PoA > 1.

Upper bound. We now show PoA $\le 1 + (2-\alpha)(\alpha-1)\lambda^2/(1-\lambda)$ in the inefficient regime. Denote the cost rate under equilibrium by $\mathcal{C}_{\rm eq}$, i.e., the population-average cost rate at the symmetric equilibrium $\tilde{p}$. When $\tilde{p} = 0$ (pure strategy), $\mathcal{C}_{\rm eq} = \mathcal{C}(0)$. When the equilibrium is mixed, a fraction $\tilde{q}$ of patients permanently adopt $\overline{p}$ and the rest permanently adopt 0, producing the population-split environment described in the proof of Theorem 2. Then, the aggregate switching probability is $\tilde{q}\overline{p}$. In the population-split steady state, the dedicated traffic rate is $r_{\rm PS}(\tilde{q}) = (1-\tilde{q})\lambda + \tilde{q}(1-\overline{p})\lambda(1 + (\alpha-1)\overline{p}\hat{\lambda}(\tilde{q}\overline{p}))$, while in a homogeneous system at $\tilde{q}\overline{p}$ it would be $r_{\rm MA}(\tilde{q}\overline{p}) = (1-\tilde{q}\overline{p})\hat{\lambda}(\tilde{q}\overline{p})$. Because permanent stayers have individual arrival rate $\lambda \le \hat{\lambda}(\tilde{q}\overline{p})$, we have $r_{\rm PS} \le r_{\rm MA}$ and consequently the expected queue backlogs satisfy $S_{\rm PS} \le S_{\rm MA}$. It follows that each type's cost in the population-split environment is bounded above by the corresponding cost in the homogeneous environment:

$$\mathcal{C}_{\rm eq} = \tilde{q}\,\mathcal{C}_{\overline{p}}^{\rm PS} + (1-\tilde{q})\,\mathcal{C}_0^{\rm PS} \le \tilde{q}\,\mathcal{C}_i(\overline{p}, \tilde{q}\overline{p}) + (1-\tilde{q})\,\mathcal{C}_i(0, \tilde{q}\overline{p}).$$

By the concavity of $\mathcal{C}_i(\cdot\,, \tilde{q}\overline{p})$ in the first argument (Proposition 6 and Jensen's inequality),

$$\mathcal{C}_{\rm eq} \le \mathcal{C}_i(\tilde{q}\overline{p}, \tilde{q}\overline{p}) = \mathcal{C}(\tilde{q}\overline{p}). \tag{EC.23}$$

In both cases, $\mathcal{C}_{\rm eq} \le \mathcal{C}(\tilde{q}\overline{p})$ (setting $\tilde{q} = 0$ in the pure-strategy case).

Next, we compute the ratio $\mathcal{C}(0)/\mathcal{C}(p^\star)$ when $p^\star$ is interior (i.e., $C_v/C_w \in (y_l, y_u)$). Let $A(p) := 1 - p(\alpha-1)\lambda$ and $B(p) := 1 - \lambda + p(2-\alpha)\lambda$. Then, from (EC.16) in the proof of Theorem 1, the first-order condition $\mathcal{C}'(p^\star) = 0$ reads

$$\frac{C_v(\alpha-1)}{A(p^\star)^2} = \frac{C_w(2-\alpha)}{B(p^\star)^2} \quad \Leftrightarrow \quad \frac{C_v}{C_w} = \frac{(2-\alpha)\,A(p^\star)^2}{(\alpha-1)\,B(p^\star)^2}. \tag{EC.24}$$

Now,

$$\frac{\mathcal{C}(0)}{\mathcal{C}(p^\star)} = \frac{C_v + C_w/(1-\lambda)}{C_v/A(p^\star) + C_w/B(p^\star)}.$$

Substituting C_v/C_w from (EC.24) and multiplying numerator and denominator by $(\alpha-1)B(p^\star)^2$, the denominator factor satisfies

$$(2-\alpha)A(p^\star) + (\alpha-1)B(p^\star) = 1 - (\alpha-1)\lambda,$$

which is a constant independent of $p^\star$ (the $p^\star$-dependent terms cancel). For the numerator factor, expanding $A(p^\star)^2$ and $B(p^\star)^2$ yields

$$(2-\alpha)A(p^\star)^2(1-\lambda) + (\alpha-1)B(p^\star)^2 = \big[1-(\alpha-1)\lambda\big]\big[(1-\lambda) + (2-\alpha)(\alpha-1)(p^\star)^2\lambda^2\big],$$

where the linear terms in $p^\star$ cancel and only the $(p^\star)^2$ term survives. Dividing the two yields

$$\frac{\mathcal{C}(0)}{\mathcal{C}(p^\star)} = 1 + \frac{(2-\alpha)(\alpha-1)\,(p^\star)^2\lambda^2}{1-\lambda}. \tag{EC.25}$$

Finally, in the inefficient regime, $\mathcal{C}_{\text{eq}} \le \mathcal{C}(\tilde{q}\overline{p})$ by (EC.23), and $\mathcal{C}(\tilde{q}\overline{p}) \le \mathcal{C}(0)$ since $\mathcal{C}$ is convex (Proposition 5) with minimizer $p^\star \in [0, \overline{p}]$ and $\tilde{q}\overline{p}$ lies between 0 and $\overline{p}$. Hence,

$$\text{PoA} = \frac{\mathcal{C}_{\text{eq}}}{\mathcal{C}(p^\star)} \le \frac{\mathcal{C}(0)}{\mathcal{C}(p^\star)}.$$

When $p^\star$ is interior, from (EC.25) and the fact that $p^\star \le 1$, we have

$$\text{PoA} \le 1 + \frac{(2-\alpha)(\alpha-1)\,\lambda^2}{1-\lambda}.$$

When $p^\star = \overline{p}$ (if $C_v/C_w < y_l$), $\tilde{q}\overline{p}$ is close to $\overline{p}$ and the PoA is close to 1. Specifically, $\mathcal{C}(\tilde{q}\overline{p}) \le \mathcal{C}(0) \le (1 + (2-\alpha)(\alpha-1)\lambda^2/(1-\lambda)) \cdot \mathcal{C}(\overline{p}) \le (1 + (2-\alpha)(\alpha-1)\lambda^2/(1-\lambda)) \cdot \mathcal{C}(p^\star)$, where the last inequality uses $p^\star = \overline{p}$. This establishes the bound stated in part (ii).

(ii-2): When $\alpha \in [2, \frac{2-\overline{p}\lambda}{1-\overline{p}\lambda})$, Theorem 1(iii) gives $p^\star = 0$ for all C_v/C_w, and Theorem 2(ii) gives $\tilde{p} = 0$ when $C_v/C_w \ge z_u$ (so PoA $= 1$), and $\tilde{p}$ is a mixed strategy when $C_v/C_w < z_u$. In the latter case, the equilibrium has a positive fraction of patients switching to $\overline{p}$, inflating demand. Since $p^\star = 0$ is optimal and $\tilde{p} \ne 0$, the equilibrium cost exceeds $\mathcal{C}(0)$, giving PoA > 1.

Upper bound. We show $\text{PoA} \le 1+(\alpha-2)\lambda$. Since $p^\star=0$, $\mathcal{C}(p^\star)=\mathcal{C}(0)$. In the mixed equilibrium, by indifference, the population-average cost rate equals the stayer cost rate. Since $S_{\text{PS}} \le S_{\text{MA}}$ (as argued in part (ii-1)), the stayer cost in the population-split environment satisfies

$$\mathcal{C}_{\text{eq}} = \lambda(C_v + C_w(1+S_{\text{PS}})) \;\le\; \lambda(C_v + C_w(1+S_{\text{MA}})) = \mathcal{C}_i(0,\, \tilde{q}\overline{p}). \tag{EC.26}$$

From (EC.20) with $p_i=0$ and population switching probability p, $\mathcal{C}_i(0,p) = \lambda(C_v+C_w) + \lambda^2 C_w/D_2(p)$ where $D_2(p)=1-\lambda-p(\alpha-2)\lambda$. Since $\mathcal{C}(0)=\mathcal{C}_i(0,0)=\lambda(C_v+C_w)+\lambda^2 C_w/(1-\lambda)$,

$$\text{PoA}-1 \;=\; \frac{\mathcal{C}_i(0,\tilde{q}\overline{p})-\mathcal{C}(0)}{\mathcal{C}(0)} \;=\; \frac{\tilde{q}\overline{p}(\alpha-2)\lambda^2}{\big(1-\lambda-\tilde{q}\overline{p}(\alpha-2)\lambda\big)\big((1-\lambda)\,C_v/C_w+1\big)}. \tag{EC.27}$$

Since $\tilde{q}\le 1$ and $(1-\lambda)\,C_v/C_w \ge 0$, we obtain

$$\text{PoA}-1 \le \frac{\overline{p}(\alpha-2)\lambda^2}{1-\lambda-\overline{p}(\alpha-2)\lambda},$$

where $1-\lambda-\overline{p}(\alpha-2)\lambda>0$. The stability condition $\overline{p}(\alpha-1)\lambda \le 1-\lambda$ gives $\overline{p}(\alpha-2)\lambda = \overline{p}(\alpha-1)\lambda - \overline{p}\lambda \le (1-\lambda)-\overline{p}\lambda$, so $1-\lambda-\overline{p}(\alpha-2)\lambda \ge \overline{p}\lambda$. Therefore,

$$\text{PoA}-1 \;\le\; \frac{\overline{p}(\alpha-2)\lambda^2}{\overline{p}\lambda} \;=\; (\alpha-2)\lambda.$$

Combining **(ii-1)** and **(ii-2)**, $\text{PoA}=1$ when $C_v/C_w \le z_l$ or $C_v/C_w \ge \max\{y_u, z_u\}$, and $1<\text{PoA}\le 1+\max\left\{\frac{(2-\alpha)(\alpha-1)\,\lambda^2}{1-\lambda},\,(\alpha-2)\lambda\right\}$ otherwise.

(iii): When $\alpha \ge \frac{2-\overline{p}\lambda}{1-\overline{p}\lambda}$, both Theorem 1(iii) (since $\alpha\ge 2$) and Theorem 2(iii) give $p^\star=\tilde{p}=0$. Hence, $\text{PoA}=\mathcal{C}(0)/\mathcal{C}(0)=1$.

∎

To complete the proof, we verify Claim EC.3 below.

Proof of Claim EC.3. Recall that $y_l = \frac{(2-\alpha)(1-\overline{p}(\alpha-1)\lambda)^2}{(\alpha-1)(1-\lambda-\overline{p}(\alpha-2)\lambda)^2}$ and $z_l = \frac{(2-\alpha)(1-\overline{p}(\alpha-1)\lambda)}{(\alpha-1)(1-\lambda-\overline{p}(\alpha-2)\lambda)}$. It is clear that

$$y_l = z_l \cdot \frac{1-\overline{p}(\alpha-1)\lambda}{1-\lambda-\overline{p}(\alpha-2)\lambda}.$$

Note that $1-\lambda-\overline{p}(\alpha-2)\lambda = 1-\overline{p}(\alpha-1)\lambda-(1-\overline{p})\lambda \le 1-\overline{p}(\alpha-1)\lambda$ for all $\overline{p}\in(0,1]$. Hence, the fraction in the above display satisfies $\frac{1-\overline{p}(\alpha-1)\lambda}{1-\lambda-\overline{p}(\alpha-2)\lambda}\ge 1$. Therefore, $y_l \ge z_l$.

∎

EC.4. Proofs from Section 6

EC.4.1. Proof of Lemma 2

We follow the proof of Lemma 1. Under the stability condition (22), as $n\to\infty$, a positive fraction of GPs is idle, so the minimum queue length is 0 with probability approaching 1. For any fixed $i\ge 1$, $s_i^{(\infty)}(t;p_1,p_2)$ increases when an arrival causes a queue-length increase from $i-1$ to i and decreases when a service completion reduces a queue from i to $i-1$.

Arrivals that increase $s_i^{(\infty)}$. For $i=1$, consider a type k arrival. The usual GP is idle with probability $1-s_1^{(\infty)}$, in which case the patient is routed there. If the usual GP is busy (probability $s_1^{(\infty)}$), the patient switches to an idle shortest-queue GP with probability p_k and stays with probability $1-p_k$. Because the minimum queue length is 0, every switcher joins an idle GP. Hence the probability that a type k arrival is routed to an idle GP is

$$\left(1-s_1^{(\infty)}\right)+s_1^{(\infty)}\cdot p_k = 1-(1-p_k)\,s_1^{(\infty)}.$$

Multiplying by $\hat{\lambda}_k^{(\infty)}$ and summing over $k\in\{1,2\}$ gives the total rate $\sum_{k=1}^{2}(1-(1-p_k)\,s_1^{(\infty)})\hat{\lambda}_k^{(\infty)}$.

For $i\geq 2$, any switcher joins an idle GP and thus cannot increase $s_i^{(\infty)}$. A patient joins a queue of length $i-1\geq 1$ only by staying with the usual GP (probability $1-p_k$) when that GP has exactly $i-1$ patients (probability $s_{i-1}^{(\infty)}-s_i^{(\infty)}$), yielding total rate $\sum_{k=1}^{2}(1-p_k)(s_{i-1}^{(\infty)}-s_i^{(\infty)})\hat{\lambda}_k^{(\infty)}$.

Departures that decrease $s_i^{(\infty)}$. A service completion at a GP with exactly i patients decreases $s_i^{(\infty)}$. The fraction of such GPs is $s_i^{(\infty)}-s_{i+1}^{(\infty)}$, each completing service at rate 1. This is identical to the homogeneous case.

Combining the arrival and departure rates establishes (23)–(24).

∎

EC.4.2. Proof of Proposition 8

Write $\hat{\lambda}:=\hat{\lambda}^{(\infty)}(p_1,p_2)$, $\hat{\lambda}_k:=\hat{\lambda}_k^{(\infty)}(p_1,p_2)$, and $r:=(1-p_1)\hat{\lambda}_1+(1-p_2)\hat{\lambda}_2$. Setting the right-hand side of (24) to zero for $i\geq 2$:

$$s_i^{(\infty)}(\infty)-s_{i+1}^{(\infty)}(\infty)=r\left(s_{i-1}^{(\infty)}(\infty)-s_i^{(\infty)}(\infty)\right).$$

Define $\Delta_i:=s_i^{(\infty)}(\infty)-s_{i+1}^{(\infty)}(\infty)$ for $i\geq 1$. Then $\Delta_i=r\,\Delta_{i-1}$ for $i\geq 2$, whence $\Delta_i=\Delta_1\,r^{i-1}$ for all $i\geq 1$. Since $s_i^{(\infty)}(\infty)\to 0$ as $i\to\infty$, summing the geometric increments gives

$$s_i^{(\infty)}(\infty)=\frac{\Delta_1\,r^{i-1}}{1-r},\quad i\geq 1. \tag{EC.28}$$

Setting the right-hand side of (23) to zero:

$$\Delta_1=\hat{\lambda}-r\,s_1^{(\infty)}(\infty). \tag{EC.29}$$

Substituting (EC.29) into (EC.28) with $i=1$ and solving: $s_1^{(\infty)}(\infty)=\hat{\lambda}$. Then $\Delta_1=\hat{\lambda}(1-r)$, and (EC.28) yields

$$s_i^{(\infty)}(\infty;p_1,p_2)=r^{i-1}\,\hat{\lambda}^{(\infty)}(p_1,p_2),\quad i\geq 1.$$

∎

EC.4.3. Proof of Proposition 9

Derivation of $\sigma^{(\infty)}$, $\hat{\lambda}^{(\infty)}$, and $\hat{\lambda}_k^{(\infty)}$. Under (22), the shortest queue length is 0 in the mean-field limit, so $\sigma^{(\infty)}(p_1,p_2) = s_1^{(\infty)}(\infty;p_1,p_2) = \hat{\lambda}^{(\infty)}(p_1,p_2)$ (the last equality by Proposition 8). Substituting $\sigma^{(\infty)} = \hat{\lambda}^{(\infty)}$ into (21):

$$\hat{\lambda}_k^{(\infty)} = \big(1 + (\alpha_k - 1)\, p_k\, \hat{\lambda}^{(\infty)}\big)\lambda_k, \quad k \in \{1,2\}.$$

Summing over k and using $\alpha_2 = 1$, we have $\hat{\lambda}^{(\infty)} = \lambda_1(1 + (\alpha_1 - 1)p_1\hat{\lambda}^{(\infty)}) + \lambda_2$. Since this equation is linear in $\hat{\lambda}^{(\infty)}$, it has the unique solution $\hat{\lambda}^{(\infty)} = (\lambda_1+\lambda_2)/(1 - p_1(\alpha_1-1)\lambda_1)$. Substituting back into the above display yields $\hat{\lambda}_k^{(\infty)}(p_1,p_2) = \big(1 + (\alpha_k - 1)\, p_k\, \hat{\lambda}^{(\infty)}(p_1,p_2)\big)\lambda_k$, for $k \in \{1,2\}$. In particular, $\hat{\lambda}_2^{(\infty)} = \lambda_2$ and $\hat{\lambda}_1^{(\infty)} = \lambda_1(1 + (\alpha_1-1)p_1\lambda_2)/(1 - p_1(\alpha_1-1)\lambda_1)$.

Write $\hat{\lambda} := \hat{\lambda}^{(\infty)}(p_1,p_2)$, $\hat{\lambda}_k := \hat{\lambda}_k^{(\infty)}(p_1,p_2)$, and $r := (1-p_1)\hat{\lambda}_1 + (1-p_2)\hat{\lambda}_2$.

Derivation of $W_k^{(\infty)}$. An arriving type k patient enters service immediately if the usual GP is idle or if the patient switches to an idle GP. Waiting beyond the unit service time arises only when the patient stays with a busy usual GP (which occurs with probability $1 - p_k$, conditional on the GP being busy), in which case the expected number of customers ahead is $\sum_{i\geq 1} s_i^{(\infty)}(\infty)$. Hence

$$W_k^{(\infty)}(p_1,p_2) = 1 + (1-p_k)\sum_{i\geq 1} s_i^{(\infty)}(\infty;p_1,p_2).$$

By Proposition 8, $\sum_{i\geq 1} s_i^{(\infty)}(\infty;p_1,p_2) = \hat{\lambda}/(1-r)$, so $W_k^{(\infty)}(p_1,p_2) = 1 + (1-p_k)\hat{\lambda}/(1-r)$. To reduce to primitive parameters, write $A := 1 - p_1(\alpha_1-1)\lambda_1$ so that $\hat{\lambda} = (\lambda_1+\lambda_2)/A$. Expanding r using the expressions for $\hat{\lambda}_1$ and $\hat{\lambda}_2$:

$$A(1-r) = 1 - \big(p_1(\alpha_1-1) + (1-p_1)\big)\lambda_1 - (1-p_2)\lambda_2 + (\alpha_1-1)p_1\lambda_1\lambda_2(p_1-p_2).$$

Since $(1-p_k)\hat{\lambda}/(1-r) = (1-p_k)(\lambda_1+\lambda_2)/\big(A(1-r)\big)$, where $A(1-r)$ can be substituted using the right-hand side of the above display, it follows that

$$W_k^{(\infty)}(p_1,p_2) = 1 + \frac{(1-p_k)(\lambda_1+\lambda_2)}{1 - \big(p_1(\alpha_1-1) + (1-p_1)\big)\lambda_1 - (1-p_2)\lambda_2 + (\alpha_1-1)p_1\lambda_1\lambda_2(p_1-p_2)}.$$

∎

EC.4.4. Proof of Theorem 3

Define $a := (\alpha_1-1)\lambda_1$, $b := (2-\alpha_1)\lambda_1 + \lambda_2$, $\Lambda := \lambda_1 + \lambda_2$, and

$$A(p_a) := 1 - a\,p_a, \qquad B(p_a) := (1-\Lambda) + b\,p_a.$$

Note that $a + b = \Lambda$. Under the across-the-board policy $p_1 = p_2 = p_a$, the cross-type interaction term in Proposition 9 vanishes ($p_1 - p_2 = 0$), and the identity $A - B = \Lambda(1-p_a)$ yields $W_k(p_a,p_a) =$

$A(p_a)/B(p_a)$ for both types. Define $F(p_a) := 1 + (\alpha_1 - 1)p_a\lambda_2$. From Proposition 9, $\hat{\lambda}_1 = \lambda_1 F/A$ and $\hat{\lambda}_2 = \lambda_2$, so (25) becomes

$$\mathcal{C}_a(p_a) = \frac{\lambda_1 F\, C_{v,1}}{A} + \frac{\lambda_1 F\, C_{w,1}}{B} + \lambda_2 C_{v,2} + \frac{\lambda_2 C_{w,2}\, A}{B}. \tag{EC.30}$$

Every $p_a \in [0, \overline{p}_1]$ satisfies $ap_a < 1 - \Lambda$, so $A > \Lambda > 0$ and $B > 0$ on $[0, \overline{p}_1]$.

Differentiating (EC.30) and using $a + b = \Lambda$:

$$\mathcal{C}_a'(p_a) = \Lambda \left(\frac{a\, C_{v,1}}{A(p_a)^2} - \frac{\Phi}{B(p_a)^2} \right), \tag{EC.31}$$

$$\mathcal{C}_a''(p_a) = 2\Lambda \left(\frac{a^2\, C_{v,1}}{A(p_a)^3} + \frac{\Phi\, b}{B(p_a)^3} \right), \tag{EC.32}$$

where $\Phi := \lambda_1 C_{w,1}\big[(2-\alpha_1) + (\alpha_1 - 1)\lambda_2\big] + \lambda_2 C_{w,2}\big[1 - (\alpha_1 - 1)\lambda_1\big] = C_{w,1}\,\varphi$.

Using $A(0) = 1$, $B(0) = 1 - \Lambda$, $A(\overline{p}_1) = \Lambda$, $B(\overline{p}_1) = (1-\Lambda)\Lambda/a$, (EC.31) gives

$$\mathcal{C}_a'(0) = \Lambda\Big(a\, C_{v,1} - \tfrac{\Phi}{(1-\Lambda)^2}\Big), \qquad \mathcal{C}_a'(\overline{p}_1) = \tfrac{a}{\Lambda}\Big(C_{v,1} - \tfrac{a\,\Phi}{(1-\Lambda)^2}\Big).$$

Hence, $\mathcal{C}_a'(0) \geq 0$ if and only if

$$C_{v,1}/C_{w,1} \geq \frac{\varphi}{a(1-\Lambda)^2} =: y_u^a,$$

and $\mathcal{C}_a'(\overline{p}_1) \leq 0$ if and only if

$$C_{v,1}/C_{w,1} \leq \frac{a\varphi}{(1-\Lambda)^2} =: y_l^a.$$

It is clear that $y_l^a = a^2\, y_u^a$.

(i): If $\alpha_1 = 1$, then $a = 0$ and $\Phi = \lambda_1 C_{w,1} + \lambda_2 C_{w,2} > 0$, so $\mathcal{C}_a' = -\Lambda\Phi/B^2 < 0$ for all p_a, giving $p_a^\star = 1$.

(ii): If $\alpha_1 \in (1,2)$, then $a > 0$, $b > 0$, and $\Phi > 0$, so (EC.32) gives $\mathcal{C}_a'' \geq 0$ (convexity). By convexity, $\mathcal{C}_a'$ is non-decreasing, and the sign at the endpoints determines the regime. Moreover, when $\alpha_1 \in (1,2)$, $a = (\alpha_1 - 1)\lambda_1 \in (0,1)$. Hence, $0 < y_l^a < y_u^a$.

(a) If $C_{v,1}/C_{w,1} \leq y_l^a$, then $\mathcal{C}_a' \leq 0$ on $[0, \overline{p}_1]$. Hence, $p_a^\star = \overline{p}_1$.

(b) If $C_{v,1}/C_{w,1} \in (y_l^a, y_u^a)$, there is a unique $p_a^\star \in (0, \overline{p}_1)$ satisfying $\mathcal{C}_a'(p_a^\star) = 0$. Setting (EC.31) to zero and taking square roots ($A, B > 0$):

$$p_a^\star = \frac{\sqrt{\Phi} - (1-\Lambda)\sqrt{a\, C_{v,1}}}{a\sqrt{\Phi} + b\sqrt{a\, C_{v,1}}}.$$

Substituting $\Phi = C_{w,1}\varphi$ and dividing by $\sqrt{C_{w,1}}$ yields

$$p_a^\star = \frac{\sqrt{\varphi} \;-\; (1-\lambda_1-\lambda_2)\,\sqrt{(\alpha_1-1)\lambda_1\, C_{v,1}/C_{w,1}}}{(\alpha_1-1)\lambda_1\,\sqrt{\varphi} \;+\; \big[(2-\alpha_1)\lambda_1+\lambda_2\big]\,\sqrt{(\alpha_1-1)\lambda_1\, C_{v,1}/C_{w,1}}}.$$

(c) If $C_{v,1}/C_{w,1} \geq y_u^a$, then $\mathcal{C}_a' \geq 0$ on $[0, \overline{p}_1]$. Hence, $p_a^\star = 0$.

(iii): If $\alpha_1 \geq 2$:

(a) If $\varphi \leq 0$, then $\Phi \leq 0$ and (EC.31) gives $\mathcal{C}_a' > 0$ on $(0, \overline{p}_1)$, so $p_a^\star = 0$.

(b) If $\varphi > 0$ and $b \geq 0$, then $\Phi > 0$ and (EC.32) gives $\mathcal{C}_a'' \geq 0$. Moreover, $b \geq 0$ requires $(\alpha_1 - 2)\lambda_1 \leq \lambda_2$, so $a = \lambda_1 + (\alpha_1 - 2)\lambda_1 \leq \lambda_1 + \lambda_2 = \Lambda < 1$. Hence, $y_l^a = a^2\, y_u^a < y_u^a$, and the three-case characterization of **(ii)** applies.

(c) If $\varphi > 0$ and $b < 0$, then $\mathcal{C}_a$ need not be convex, so the optimum is found by comparing $\mathcal{C}_a$ at the boundaries and at any interior critical points of (EC.31).

∎

EC.4.5. Proof of Theorem 4

Throughout, write $\Lambda := \lambda_1 + \lambda_2$ and $r := (1 - p_1)\hat{\lambda}_1 + (1 - p_2)\hat{\lambda}_2$.

Step 1: Optimality of $p_2^\star = 1$. From Proposition 9,

$$W_k(p_1, p_2) = 1 + \frac{(1 - p_k)\hat{\lambda}}{1 - r}, \quad k \in \{1, 2\}.$$

Differentiating with respect to p_2 and using $\partial r / \partial p_2 = -\lambda_2$:

$$\frac{\partial W_1(p_1, p_2)}{\partial p_2} = -\frac{(1-p_1)\hat{\lambda}\,\lambda_2}{(1-r)^2} \leq 0, \qquad \frac{\partial W_2(p_1, p_2)}{\partial p_2} = -\frac{\hat{\lambda}\big[(1-r) + (1-p_2)\lambda_2\big]}{(1-r)^2} < 0.$$

Since $\hat{\lambda}_k$ is independent of p_2 and $C_{w,k} > 0$,

$$\frac{\partial \mathcal{C}(p_1, p_2)}{\partial p_2} = \hat{\lambda}_1 C_{w,1} \frac{\partial W_1(p_1, p_2)}{\partial p_2} + \lambda_2 C_{w,2} \frac{\partial W_2(p_1, p_2)}{\partial p_2} < 0,$$

for all $(p_1, p_2) \in [0, \overline{p}_1] \times [0, 1]$. Hence, $p_2^\star = 1$.

With $p_2 = 1$ fixed, define the following auxiliary functions of p_1:

$$D(p_1) := 1 - (\alpha_1 - 1)p_1\lambda_1, \qquad F(p_1) := 1 + (\alpha_1 - 1)p_1\lambda_2,$$
$$E(p_1) := (1-\lambda_1) + p_1\lambda_1\big[(2-\alpha_1) - (\alpha_1 - 1)\lambda_2\big] + (\alpha_1 - 1)p_1^2\lambda_1\lambda_2, \quad \phi(p_1) := (2 - \alpha_1) + (\alpha_1 - 1)\lambda_2(2p_1 - 1).$$

From Lemma 2 and Proposition 9, with $p_2 = 1$, $\hat{\lambda}_1 = \lambda_1 F/D$, $\hat{\lambda}_2 = \lambda_2$, $\hat{\lambda} = \Lambda/D$, $W_1 = 1 + (1 - p_1)\Lambda/E$, and $W_2 = 1$. Then, $\mathcal{C}(p_1, 1)$ is given by

$$\begin{aligned}
\mathcal{C}(p_1, 1) =& \hat{\lambda}_1 (C_{v,1} + C_{w,1} W_1) + \hat{\lambda}_2 (C_{v,2} + C_{w,2} W_2) \\
=& \frac{\lambda_1 F C_{v,1}}{D} + \frac{\lambda_1 F \left(1 + \frac{(1-p_1)\Lambda}{E}\right) C_{w,1}}{D} + \lambda_2(C_{v,2} + C_{w,2}) \\
=& \frac{\lambda_1 F C_{v,1}}{D} + \frac{\lambda_1 F \,(E + (1-p_1)\Lambda)\, C_{w,1}}{D\,E} + \lambda_2(C_{v,2} + C_{w,2}) \\
=& \frac{\lambda_1 F C_{v,1}}{D} + \frac{\lambda_1 F \big[1 + \lambda_2(1 - p_1)\big]\, C_{w,1}}{E} + \lambda_2(C_{v,2} + C_{w,2}) \\
=& \frac{\lambda_1 F C_{v,1}}{D} + \frac{\lambda_1 G C_{w,1}}{E} + \lambda_2(C_{v,2} + C_{w,2}),
\end{aligned} \tag{EC.33}$$

where $G(p_1) := F(p_1)\big[1+\lambda_2(1-p_1)\big]$, and the last equality follows from the identity below:

$$F(p_1)\big[E(p_1)+(1-p_1)\Lambda\big] = D(p_1)\,G(p_1).$$

By the stability condition (22), we have $D(p_1) \geq \Lambda > 0$ and $E(p_1) = D(p_1)(1-r) > 0$.

Differentiating (EC.33) with respect to p_1:

$$\begin{aligned}\frac{d}{dp_1}\mathcal{C}(p_1,1) =& \lambda_1 C_{v,1}\cdot\frac{d}{dp_1}\left(\frac{F}{D}\right)+\lambda_1 C_{w,1}\cdot\frac{d}{dp_1}\left(\frac{G}{E}\right)\\ =& \lambda_1 C_{v,1}\cdot\frac{F'D-FD'}{D^2}+\lambda_1 C_{w,1}\cdot\frac{G'E-GE'}{E^2}\\ =& \lambda_1 C_{v,1}\cdot\frac{(\alpha_1-1)\Lambda}{D^2}+\lambda_1 C_{w,1}\cdot\frac{-\phi(\lambda_2 E+\lambda_1 G)}{E^2},\end{aligned} \tag{EC.34}$$

where $E'(p_1) = \lambda_1\,\phi(p_1)$ and $G'(p_1) = -\lambda_2\,\phi(p_1)$.

Assume that we can establish the following result, whose proof is delayed until the end.

Claim EC.4. *$\lambda_2\,E(p_1)+\lambda_1\,G(p_1) = \Lambda$ for all p_1.*

Then, applying Claim EC.4 to (EC.34) yields

$$\mathcal{C}'(p_1,1) = \Lambda\,\lambda_1\left[\frac{(\alpha_1-1)\,C_{v,1}}{D(p_1)^2}-\frac{C_{w,1}\,\phi(p_1)}{E(p_1)^2}\right]. \tag{EC.35}$$

Step 2: Boundary derivatives and threshold ordering. At $p_1 = 0$, we have $D(0) = 1$, $E(0) = 1-\lambda_1$, $\phi(0) = (2-\alpha_1)-(\alpha_1-1)\lambda_2$. Then, from (EC.35),

$$\mathcal{C}'(0,1) = \Lambda\,\lambda_1\left[(\alpha_1-1)\,C_{v,1}-\frac{C_{w,1}\,\phi(0)}{(1-\lambda_1)^2}\right].$$

Hence $\mathcal{C}'(0,1)\geq 0$ if and only if

$$\frac{C_{v,1}}{C_{w,1}} \geq \frac{\phi(0)}{(\alpha_1-1)(1-\lambda_1)^2} = \frac{(2-\alpha_1)-(\alpha_1-1)\lambda_2}{(\alpha_1-1)(1-\lambda_1)^2} =: y'_u. \tag{EC.36}$$

At $p_1 = \overline{p}_1$, $\mathcal{C}'(\overline{p}_1,1)\leq 0$ if and only if

$$\begin{aligned}\frac{C_{v,1}}{C_{w,1}} \leq& \frac{\phi(\overline{p}_1)\,D(\overline{p}_1)^2}{(\alpha_1-1)\,E(\overline{p}_1)^2}\\ =& \frac{\big[(2-\alpha_1)+(\alpha_1-1)\lambda_2(2\overline{p}_1-1)\big]\,\big(1-\overline{p}_1(\alpha_1-1)\lambda_1\big)^2}{(\alpha_1-1)\big[(1-\lambda_1)+\overline{p}_1\lambda_1\big((2-\alpha_1)-(\alpha_1-1)\lambda_2\big)+(\alpha_1-1)\overline{p}_1^2\lambda_1\lambda_2\big]^2} =: y'_l.\end{aligned} \tag{EC.37}$$

Step 3: Optimal $p_1^\star$. Define

$$T(p_1) := \frac{\phi(p_1)\,D(p_1)^2}{(\alpha_1-1)\,E(p_1)^2}. \tag{EC.38}$$

Then, it is clear that $T(0) = y'_u$ and $T(\overline{p}_1) = y'_l$. One important observation is that, from (EC.35),

$$\mathcal{C}'(p_1,1)\leq 0 \quad \text{if and only if} \quad C_{v,1}/C_{w,1}\leq T(p_1). \tag{EC.39}$$

To proceed, we need the following auxiliary claims, whose proofs are delayed until the end.

CLAIM EC.5. *For all $\alpha_1 > 1$, $\mathcal{C}'(p_1,1)$ has at most two zeros on $[0,\overline{p}_1]$. Equivalently, $\mathcal{C}(p_1,1)$ has at most one interior local minimum and one interior local maximum on $(0,\overline{p}_1)$.*

CLAIM EC.6. *$\phi(\overline{p}_1)$ (equivalently, y_l') is strictly decreasing in α_1 on $(1,\infty)$, with $\lim_{\alpha_1\to 1^+}\phi(\overline{p}_1) = 1$ and $\lim_{\alpha_1\to\infty}\phi(\overline{p}_1) = -\infty$. Consequently, there exists a unique $\overline{\alpha}_1 > 1$ at which $y_l' = 0$, with $\phi(\overline{p}_1) > 0$ (equivalently, $y_l' > 0$) for $\alpha_1 \in (1,\overline{\alpha}_1)$ and $\phi(\overline{p}_1) < 0$ (equivalently, $y_l' < 0$) for $\alpha_1 \in (\overline{\alpha}_1,\infty)$.*

CLAIM EC.7. *Define $\alpha_1^\ddagger := (2+\lambda_2)/(1+\lambda_2)$. Then:*

(a) $y_u' > 0$ if and only if $\alpha_1 < \alpha_1^\ddagger$;

(b) $y_l' > 0$ for all $\alpha_1 \in (1,\alpha_1^\ddagger)$;

(c) y_u'/y_l' is strictly decreasing in α_1 on $(1,\alpha_1^\ddagger)$, with $\lim_{\alpha_1\to 1^+} y_u'/y_l' = 1/(1-\lambda_1)^2 > 1$ and $\lim_{\alpha_1\to\alpha_1^\ddagger} y_u'/y_l' = 0$. Consequently, there exists a unique $\alpha_1^\dagger \in (1,\alpha_1^\ddagger)$ at which $y_u' = y_l'$, with $y_u'/y_l' > 1$ for $\alpha_1 \in (1,\alpha_1^\dagger)$ and $y_u'/y_l' \le 1$ for $\alpha_1 \in [\alpha_1^\dagger,\alpha_1^\ddagger)$.

CLAIM EC.8. *For $\alpha_1 \in (1,\alpha_1^\ddagger)$, $T(p_1) \ge \min(y_u', y_l')$ for all $p_1 \in [0,\overline{p}_1]$.*

(i): If $\alpha_1 = 1$, then $(\alpha_1 - 1) = 0$ and $\phi(p_1) = 1 + \lambda_2(2p_1 - 1) > 0$ for all $p_1 \in [0,1]$. The first term of (EC.35) vanishes and the second is negative, so $\mathcal{C}'(p_1,1) < 0$, implying that $\mathcal{C}'(p_1,1)$ is strictly decreasing in $p_1 \in [0,\overline{p}_1]$ (with $\overline{p}_1 = 1$). Hence, $p_1^\star = 1$.

(ii): If $\alpha_1 \in (1,\overline{\alpha}_1)$, then $y_l' > 0$ by Claim EC.6.

(a) If $C_{v,1}/C_{w,1} \le \min(y_l', y_u')$, then $\mathcal{C}'(0,1) \le 0$ and $\mathcal{C}'(\overline{p}_1,1) \le 0$. When $\alpha_1 \ge \alpha_1^\ddagger$, $y_u' \le 0$ (Claim EC.7(a)), so the condition $C_{v,1}/C_{w,1} \le \min(y_l', y_u')$ is vacuous for positive cost ratios. When $\alpha_1 < \alpha_1^\ddagger$, Claim EC.8 gives $T(p_1) \ge \min(y_u', y_l') \ge C_{v,1}/C_{w,1}$ on $[0,\overline{p}_1]$, so $\mathcal{C}'(p_1,1) \le 0$ everywhere and $p_1^\star = \overline{p}_1$.

(b) If $C_{v,1}/C_{w,1} \in (y_l', y_u')$ (non-empty only when $y_l' < y_u'$, i.e., $\alpha_1 < \alpha_1^\dagger$), then $\mathcal{C}'(0,1) < 0$ and $\mathcal{C}'(\overline{p}_1,1) > 0$. By the intermediate value theorem, $\mathcal{C}'(p_1,1)$ has at least one zero in $(0,\overline{p}_1)$. By Claim EC.5, it has at most two, so the count is odd and thus exactly one. Since $\mathcal{C}(p_1,1)$ transitions from decreasing to increasing, this zero is the unique interior minimum. Hence, $p_1^\star \in (0,\overline{p}_1)$ is the unique solution to $\mathcal{C}'(p_1^\star,1) = 0$.

(c) If $C_{v,1}/C_{w,1} \in [y_u', y_l']$ (non-empty only when $y_u' \le y_l'$, i.e., $\alpha_1 \ge \alpha_1^\dagger$), then $\mathcal{C}'(0,1) \ge 0$ and $\mathcal{C}'(\overline{p}_1,1) \le 0$. By the same parity argument as (b), $\mathcal{C}'(p_1,1)$ has exactly one interior zero. Since $\mathcal{C}(p_1,1)$ transitions from increasing to decreasing, this zero is an interior maximum. Hence, $p_1^\star = \arg\min_{p_1\in\{0,\overline{p}_1\}} \mathcal{C}(p_1,1)$.

(d) If $C_{v,1}/C_{w,1} \ge \max(y_l', y_u')$, then $\mathcal{C}'(0,1) \ge 0$ and $\mathcal{C}'(\overline{p}_1,1) \ge 0$. By Claim EC.5, $\mathcal{C}(p_1,1)$ has at most one interior local minimum on $(0,\overline{p}_1)$. If one exists, $p_1^\star$ is determined by comparing $\mathcal{C}(0,1)$ with $\mathcal{C}(p_1,1)$ at that interior local minimum; if none exists, $p_1^\star = 0$.

(iii): If $\alpha_1 \geq \overline{\alpha}_1$, then $\phi(\overline{p}_1) \leq 0$ by Claim EC.6. Since $\phi(p_1)$ is increasing in p_1, $\phi(p_1) \leq 0$ on $[0, \overline{p}_1]$. As a result, both terms of (EC.35) are non-negative, so $\mathcal{C}'(p_1, 1) \geq 0$ for all $p_1 \in [0, \overline{p}_1]$ and for all cost ratios. Hence, $p_1^\star = 0$.

∎

To complete the proof, we verify Claims EC.4–EC.8 below.

Proof of Claim EC.4. Note that both $\lambda_2 E + \lambda_1 G$ and the constant Λ are polynomials in p_1 of degree at most 2. It suffices to verify equality at three points. At $p_1 = 0$, we have $\lambda_2 E(0) + \lambda_1 G(0) = \lambda_2(1-\lambda_1) + \lambda_1(1+\lambda_2) = \Lambda$. At $p_1 = 1$, we have $E(1) = 1-(\alpha_1-1)\lambda_1$, $G(1) = 1+(\alpha_1-1)\lambda_2$, so $\lambda_2 E(1) + \lambda_1 G(1) = \Lambda$. At $p_1 = 1/2$, it can be verified by direct substitution. ∎

Proof of Claim EC.5. Write $c := C_{v,1}/C_{w,1} > 0$ and $a := \alpha_1 - 1 > 0$. From (EC.35), the zeros of $\mathcal{C}'(p_1, 1)$ on $[0, \overline{p}_1]$ are the zeros of

$$h(p_1) := a\, c\, E(p_1)^2 - \phi(p_1)\, D(p_1)^2.$$

Observe that $\phi(p_1) \leq 0$ implies $h > 0$ (since $a\, c\, E^2 > 0$), so h has no zeros in the region $\{\phi \leq 0\}$. It remains to bound the zeros in the open set $\{\phi > 0\}$, which is an interval because ϕ is linear. On this interval, factor

$$h = \big(\sqrt{ac}\, E + \sqrt{\phi}\, D\big)\big(\sqrt{ac}\, E - \sqrt{\phi}\, D\big). \tag{EC.40}$$

The first factor is strictly positive ($E, D, \phi > 0$). Hence, $h = 0$ if and only if $\sqrt{ac}\, E = \sqrt{\phi}\, D$, equivalently

$$\underbrace{\frac{E(p_1)}{D(p_1)}}_{=:\mathcal{F}_1(p_1)} = \underbrace{\sqrt{\frac{\phi(p_1)}{ac}}}_{=:\mathcal{F}_2(p_1)}. \tag{EC.41}$$

We show that $\mathcal{F}_1$ is strictly convex and $\mathcal{F}_2$ is strictly concave on the interval $\{\phi > 0\} \cap [0, \overline{p}_1]$, so that $\mathcal{F}_1 - \mathcal{F}_2$ is strictly convex and has at most two zeros.

$\mathcal{F}_1$ is strictly convex. Using $E' = \lambda_1 \phi$ and $D' = -a\lambda_1$:

$$\mathcal{F}_1' = \frac{E'D - ED'}{D^2} = \frac{\lambda_1(\phi\, D + a\, E)}{D^2} > 0.$$

Differentiating again (noting $(\phi D + aE)' = \phi' D + \phi D' + aE' = 2a\lambda_2 D - a\lambda_1\phi + a\lambda_1\phi = 2a\lambda_2 D$ and $(D^2)' = -2a\lambda_1 D$):

$$\mathcal{F}_1'' = \frac{2a\lambda_1\big[\lambda_2 D^2 + \lambda_1(\phi\, D + a\, E)\big]}{D^3} > 0, \tag{EC.42}$$

since every factor in the numerator is positive on $\{\phi > 0\} \cap [0, \overline{p}_1]$.

$\mathcal{F}_2$ is strictly concave. $\mathcal{F}_2(p_1)=\sqrt{\phi(p_1)/(ac)}$ is the composition of $t\mapsto\sqrt{t/(ac)}$ (strictly concave for $t>0$) with the linear function $p_1\mapsto\phi(p_1)$, hence strictly concave on $\{\phi>0\}$.

In conclusion, $\mathcal{F}_1-\mathcal{F}_2$ is strictly convex on $\{\phi>0\}\cap[0,\overline{p}_1]$. A strictly convex function on an interval has at most two zeros. Combined with the absence of zeros where $\phi\le 0$, h has at most two zeros on $[0,\overline{p}_1]$. ∎

Proof of Claim EC.6. Write $a:=\alpha_1-1$.

Case (I): If $\overline{p}_1=1$ (equivalently, $a\le(1-\Lambda)/\lambda_1$), then $\phi(1)=(2-\alpha_1)+(\alpha_1-1)\lambda_2=1-a(1-\lambda_2)$, so $d\phi(1)/da=-(1-\lambda_2)<0$.

Case (II): If $\overline{p}_1<1$ (equivalently, $a>(1-\Lambda)/\lambda_1$), then, substituting $\overline{p}_1=(1-\Lambda)/(a\lambda_1)$ gives

$$\phi(\overline{p}_1)=(1-a)+a\lambda_2\left(\frac{2(1-\Lambda)}{a\lambda_1}-1\right)=1-a(1+\lambda_2)+\frac{2\lambda_2(1-\Lambda)}{\lambda_1},$$

which is affine in a with slope $-(1+\lambda_2)<0$. At $a=(1-\Lambda)/\lambda_1$, both expressions give $\phi=1-(1-\Lambda)(1-\lambda_2)/\lambda_1$, so $\phi(\overline{p}_1)$ is continuous. As $a\to 0^+$, $\phi(\overline{p}_1)\to 1>0$; as $a\to\infty$, $\phi(\overline{p}_1)\to-\infty$. By the intermediate value theorem, the unique root $\overline{\alpha}_1$ exists. ∎

Proof of Claim EC.7.

(a) From (EC.36), $y_u'>0$ if and only if $\phi(0)>0$, i.e., $\alpha_1<(2+\lambda_2)/(1+\lambda_2)=\alpha_1^\ddagger$.

(b) For $\alpha_1\in(1,\alpha_1^\ddagger)$, ϕ is increasing (because $\phi'=2(\alpha_1-1)\lambda_2>0$), so $\phi(\overline{p}_1)\ge\phi(0)>0$. Since $D(\overline{p}_1)>0$ and $E(\overline{p}_1)>0$, we have $y_l'>0$.

(c) Write $a:=\alpha_1-1$ and

$$R(a):=\frac{y_u'}{y_l'}=\frac{\phi(0)\,E(\overline{p}_1)^2}{\phi(\overline{p}_1)\,D(\overline{p}_1)^2\,(1-\lambda_1)^2}.$$

It suffices to show that R is strictly decreasing in a on $(0,\alpha_1^\ddagger-1)$.

Case (I): If $\overline{p}_1=1$ (equivalently, $a\le(1-\Lambda)/\lambda_1$), then $D(1)=E(1)$, so $R=\phi(0)/[\phi(1)(1-\lambda_1)^2]$. Writing $\phi(0)=1-a(1+\lambda_2)$ and $\phi(1)=1-a(1-\lambda_2)$,

$$\frac{d}{da}\frac{\phi(0)}{\phi(1)}=\frac{-2\lambda_2}{\phi(1)^2}<0,$$

so R is strictly decreasing.

Case (II): If $\overline{p}_1<1$ (equivalently, $a>(1-\Lambda)/\lambda_1$), then $\overline{p}_1=(1-\Lambda)/(a\lambda_1)$, $D(\overline{p}_1)=\Lambda$, and $F(\overline{p}_1)=1+(1-\Lambda)\lambda_2/\lambda_1$. In particular, $D(\overline{p}_1)$ and $F(\overline{p}_1)$ are constant in a. Since $dE/da=-(1-\Lambda)F(\overline{p}_1)/a^2<0$,

$$\frac{d\log R}{da}=-(1+\lambda_2)\left(\frac{1}{\phi(0)}-\frac{1}{\phi(\overline{p}_1)}\right)+\frac{2}{E(\overline{p}_1)}\frac{dE}{da}<0,$$

where the first term is negative because $\phi(\overline{p}_1)>\phi(0)>0$, and the second term is also negative.

At $a=(1-\Lambda)/\lambda_1$, both expressions for R coincide since $D(1)=E(1)=\Lambda$; hence R is continuous and strictly decreasing on all of $(0,\alpha_1^\ddagger-1)$. As $a\to 0^+$, we have $\overline{p}_1\to 1$, $D(\overline{p}_1)/E(\overline{p}_1)\to 1$, and

$\phi(0)/\phi(\overline{p}_1) \to 1$, so $R \to 1/(1-\lambda_1)^2 > 1$. As $a \to \alpha_1^{\ddagger} - 1$, we have $\phi(0) \to 0$ while $\phi(\overline{p}_1) > 0$ and $E(\overline{p}_1) > 0$, so $R \to 0$. By the intermediate value theorem and strict monotonicity, the unique $\alpha_1^{\dagger}$ exists.

∎

Proof of Claim EC.8. Since $\alpha_1 < \alpha_1^{\ddagger}$, $\phi(0) > 0$. In addition, since $\phi' > 0$, $\phi > 0$ on all of $[0, \overline{p}_1]$. Set $c := \min(y'_u, y'_l) > 0$ (positive by Claim EC.7(a)(b)) and define $g := \mathcal{F}_1 - \mathcal{F}_2$ as in (EC.41) with this c. By (EC.42) and the concavity of $\mathcal{F}_2$, g is strictly convex on $[0, \overline{p}_1]$.

At $p_1 = 0$: since $c \le y'_u = \phi(0)/\big[a(1-\lambda_1)^2\big]$, we have $\phi(0)/(ac) \ge (1-\lambda_1)^2$, so $\mathcal{F}_2(0) = \sqrt{\phi(0)/(ac)} \ge 1 - \lambda_1 = \mathcal{F}_1(0)$ and $g(0) \le 0$.

At $p_1 = \overline{p}_1$: since $c \le y'_l = \phi(\overline{p}_1) D(\overline{p}_1)^2 / \big[aE(\overline{p}_1)^2\big]$, we have $\phi(\overline{p}_1)/(ac) \ge E(\overline{p}_1)^2 / D(\overline{p}_1)^2$, so $\mathcal{F}_2(\overline{p}_1) \ge E(\overline{p}_1)/D(\overline{p}_1) = \mathcal{F}_1(\overline{p}_1)$ and $g(\overline{p}_1) \le 0$.

By strict convexity, for any $p_1 \in [0, \overline{p}_1]$ with $p_1 = t\,\overline{p}_1$:

$$g(p_1) \le (1-t)\, g(0) + t\, g(\overline{p}_1) \le 0.$$

Hence, $\mathcal{F}_1 \le \mathcal{F}_2$ on $[0, \overline{p}_1]$, equivalently $h(p_1) = ac\,E^2 - \phi\,D^2 \le 0$, so $c \le \phi\,D^2/(a\,E^2) = T(p_1)$. ∎

EC.4.6. Proof of Theorem 5

Throughout, we write $\Lambda := \lambda_1 + \lambda_2$, $r := (1-p_1)\hat{\lambda}_1 + (1-p_2)\hat{\lambda}_2$, $D(p_1) := 1 - (\alpha_1 - 1)p_1\lambda_1$, $E(p_1) := (1-\lambda_1) + p_1\lambda_1\big[(2-\alpha_1) - (\alpha_1-1)\lambda_2\big] + (\alpha_1-1)p_1^2\lambda_1\lambda_2$, and $F(p_1) := 1 + (\alpha_1-1)p_1\lambda_2$. From Lemma 2 and Proposition 9, when the population plays $(p_1, p_2{=}1)$, we have $\hat{\lambda} = \Lambda/D$, $\hat{\lambda}_1 = \lambda_1 F/D$, $\hat{\lambda}_2 = \lambda_2$, $\sigma = \hat{\lambda}$, $r = (1-p_1)\hat{\lambda}_1$, $1 - r = E/D$, and $W_1 = 1 + (1-p_1)\Lambda/E$.

A focal type-k patient choosing p'_k while all others play (p_1, p_2) incurs cost $\mathcal{C}_k(p'_k; p_1, p_2) = \hat{\lambda}_k(p'_k; p_1, p_2)\big(C_{v,k} + C_{w,k} W_k(p'_k; p_1, p_2)\big)$. Since $\alpha_2 = 1$, $\hat{\lambda}_2 = \lambda_2$ regardless of p'_2, so

$$\frac{\partial \mathcal{C}_2(p'_2; p_1, p_2)}{\partial p'_2} = -\lambda_2 C_{w,2} \frac{\hat{\lambda}}{1-r} < 0,$$

implying that every equilibrium has $\tilde{p}_2 = 1$. With $p_2 = 1$ fixed, the focal type-1 cost is given by

$$\mathcal{C}_1(p'_1; p_1, 1) = \lambda_1 \left(1 + (\alpha_1 - 1)p'_1 \frac{\Lambda}{D}\right) \left(C_{v,1} + C_{w,1}\left[1 + \frac{(1-p'_1)\Lambda}{E}\right]\right), \tag{EC.43}$$

which is quadratic in p'_1 with $(p'_1)^2$ coefficient $-\lambda_1(\alpha_1 - 1)C_{w,1}\Lambda^2/(DE) < 0$, hence strictly concave. Every best response is therefore bang-bang, namely, $\arg\min_{p'_1 \in [0, \overline{p}_1]} \mathcal{C}_1(p'_1; p_1, 1) \subseteq \{0, \overline{p}_1\}$.

(i): If $\alpha_1 = 1$, then $(\alpha_1 - 1) = 0$ and $\hat{\lambda}_1 = \lambda_1$ regardless of p'_1, so $\mathcal{C}_1$ is given by

$$\mathcal{C}_1(p'_1; p_1, 1) = \lambda_1 \left(C_{v,1} + C_{w,1}\left[1 + \frac{(1-p'_1)\Lambda}{1-\lambda_1}\right]\right),$$

which is strictly decreasing in p'_1. Hence $\tilde{p}_1 = 1$.

Next, we study pure-strategy equilibria first and then followed by mixed-strategy equilibria.

Pure-strategy equilibria for type 1. By the concavity above, only 0 and $\overline{p}_1$ can be symmetric pure equilibria. We examine each candidate in turn.

First, suppose all type-1 patients adopt $\tilde{p}_1 = \overline{p}_1$ (maximal stable pooling). This is a Nash equilibrium if and only if $\mathcal{C}_1(\overline{p}_1;\overline{p}_1,1) \le \mathcal{C}_1(0;\overline{p}_1,1)$. Write $\bar{D} = D(\overline{p}_1)$, $\bar{E} = E(\overline{p}_1)$, $\bar{F} = F(\overline{p}_1)$. From (EC.43):

$$\mathcal{C}_1(0;\overline{p}_1,1) = \lambda_1\left(C_{v,1} + C_{w,1}\frac{\bar{E}+\Lambda}{\bar{E}}\right),$$
$$\mathcal{C}_1(\overline{p}_1;\overline{p}_1,1) = \frac{\lambda_1\bar{F}}{\bar{D}}\left(C_{v,1} + C_{w,1}\frac{\bar{E}+(1-\overline{p}_1)\Lambda}{\bar{E}}\right).$$

Subtracting and using $\bar{F}-\bar{D} = (\alpha_1-1)\overline{p}_1\Lambda$ and $\bar{E}+(1-\overline{p}_1)\Lambda = \bar{D}\bar{G}/\bar{F}$, together with $\lambda_2 E + \lambda_1 G = \Lambda$ (Claim EC.4):

$$\mathcal{C}_1(\overline{p}_1;\overline{p}_1,1) - \mathcal{C}_1(0;\overline{p}_1,1) = \frac{\lambda_1\overline{p}_1\Lambda(\alpha_1-1)C_{w,1}}{\bar{D}}\left(\frac{C_{v,1}}{C_{w,1}} - z_l'\right), \tag{EC.44}$$

where

$$z_l' := \frac{\big[(2-\alpha_1)+(\alpha_1-1)\lambda_2(\overline{p}_1-1)\big]\bar{D}}{(\alpha_1-1)\bar{E}}. \tag{EC.45}$$

Hence, $\tilde{p}_1 = \overline{p}_1$ is Nash if and only if $C_{v,1}/C_{w,1} \le z_l'$.

Next, suppose all type-1 patients adopt $\tilde{p}_1 = 0$ (full continuity). This is a Nash equilibrium if and only if $\mathcal{C}_1(0;0,1) \le \mathcal{C}_1(\overline{p}_1;0,1)$. At population strategy profile $(0,1)$, we have $D(0)=1$, $E(0)=1-\lambda_1$, $\hat{\lambda}=\Lambda$, $r=\lambda_1$. From (EC.43):

$$\mathcal{C}_1(0;0,1) = \lambda_1\left(C_{v,1} + C_{w,1}\frac{1+\lambda_2}{1-\lambda_1}\right),$$
$$\mathcal{C}_1(\overline{p}_1;0,1) = \lambda_1\big(1+(\alpha_1-1)\overline{p}_1\Lambda\big)\left(C_{v,1} + C_{w,1}\frac{1+\lambda_2-\overline{p}_1\Lambda}{1-\lambda_1}\right).$$

Subtracting and factoring imply

$$\mathcal{C}_1(0;0,1) - \mathcal{C}_1(\overline{p}_1;0,1) = -\lambda_1\overline{p}_1\Lambda(\alpha_1-1)C_{w,1}\left(\frac{C_{v,1}}{C_{w,1}} - z_u'\right), \tag{EC.46}$$

where

$$z_u' := \frac{(2-\alpha_1)+(\alpha_1-1)(\overline{p}_1\Lambda-\lambda_2)}{(\alpha_1-1)(1-\lambda_1)}. \tag{EC.47}$$

Hence, $\tilde{p}_1 = 0$ is Nash if and only if $C_{v,1}/C_{w,1} \ge z_u'$.

We further establish the following ordering result, whose proof is delayed until the end.

Claim EC.9. *If $z_u' > 0$, then $z_l' < z_u'$.*

Mixed-strategy equilibrium and regime classification. For $C_{v,1}/C_{w,1} \in (z_l', z_u')$, neither pure strategy is Nash. Consider a mixed strategy where each type-1 patient plays $\overline{p}_1$ with probability $\tilde{q}_1$ and 0 otherwise, while all type-2 patients play 1. In the mean-field limit, a fraction $\tilde{q}_1$ of type-1 patients permanently adopts the "always-switch" strategy $\overline{p}_1$ and the remainder permanently adopts 0.

Type-1 stayers never disrupt continuity, so their arrival rate remains λ_1. Type-1 switchers disrupt continuity on fraction $\overline{p}_1\sigma$ of visits, where $\sigma = \hat{\lambda}/\mu$ is the busy probability, so each switcher's effective rate is $\lambda_1(1+(\alpha_1-1)\overline{p}_1\hat{\lambda})$ (with $\mu = 1$). Type-2 patients always switch with $\alpha_2 = 1$, so $\hat{\lambda}_2 = \lambda_2$. The aggregate effective arrival rate is

$$\hat{\lambda}(\tilde{q}_1) = \frac{\Lambda}{1-(\alpha_1-1)\overline{p}_1\lambda_1\tilde{q}_1} = \frac{\Lambda}{D(\tilde{q}_1\overline{p}_1)}.$$

Only patients who stay with a busy GP contribute to queues of length two or more. The dedicated traffic rate is

$$r(\tilde{q}_1) = (1-\tilde{q}_1)\lambda_1 + \tilde{q}_1(1-\overline{p}_1)\lambda_1\big(1+(\alpha_1-1)\overline{p}_1\hat{\lambda}(\tilde{q}_1)\big),$$

and $S(\tilde{q}_1) = \hat{\lambda}(\tilde{q}_1)/(1-r(\tilde{q}_1))$. A focal type-1 stayer costs $\lambda_1(c_1+1+S)$ and a focal type-1 switcher costs $\lambda_1(1+(\alpha_1-1)\overline{p}_1\hat{\lambda})(c_1+1+(1-\overline{p}_1)S)$, where $c_1 := C_{v,1}/C_{w,1}$. Setting these equal and eliminating $\tilde{q}_1$ via $\tilde{q}_1 = (\hat{\lambda}-\Lambda)/((\alpha_1-1)\overline{p}_1\lambda_1\hat{\lambda})$ yields the quadratic equation (31) in $\hat{\lambda}$.

Define $H_1(\hat{\lambda})$ to be the left-hand side of (31). Then H_1 is convex, and

$$H_1(\Lambda) = (\alpha_1-1)\Lambda(1-\lambda_1)(z_u'-c_1), \qquad H_1\left(\frac{\Lambda}{D(\overline{p}_1)}\right) = \frac{(\alpha_1-1)\Lambda\,E(\overline{p}_1)}{D(\overline{p}_1)^2}(z_l'-c_1).$$

For $c_1 \in (z_l', z_u')$, we have $H_1(\Lambda) > 0$ and $H_1(\Lambda/D(\overline{p}_1)) < 0$, so by convexity there is a unique root $\hat{\lambda}_1^\dagger$ in $(\Lambda, \Lambda/D(\overline{p}_1))$, yielding a unique $\tilde{q}_1 \in (0,1)$. This proves part (2b).

We now classify the equilibrium regimes. From (EC.45) and (EC.47), write $\psi := (2-\alpha_1) + (\alpha_1-1)\lambda_2(\overline{p}_1-1)$ and $\eta := (2-\alpha_1) + (\alpha_1-1)(\overline{p}_1\Lambda-\lambda_2)$ for the numerators of z_l' and z_u' respectively. Since $\bar{D} > 0$, $\bar{E} > 0$, $\alpha_1-1 > 0$, and $1-\lambda_1 > 0$, we have $\mathrm{sign}(z_l') = \mathrm{sign}(\psi)$ and $\mathrm{sign}(z_u') = \mathrm{sign}(\eta)$.

Write $\alpha_1^\circ := (1-\lambda_2)/\lambda_1$, so that $\overline{p}_1 = 1$ for $\alpha_1 \le \alpha_1^\circ$ and $\overline{p}_1 = (1-\Lambda)/[(\alpha_1-1)\lambda_1] < 1$ for $\alpha_1 > \alpha_1^\circ$. In the $\overline{p}_1$=1 regime, $d\psi/d\alpha_1 = -1$ and $d\eta/d\alpha_1 = -(1-\lambda_1)$. In the $\overline{p}_1$<1 regime, $d\psi/d\alpha_1 = d\eta/d\alpha_1 = -(1+\lambda_2)$. Both functions are continuous, piecewise linear with strictly negative slopes, and satisfy $\psi(1) = \eta(1) = 1 > 0$ with both tending to $-\infty$. Hence, each has a unique root.

Setting $\psi = 0$ and solving for α_1:

$$(2-\alpha_1) + (\alpha_1-1)\lambda_2(\overline{p}_1-1) = 0 \Leftrightarrow \alpha_1 = \frac{2+(1-\overline{p}_1)\lambda_2}{1+(1-\overline{p}_1)\lambda_2} =: \alpha_1^{\dagger\prime}.$$

Setting $\eta = 0$:

$$(2-\alpha_1) + (\alpha_1-1)(\overline{p}_1\Lambda-\lambda_2) = 0 \Leftrightarrow \alpha_1 = \frac{2-\overline{p}_1\lambda_1+(1-\overline{p}_1)\lambda_2}{1-\overline{p}_1\lambda_1+(1-\overline{p}_1)\lambda_2} =: \alpha_1^{\ddagger\prime}.$$

Moreover, $\eta(\alpha_1)-\psi(\alpha_1)=(\alpha_1-1)\overline{p}_1\lambda_1>0$ for all $\alpha_1>1$, so $\eta(\alpha_1^{\dagger\prime})>\psi(\alpha_1^{\dagger\prime})=0$. Since η is strictly decreasing, $\alpha_1^{\ddagger\prime}>\alpha_1^{\dagger\prime}$, confirming $\alpha_1^{\dagger\prime}\in(1,\alpha_1^{\ddagger\prime})$. When $\lambda_2=0$, $\alpha_1^{\dagger\prime}=2$ and $\alpha_1^{\ddagger\prime}=(2-\overline{p}_1\lambda_1)/(1-\overline{p}_1\lambda_1)$, recovering the single-type boundaries in Theorem 2.

Combining the above, the equilibrium for type 1 is as follows.

(ii): If $\alpha_1\in(1,\alpha_1^{\ddagger\prime})$, then $\eta>0$, hence $z_u'>0$. By Claim EC.9, $z_l'<z_u'$. Parts (a)–(c) follow from the pure- and mixed-strategy analyses. When $\psi\le 0$ (equivalently $\alpha_1\ge\alpha_1^{\dagger\prime}$), $z_l'\le 0$ and sub-case (a) is vacuous.

(iii): If $\alpha_1\ge\alpha_1^{\ddagger\prime}$, then $\eta\le 0$ (so $z_u'\le 0$), and $\psi<\eta\le 0$ (so $z_l'<0$). Since $z_u'\le 0$, the condition $C_{v,1}/C_{w,1}\ge z_u'$ is automatically satisfied, so $\tilde{p}_1=0$ is always Nash. Since $z_l'<0$, $\tilde{p}_1=\overline{p}_1$ is not Nash. Hence $\tilde{p}_1=0$ is the unique equilibrium. ∎

To complete the proof, we verify Claim EC.9 below.

Proof of Claim EC.9. From (EC.45) and (EC.47), define $\psi:=(2-\alpha_1)+(\alpha_1-1)\lambda_2(\overline{p}_1-1)$ and $\eta:=(2-\alpha_1)+(\alpha_1-1)(\overline{p}_1\Lambda-\lambda_2)$, so $z_l'=\psi\bar{D}/[(\alpha_1-1)\bar{E}]$ and $z_u'=\eta/[(\alpha_1-1)(1-\lambda_1)]$. Since $z_u'>0$ implies $\eta>0$ and $\eta-\psi=(\alpha_1-1)\overline{p}_1\lambda_1>0$, if $\psi\le 0$ then $z_l'\le 0<z_u'$ and the inequality is immediate. It remains to show $z_u'>z_l'$ when $\psi>0$. This is equivalent to $\eta\bar{D}/(\alpha_1-1)(1-\lambda_1)>\psi\bar{D}/(\alpha_1-1)\bar{E}$, i.e., $\eta(1-\bar{r})>\psi(1-\lambda_1)$, where $\bar{r}:=(1-\overline{p}_1)\lambda_1\bar{F}/\bar{D}$ and $1-\bar{r}=\bar{E}/\bar{D}$.

Since $\psi>0$ forces $\alpha_1<2$. We discuss the following two cases.

Case (I): If $\overline{p}_1=1$, then $\bar{r}=0$, and thus $\eta(1-\bar{r})-\psi(1-\lambda_1)=\eta-\psi(1-\lambda_1)=(2-\alpha_1)\lambda_1+(\alpha_1-1)\lambda_1=\lambda_1>0$, as desired.

Case (II): If $\overline{p}_1<1$, then $(\alpha_1-1)\overline{p}_1\lambda_1=1-\Lambda$, so $\eta=\psi+(1-\Lambda)$. We can rewrite $\eta(1-\bar{r})-\psi(1-\lambda_1)=\psi(\lambda_1-\bar{r})+(1-\Lambda)(1-\bar{r})$. It is clear that the second term $(1-\Lambda)(1-\bar{r})>0$. For the first term $\psi(\lambda_1-\bar{r})$, since $\alpha_1<2$, one can verify that $(1-\overline{p}_1)\bar{F}\le\Lambda$ (equivalently, $a\lambda_1\Lambda-(1-\overline{p}_1)\bar{F}\cdot a\lambda_1=(1-\Lambda)[1-(\alpha_1-1)\Lambda+(1-\Lambda)\lambda_2/\lambda_1]\cdot\lambda_1>0$ under $(\alpha_1-1)\Lambda<1$), implying that $\bar{r}\le\lambda_1$ and $\psi(\lambda_1-\bar{r})\ge 0$. Since both terms are positive, we conclude that $\eta(1-\bar{r})-\psi(1-\lambda_1)$, as desired.

∎

EC.5. Case Study and Sensitivity Analysis

We document the empirical basis for the visit cost C_v and waiting cost C_w calibrated in Section 8.1, and report a sensitivity analysis over the range of plausible values implied by the underlying data sources.

EC.5.1. Cost Parameter Calibration

EC.5.1.1. Visit Cost C_v

Part (I) NHS resource component.

In the main body, we adopt £42 as the unit cost of a GP surgery consultation (average duration 10 minutes), taken from Jones et al. (2024) (Table 9.4.2) in 2023/24 prices. This estimate includes direct care staff costs but excludes qualification costs. For the sensitivity analysis we consider both components separately:

- Baseline unit cost per 10-minute GP surgery consultation (excluding direct care staff and qualification costs): £38.
- Direct care staff costs per 10-minute consultation: £4. These reflect the cost of nurses (practice nurses, advanced-level nurses, and extended-role/specialist nurses) who routinely work alongside GPs.
- Qualification costs per 10-minute consultation: £7. These represent the equivalent annual cost of pre-registration and postgraduate medical education; see Jones et al. (2024) (Table 12.4.1) for detail.

Including or excluding each component yields an NHS resource cost ranging from £38 to £49 per consultation in 2023/24 prices.

Part (II) Patient-borne component.

From NIHR (2023) (Section 3), the private costs of a GP appointment (2016 prices) comprise:

- *Travel cost*: £8 on average, ranging from £0 (if walked) to £14.92 (if public transport).
- *Opportunity cost of time*: £7.28 on average, ranging from £6.37 to £8.26 across study arms. Approximately 25% of GP appointments are accompanied; for these, the carer's time is valued at £10.07 on average (range £7.36–£13.25). Combining patient and carer yields an average of £7.28 + 0.25 × £10.07 = £9.80, with a range from £6.37 + 0.25 × £7.36 = £8.21 to £8.26 + 0.25 × £13.25 = £11.57.

The total patient-borne cost in 2016 prices therefore lies between £8.21 (walk, control arm, unaccompanied) and £26.49 (public transport, intervention arm, accompanied). Uprating via the Consumer Prices Index including owner occupiers' housing costs (CPIH Index 128.6 for 2023/24, 101.0 for 2016) yields a range of £10.45–£33.73 in 2023/24 prices.

Combining the NHS resource component and the patient-borne component, the unit cost of a 10-minute GP surgery appointment ranges from $C_v^{\min} = £48.45$ to $C_v^{\max} = £82.73$, with a baseline of $C_v \approx £64.67$.

EC.5.1.2. Waiting Cost C_w We calibrate C_w from the stated-preference evidence of Cheraghi-Sohi et al. (2008) (Table 5), who estimate a discrete-choice model of GP consultation preferences. Their mean willingness to pay for a one-day reduction in waiting time is $7.22 (2008 prices). The study reports separate estimates across two questionnaire formats (generic and patient-centred) and

three clinical scenarios (ambiguous, minor, and urgent), giving six point estimates ranging from \$3.07 to \$18.48 per day. Converting to sterling at the 2008 exchange rate (£1 = \$2.033) and uprating to 2023/24 prices via CPIH (CPIH Index 128.6 for 2023/24, 86.2 for 2008) yields a range of £2.25 to £13.56 per day, with a baseline of £5.30 per day. With 10 working hours per day Jones et al. (2024) (Table 9.4.1), this corresponds to a baseline of $C_w \approx$ £0.53 per working hour, ranging from $C_w^{\min} =$ £0.225 to $C_w^{\max} =$ £1.356 per working hour.

EC.5.2. Simulation Methodology

The mean-field results of Sections 4–6 provide closed-form characterizations for $n \to \infty$. The NHS case study, however, involves practices with $n \in \{2, \ldots, 14\}$ GPs, where finite-size effects are non-negligible. We therefore evaluate all three policy regimes—across-the-board, type-dependent, and decentralized equilibrium—via discrete-event simulation of the finite-n queueing system described in Section 3, using log-normal inter-arrival and service times as specified in Section 8.1.

EC.5.2.1. Across-the-board and type-dependent policies For the across-the-board policy, we sweep p_a over a fine grid and, for each value, simulate the two-type JSQ system with $p_1 = p_2 = p_a$. Let $\hat{\lambda}_k^{(n)}$ denote the numerically computed effective arrival rate for type k (obtained by iterating the finite-n fixed-point equation for the occupancy distribution), and let $\bar{W}_k^{(n)}$ denote the mean time in system for type k averaged over multiple independent replications. The simulation-based cost is

$$\mathcal{C}^{(n)}(p_a) \;=\; \hat{\lambda}_1^{(n)}(C_v + C_w\,\bar{W}_1^{(n)}) + \hat{\lambda}_2^{(n)}(C_v + C_w\,\bar{W}_2^{(n)}),$$

and the across-the-board optimum is $\arg\min_{p_a} \mathcal{C}^{(n)}(p_a)$. The type-dependent policy is identified similarly.

EC.5.2.2. Equilibrium classification via deviation analysis The candidates for a type-1 pure-strategy equilibrium are $\tilde{p}_1 = 0$ and $\tilde{p}_1 = \overline{p}_1$. To test each candidate, we simulate the system under the candidate population strategy, collect the empirical steady-state occupancy distribution $s = (s_0, s_1, s_2, \ldots)$ (where s_k is the fraction of GPs with queue length $\geq k$), and compute the cost-ratio threshold at which a focal type-1 patient is exactly indifferent between staying and switching.

Derivation of the deviation threshold. Consider a focal type-1 patient arriving to a system in occupancy state $\mathbf{s}$, where the remaining players adopt $(p_1, p_2{=}1)$. Her usual GP is selected uniformly at random from the n GPs. Let $m := m(s) = \min\{k \geq 0 : s_k > s_{k+1}\}$ denote the minimum queue length in the system. By the argument of Claim EC.1 (in the proof of Proposition 3), the routing outcome depends only on whether the usual GP's queue length J exceeds m. If $J = m$, the usual GP is among the shortest queues and the patient stays; if $J > m$, all GPs at length m are among the $n-1$ non-usual GPs, so the patient switches to a queue of length m with probability p_1' and

stays with probability $1-p_1'$. The probability that the usual GP is not among the shortest queues is therefore $\sigma(s)=s_{m+1}$, the fraction of GPs with queue length strictly above the minimum.

If the focal patient stays, she always joins her usual GP (a uniformly random GP). Her expected time in system is $W_{\text{stay}}(s)=(1+\sum_{k\geq 1}s_k)/\mu$, because $\sum_{k\geq 1}s_k$ is the expected queue length at a uniformly random GP (tail-sum formula). If the focal patient switches, she joins a queue of length m whenever $J>m$ and stays at a queue of length m whenever $J=m$. In either case, the queue she joins has length exactly m. Her expected time in system is therefore $W_{\text{switch}}(s)=(1+m)/\mu$.

Setting the stayer's cost rate $\lambda_1(C_v+C_w W_{\text{stay}})$ equal to the switcher's cost rate $\lambda_1\big(1+(\alpha_1-1)\sigma\big)(C_v+C_w W_{\text{switch}})$ and solving for C_v/C_w yields the deviation threshold in state $\mathbf{s}$:

$$\text{rhs}(s)=\frac{W_{\text{stay}}(s)-W_{\text{switch}}(s)}{(\alpha_1-1)\sigma(s)}-W_{\text{switch}}(s)=\frac{\sum_{k\geq 1}s_k-m}{(\alpha_1-1)s_{m+1}}-\frac{1+m}{\mu}. \tag{EC.48}$$

In practice, we evaluate (EC.48) on the time-averaged empirical occupancy distribution collected from the simulation. Then $\tilde{p}_1=0$ is a Nash equilibrium if and only if $C_v/C_w\geq \text{rhs}(s|_{p_1=0})$, and $\tilde{p}_1=\overline{p}_1$ is a Nash equilibrium if and only if $C_v/C_w\leq \text{rhs}(s|_{p_1=\overline{p}_1})$. We denote the two resulting thresholds by

$$\text{rhs}_0:=\text{rhs}\big(s|_{(p_1=0,\,p_2=1)}\big),\qquad \text{rhs}_1:=\text{rhs}\big(s|_{(p_1=\overline{p}_1,\,p_2=1)}\big).$$

Convergence to the mean-field thresholds. As $n\to\infty$, the occupancy distribution converges to the mean-field fixed point (Proposition 8), that is, $s_k\to r^{k-1}\hat{\lambda}$ for a geometric parameter $r<1$. In the mean-field limit, the minimum queue length is $m=0$ with probability one (a positive fraction of GPs is idle under stability), so $\sigma\to s_1=\hat{\lambda}$, $W_{\text{switch}}\to 1/\mu$, and $W_{\text{stay}}\to(1+\hat{\lambda}/(1-r))/\mu$. Substituting into (EC.48) recovers the analytical thresholds:

$$\text{rhs}_0\to z_u',\qquad \text{rhs}_1\to z_l',$$

as defined in Theorem 5.

Dual equilibria in finite-n systems. A key difference between the finite-n and mean-field analyses is the ordering of these thresholds. In the mean-field limit, $z_l'<z_u'$ (Claim EC.9), so the two pure-strategy Nash regions do not overlap and the intermediate range (z_l',z_u') is filled by the population-split mixed equilibrium of Theorem 5. In finite-n systems, however, we can have $\text{rhs}_0<\text{rhs}_1$, creating an overlap; that is, for $C_v/C_w\in(\text{rhs}_0,\,\text{rhs}_1)$, *both* pure strategies are simultaneously Nash equilibria.

This reversal arises because finite-n effects dampen the incentive to deviate from either pure strategy. The dominant mechanism is that the shortest queue is often non-empty in a finite system (i.e., $m>0$ with positive probability), unlike the mean-field limit where $m=0$ always. A focal switcher therefore faces positive queueing delay even after switching ($W_{\text{switch}}=(1+m)/\mu>1/\mu$),

which reduces the switching benefit relative to the mean-field prediction. This widens the cost-ratio range that sustains $\tilde{p}_1 = 0$ as Nash (pushing rhs_0 below z'_u). A symmetric dampening operates on the all-switch side, widening the range that sustains $\tilde{p}_1 = \overline{p}_1$ (pushing rhs_1 above z'_l). When the two expanded Nash regions overlap, the result is a coordination game with two coexisting pure-strategy equilibria (an efficient one and an inefficient one) rather than the unique mixed equilibrium predicted by the mean-field model.

EC.5.3. Sensitivity Analysis

All policy outcomes depend on (C_v, C_w) only through their ratio $r = C_v/C_w$. With C_w expressed per working hour, the ratio $r = C_v/C_w$ ranges from $C_v^{\min}/C_w^{\max} = 35.7$ working hours per visit to $C_v^{\max}/C_w^{\min} = 367.7$ working hours per visit, a more than tenfold range around the baseline of $r = 122.0$ working hours per visit.

We evaluate the robustness of the three qualitative conclusions of Section 8.2 by sweeping over eight representative values of this ratio, equally spaced between the two endpoints:

$$r \in \{35.7,\ 83.1,\ 130.6,\ 178.0,\ 225.4,\ 272.8,\ 320.2,\ 367.7\} \text{ working hours per visit.}$$

Tables EC.1-EC.8 report the results for each of the eight representative clusters.

C_v/C_w (working hours per visit)	Type-dependent policy				Across-the-board policy				Decentralized equilibrium		
	$p_1^\star$	$p_2^\star$	Continuity (%)	Avg wait (days)	$p_a^\star$	Continuity (%)	Avg wait (days)	Optimality gap (%)	$\tilde{p}_1$	$\tilde{p}_2$	PoA
35.7	0	1	83.98	0.92	0.17	86.80	1.02	2.96	0	1	1
83.2	0	1	83.98	0.92	0.12	90.65	1.07	2.08	0	1	1
130.6	0	1	83.98	0.92	0.10	92.04	1.13	1.74	0	1	1
178.0	0	1	83.98	0.92	0.04	96.75	1.40	1.56	0	1	1
225.4	0	1	83.98	0.92	0.04	96.75	1.40	1.32	0	1	1
272.8	0	1	83.98	0.92	0.04	96.75	1.40	1.17	0	1	1
320.3	0	1	83.98	0.92	0.02	98.40	1.54	1.04	0	1	1
367.7	0	1	83.98	0.92	0.02	98.40	1.54	0.93	0	1	1

Table EC.1 Sensitivity results for Cluster 1.

C_v/C_w (working hours per visit)	Type-dependent policy				Across-the-board policy				Decentralized equilibrium		
	$p_1^\star$	$p_2^\star$	Continuity (%)	Avg wait (days)	$p_a^\star$	Continuity (%)	Avg wait (days)	Optimality gap (%)	$\tilde{p}_1$	$\tilde{p}_2$	PoA
35.7	0	1	89.55	1.09	0.16	88.49	1.11	2.79	0	1	1
83.2	0	1	89.55	1.09	0.10	92.66	1.17	1.89	0	1	1
130.6	0	1	89.55	1.09	0.08	94.12	1.23	1.58	0	1	1
178.0	0	1	89.55	1.09	0.02	98.49	1.51	1.38	0	1	1
225.4	0	1	89.55	1.09	0.02	98.49	1.51	1.14	0	1	1
272.8	0	1	89.55	1.09	0.02	98.49	1.51	0.98	0	1	1
320.3	0	1	89.55	1.09	0.02	98.49	1.51	0.87	0	1	1
367.7	0	1	89.55	1.09	0.00	100.00	1.69	0.78	0	1	1

Table EC.2 Sensitivity results for Cluster 2.

C_v/C_w (working hours per visit)	Type-dependent policy				Across-the-board policy				Decentralized equilibrium		
	$p_1^\star$	$p_2^\star$	Continuity (%)	Avg wait (days)	$p_a^\star$	Continuity (%)	Avg wait (days)	Optimality gap (%)	$\tilde{p}_1$	$\tilde{p}_2$	PoA
35.7	0	1	90.12	1.56	0.04	98.10	1.75	3.10	0	1	1
83.2	0	1	90.12	1.56	0.04	98.10	1.75	1.58	0	1	1
130.6	0	1	90.12	1.56	0.04	98.10	1.75	1.12	0	1	1
178.0	0	1	90.12	1.56	0.04	98.10	1.75	0.89	0	1	1
225.4	0	1	90.12	1.56	0.04	98.10	1.75	0.76	0	1	1
272.8	0	1	90.12	1.56	0.04	98.10	1.75	0.67	0	1	1
320.3	0	1	90.12	1.56	0.04	98.10	1.75	0.61	0	1	1
367.7	0	1	90.12	1.56	0.04	98.10	1.75	0.57	0	1	1

Table EC.3 Sensitivity results for Cluster 3.

C_v/C_w (working hours per visit)	Type-dependent policy				Across-the-board policy				Decentralized equilibrium		
	$p_1^\star$	$p_2^\star$	Continuity (%)	Avg wait (days)	$p_a^\star$	Continuity (%)	Avg wait (days)	Optimality gap (%)	$\tilde{p}_1$	$\tilde{p}_2$	PoA
5.7	0	1	85.54	0.79	0.20	83.39	0.79	2.85	0	1	1
83.2	0	1	85.54	0.79	0.07	94.16	1.01	2.12	0	1	1
130.6	0	1	85.54	0.79	0.07	94.16	1.01	1.67	0	1	1
178.0	0	1	85.54	0.79	0.07	94.16	1.01	1.45	0	1	1
225.4	0	1	85.54	0.79	0.04	96.54	1.17	1.25	0	1	1
272.8	0	1	85.54	0.79	0.04	96.54	1.17	1.12	0	1	1
320.3	0	1	85.54	0.79	0.04	96.54	1.17	1.03	0	1	1
367.7	0	1	85.54	0.79	0.00	100.00	1.58	0.93	0	1	1

Table EC.4 Sensitivity results for Cluster 4.

C_v/C_w (working hours per visit)	Type-dependent policy				Across-the-board policy				Decentralized equilibrium		
	$p_1^\star$	$p_2^\star$	Continuity (%)	Avg wait (days)	$p_a^\star$	Continuity (%)	Avg wait (days)	Optimality gap (%)	$\tilde{p}_1$	$\tilde{p}_2$	PoA
35.7	0	1	75.06	0.65	0.17	86.38	0.85	4.02	0	1	1
83.2	0	1	75.06	0.65	0.13	89.50	0.91	2.60	0	1	1
130.6	0	1	75.06	0.65	0.06	95.10	1.07	1.95	0	1	1
178.0	0	1	75.06	0.65	0.06	95.10	1.07	1.58	0	1	1
225.4	0	1	75.06	0.65	0.06	95.10	1.07	1.37	0	1	1
272.8	0	1	75.06	0.65	0.04	96.64	1.17	1.20	0	1	1
320.3	0	1	75.06	0.65	0.00	100.00	1.45	1.06	0	1	1
367.7	0	1	75.06	0.65	0.00	100.00	1.45	0.93	0	1	1

Table EC.5 Sensitivity results for Cluster 5.

C_v/C_w (working hours per visit)	Type-dependent policy				Across-the-board policy				Decentralized equilibrium		
	$p_1^\star$	$p_2^\star$	Continuity (%)	Avg wait (days)	$p_a^\star$	Continuity (%)	Avg wait (days)	Optimality gap (%)	$\tilde{p}_1$	$\tilde{p}_2$	PoA
35.7	0	1	85.66	0.75	0.19	85.60	0.78	2.44	0	1	1
83.2	0	1	85.66	0.75	0.12	90.78	0.87	1.95	0	1	1
130.6	0	1	85.66	0.75	0.05	95.99	1.07	1.59	0	1	1
178.0	0	1	85.66	0.75	0.05	95.99	1.07	1.32	0	1	1
225.4	0	1	85.66	0.75	0.00	100.00	1.32	1.07	0	1	1
272.8	0	1	85.66	0.75	0.00	100.00	1.32	0.89	0	1	1
320.3	0	1	85.66	0.75	0.00	100.00	1.32	0.76	0	1	1
367.7	0	1	85.66	0.75	0.00	100.00	1.32	0.66	0	1	1

Table EC.6 Sensitivity results for Cluster 6.

C_v/C_w (working hours per visit)	Type-dependent policy				Across-the-board policy				Decentralized equilibrium		
	$p_1^\star$	$p_2^\star$	Continuity (%)	Avg wait (days)	$p_a^\star$	Continuity (%)	Avg wait (days)	Optimality gap (%)	$\tilde{p}_1$	$\tilde{p}_2$	PoA
35.7	0	1	92.13	1.23	0.11	94.99	1.26	1.58	0	1	1
83.2	0	1	92.13	1.23	0.04	98.11	1.35	1.07	0	1	1
130.6	0	1	92.13	1.23	0.04	98.11	1.35	0.79	0	1	1
178.0	0	1	92.13	1.23	0.04	98.11	1.35	0.65	0	1	1
225.4	0	1	92.13	1.23	0.04	98.11	1.35	0.57	0	1	1
272.8	0	1	92.13	1.23	0.04	98.11	1.35	0.51	0	1	1
320.3	0	1	92.13	1.23	0.04	98.11	1.35	0.48	0	1	1
367.7	0	1	92.13	1.23	0.04	98.11	1.35	0.45	0	1	1

Table EC.7 Sensitivity results for Cluster 7.

C_v/C_w (working hours per visit)	Type-dependent policy				Across-the-board policy				Decentralized equilibrium		
	$p_1^\star$	$p_2^\star$	Continuity (%)	Avg wait (days)	$p_a^\star$	Continuity (%)	Avg wait (days)	Optimality gap (%)	$\tilde{p}_1$	$\tilde{p}_2$	PoA
35.7	0	1	88.14	1.12	0.18	91.93	1.15	1.49	0	1	1
83.2	0	1	88.14	1.12	0.10	95.42	1.21	1.10	0	1	1
130.6	0	1	88.14	1.12	0.02	99.08	1.34	0.87	0	1	1
178.0	0	1	88.14	1.12	0.01	99.56	1.37	0.66	0	1	1
225.4	0	1	88.14	1.12	0.01	99.56	1.37	0.54	0	1	1
272.8	0	1	88.14	1.12	0.01	99.56	1.37	0.46	0	1	1
320.3	0	1	88.14	1.12	0.01	99.56	1.37	0.40	0	1	1
367.7	0	1	88.14	1.12	0.01	99.56	1.37	0.35	0	1	1

Table EC.8 Sensitivity results for Cluster 8.

The sensitivity results confirm that all three policy conclusions in Section 8.2 are robust across the full range of empirically supported cost parameters ($C_v/C_w \in [35.7, 367.7]$ working hours per visit). Specifically: (i) the type-dependent optimum prescribes $p_1^\star = 0$ and $p_2^\star = 1$ for every cluster at every cost ratio in the grid; (ii) the across-the-board policy leaves value on the table, with the welfare gap widening as C_v/C_w decreases; and (iii) the decentralized equilibrium replicates the social optimum (PoA $= 1$) for every cluster at every cost ratio.

Although Tables EC.1–EC.8 evaluate eight discrete cost ratios and report zero welfare loss at every grid point, the equilibrium structure can be examined more finely. In particular, for cost ratios sufficiently below the empirical range, the system can admit a second, inefficient equilibrium, in which all type 1 patients also switch ($\tilde{p}_1 = 1$, $\tilde{p}_2 = 1$), that differs from the social optimum. Two conditions must hold simultaneously for this to occur: (i) the all-switch state must be feasible, meaning $\overline{p}_1 = 1$; and (ii) the individual benefit of switching must increase when more type 1 patients switch, creating a positive-feedback loop that sustains the all-switch state as a self-fulfilling equilibrium. When condition (i) fails, the system would become unstable if all type 1 patients switched, so the inefficient equilibrium cannot physically exist. When condition (ii) fails, congestion from aggregate switching overwhelms the pooling benefit, making the efficient equilibrium the unique outcome for every cost ratio.

Computing these thresholds for the eight representative clusters reveals that four practices—Clusters 1, 2, 3, and 4—admit only a unique equilibrium across the entire C_v/C_w spectrum. In all four, condition (i) fails: these practices operate at utilization levels high enough (96–97%) that universal type 1 switching would destabilize the system ($\overline{p}_1 < 1$). The remaining four clusters—Cluster 5 ($N = 9$), Cluster 6 ($N = 7$), Cluster 7 ($N = 2$), and Cluster 8 ($N = 2$), all operating at moderate utilization (92–94%)—do admit a dual-equilibrium range, but only at cost ratios far below the empirically relevant region (Table 4): $C_v/C_w \in (28.6, 52.0)$ for Cluster 5, (35.3, 61.7) for Cluster 6, (59.3, 87.6) for Cluster 7, and (58.3, 79.5) working hours per visit for Cluster 8. All four

ranges lie below the baseline calibration of 122 working hours per visit. Within these ranges, the inefficient all-switch equilibrium reduces continuity to 35–67% and raises the price of anarchy to 1.08–1.15. However, dual equilibria would require waiting costs substantially larger than even the most extreme empirical estimate, which is a scenario unsupported by the evidence documented above. Within the full range of empirically plausible cost parameters, the decentralized equilibrium is therefore unique and efficient for every representative practice, and the delegation result is immune to coordination-failure concerns.