Universal Healthcare with a Private Option: Investment and Congestion in Diagnostic Imaging*

Nicolò Generoso† Miho Hong‡

November 2, 2026

Please click here for the latest version

Abstract

We study patient welfare, capacity investment, and wait times in a universal healthcare system where patients can either wait for publicly funded care or pay for faster private treatment. Allowing private-pay care can stimulate capacity investment and relieve public queues, but market power can distort prices and the allocation of treatment. We build a model of patient choice, wait times, pricing, and capacity investment. Estimation combines administrative data on publicly funded care and scraped private prices and appointment availability. We find that patients have heterogeneous preferences for timely care. Moreover, pricing and capacity choices affect equilibrium wait times, but profit-maximizing providers do not internalize the resulting congestion externality. The socially efficient allocation would lower private prices and expand capacity. Disallowing private-pay care would reduce patient welfare through the loss of sorting on willingness to pay for timeliness, increased congestion in public queues, and reduced capacity investment. Attaining the status quo welfare under public-only provision would require a 35%-higher public budget to fund additional publicly provided capacity. Private-pay care benefits patients in nearly all urgency, age, and income groups, including those unlikely to use it themselves, who gain from reduced congestion in public queues.

*We are grateful to Steve Berry, Phil Haile, Charles Hodgson, Victoria Marone, and Katja Seim for advice and guidance. Many thanks to Rebekah Dix, Yunmi Kong, Pearl Li, Nano Ochoa, Alejandro Sabal, Pietro Tebaldi, and Jingyuan Wang for helpful discussions; to Michele Ercolanoni, Olivia Leoni, and Marco Villa for making this project possible; to Giulio Turola and Stefan Vrabie for their help and patience in handling the data; and to the Yale Center for Research Computing for assistance in computations run on the Bouchet cluster. All errors are our own. The conclusions drawn from the data are those of the authors and do not reflect the views of Regione Lombardia.

†Yale University, nicolo.generoso@yale.edu. *(Job Market Paper)*

‡Yale University, miho.hong@yale.edu.

1 Introduction

Public systems often rely on non-price mechanisms like waitlists or lotteries to allocate essential services such as healthcare, housing, and education. The allocation induced by such mechanisms is typically less efficient than what would be achieved by a market that rations access using prices. In addition, private provision of services might increase capacity investment if the private returns to investment are high. Relative to a public monopoly, a mixed system in which consumers can select between a free public option and a private option may generate efficiency gains by letting the allocation reflect consumers' willingness to pay and by expanding capacity. However, provider market power can introduce inefficient markups and misallocation of productive capacity relative to the socially optimal mixed provision scheme.

We study how the interaction of a public option and private markets shapes installed capacity, welfare, and timeliness of care in the market for diagnostic imaging in Italy. As in other mixed healthcare systems, a tax-funded scheme provides coverage to all citizens at modest co-pays and longer wait times, while private-pay care represents a more expensive but timelier option. Providers receive a regulated fee per publicly funded scan, sell private-pay care at prices they set, and divide their capacity between the two segments.

These features generate several competing forces. Universal coverage promotes access, but each patient who joins a congested public queue increases the wait time faced by others. Private care can improve welfare through three channels. First, it improves sorting by allocating faster care to patients with higher willingness to pay for timeliness. Second, it diverts some patients to uncongested private queues, where they impose smaller wait-time externalities than they would in congested public queues, thereby reducing congestion and the externalities borne by public users. Third, patient spending in the private market raises providers' returns to diagnostic equipment to support additional treatment capacity and expands the delivery of care beyond what could be financed by the public budget. These gains need not be fully realized because providers choose prices and capacity according to profits rather than social returns. Private provision therefore creates a central tradeoff: markups restrict access to private-pay care, limiting patient sorting and congestion relief, while the resulting profits sustain the capacity that allows providers to treat patients outside the public budget. We quantify the distortions created by market provision and the gains from private access through patient sorting, congestion relief, and the financing of capacity.

Our setting is well suited to study these tradeoffs. Publicly funded care is often congested and the standardized nature of diagnostic imaging lets us focus on the stark price-timeliness tradeoff faced by patients. For example, the average private-pay price of an MRI is about €295 for next-day service, whereas publicly funded access is available at a €36 co-pay with an average wait of 100 days for non-urgent patients. With available evidence suggesting that around 25% of diagnostic procedures are fully paid by patients (Fondazione Censis and Intesa

Sanpaolo RBM Salute, 2021), the private-pay market is sizable and substitution into it can be consequential for congestion in the public system. Finally, the two segments generate sharply different returns to capacity and therefore different incentives for its allocation. For instance, publicly funded MRI procedures are reimbursed at roughly €140 on average, but the same scans are priced at double that amount on the private-pay market.

Empirical research on mixed systems is typically hindered by data limitations: while administrative sources can provide rich information on publicly funded care, data on private-market activity are typically lacking. We address this gap by combining detailed patient-level administrative data on referrals and publicly funded procedures with web-scraped prices and appointment availabilities for a large sample of private providers. Referrals identify the pool of patients needing care, publicly funded procedure records show where treatment is ultimately delivered within the public system, and the scraped data describe the prices and timeliness of market alternatives.

We develop a model of patient choice, congestion, and provider behavior to evaluate how alternative regulations affect wait times, welfare, and installed capacity. Providers maximize profits by deciding (i) how many scanners to operate, (ii) how many slots to allocate to publicly funded and private-pay care, and (iii) what prices to charge for private-pay care. Over time, patients receive referrals and choose which provider to seek care from, and which channel to use – publicly funded or private-pay. They trade off distance, price, and wait time, with preferences that depend on demographics. Service follows a first-come, first-served queuing protocol, which pins down the evolution of wait times as a function of the available slots and the random processes of referral generation and patient choice.

We estimate the model in two steps. First, we use patients' choices, providers' pricing decisions, and the evolution of wait times implied by our queuing model to reveal patient preferences, the marginal costs of each scan, and the slots available for publicly funded and private-pay care at different providers. Second, we use providers' capacity and scanner choices to estimate the marginal costs of opening additional slots and bound scanner fixed costs.

The first step leverages three restrictions imposed by the model: (i) patients choose optimally; (ii) providers set prices to maximize profits; and (iii) wait times are generated by the queuing model. On the demand side, administrative records reveal patients' publicly funded choices but not private-pay ones. Motivated by the standardized nature of imaging, we assume that the publicly funded and private-pay options offered by the same provider share a common unobserved quality. We show that this allows observed publicly funded choices to reveal provider attractiveness, while variation in distance, wait times, and lapsed referrals across patients and demographics disciplines substitution into private-pay care and foregoing treatment. Since regulated co-pays provide little direct variation in price, we use private-pay pricing decisions to inform price sensitivity. Finally, the queuing model provides a map from observed activity – performed scans and wait times – to measures of the latent number of slots

devoted to private-pay and publicly funded care by each provider. Under persistent congestion of publicly funded care, our queuing model implies that almost no public slots will remain idle, so the capacity each provider devotes to public patients coincides with the number of scans they performed. Since private queues can have idle slots, we instead identify private capacity as the number of slots under which the patient choice model and the queuing protocol reproduce each provider's distribution of scraped first-availability private waits.

The second step leverages two optimality conditions: (i) the marginal cost of opening an additional slot on an existing scanner equals the marginal operating profit it generates; (ii) the fixed cost savings from operating one fewer scanner cannot exceed forgone operating profits net of slot costs, and vice versa for an additional scanner. Under a constant-returns capacity technology, we obtain marginal slot costs as a function of slots per scanner using condition (i). Given this estimated technology, we obtain provider-specific bounds on scanner fixed costs by evaluating profits at alternative scanner counts and imposing condition (ii).

Our estimates reveal heterogeneity in patient preferences, as more urgent, younger, and richer patients feature higher willingness to pay for faster care. On the supply side, providers charge substantial markups, with MRIs being the most lucrative scans. Across imaging modalities (i.e., X-ray, CT, or MRI), the median provider devotes 16–40% of their capacity to private-pay patients and 12–18% of such slots go unfilled. While this capacity allocation allows providers to keep their own private-pay wait times consistently low, each provider's price and capacity choices affect patient sorting and wait times at other providers. This results in large congestion spillovers that individual providers do not fully internalize. For example, a 1% reduction in private MRI prices across providers would shorten public patients' average wait by approximately 0.3 days, worth about €0.50 per patient. Across roughly 12,600 public MRI patients per week, these benefits amount to about €6,100 in weekly patient surplus. Moreover, much of the gains generated by a provider's price cut accrues to patients queuing at other providers.

Our counterfactuals assess (i) the value of allowing private-pay care relative to a system offering only publicly funded care, (ii) the inefficiencies arising from decentralized pricing and capacity choices, and (iii) the distributional impact of allowing private-pay care.

We first measure the welfare effects of private-pay care by removing private-pay options and having all providers exclusively serve publicly funded patients. This scenario would generate a large welfare loss, because patients could no longer be allocated shorter waits according to their willingness to pay, total capacity would fall, and congestion in public care would therefore increase. Providers would also reallocate their capacity towards profitable MRI scans, amplifying the welfare loss for X-ray and CT patients and partially mitigating it for MRI patients. At existing reimbursement rates, the status quo government budget finances a given number of public slots: holding the public budget and reimbursement rates fixed, a planner could recover roughly 60% of the welfare loss by reallocating publicly funded slots across providers

within each modality. Increasing the public budget to finance more slots and allocating them to maximize welfare would further mitigate congestion, but restoring the status quo welfare would require a budget increase of about 35%. The key insight is that, although the planner could direct capacity toward providers where it generates the greatest welfare gains, shorter waits could no longer be allocated selectively to patients who value timeliness most. Replacing the sorting mechanism allowed by private-pay care would require reducing public waits much more broadly, making the public-only alternative substantially more costly.

Despite improving welfare relative to a fully public system at the status quo budget, private provision also introduces inefficiencies from market power and decentralized pricing and capacity decisions. We quantify these distortions by comparing the status quo with a first-best allocation in which a planner maximizes welfare by jointly choosing private-pay prices and public and private capacity. The planner would substantially reduce private prices and expand private capacity. The main distortion is in pricing: high prices limit take-up of private-pay care even though existing scanners could accommodate capacity expansions at relatively low cost and without reducing public slots. A second-best planner with full control only over total capacity and its public-private split would recover only about 41% of the first-best welfare gains if providers continued to set private-pay prices.

Finally, we examine how the welfare effects of private-pay care are distributed across patients. Disallowing private-pay care would reduce patient surplus across nearly all demographic groups, with particularly large losses among higher-income patients. Correcting the inefficiencies of private provision would also generate widespread gains. Importantly, even patients unlikely to use private-pay care would be worse off without it. This result underscores how the social benefits of private provision extend beyond its direct users, through both the congestion it relieves in public care and its effects on providers' incentives to supply capacity.

Related literature This project is closely related to some recent papers on mixed healthcare systems. Markkula (2025) uses a model similar to ours to evaluate expansions of public capacity in Finnish dental care. We complement his work by studying distortions in mixed provision relative to a public-only system. Russo (2024) studies substitution between congested VA care and subsidized private alternatives to assess the efficiency and distributional consequences of rationing public care through wait times rather than prices. Her counterfactuals hold supply fixed, while we endogenize provider behavior. Burn (2023) studies the expansion of publicly funded care through private hospitals in England and finds that private-sector participation reduces congestion in public hospitals. We instead study private-pay care alongside publicly funded care, allowing providers to choose prices and capacity in both channels.

A theoretical literature in health economics studies wait-time rationing in mixed systems, including sorting between public and private care (Hoel and Sæther, 2003), optimal waits (Gravelle and Siciliani, 2008), and how private provision and congestion shape public waits

and redistribution (Iversen, 1997; Canta and Leroux, 2016). Related work studies shared capacity and provider incentives across public and private care (Biglaiser and Ma, 2007; Brekke and Sørgard, 2007; Cuff et al., 2012). Empirical work on mixed systems estimates aggregate demand and supply responses to waits, differentiated provider choice, public-private substitution, and socioeconomic differences in access (Gravelle et al., 2002; Martin et al., 2007; Fabbri and Monfardini, 2009; Siciliani and Verzulli, 2009; Moscelli et al., 2016; Riganti et al., 2017). We bring differentiated-provider demand methods (Berry et al., 1995; Berry and Haile, 2014, 2024) to this setting, allowing substitution across providers and access type to discipline equilibrium counterfactuals.

The paper also relates to work on redistribution through in-kind provision, where public programs induce sorting between subsidized and market alternatives (Blackorby and Donaldson, 1988; Besley and Coate, 1991; Gahvari and Mattos, 2007; Currie and Gahvari, 2008).[1] A recent theoretical literature builds on Nichols and Zeckhauser (1982) to study redistribution and ordeals from a mechanism-design perspective (Condorelli, 2013; Akbarpour et al., 2024; Yang et al., 2025; Kang and Dworczak, 2026; Kang and Watt, 2026). The underlying tradeoffs are similar, but rather than characterize an optimal mechanism, we quantify how they operate in equilibrium under existing institutional constraints. Related work studies waitlists when waiting affects sorting and allocation, both theoretically (Naor, 1969; Mendelson and Whang, 1990; Bloch and Cantala, 2017; Leshno, 2022; Che and Tercieux, 2026) and empirically (Agarwal et al., 2021; Waldinger, 2021; Agarwal et al., 2025; Huitfeldt et al., 2026).

Outside healthcare, Cook and Li (2025) study an analogous sorting margin between monetary cost and timeliness in highway toll lanes. Recent papers evaluate investment with spillovers in mobile telecommunications (Elliott et al., 2025), ports (Brancaccio et al., 2025), subsea internet cables (Jeon and Rysman, 2025), and transportation (Almagro et al., 2026). In addition to sharing with these papers a focus on investment externalities in networked systems, our counterfactuals evaluate idealized centralized-planning benchmarks in a spirit particularly close to Almagro et al. (2026).

Roadmap Section 2 provides details on the institutions and data. Section 3 discusses some descriptive facts. Section 4 describes the model. Sections 5 and 6 describe our estimation and results. Section 7 discusses counterfactuals. Appendices collect details; for the reader's convenience, they expose numerous links to return to related parts of the main text.

[1] Related empirical work studies equilibrium responses to public provision and subsidies in education (Epple and Romano, 1998; Dinerstein and Smith, 2021; Dinerstein et al., 2023) and pharmaceutical markets (Atal et al., 2024).

2 Setting and data

2.1 Institutional details

Italy's National Health Service provides universal coverage through a system administered largely by regional governments. These governments finance care delivered by both government-owned and accredited private providers. In the following, we discuss (i) how publicly funded care is financed, (ii) the agreements governing its delivery by providers, (iii) how patients access publicly funded and private-pay care, and (iv) how publicly funded appointments are allocated according to urgency.

Financing care Article 32 of the Italian Constitution states that "the Republic shall safeguard health as a fundamental right of the individual and a collective interest and shall ensure free medical care for the indigent." The National Health Service was established in its current form in 1978, and is financed through income taxation collected by the central government and regional taxes on productive activities. The central government allocates resources to regional governments to finance policy areas under their authority, with healthcare being the main such domain. While central bodies like the Ministry of Health oversee the system, regional governments have discretion in choosing how to organize and finance publicly funded care.

Delivering care Policy reforms in the 1990s allowed private providers to deliver care on behalf of the National Health Service to expand patient choice and available capacity. Healthcare delivery involves government-operated providers as well as private providers (e.g., hospitals, clinics, diagnostic centers). When referring to providers, we label them as "public" or "private" based on their ownership. Private providers may deliver publicly funded services only if they are *accredited*. Accreditation is a facility-level qualification and is granted upon demonstrated compliance with service standards. The vast majority of private facilities are accredited.

Contractual agreements All providers, both public and private, are paid a *reimbursement* for each completed publicly funded procedure. Reimbursements are procedure-specific, set by the regional government, and apply uniformly across providers.[2] In each calendar year, the regional government allocates its healthcare budget across providers delivering publicly funded care. Each such provider is assigned a *cap* on the total reimbursement it can collect from outpatient procedures performed during the year. Contractual agreements also require providers to deliver publicly funded care in a "continuous way," meaning that providers are

[2]The uniformity of this reimbursement scheme was justified by the policymakers' desire to create a quasi-market for publicly funded care where private and public providers would compete for publicly funded patients on equal footing.

expected to spread activity over the calendar year rather than exhaust the cap early.[3] Since the cap applies broadly to outpatient activity – including, for example, both specialist visits and diagnostics – providers retain discretion over the mix of publicly funded procedures they provide.[4] Caps are updated over time according to mechanical rules. For example, providing volumes of selected procedures above the given thresholds results in future caps being adjusted upward by a few percentage points.

Accessing care Patients can pursue publicly funded care or choose to pay privately. In addition to delivering publicly funded care, private providers can treat patients who are willing to pay out of pocket or are privately insured.[5,6] Publicly funded care involves cost sharing: the provider receives the procedure-specific reimbursement; the patient pays at most €36 of that amount, and the regional government pays the remainder. By contrast, private-pay access requires the patient to pay the full price posted by the provider. In practice, publicly funded outpatient care is associated with long waits, while appointments through market options are more readily available, albeit more costly.

Patients need a referral to be able to access publicly funded care. Referrals can only be issued by doctors working within the National Health Service, including both general practitioners and specialists.[7] A referral is not formally required to access private-pay care.

[3] Templates of such agreements are publicly available. The template used by the local health agency for the Milan metro area – one of many territorial entities that purchase care on behalf of the government – states that "services shall be delivered so as to ensure continuity of service throughout the relevant year, with activities planned and productive capacity adjusted in line with the allocated budget."

[4] One might wonder why the government does not contract directly on volumes of specific classes of procedures. The status quo arrangement can be rationalized as follows. On the government side, the caps ensure that the government's healthcare expenditure stays within its budget. On the provider side, the contract allows them to choose to supply the most profitable procedure mix. This allows the provider to avoid losses if reimbursements for some procedures are below their average cost. This flexibility can of course result in different procedure mixes being offered than what the government would like. However, if providers have heterogeneous and privately known costs, *some* flexibility is needed to make the providers willing to contract – in the status quo, regulators decided to impose uniform reimbursement and to load such flexibility on the procedure mix choice rather than, say, negotiate volumes and reimbursements separately with each provider.

[5] Public providers are allowed to raise additional revenues by effectively renting some of their space to moonlighting doctors. These doctors may work as public employees in the provider and see paying patients in the same provider after hours. There are limits to the extent of these arrangements: for instance, the volume of moonlighting visits cannot exceed that of publicly funded visits delivered by the same provider. Anecdotal evidence suggests that patients routinely resort to this channel for specialist visits. Since delivery of diagnostic services through these arrangements seems more limited, we neglect this channel in our analysis.

[6] Survey data collected by the National Institute of Statistics suggest that only about 6% of all diagnostic procedures completed in 2023 were covered by insurance, whereas 70% were publicly funded and 24% were paid entirely by patients. We will mostly neglect insured patients in our analysis.

[7] This means that a specialist seeing a patient in their private practice cannot issue a referral for publicly funded care. Specialist doctors can have "dual" careers: they can work as employees of a public provider while also running a private practice, e.g., on weekends. These doctors can issue a referral that allows access to publicly funded care only when they are seeing patients in the former capacity. See also footnote 5. When receiving a treatment recommendation upon a private-practice specialist visit, it is not uncommon for patients to ask their general practitioners to issue them a referral for the recommended treatment to be able to pursue

However, private diagnostic services are typically obtained following a medical prescription and most occur within care pathways that involve the National Health Service (Fondazione Censis and Intesa Sanpaolo RBM Salute, 2021).

Allocating publicly funded care Wait times for publicly funded outpatient care are arguably the most visible and politically salient issue affecting the public healthcare system. To ensure faster access for the most urgent patients and to improve wait times monitoring, a policy introduced in 2019 by the Ministry of Health mandates referring doctors to record an urgency tier on referrals for outpatient services, and all providers to deliver publicly funded care within tier-specific deadlines. For diagnostic imaging, there are four urgency tiers, which are associated with deadlines of 3, 10, 60, and 120 days. To meet these deadlines, providers reserve part of their appointment slots for higher-priority patients. Patients book appointments from the slots available for their urgency tier. They can book either through a centralized system or by contacting individual providers.

2.2 Data

We combine three sources of data: (i) administrative records of referrals and publicly funded procedures, (ii) scraped private-pay prices and appointment availability, and (iii) publicly available data on income, installed medical equipment, and local demographics.

Administrative referral and procedure data We obtained access to an administrative dataset maintained by the regional government that covers the universe of issued referrals and publicly funded performed procedures. The data include all referrals issued in 2023 and 2024, and all publicly funded care linked to those referrals. Each referral is associated with an issuance date and patient demographic information, including age, urgency tier, and town of residence. Each completed procedure records the procedure date and the provider that performed it. Coverage of publicly funded procedures is complete, while private-pay procedures are almost entirely unobserved. Referrals that are not associated with any subsequent procedure correspond to patients that decided to either use a private-pay option or forego care. We classify as "lapsed" all referrals in our sample which are not followed by any procedure as of February 2026.[8] Details on the data and on the procedure followed to export summary statistics are in Appendix A.

Scraped private prices We combine this administrative dataset with web-scraped data on private-pay prices and availabilities at a large sample of private providers. We selected all

care within the publicly funded scheme.

[8] Receiving care more than a year after the referral issuance is very uncommon and referrals themselves expire, so this classification seems reasonable.

providers for which an online booking system was available.[9] We submitted queries to book appointments for a wide range of diagnostic procedures on a weekly basis throughout 2024, and we recorded each query's posted price and first availability date.

Public use data We also make use of publicly available data. First, the Ministry of Finances provides data on the town-level distribution of income. Second, the Ministry of Health maintains a registry of installed large medical machines at the provider level, covering all CT scanners and MRIs and all providers. Third, we employ public-use datasets maintained by the National Statistics Institute to retrieve town-level demographic features such as the level of urbanization and local wages.

3 Descriptives

This section presents the data patterns that are informative of patients' preferences. In what follows, an imaging *modality* is the technology or type of equipment used to capture pictures of the body – in our case, X-ray, CT, or MRI. A *scan* is the specific procedure or individual test performed using that technology, such as an MRI of the brain or a CT scan of the chest.

Summary statistics Our data include all referrals issued in 2023 and 2024, linked with subsequent publicly funded scans. The analysis is restricted to the referrals issued for selected scans to individuals who are registered as residing in the region.[10] Table B1 presents summary statistics of the sample.[11] X-rays account for the majority (56%) of the 5.6 million referrals issued, followed by MRI (23%) and CT (22%). A significant share (39%) of 2024 referrals is not associated with any publicly funded scan as of the beginning of 2026, consistent with patients choosing private-pay options or foregoing care. The share of referrals leading to a publicly funded scan varies by modality (63% for X-ray, 67% for CT, and 50% for MRI). Delivery of publicly funded scans by private providers is quantitatively important and heterogeneous across modalities (53% for CTs versus 71% for MRIs).

Patient tradeoffs Patients face meaningful variation in the monetary, temporal, and travel costs of obtaining care. Table 1 documents the magnitudes of the entailed tradeoffs for lowest-

[9] While larger providers developed their own in-house booking systems, there are platforms (see, e.g., here) that allow smaller providers to post their availabilities online. As a result, our scraped data cover both large provider systems and smaller clinics.

[10] Individuals can temporarily live in the region for study or work reasons, while still appearing as residents of their out-of-region hometown. We focus on registered residents because knowledge of their location allows us to reconstruct accessibility to different providers. We also restrict attention to common scans and exclude some that are part of cancer screening campaigns (e.g., mammograms). Appendix A provides details.

[11] For the (digital) reader's convenience, appendix materials expose hyperlinks to make it convenient to return to the main text.

Table 1 Patient tradeoffs

	Mean	25th	50th	75th	Observation
Prices[a] (€)					
X-ray					
public	23.88	17.25	20.60	36.00	scan
private	71.29	50.03	63.42	81.11	provider
CT					
public	36.00	36.00	36.00	36.00	scan
private	214.08	172.12	188.72	224.62	provider
MRI					
public	36.00	36.00	36.00	36.00	scan
private	294.56	179.96	245.65	358.96	provider
Wait times[b] (days)					provider-week
X-ray					
public	61.33	34.87	52.08	78.54	
private	5.06	2.00	3.00	5.00	
CT					
public	82.67	57.75	81.67	106.16	
private	5.44	1.00	3.00	6.00	
MRI					
public	81.83	55.07	77.18	106.23	
private	7.12	2.00	3.00	9.00	
Distances[c] (minutes)					patient
1(has provider in town)	0.56				
to closest out-of-town provider	9.28	6.00	8.00	11.00	
Nearby providers[d] (count)					patient
X-ray	94.52	39.43	108.98	148.00	
CT	68.77	32.00	78.85	106.00	
MRI	65.47	29.07	73.00	100.00	

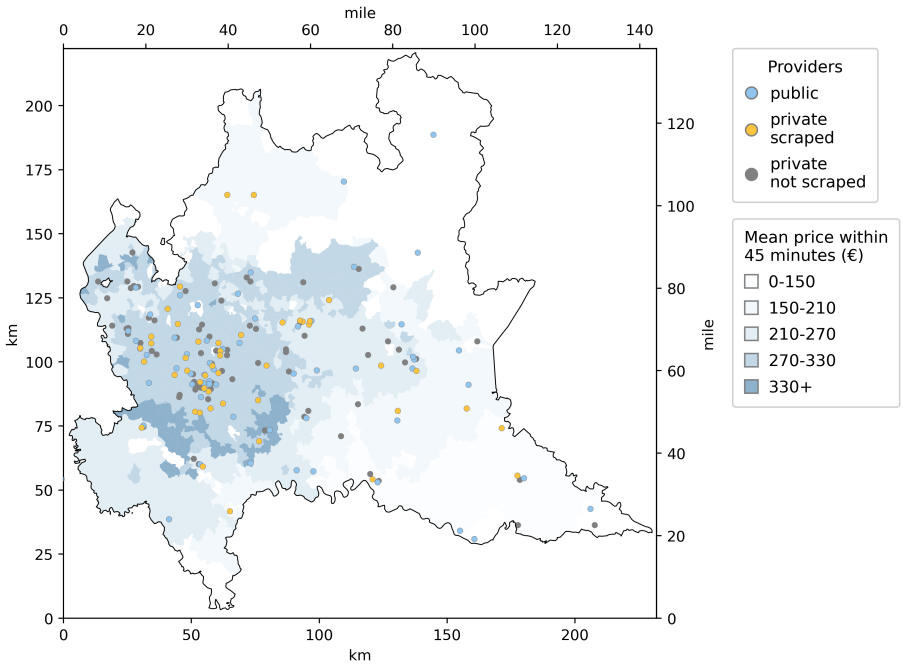
Notes: Table summarizes the distributions of the key variables factoring into patients' decisions.
[a]Public prices are the co-pays facing public patients. Private prices are reported as provider-specific means of private prices from our scraped data.
[b]Public provider-week mean wait times are computed as the time elapsing between referral and scan among patients who received a referral in that week and received treatment at that provider. Public wait times are reported for lowest-urgency patients only. See Table 2 for urgency-specific wait times. Private wait times are the days between the first available slot and the date when the scraping query was submitted in each week. Since our scraping routine submitted queries during the weekend, wait times of below three days effectively mean that an appointment was available on the first business day following the query.
[c]Distances are computed as the time required to travel from a town's centroid to the provider location.
[d]A provider is "nearby" if it is located within 45 minutes of a patient's hometown. A provider is classified as offering a given modality if it delivered at least 100 publicly funded scans in that modality in 2024.

Table 2 Mean wait times by urgency tier

	Urgency tier			
	1	2	3	4
X-ray	1.63	10.55	31.49	43.06
CT	5.04	15.11	44.53	92.66
MRI	7.97	20.14	56.91	92.30

Notes: Table reports the mean realized wait times by urgency and modality among publicly funded patients. Wait times are computed as the number of days between referral date and delivery date. Lower-numbered tiers correspond to higher urgency.

Figure 1 Private prices



Notes: Figure presents the mean price of spine MRIs (the most common MRI scan) offered by scraped providers located within 45 minutes of each town. Distances are computed as the time required to travel from each town's centroid to the provider location. Towns with no scraped provider within 45 minutes are left blank. Providers are colored based on ownership and scraped status.

urgency patients. Publicly funded and private-pay options differ sharply along the price-wait dimension. The former is inexpensive but slow, while the latter is costly but typically available within days. The tradeoff is particularly steep for MRI scans.[12] Prioritization is enforced among publicly funded patients, and Table 2 shows that the current system successfully allocates shorter waits to more urgent patients.[13]

The data reveal different outcomes of the patient choice problem with different degrees of detail. Observed choices of publicly funded patients are informative of preferences for wait times and distance because nearby providers differ across patients and public wait times vary across providers and weeks. For instance, the interquartile range spans 29–100 MRI providers within 45 minutes of a patient's hometown and 55–106 days for provider-week mean wait times among lowest-urgency MRI patients (cf. Table 1). On the other hand, most price variation is concentrated in private-pay care. Although we do not observe the provider chosen by private patients, the attractiveness of private-pay care varies across towns and scans because nearby private providers, posted prices, and private availability differ across locations. Scraped price data represent the most novel piece of information used in this project, and allow us to infer the monetary dimension of private-pay access to care. Variation in the share of lapsed referrals with prices, waits, and provider proximity is informative of substitution away from publicly funded care, although we cannot directly distinguish private-pay treatment from foregoing care. Since differentiated providers optimize against heterogeneous residual demands shaped by local demographics, posted prices themselves provide additional direct information about patient price sensitivity. The substantial spatial heterogeneity of prices documented by Figure 1 aids identification of patients' distaste for prices through these channels.

4 Model

At the beginning of the year, providers' qualities and costs are publicly observed. Then, private providers simultaneously and irreversibly choose (i) how many scanners to operate, (ii) the average number of slots to offer to public and private patients each week, and (iii) the prices for private-pay patients. Public providers are inert.

Within the year, the model proceeds in discrete time. Each week, patients randomly receive health shocks and seek treatment at the provider-access combination that maximizes their indirect utility. Weekly slot availability may fluctuate for exogenous reasons (e.g., holidays) around the capacity levels chosen by providers, but providers do not adjust capacity in

[12] Of course, prices and wait times are equilibrium outcomes. Each patient takes them as given when making their choices, and Table 1 summarizes their magnitudes as observed in the data.

[13] The magnitudes of waits also line up well with the intended target. Priority guidelines require treatment within 3, 10, 60, and 120 days for urgency tiers 1 through 4, respectively. Our wait times are defined as days elapsed between referral and treatment, and as such are likely to overestimate the wait times faced by a patient at the moment of booking.

response to realized demand or wait times.[14] Public providers can only be accessed via the public queue paying the regulated co-pay. Wait times emerge as a function of patient inflow and allocated capacity.

Notation *Providers* are indexed by $h \in \mathcal{H}$ (a mnemonic for "hospital"). *Access* can be publicly funded or private-pay and is indexed by $e \in \{\text{pub}, \text{pvt}\}$. An *alternative* available to patients is a provider-access combination $j = he \in \mathcal{H} \times \{\text{pub}, \text{pvt}\}$. The provider and access implied by alternative j are denoted $h(j)$ and $e(j)$. *Patients* are indexed by i and are associated with *scan* $s(i)$, *demographic group* $g(i)$, and *location* $\ell(i)$. Demographics include *urgency* $u(i)$ and *modality* $m(i) \in \{\text{XR}, \text{CT}, \text{MRI}\}$. A scan is a modality-body part combination, and we use $m(s)$ to denote the modality of scan s.

Each patient choosing an alternative is assigned to a queue within that provider. Queues are specific to access and modality, and publicly funded queues additionally differ by urgency. Since all scans within a modality use the same equipment, a patient's scan is irrelevant in determining the queue they join.[15] We index queues by $q \in \mathcal{Q}$ and write $q(i, j)$ for the queue joined by patient i after choosing alternative j.[16]

4.1 Patient choice

Upon receiving a health shock, a patient books an appointment at their preferred alternative. Prices are fixed across periods and depend on the chosen alternative and the patient's scan. Wait times depend on the queue a patient joins and evolve stochastically over time. At the moment of choice in week t, patient i is assumed to know all the relevant prices and wait times $(p_{js(i)}, w_{q(i,j)t})_{j \in \mathcal{J}}$. Reaching provider h from location ℓ takes $d_{\ell h}$ minutes.

A patient's choice set is defined as the set of alternatives located within 60 minutes of their town. Patient i with referral issued in week t chooses the alternative $j \in \mathcal{J}_i := \{j : d_{\ell(i)h(j)} \leqslant 60\}$ that maximizes their indirect utility

$$v_{ijt} = \delta_{jm(i)} - \beta^{\text{price}}_{g(i)} p_{js(i)} - \beta^{\text{wait}}_{g(i)} w_{q(i,j)t} - \beta^{\text{dist}}_{g(i)} d_{\ell(i)h(j)} + \varepsilon_{ijt}. \tag{1}$$

Foregoing care yields indirect utility normalized to $v_{i0t} = \varepsilon_{i0t}$. Idiosyncratic tastes are captured by preference parameters $\boldsymbol{\beta}_{g(i)}$, which depend on i's demographic group $g(i)$, and type-1 extreme value errors ε_{ijt} independently distributed across patients and alternatives. Preference

[14] Appendix B.2 presents suggestive evidence that capacity does not adjust in response to past wait times.

[15] When scraped on the same date, pairs of scans offered by a provider within a modality have identical waits in 67–83% of cases, depending on the modality. Waits differ by at most 3 days in 83–91% of cases. These results justify our common-queue approximation and are reported in Appendix B.3.

[16] For $j = h\,\text{pub}$, $q(i, j)$ is indexed by provider h, modality $m(i)$, and urgency $u(i)$. For $j = h\,\text{pvt}$, it is indexed by provider h and modality $m(i)$.

coefficients depend on demographics according to the following parameterization:

$$\beta^{\text{price}}_{g(i)} = \frac{\beta^{\text{price}}_{m(i)}}{\text{inc}(i)^{1-\theta}}, \quad \beta^{\text{wait}}_{g(i)} = \beta^{\text{wait}}_{m(i)u(i)} + \beta^{\text{wait}}_{m(i)\text{age}(i)}, \quad \beta^{\text{dist}}_{g(i)} = \beta^{\text{dist}}_{m(i)u(i)} + \beta^{\text{dist}}_{m(i)\text{age}(i)} + \beta^{\text{dist}}_{m(i)\text{urb}(i)}, \tag{2}$$

where $\text{urb}(i)$ is the urbanization level of i's town and $\text{inc}(i)$ is i's income.

In (1), the terms denoted as δ capture *quality* of service at a given alternative-modality. Since imaging is highly standardized, we assume that quality is shared by the private-pay and publicly funded options offered by the same provider. Public and private patients receiving the same modality at the same provider may therefore face different prices and wait times, but share the same underlying quality of care.

Assumption 1 (Provider quality)
Quality is provider-modality-specific: $h(j) = h(j')$ implies $\delta_{jm} = \delta_{j'm}$ for all m.

Denote the probability that patient i chooses alternative j by $\sigma_{ij}(\boldsymbol{p}, \boldsymbol{w}_t)$. Let $\boldsymbol{W} = (\boldsymbol{w}_t)_t$ denote a sequence of realized wait vectors. Aggregating these probabilities over patients and weeks and making dependence on idiosyncratic preference parameters $\boldsymbol{\beta}$ and quality levels $\boldsymbol{\delta}$ explicit, let $\sigma_{jm}(\boldsymbol{p}, \boldsymbol{W}; \boldsymbol{\beta}, \boldsymbol{\delta})$ denote the share of alternative j among patients needing modality m. As in Berry (1994) and Berry et al. (1995), we invert this demand system to recover provider qualities $\boldsymbol{\delta}$ from observed shares, conditional on preference parameters $\boldsymbol{\beta}$, prices, and wait times. Although private-pay shares are unobserved in our setting, the common-quality restriction ensures that publicly funded shares alone uniquely determine provider qualities whenever a solution exists.[17] The proof is in Appendix G.2.

Proposition 1 (Inversion)
Under Assumption 1, the map $\boldsymbol{\delta} \mapsto (\sigma_{jm}(\boldsymbol{p}, \boldsymbol{W}; \boldsymbol{\beta}, \boldsymbol{\delta}))_{j:e(j)=pub}$ is injective for all $(\boldsymbol{\beta}, \boldsymbol{p}, \boldsymbol{w})$.

The key intuition is that Assumption 1 restricts the vector of qualities to be of the same dimensionality as the vector of observed aggregate market shares, and aggregate choice probabilities are invertible over this restricted domain.[18] Hence, conditional on preference parameters $\boldsymbol{\beta}$, prices, and waits, the observed public shares uniquely pin down $\boldsymbol{\delta}$ whenever a solution exists, and the same quality vector determines the model-implied private shares.

[17] Of course, the restriction would not be needed if all choices were observed. Given the data limitation, Assumption 1 allows us to apply standard inversion methods and is well motivated by our setting.

[18] The result does not rely on the parameterization presented in this section, and can be generalized to a broader class of semi-parametric demand models arising from maximization of random utility where quality enters indirect utility linearly. Unlike the standard results of Berry (1994), Proposition 1 ensures uniqueness but not existence of a vector $\boldsymbol{\delta}$ that perfectly rationalizes observed market shares: since some products share the same quality terms, our model might be unable to generate some vectors of market shares. Appendix G.2 provides additional discussion.

4.2 Queues

For simplicity of exposition, we suppress within-year capacity fluctuations and let each queue q have constant weekly *capacity* k_q, defined as the number of appointment *slots* available each week.[19]

Let b_{qt} denote the *backlog* (i.e., the mass of patients waiting for treatment) at the beginning of week t and a_{qt} the mass of patients joining queue q during that week. Weekly *throughput* y_{qt} and the evolution of the backlog satisfy

$$y_{qt} := \min\{b_{qt} + a_{qt}, k_q\}, \qquad b_{q,t+1} = b_{qt} + a_{qt} - y_{qt}.$$

Under first-come, first-served treatment, the first-availability wait is determined by the backlog relative to available capacity: $w_{qt} := b_{qt}/k_q$.

Throughput at a given queue is either equal to the mass of waiting and arriving patients, or to the queue's capacity. When a queue is slack, some slots go unused ($y_{qt} < k_q$), while a sufficiently congested queue uses all available slots ($y_{qt} = k_q$).

Let $\mathcal{I}_t$ denote the random set of patients needing treatment in week t, whose realizations are independent across weeks and independent of prices and wait times. The probability that i joins queue q is $\sigma_{iq}(\boldsymbol{p}, \boldsymbol{w}_t) = \sum_j \mathbf{1}\{q(i,j) = q\}\sigma_{ij}(\boldsymbol{p}, \boldsymbol{w}_t)$. Aggregating individual choices, arrivals to queue q are $a_{qt} = \int_{\mathcal{I}_t} \sigma_{iq}(\boldsymbol{p}, \boldsymbol{w}_t)\,\mathrm{d}i$.

Together with the backlog transition above, patient choices make $(\boldsymbol{b}_t)_{t=1,2,\ldots}$ a Markov process: current backlogs determine waits, waits affect patient choices and arrivals, and arrivals determine subsequent backlogs. The resulting process admits a stationary distribution when patients strictly dislike waiting.[20]

Proposition 2 (Stationary backlogs)
If $\beta^{wait}_{g(i)} > 0$ almost surely, there exists a stationary distribution for $(\boldsymbol{b}_t)_{t=1,2,\ldots}$.

Wait times therefore also admit a stationary distribution. Under the stationary distribution, expected arrivals equal expected throughput, $\mathbb{E}[a_{qt}] = \mathbb{E}[y_{qt}]$. Moreover, queues that remain congested use all available slots, so their expected throughput equals capacity. These relationships connect the weekly queue dynamics to the capacity decisions made by providers.

[19] While maintaining that providers set the mean level of capacity, our estimation does incorporate the fact that the realized number of slots might vary within the year because of seasonality (see Section D.3). For completeness, Appendix C.1 generalizes the model to allow capacity to vary exogenously within the year.

[20] State-dependent arrivals generally rule out standard queuing formulas for stationary waits. Appendix C.4 discusses the queuing model and its relation to prior work, while Appendix G.1 establishes the result below.

4.3 Supply

Providers maximize profits under the stationary distribution of wait times by deciding (i) how many scanners to use, (ii) how many slots to dedicate to publicly funded and private-pay care,[21] and (iii) each scan's posted price for private-pay access. Private providers make such choices simultaneously under complete information; public providers are inert. The slots committed by provider h to access $e \in \{\text{pub}, \text{pvt}\}$ in modality m are denoted by k_{hem}. Each slot allows one patient to receive a scan in that modality.

Denote by $a_{js}(\boldsymbol{p}, \boldsymbol{k})$ the expected arrivals of patients needing scan s at alternative j as a function of prices and capacities, obtained by integrating over possible realizations of wait times and patient sets in the stationary environment.[22] Publicly funded patients result in regulated reimbursements $\boldsymbol{r} = (r_s)_{s \in \mathcal{S}}$, private-pay patients yield revenues equal to posted prices $\boldsymbol{p}_h = (p_{h\,\text{pvt}\,s})_{s \in \mathcal{S}}$, and scans have marginal costs $\boldsymbol{c}_h = (c_{hs})_{s \in \mathcal{S}}$. Provider h's profits are

$$\Pi_h(\boldsymbol{p}, \boldsymbol{k}) = \sum_{s \in \mathcal{S}} (p_{h\,\text{pvt}\,s} - c_{hs}) a_{h\,\text{pvt}\,s}(\boldsymbol{p}, \boldsymbol{k}) + (r_s - c_{hs}) a_{h\,\text{pub}\,s}(\boldsymbol{p}, \boldsymbol{k}).$$

Scan marginal costs are parameterized as

$$\ln c_{hs} = \gamma_s + \gamma_{m,\text{wage}} \ln \text{wage}_{\ell(h)} + \omega_{hs}, \tag{3}$$

where γ_s is a scan fixed effect, $\text{wage}_{\ell(h)}$ is the median wage in provider h's location, and ω_{hs} is a provider-scan cost shock.

Providers use scanners and divisible inputs, such as staff time, to make appointment slots available. These inputs are committed before patients arrive, so their costs are incurred even when some slots go unused. For provider h and modality m, the cost of offering $K_{hm} = k_{h\,\text{pub}\,m} + k_{h\,\text{pvt}\,m}$ total slots using n_{hm} scanners is $n_{hm}\phi_{hm}(K_{hm}/n_{hm})$, where ϕ_{hm} is differentiable and satisfies $\phi_{hm}(0) = 0$.[23] We parameterize slot marginal costs as

$$\ln \phi'_{hm}(v) = \boldsymbol{\zeta}_x \cdot \boldsymbol{x}_{hm} + \zeta_{\text{input}} \log v + \eta_{hm}, \tag{4}$$

[21] To focus on the more interesting public-versus-private margin, we assume that each provider splits each modality's publicly funded capacity across urgency tiers according to exogenously fixed weights. We elaborate in Appendix C.1.

[22] Since indices run over different (although related) sets, the formal definitions of different sets of arriving patients are notationally tedious. For the sake of exposition, we provide them in Appendix H.1 and rely on intuitive notation.

[23] We assume that the production technology mapping scanners and divisible inputs into slots separates across modalities and satisfies constant returns to scale. This captures the idea that adding another scanner with the same staffing and operating hours would double the number of appointment slots available. Appendix C.2 follows standard producer theory arguments to derive $\phi_{hm}(v)$ as the minimum divisible-input cost of offering v slots with one scanner.

where η_{hm} is a provider-modality slot cost shock and $\boldsymbol{x}_{hm}$ collects provider-modality characteristics. Additionally, provider h incurs fixed cost χ_{hm} for each of its modality-m scanners.

Finally, providers cannot collect reimbursement above provider-specific caps R_h. Denoting as $\bar{r}_m$ the mean reimbursement in modality m, we capture this constraint by requiring[24]

$$\sum_m k_{h\,\mathrm{pub}\,m}\,\bar{r}_m \leqslant R_h. \tag{5}$$

Denoting total slot costs by $\Phi_h(\boldsymbol{k}_{h\,\mathrm{pub}} + \boldsymbol{k}_{h\,\mathrm{pvt}}, \boldsymbol{n}) = \sum_m n_{hm}\phi_{hm}((k_{h\,\mathrm{pub}\,m} + k_{h\,\mathrm{pvt}\,m})/n_{hm})$, prices, capacity, and scanners $(\boldsymbol{p}^*, \boldsymbol{k}^*, \boldsymbol{n}^*)$ form an *equilibrium* if, for all h,

$$(\boldsymbol{p}_h^*, \boldsymbol{k}_h^*, \boldsymbol{n}_h^*) \in \arg\max_{\boldsymbol{p}_h, \boldsymbol{k}_h, \boldsymbol{n}_h} \Pi_h(\boldsymbol{p}_h, \boldsymbol{p}_{-h}^*, \boldsymbol{k}_h, \boldsymbol{k}_{-h}^*) - \Phi_h(\boldsymbol{k}_{h\,\mathrm{pub}} + \boldsymbol{k}_{h\,\mathrm{pvt}}, \boldsymbol{n}_h) - \boldsymbol{\chi}_h \cdot \boldsymbol{n}_h \quad \text{sub (5)}.$$

Appendix C.3 derives the corresponding optimality conditions, whose economic content is standard. First, private-pay prices balance the additional revenue collected from patients who continue to buy against the profit lost when higher prices reduce private demand, accounting for patients diverted toward publicly funded care and for the resulting changes in wait times. Second, total capacity is chosen so that the marginal cost of an extra slot equals the marginal operating profits it produces. Third, the allocation of capacity between public and private care similarly equalizes the marginal returns to a slot across the two uses, except that publicly funded capacity also uses scarce reimbursement budget when the cap binds. Fourth, optimal scanner choice requires that the gain from adding scanners be no larger than their fixed cost and that the loss from removing scanners be no smaller than the fixed cost saved.

5 Estimation

Estimation proceeds in three steps. We first recover capacity in persistently congested public queues directly from observed throughput. We then jointly estimate patient preferences and scan marginal costs, recovering private capacities within the estimation routine. Finally, we estimate slot costs and bound scanner fixed costs.

5.1 Queuing model

For estimation, we extend the model of Section 4.2 to allow the capacity of public queues to vary across weeks. We reconstruct weekly backlogs at publicly funded queues from observed

[24] Since reimbursements are paid on performed scans ($a_{h\,\mathrm{pub}}$) rather than allocated slots ($k_{h\,\mathrm{pub}}$), we can view (5) as a "relaxed" constraint. Whenever the resulting equilibrium features persistent congestion on the public side (as the status quo does), (5) ensures that the reimbursement cap is not exceeded at the realized arrivals. Using a modality's mean reimbursement $\bar{r}_m$ is a shortcut to avoid explicit accounting of the scan composition of arriving patients.

arrivals and throughput.[25] Given a reconstructed backlog and the subsequent weekly capacity path, the first-availability wait is the time required for available slots to serve the patients already waiting in the queue. These reconstructed waits are used as the wait attributes of publicly funded alternatives in the patient choice model. Appendix D.1 provides the exact construction and shows that the model-implied wait times track the realized ones very closely, particularly for the congested lower-urgency queues accounting for the majority of patients.

Provider optimality conditions require expectations over the stationary distribution of waits and derivatives of these expectations with respect to prices and capacity. We approximate the stationary wait distribution using its first two moments. This allows us to frame stationarity as a fixed point, from which the responses of wait times to provider choices can be obtained. Appendix D.2 provides details on these approximations and their implementation.

5.2 Capacity of public queues

As mentioned in Section 4.2, the weekly number of slots available in almost-surely-congested queues coincides with throughput. Define $\mathcal{B}_{h\,\mathrm{pub}\,mu}$ as the subset of weeks in which publicly funded care at h is backlogged in modality m and urgency u, measured by positive realized waits or positive backlog as reconstructed by our queuing model with the procedure mentioned in Section 5.1 and detailed in Appendix D.1. We estimate public capacity at the provider-modality-urgency level as the mean total throughput across those busy weeks:

$$\hat{k}_{h\,\mathrm{pub}\,mu} = \frac{1}{|\mathcal{B}_{h\,\mathrm{pub}\,mu}|} \sum_{t \in \mathcal{B}_{h\,\mathrm{pub}\,mu}} y_{h\,\mathrm{pub}\,mut}.$$

5.3 Private capacity, patient demand, and scan costs

Private capacity, patient preferences, and scan marginal costs are estimated jointly. We first describe the estimation procedure and then discuss identification.

Under the chosen parameterization, scan cost parameters $\boldsymbol{\gamma}$ can be concentrated out as in Berry et al. (1995). Provider qualities $\boldsymbol{\delta}$ are pinned down by preference parameters $\boldsymbol{\beta}$. Private capacities are inferred from patient demand, which is completely summarized by $(\boldsymbol{\beta}, \boldsymbol{\delta})$. The joint GMM estimator therefore searches only over $\boldsymbol{\beta}$, with the objective constructed as follows. In the following, script lowercase letters (e.g., ω) denote observed quantities we match in estimation.

1. *Provider quality* For each candidate $\boldsymbol{\beta}$, we recover provider qualities $\boldsymbol{\delta}(\boldsymbol{\beta})$ by solving

[25] Because we observe completed procedures rather than bookings, we measure arrivals to queue q in week t as patients who receive a referral in week t and are subsequently observed completing care in that queue. This assumes that patients who eventually use publicly funded care book in the week in which they receive their referral, and that patients do not subsequently renege from the queue. Patients who never complete publicly funded care are treated as never having joined a public queue.

$\sigma_{h\,\mathrm{pub}\,m}(\boldsymbol{p}, \boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\delta}) = s_{h\,\mathrm{pub}\,m}$ for all h, m. The left-hand side is the model-predicted aggregate share of publicly funded care at provider h in modality m, and $s_{h\,\mathrm{pub}\,m}$ is the corresponding observed value. Prices $\boldsymbol{p}$ and the realized sequence of waits $\boldsymbol{W} = (\boldsymbol{w}_t)_t$ are held fixed.[26] Proposition 1 ensures uniqueness whenever a solution exists, and we implement the inversion using the contraction map of Berry et al. (1995).[27]

2. *Patient choice* Given $(\boldsymbol{\beta}, \boldsymbol{\delta}(\boldsymbol{\beta}))$, we match (i) average waits and distances among patients completing publicly funded care and (ii) the fraction of referrals without a publicly funded treatment, for different combinations of demographics. Appendix D.5 lists the targeted moments. Writing $x \in \{w, d, \mathrm{np}\}$ for wait, distance, or an indicator for no publicly funded treatment, each sample moment takes the form

$$M^{\mathrm{D},x}(\boldsymbol{\beta}) = \frac{\sum_{i,j} \sigma_{ij}(\boldsymbol{p}, \boldsymbol{w}_{t(i)}; \boldsymbol{\beta}, \boldsymbol{\delta}(\boldsymbol{\beta}))\, x_{ij\,t(i)}}{\sum_{i,j} \sigma_{ij}(\boldsymbol{p}, \boldsymbol{w}_{t(i)}; \boldsymbol{\beta}, \boldsymbol{\delta}(\boldsymbol{\beta}))} - \bar{x},$$

where $t(i)$ is patient i's referral week and $\bar{x}$ is the corresponding observed sample average. The sums run over patients satisfying the relevant demographic restrictions. For waits and distance, they are further restricted to publicly funded alternatives. We stack these sample moments into $M^{\mathrm{D}}(\boldsymbol{\beta})$.

3. *Private capacity* For each private queue, we find the mean capacity that reproduces its observed average wait. Parameters $(\boldsymbol{\beta}, \boldsymbol{\delta}(\boldsymbol{\beta}))$ pin down patient arrivals. Given a candidate mean capacity $k_{h\,\mathrm{pvt}\,m}$, we simulate weekly arrivals and backlogs using the realized referrals, prices, and waits at other providers. The queue's own wait evolves with its backlog and affects subsequent patient choices.[28] Denoting the simulated week-t wait by $w_{h\,\mathrm{pvt}\,mt}(k_{h\,\mathrm{pvt}\,m}; \boldsymbol{\beta}, \boldsymbol{\delta}(\boldsymbol{\beta}))$ and the observed average scraped first-availability

[26] We use scraped data for private-pay waits. Since data on publicly funded care only report realized waits, we reconstruct public waits predicted by our queuing model as described in Section 5.1.

[27] In our model, a solution may fail to exist, as discussed in Appendix G.2. In practice, we were always able to invert the demand system.

[28] Recovering mean capacity from average waits requires accounting for when slots are available. Consider two queues facing the same patient arrivals. One offers a steady number of slots and keeps patients waiting a few days throughout the year, while the other offers more slots on average but sharply reduces availability during holidays. The latter can have almost no waits during most of the year and long waits around holidays, producing the same average wait as the former. This is because extra slots go unused when there are too few patients waiting and cannot prevent the holiday backlog. Ignoring these fluctuations could therefore lead us to back out too little capacity from observed average waits. We use public throughput to infer how much each provider's public activity contracts when holidays occur. We then assume that private capacity follows the same within-year pattern, with its mean shifting the level of average weekly slots. Inversion through (6) recovers this mean level. Appendix D.3 provides details. Note that persistently congested queues use every slot, so average throughput directly reveals mean capacity despite weekly fluctuations. The subsequent pricing step takes recovered capacities as given and evaluates providers' expected profits across possible demand realizations, using the mean and variance of stationary waits to approximate providers' expectations.

wait by $\bar{\omega}_{h\,\mathrm{pvt}\,m}$, we solve

$$\frac{1}{T}\sum_{t} w_{h\,\mathrm{pvt}\,mt}\big(k_{h\,\mathrm{pvt}\,m};\boldsymbol{\beta},\boldsymbol{\delta}(\boldsymbol{\beta})\big) - \bar{\omega}_{h\,\mathrm{pvt}\,m} = 0. \tag{6}$$

Repeating this inversion across private queues gives $\hat{\boldsymbol{k}}_{\mathrm{pvt}}(\boldsymbol{\beta})$.

4. *Stationary queue environment* Combining $\hat{\boldsymbol{k}}_{\mathrm{pvt}}(\boldsymbol{\beta})$ with the publicly funded capacities $\hat{\boldsymbol{k}}_{\mathrm{pub}}$ estimated as described in Section 5.2, we approximate the stationary distribution of waits using its first two moments. We specify the variance as a function of mean wait and capacity, then solve for the stationary mean wait times $\bar{\boldsymbol{w}}(\boldsymbol{p},\hat{\boldsymbol{k}};\boldsymbol{\beta})$ implied by the patient choice, prices, and capacity. Appendix D.2 provides details.

5. *Pricing* We use the mean and variance of stationary waits from the preceding step to approximate providers' expected profits in the stationary distribution of waits. Their pricing optimality conditions account for how a price change affects patient choices both directly and through the resulting changes in wait times. We invert providers' pricing optimality conditions at observed private-pay prices to recover scan marginal costs $c_{hs}(\boldsymbol{\beta})$.[29] We form pricing moments by interacting residuals from the scan cost specification in (3) with instruments $\boldsymbol{z}_{hs}^{\text{scan-cost}}$. These include the cost regressors, the number of nearby providers offering the same modality, and demographics-predicted demand. We construct predicted demand holding private-pay prices, wait times, and provider qualities fixed at common values, so that variation comes only from the geographic distribution of patient characteristics. For each candidate $\boldsymbol{\beta}$, we concentrate out $\boldsymbol{\gamma}$ by minimizing the pricing-moment criterion, giving $\boldsymbol{\gamma}(\boldsymbol{\beta})$ and residuals $\omega_{hs}(\boldsymbol{\beta})$. The resulting sample moments are

$$M^{\mathrm{P}}(\boldsymbol{\beta}) = \frac{1}{N^{\mathrm{P}}}\sum_{h,s} \boldsymbol{z}_{hs}^{\text{scan-cost}}\omega_{hs}(\boldsymbol{\beta}).$$

6. *Joint criterion* Stacking moments into $M(\boldsymbol{\beta})$, compute $M(\boldsymbol{\beta})'WM(\boldsymbol{\beta})$.

The estimator minimizes this criterion over $\boldsymbol{\beta}$, repeating the steps above at each candidate. Appendix D.4 provides computational details.

Additional details We approximate integration over patient types using a simulated sample drawn from the empirical distribution of referrals over scans, referral weeks, towns, and patient demographics. Appendix A.3 describes the construction of this sample.

[29] Appendix C.3 derives the optimality conditions, and Appendix D.4 describes their inversion.

To avoid restricting private choice sets to the providers and scans for which prices and waits were successfully scraped, we impute missing private prices and wait times using observed scan, provider, geographic, and calendar characteristics. The resulting estimates are similar when missing values are instead replaced by simple averages among observed alternatives. Imputed prices and waits are used only to compute patient choice probabilities, while pricing optimality conditions and private capacity are evaluated only for providers observed in the scraped data. Appendix D.10 provides further details and validation of the imputation.

Identification Intuitively, identification of patient preferences exploits the assumption that the provider-modality qualities operate at a higher level of aggregation than the variation in prices and wait times. Private-pay prices vary across scans within provider-modality, while public wait times vary across referral weeks. This difference in aggregation allows us to exploit variation in prices and waits while avoiding the need to instrument for these attributes, as would typically be needed for identification with micro data (Berry and Haile, 2024). With complete choice data, this variation would inform both price and wait sensitivity. Since private choices are unobserved and public co-pays exhibit little variation, price sensitivity is additionally disciplined by the pricing optimality conditions. Variation across demographics provides information on patient preference heterogeneity.

The pricing restrictions provide additional information on patient preferences by requiring the marginal costs implied by the demand system to satisfy the exclusion restrictions described above. Following the insight of Waldfogel (2003), demographics-predicted demand predicts the demand and hence the markup a provider can sustain, while being excluded from the residual marginal cost shock conditional on local wages. Because candidate demand parameters imply different markups and hence different cost residuals, these restrictions help identify price sensitivity when observed public choices exhibit little price variation and private choices are not observed.[30]

Finally, identification of private capacity rests on the assumption that, holding $(\boldsymbol{\beta}, \boldsymbol{\delta})$ and the realized histories fixed, increasing $k_{h\,\mathrm{pvt}\,m}$ increases the number of available slots throughout the sample and weakly reduces the induced waits. If this relationship is strictly decreasing over the relevant range of capacities, (6) has a unique solution. Appendix G.3 elaborates on existence and uniqueness.

[30] Price sensitivity allows for income effects following the approach proposed by Miravete et al. (2026), which adds flexibility in the curvature of demand. The correlation of demographics in the population naturally induces correlations between tastes for different characteristics. As demonstrated by Goldszmidt et al. (2026), this can break the stiff elasticity patterns implied by plain logit models and by mixed logit models with uncorrelated random effects. Since we use pricing optimality conditions to help identification – especially of price sensitivity – these features mitigate the concern that functional form restrictions might drive our results. The strategy of using the supply side to identify price sensitivity has precedents in, e.g., Gentzkow (2007) and Sullivan (2026).

5.4 Capacity and scanner costs

The final estimation step recovers slot costs and bounds scanner fixed costs, taking results from the previous steps as given. We impose provider optimality conditions only on privately owned providers.

Slot costs Optimality of offered slots requires that the marginal revenue of an additional slot equal its marginal cost. A complete derivation of slot marginal revenues and optimality conditions is in Appendix C.3. Denoting the marginal operating profit from an additional private slot as MR_{hm} and using the slot cost parameterization (4), optimality requires

$$\log MR_{hm} = \underbrace{\boldsymbol{\zeta}_x \cdot \boldsymbol{x}_{hm} + \zeta_{\text{input}} \log(K_{hm}/n_{hm}) + \eta_{hm}}_{\text{slot marginal cost}}.$$

The vector $\boldsymbol{x}_{hm}$ contains modality and provider characteristics. Capacity per scanner is endogenous because providers choose both capacity and scanner counts in response to slot-cost shocks. We estimate this equation by GMM using the orthogonality conditions $\mathbb{E}[\eta_{hm}\boldsymbol{z}_{hm}^{\text{slot-cost}}] = 0$, where $\boldsymbol{z}_{hm}^{\text{slot-cost}}$ collects $\boldsymbol{x}_{hm}$ and excluded instruments for capacity per scanner.[31] As with the Waldfogel-style instruments used for scan costs, we instrument K_{hm}/n_{hm} with private demand as predicted by patient demographics. These demand differences shift the marginal return to additional slots and help predict capacity per scanner, under the assumption that the underlying demographic characteristics are unrelated to slot-cost shocks conditional on $\boldsymbol{x}_{hm}$. Appendix D.7 elaborates on identification and the construction of our instrument.

Scanner fixed costs We bound scanner fixed costs using necessary conditions for optimality. The observed choices must yield at least as much operating profit, net of slot and scanner costs, as any feasible deviation. Since providers jointly choose scanners, slots, and prices, the strongest restriction for each alternative scanner count is obtained by maximizing the deviating provider's profits over feasible slots and prices, holding competitors' choices fixed. In practice, we consider a subset of deviations that change scanner counts while (i) reoptimizing private-pay prices and private capacity at fixed public capacity, or (ii) reducing public capacity alone or both public and private capacity proportionally at fixed prices.[32] Other

[31] Note that this step is analogous to the estimation of scan marginal costs described in Section 5.3. Both pricing and capacity optimality conditions require implicit differentiation of equilibrium mean wait times. We include pricing conditions in the joint estimator because they help identify patient preferences, particularly price sensitivity. Evaluating capacity conditions at every parameter guess would add substantial computational cost, so we estimate slot costs separately.

[32] See Pakes et al. (2015) for the general revealed-preference approach. Out of computational considerations, related applications also bound fixed or sunk costs using a subset of feasible deviations, such as adding or removing individual products (Eizenberg, 2014; Wollmann, 2018; Fan and Yang, 2020). Our deviations reduce the number of variables over which we optimize and automatically satisfy the reimbursement constraint (5),

providers' choices remain fixed at the observed values, and stationary waits are recomputed. The absence of profitable deviations of these types is necessary but not sufficient for optimality of the observed scanner counts. These deviations are computationally cheaper than fully reoptimizing prices and both public and private capacity at each alternative scanner count, while still yielding informative bounds on scanner fixed costs.

For each scanner count n, define $V_{hm}(n)$ as provider h's maximum operating profits in modality m net of slot costs over the price and capacity adjustments described above. Optimality of the observed scanner count n_{hm} implies

$$\max_{n>n_{hm}} \frac{V_{hm}(n) - V_{hm}(n_{hm})}{n - n_{hm}} \leqslant \chi_{hm} \leqslant \min_{n<n_{hm}} \frac{V_{hm}(n_{hm}) - V_{hm}(n)}{n_{hm} - n}.$$

Appendix D.8 derives these outer bounds and details the deviations.

6 Results

This section highlights some insights from our estimates. Full results are in Appendix E.

Table 3 reports summary statistics of the distribution of willingness to pay for a shorter wait, marginalized against demographic dimensions. Patients exhibit heterogeneity in the value they attach to timeliness. Younger, richer, and more urgent patients appear to be more willing to pay for faster treatment. Richer individuals are likely to have a higher opportunity cost of time, or, equivalently, a lower marginal utility of money. The age gradient is consistent with older patients being more likely to receive referrals for chronic illnesses (conditional on urgency tier), which makes the need for faster treatment less of a priority. The ordering across urgency tiers is also reasonable, with tiers 2, 3, and 4 being ordered in the expected way. Highest-urgency patients appear to attach low value to timely care: the wait times they are exposed to under publicly funded care are in fact very similar to private-pay ones, and this lack of variation makes estimates of their distaste for wait quite noisy. Across modalities, MRI patients value timeliness the most. The full set of estimated patient preference parameters is reported in Table E3.

Table 4 summarizes the distribution of own-price and own-wait elasticities across patients, by modality and access type. Private-pay patients are almost entirely insensitive to wait times. This is consistent with private-pay care offering substantially shorter waits than publicly funded alternatives, so that marginal increases in wait times are unlikely to induce substantial diversion. The low price elasticities are also consistent with the implicit price discrimination afforded by mixed provision: if publicly funded care absorbs relatively price-elastic consumers and marginal scan costs are low, profit-maximizing providers optimally set prices against the

since public capacity is held fixed or reduced.

Table 3 Willingness to pay for shorter wait

	X-ray			CT			MRI		
	25th	50th	75th	25th	50th	75th	25th	50th	75th
Income									
⩽ 15k	0.21	0.26	0.41	0.10	0.11	0.27	0.73	1.08	1.35
15-26k	0.39	0.49	0.76	0.15	0.17	0.40	1.03	1.52	1.91
26-55k	0.60	0.75	1.16	0.20	0.22	0.53	1.30	1.93	2.42
⩾ 55k	0.88	1.09	1.69	0.25	0.28	0.67	1.61	2.39	2.99
Urgency									
Tier 1	0.05	0.07	0.11	0.10	0.17	0.27	0.89	1.18	1.60
Tier 2	1.92	2.51	3.72	0.23	0.40	0.67	3.04	4.00	5.37
Tier 3	0.64	0.84	1.27	0.19	0.32	0.53	1.54	2.03	2.78
Tier 4	0.37	0.49	0.68	0.09	0.15	0.22	0.85	1.13	1.55
Age									
18–29	0.62	1.02	1.49	0.95	1.24	1.68	1.52	1.93	3.47
30–39	0.79	1.31	1.91	0.97	1.27	2.03	1.22	1.55	2.20
40–49	0.68	1.13	1.65	0.53	0.69	1.11	1.06	1.34	1.91
50–59	0.41	0.68	0.76	0.17	0.22	0.30	1.74	2.20	3.12
60–69	0.36	0.60	0.68	0.15	0.20	0.27	0.67	0.85	1.22
70–79	0.35	0.57	0.64	0.07	0.09	0.12	0.89	1.13	1.61
80+	0.45	0.75	1.09	0.20	0.26	0.42	1.03	1.30	2.35

Notes: Table reports the 25th, 50th, and 75th percentiles of model-implied willingness to pay for a one-day reduction in wait time, measured in euros. Percentiles are weighted by patient population within each demographic group and modality, integrating over the remaining demographics and geographic composition.

Table 4 Own elasticities

	Price			Wait		
	25th	50th	75th	25th	50th	75th
X-ray						
Publicly funded	-0.85	-0.72	-0.59	-1.19	-0.69	-0.03
Private-pay	-2.06	-1.70	-1.42	-0.17	-0.07	-0.02
CT						
Publicly funded	-0.61	-0.55	-0.48	-0.44	-0.15	-0.04
Private-pay	-2.77	-2.34	-2.06	-0.04	-0.02	-0.01
MRI						
Publicly funded	-0.44	-0.40	-0.37	-1.26	-0.78	-0.31
Private-pay	-2.97	-2.11	-1.78	-0.13	-0.07	-0.04

Notes: Table reports the 25th, 50th, and 75th percentiles of own-price and own-wait elasticities for each modality-access combination.

Table 5 Provider costs

(a) Per-scan margins (€/scan)

	X-ray			CT			MRI		
	25th	50th	75th	25th	50th	75th	25th	50th	75th
Public	-3.01	5.15	16.95	-23.36	10.15	55.94	-64.19	69.45	120.20
Private	35.11	37.80	40.97	77.91	84.53	114.78	104.74	116.07	147.54

(b) Scanner fixed costs (1,000€/year)

	CT	MRI
Median lower bound	4.47	14.89
Median upper bound	88.52	232.22

Notes: Panel (a) reports percentiles of per-scan contribution margins. Publicly funded and private-pay margins are defined as reimbursement minus marginal scan cost and private price minus marginal scan cost, respectively. Publicly funded margin percentiles are weighted by observed publicly funded scan volumes, while private-pay margin percentiles are weighted by model-implied private scan volumes. Scan costs are obtained by inverting the pricing optimality conditions as discussed in Section 5.3. Panel (b) reports medians of the lower and upper bounds on annual fixed costs per scanner. Both medians are computed over provider-modalities with nonempty intervals. Bounds are obtained from scanner-count deviations conditional on the estimated slot-cost technology, as described in Appendix D.8.

Table 6 Private-pay capacity allocation

	X-ray			CT			MRI		
	25th	50th	75th	25th	50th	75th	25th	50th	75th
Share (%)	32.07	39.79	50.12	12.09	16.07	23.11	13.25	26.42	39.40
Utilization (%)	65.13	83.95	95.03	71.57	81.54	92.00	78.22	87.97	96.25

Notes: Table reports moments of the distribution of private-pay capacity shares and utilization rates. Shares are the ratio of private-pay slots to total slots offered by a provider in a modality (i.e., $k_{h\,\mathrm{pvt}\,m}/(k_{h\,\mathrm{pvt}\,m}+k_{h\,\mathrm{pub}\,m})$). Utilization rates are defined as the ratio between the model-implied expected throughput and the weekly mean capacity. Publicly funded capacity is obtained as described in Section 5.2. Private-pay capacity is solved for inside our estimation routine, as described in Sections 5.3 and D.3. Levels of publicly funded and private-pay capacities are in Tables E1 and E2.

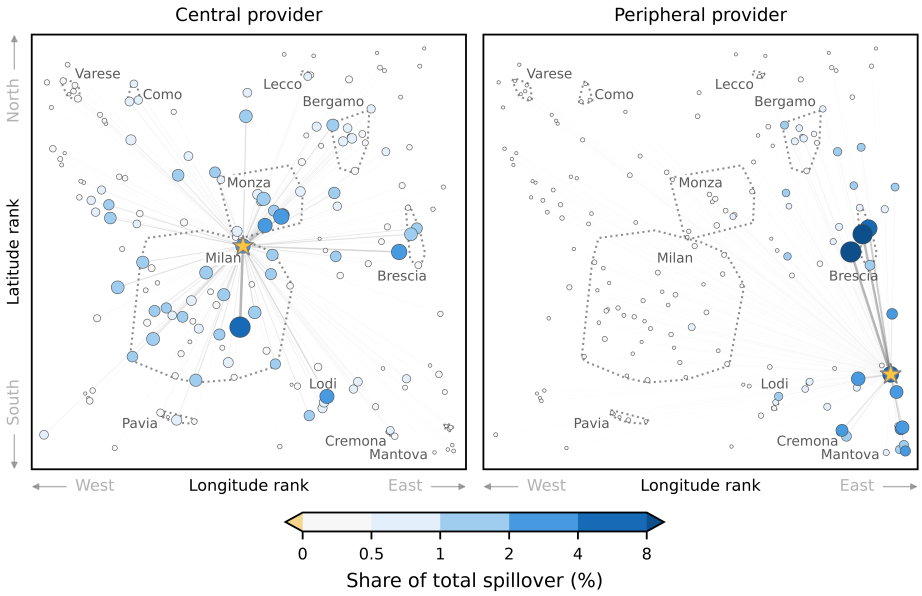
comparatively inelastic residual demand for private-pay care. Among publicly funded patients, demand is below unit elastic with respect to both prices and waits, consistent with public co-pays being set below cost and with the relatively low wait sensitivity documented in Table 3.

Table 5 reports key summaries of providers' costs, while complete parameter estimates are presented in Table E4. The ordering of modalities in terms of profitability is as expected, with MRI being the most lucrative and X-ray the least lucrative. Despite larger dispersion, publicly funded MRIs are substantially more profitable than CTs. This is consistent with the publicly funded supply mix of private providers being more skewed towards MRI than that of public providers, as documented in Figure E1. The costs of offering additional slots are negligible for X-rays and modest for CTs and MRIs. The slot cost function exhibits mild convexity, with $\hat{\zeta}_{\text{input}} = 0.56$. In equilibrium, higher slot costs are associated with higher utilization rates, as illustrated by Figure E2. Finally, our median annual scanner-cost bounds, approximately €4k–89k for CT and €15k–232k for MRI, span values below simple ownership-cost benchmarks, although their upper ends are close to the range implied by procurement evidence.[33]

Table 6 shows the estimated shares of capacity devoted to private-pay patients and the utilization rate of such capacity at private providers. Private-pay queues are operated with

[33]Recent procurement records from the region we study place installed CT investment around €0.7M and MRI investment around €0.9–1.2M. Full-risk maintenance contracts suggest recurring expenditures of roughly €30–70k per year for CT and around €80–125k for MRI. Spreading investment costs over eight to ten years therefore implies total annual ownership costs of roughly €100–170k for CT and €170–280k for MRI. The upper end of our MRI interval falls squarely within this range, while the upper end for CT is somewhat below it. These comparisons are only suggestive because our static model does not capture all of the economic returns and costs associated with scanner ownership. Lumpy and partly irreversible investment can make the cost relevant for capacity choice differ from annualized acquisition and maintenance expenditures. In the opposite direction, scanners can also support inpatient care, which is outside our model and can generate substantial additional revenue. Omitting this use lowers the value of scanner ownership recovered by the model and therefore pushes our scanner-cost estimates downward.

Figure 2 Congestion spillovers across providers



Notes: Each panel shows the distribution of patient surplus gains through shorter public waits following a 1% reduction in all private-pay MRI prices at the provider marked by the yellow star. Effects are evaluated at the status quo, holding other prices and all capacities fixed while wait times adjust endogenously. Circles represent providers offering MRIs. Their size and color indicate the share of the starred provider's total congestion benefit accruing to publicly funded patients at each provider. Lines connect the starred providers to providers receiving positive spillovers. Provider positions are based on longitude and latitude ranks. Dotted boundaries group providers located in the same metropolitan area.

some slack, with 12–18% of the slots going unfilled. This is consistent with basic insights of queuing models: in the presence of stochastic arrivals, it is often the case that low average waits can only be sustained by a positive probability of the scanner being idle. Estimates of the levels of publicly funded and private-pay capacity are reported in Tables E1 and E2.

The estimates allow us to quantify the congestion spillovers generated by private-pay care. When a provider lowers its private-pay prices, some patients switch from publicly funded treatment to private-pay care, reducing congestion in the public queues they would otherwise have joined. This shortens waiting times for patients who continue to receive publicly funded care, including at providers that have not changed their prices. Beyond the usual failure of profit maximization to account for patient surplus, private pricing therefore generates congestion spillovers across providers. The vast majority of welfare benefits accrue to patients waiting at other providers, and are therefore not internalized by the provider setting its private-pay prices. Figure 2 illustrates the geographic reach of these spillovers for a central and peripheral provider. Figure E3 and Table E5 document the broader spillovers from capacity and pricing decisions, as well as their geographic distribution across all providers.

7 Counterfactuals

We present our counterfactuals in three steps. First, we measure the welfare effects of removing private-pay care. Second, we quantify the inefficiencies in pricing and capacity decisions coming from providers' profit-maximizing behavior in the status quo. Third, we explore the distributional effects of private-pay care. Throughout these counterfactuals, we hold co-pays for publicly funded access and reimbursements fixed at the observed levels. We assign each provider their inferred upper bound for scanner costs because it aligns best with available procurement evidence.[34] Appendix F provides the relevant implementation details.

7.1 The value of market provision

We quantify the welfare value of the private-pay option, accounting jointly for how it allocates shorter waits to patients who value timeliness most and how it affects capacity provision and utilization. To measure this, we consider a counterfactual where private-pay care is unavailable and providers only decide how many slots to supply to serve publicly funded patients. Removing private care causes a large welfare loss through both removed sorting and additional congestion from diverting demand towards publicly funded care. A planner who reallocates slots across providers to maximize welfare would recover roughly 60% of the welfare loss, but it would take a significantly larger budget to restore the status quo welfare. This is due to the inferior matching allowed by the counterfactual system: absent private-pay care, surplus can only be improved by reducing public waits broadly rather than offering faster access selectively to patients who value it most.

Table 7 breaks down the result. In a decentralized public-only system, providers cannot serve private patients and can only decide how much capacity to supply. Relative to the status quo, providers would increase supply of publicly funded MRI slots and significantly cut X-rays and CTs. This is consistent with the marked differences in profitability across modalities documented in Table 5: MRI reimbursements exceed scan marginal costs and provide enough operating margin on publicly funded activity to fund capacity and to make this modality relatively more attractive. The different capacity allocation translates into fewer patients treated, longer wait times for CT and X-ray patients, and shorter waits for MRI patients. Patient surplus falls across all three modalities, as decomposed in Figure 3. Removing private access and the additional congestion generated by patients displaced from private-pay care would reduce patient surplus in all three modalities. Providers' capacity responses would amplify these losses for CT and X-ray, but substantially mitigate them for MRI. Nevertheless,

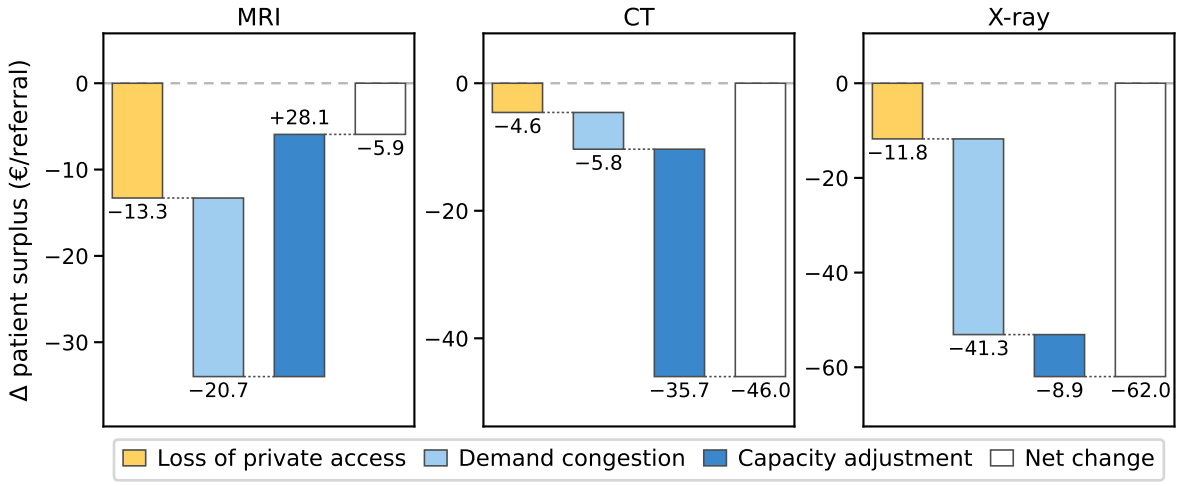
[34] See footnote 33 for a discussion of how our fixed cost estimates compare with public providers' procurement records. For counterfactuals that allow changes in the scanners installed at public providers, we assign these providers the median upper bound. Our results are robust to alternative specifications that use scanner costs calibrated on the procurement evidence.

Table 7 Disallowing private care

	Status quo	Public only			
		Decentralized	Planner		
			Current budget	Expanded budget	Direct investment
Welfare (M€/week)					
Total welfare	10.73	5.37	8.57	10.73	10.73
Total change (%)	–	-49.99	-20.13	0.00	0.00
Government expenditure (M€/week)					
Reimbursements	6.13	5.88	6.13	8.29	–
Direct scan and slot costs	–	–	–	–	2.94
Existing scanner costs	–	–	–	–	1.46
Total expenditure	6.13	5.88	6.13	8.29	4.39
Public capacity (1,000 slots/week)					
MRI	12.58	15.26	12.58	13.33	15.82
CT	15.36	10.81	15.36	25.89	26.30
X-ray	45.10	40.01	45.10	64.52	83.98
Patients treated (1,000/week)					
MRI	15.34	15.02	12.13	12.83	13.23
CT	15.77	10.65	14.81	16.42	16.42
X-ray	50.44	30.81	36.48	51.91	50.90
Public wait times (days)					
MRI	59.92	47.58	21.45	26.16	16.12
CT	61.84	233.54	34.91	13.75	13.55
X-ray	42.83	107.00	17.87	9.41	6.07

Notes: Table compares the status quo mixed system with four scenarios where private care is removed. Public co-pays and reimbursements are held fixed at the observed levels. *Decentralized* provision allows private providers to adjust their capacity in response to the removal of private-pay care while public providers are inert. *Current budget* allows a planner to reallocate publicly funded capacity across providers to maximize welfare, subject to the status quo reimbursement budget within each modality. *Expanded budget* allows the planner to raise resources to fund additional scans until the status quo welfare is reached. Such additional scans are still reimbursed at the status quo rates and are allocated optimally. *Direct investment* allows the planner to fund capacity expansions by directly incurring slot and scanner costs. This planner optimally allocates capacity starting from the decentralized allocation and proceeds until the status quo welfare is reached. Public waits are patient-weighted means of equilibrium waits, across Tier 2–4 urgency levels. Treatment volumes under the status quo include both publicly funded and private-pay care.

Figure 3 Decomposing patient surplus changes



Notes: Figure decomposes the change in patient surplus per referral between the status quo and the equilibrium without private-pay care in which providers adjust their capacity to maximize profits. *Private access* measures the loss from removing private-pay options and recomputing patient choices while holding public waits fixed at their status quo levels. *Demand congestion* captures the additional loss from patients re-sorting and public waits readjusting upon removal of private care, holding public capacity fixed. *Capacity adjustment* measures the further change when providers adjust their public capacity and queues re-equilibrate. *Net change* is the sum of these three effects. Negative values indicate losses in patient surplus.

MRI patients would still be worse off despite facing shorter public waits, reflecting the loss of sorting according to willingness to pay for timeliness.

Reallocating slots within each modality could afford large welfare gains relative to the decentralized benchmark. This scenario allows a planner to reallocate each modality's status quo public slots across providers in a way that maximizes welfare. Since slots are reimbursed at the status quo rate, reshuffling them creates no additional expenditure.[35] This planner would be able to achieve reductions in wait times between 25 and 40 days across modalities. Despite the shorter waits, welfare would still be roughly 20% lower than in the status quo. This counterfactual offers two insights. First, centralized planning can be socially valuable because it accounts precisely for those spillovers across providers that are not fully internalized by profit maximizing behavior. Second, the welfare gains achievable by large but untargeted wait reductions can still fall short of the ones generated by the sorting mechanism introduced by private-pay care.

Building on this, we ask how large of an additional budget would be required to restore the status quo welfare without private care. We allow the planner to increase the level of public expenditures to fund more slots to allocate optimally, while still reimbursing them at

[35]The fixed costs of operating scanners are still borne by providers.

the status quo rate. Despite enabling capacity expansions and reductions in congestion, it would take a budget increase of about 35% to restore the status quo welfare. The exercise assumes no cost of raising additional resources, and an even larger budget expansion would be needed if raising funds was costly (e.g., because of distortionary taxation). This magnitude highlights how costly it can be to deliver welfare gains without leveraging patient sorting.

Reimbursing scans at the status quo rate is the most natural counterfactual comparison, but it is generally inefficient because providers that differ in how efficiently they can supply capacity are paid at flat rate. We consider an alternative scenario where a planner expands capacity from the public-only system by directly financing scanner costs and capacity marginal costs. Direct investment would allow the planner to attain the status quo welfare under a public-only system with 28% *less* public expenditure than the status quo. The intuition behind this result is that, with a convex cost of supplying capacity, a flat reimbursement grants providers rents on inframarginal slots. When each slot is instead supplied at its marginal cost, public funds can be used to allocate capacity in a way that reflects its social value.[36] Appendix F.2 describes the planner's problems formally.

7.2 Inefficiencies in market provision

Private-pay care is welfare-improving, but profit motives and market power can introduce inefficiencies if providers fail to internalize the benefits that shifting patients into private-pay care generates for public users. We quantify pricing and capacity distortions by comparing the status quo with the first best allocation implemented by a budget-constrained planner that jointly chooses private prices and public and private capacity. The planner would substantially reduce private prices, shifting patients away from congested publicly funded care, and adjust private capacity to accommodate the resulting demand. This reveals that the status quo features both inefficiently high prices and underinvestment in capacity. The former is the dominant distortion: a second-best planner that controls both total capacity and its allocation across public and private care, but leaves private prices to providers, only achieves about 40% of the first-best welfare gains. The key insight is that expanding and reallocating capacity cannot substitute for correcting the prices that govern patient sorting.

The first-best planner maximizes total surplus, which accounts for patient surplus and the resource costs of scans, scanners, and slots. The planner jointly chooses private prices and public and private capacity[37] at all providers under the constraints that public expenditure cannot exceed the status quo budget within each modality, including the cost of capacity

[36] This highlights how procurement schemes that more closely reflect production costs might create large savings of public resources, potentially enabling further capacity expansion. How to optimally procure publicly funded capacity is an interesting question, both theoretically and empirically.

[37] They can choose both the number of scanners operated and the slots offered.

Table 8 Efficiency

	MRI			CT			X-ray		
	Status quo	First best	Second best	Status quo	First best	Second best	Status quo	First best	Second best
Welfare (M€/week)									
Total welfare	2.37	3.61	3.00	1.76	2.18	1.89	6.61	7.63	6.94
Total change (%)	–	52.55	26.92	–	23.53	7.50	–	15.49	5.08
Prices (€)									
Private price	209.67	59.69	201.70	170.33	58.05	165.49	53.25	19.05	52.81
Median change (%)	–	-80.01	-1.85	–	-72.60	-0.97	–	-65.48	-2.39
Capacity (1,000 slots/week)									
Public capacity	12.70	12.53	12.71	15.36	15.35	15.46	45.10	45.11	45.00
Private capacity	3.92	9.63	5.27	1.81	6.53	2.43	21.60	34.24	35.73
Total change (%)	–	33.39	8.23	–	27.38	4.19	–	18.96	21.03
Wait times (days)									
Public wait	59.92	13.17	15.93	61.84	16.28	24.05	42.83	6.75	14.94
Private wait	3.35	3.41	2.41	6.37	3.81	2.77	3.50	2.53	2.24
Patients treated (1,000/week)									
Public	12.58	11.04	11.90	14.55	12.07	14.15	36.35	33.48	35.07
Private	2.76	7.73	2.74	1.22	5.16	1.30	14.09	20.16	15.66
Total	15.34	18.78	14.63	15.77	17.23	15.45	50.44	53.64	50.73

Notes: Table compares the status quo with two alternative allocations of the mixed system. We keep each modality's budget fixed at the observed public expenditures. Reimbursement rates and public co-pays are also held fixed. The first-best planner maximizes welfare by jointly choosing private-pay prices, treatment capacity, and scanner investment. The second-best planner chooses treatment capacity and its allocation between public and private care, while private providers set private-pay prices to maximize profits. Welfare is normalized to a common status quo level for each modality, and its changes account for scan costs, capacity costs, and changes in installed scanners. Private prices are weighted by status quo private treatment volumes. The median percentage change is calculated across providers, comparing each provider's average private price with its status quo value. Wait times are patient-weighted equilibrium means in days, across Tier 2–4 urgency levels.

investment.[38] The second-best planner maximizes the same welfare objective but can only choose public and private capacity at each provider, while private-pay prices remain determined by providers. Appendix F.3 describes the problems formally.

Table 8 presents the results. The first best would reduce median private prices by 65–80% across modalities and substantially expand private capacity, while keeping publicly funded capacity near its status quo level. The additional private slots would accommodate a sharp increase in private-pay care without longer private waits. As more patients shift away from congested public queues, public waits would fall by roughly 35–50 days across modalities. The welfare gains would therefore arise primarily from expanding total capacity and improving patient sorting, rather than diverting existing public slots toward private-pay care.

The second best allocation illustrates the limits of capacity controls alone. Private prices would remain close to their status quo levels even as the planner expands private capacity. At these prices, additional private slots would attract relatively few patients, leaving private volumes almost unchanged for MRI and CT despite capacity expansions of roughly 35%. For X-ray, the planner would expand private capacity more than under the first best, but substantially fewer patients would use it. Reallocating public capacity across providers would still generate appreciable welfare gains by reducing public waits by 25–45 days across modalities, but could not replicate the patient sorting achieved through lower private prices.

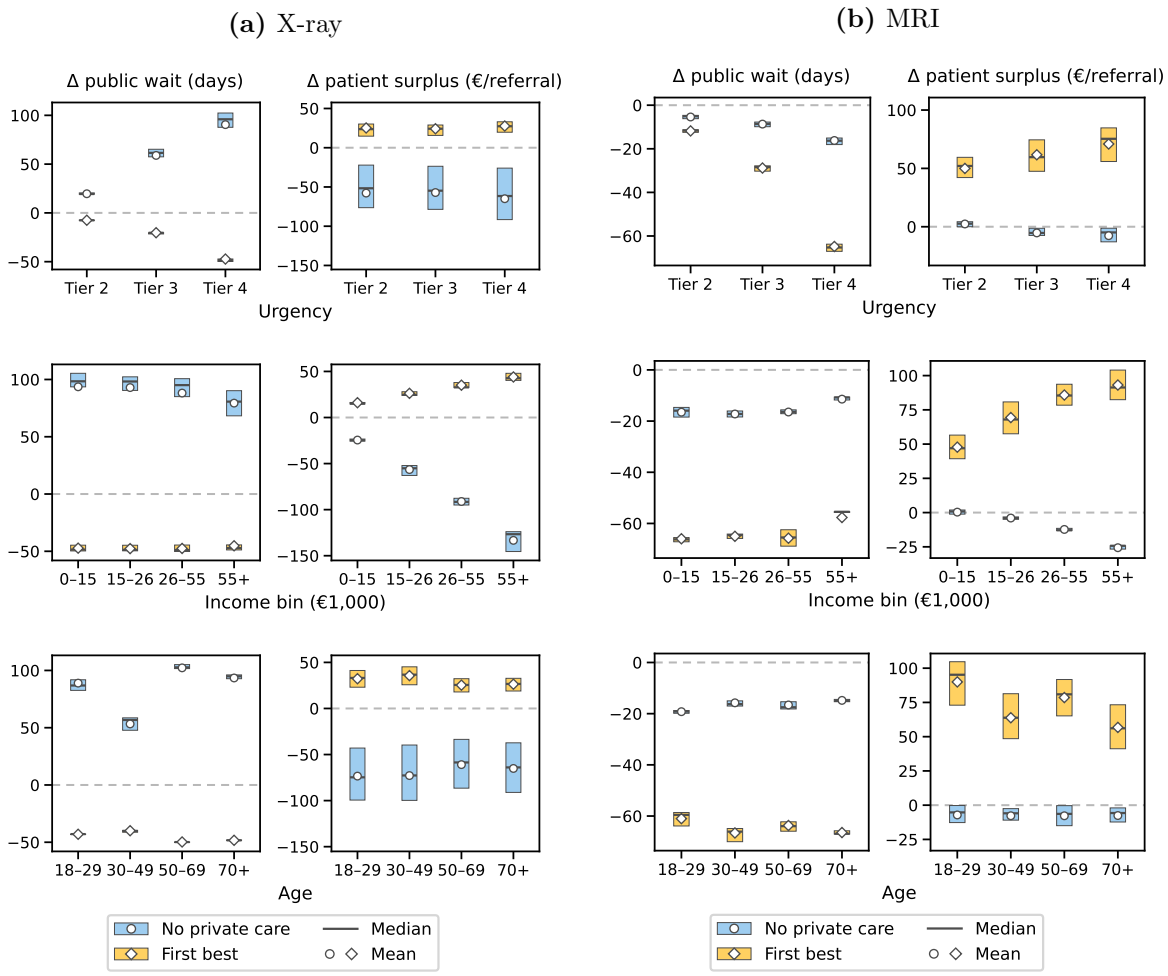
7.3 Distributional effects of private care

Given the heterogeneity in patient preferences, the welfare gains from private-pay care can be unevenly distributed – both across demographics and patients' propensity to use private-pay care. We examine this distributional tradeoff by measuring how the welfare of different patient groups would change in the decentralized equilibrium described in Section 7.1, where private-pay care is removed and providers adjust their capacity. Average patient surplus would fall across all urgency, income, and age groups, although with different magnitudes. Perhaps more importantly, removing private-pay care would hurt also those patients who are *unlikely* to use private-pay in the status quo. The result connects to the insight that diverting people away from congested public care can generate substantial welfare gains by reducing the wait-time externalities borne by public users.

Figure 4 shows how patient surplus and public wait times would change across urgency, income, and age groups when (i) private-pay options are removed, and (ii) the first best described in Section 7.2 is introduced. For ease of exposition, we focus on MRI and X-ray

[38] The budget is the total public spending observed in the status quo. Modality-specific budgets prevent the planner from reallocating resources across modalities. This is motivated by a "public health" institutional constraint enforcing that all modalities be worth offering. It prevents the perverse incentive to defund expensive modalities (e.g., MRI) to treat relatively more patients in cheaper ones (e.g., X-ray), which arises if patients across modalities contribute similarly to welfare.

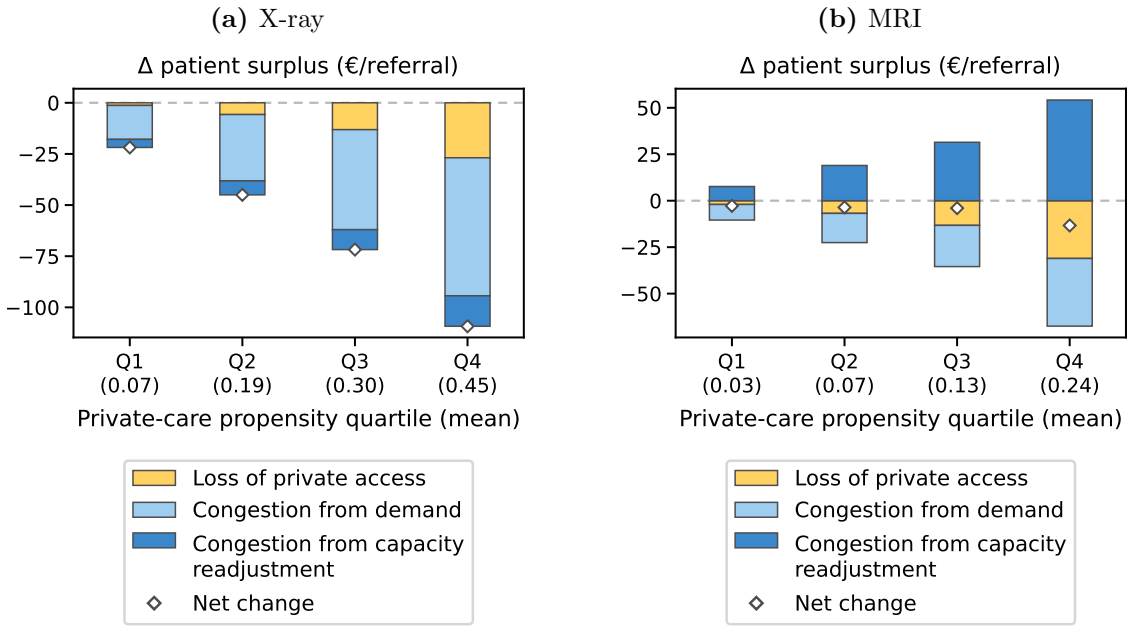
Figure 4 Distributional effects by demographics



Notes: Figure compares changes relative to the status quo under two counterfactual allocations: (i) removing private-pay care and allowing providers to readjust public capacity as described in Section 7.1 and reported in Table 7; and (ii) the first-best mixed provision allocation described in Section 7.2 and reported in Table 8. Panels (a) and (b) report X-ray and MRI results, respectively. The left column reports changes in expected public wait times, conditional on receiving publicly funded treatment. The right column reports changes in patient surplus per referral. Results are grouped by urgency, patient income, and age. Colored boxes span the referral-weighted 25th to 75th percentiles of changes across demographic subgroups. Horizontal lines indicate medians. To denote group means, circles are used for counterfactual (i) and diamonds for counterfactual (ii).

because removing private-pay care would induce opposite responses in public capacity, as discussed in Section 7 and illustrated by Figure 3. For X-ray, the contraction in capacity would increase public waits and reduce patient surplus across all demographic groups. For MRI, public waits would instead fall, but the resulting gains would only partially offset the loss of private-pay care, leaving patient surplus lower or nearly unchanged across most groups. The losses exhibit a pronounced income gradient, while differences across urgency and age groups

Figure 5 Distributional effects by private care usage



Notes: Figure decomposes the patient surplus changes from removing private-pay care by patients' propensity to use private care under the status quo. Patients are divided into four equally sized groups by their modeled probability of choosing private-pay treatment, with the mean probability reported in parentheses below each quartile. Bars decompose the change in patient surplus per referral into three components: (i) the direct loss from removing private-pay options keeping public waits at the status quo levels; (ii) the additional effect of congestion as patients re-sort and public queues re-equilibrate at the status quo capacity; and (iii) the additional changes in equilibrium waits due to providers readjusting the supplied capacity. Positive and negative contributions are stacked separately. Diamonds report the net change in patient surplus between the status quo mixed system and decentralized public-only provision.

are smaller. The first-best allocation would raise patient surplus and reduce public waits across all demographic groups, with particularly large gains among higher-income patients who would gain the most from the possibility of matching waits with willingness to pay.

Figure 5 shows how removing private care would affect patients based on their propensity to use private-pay options in the status quo. For X-ray, losses would rise sharply with this propensity, but even the patients who are the least likely to use private-pay options would lose around €20 per referral. Most of this loss would arise from additional congestion as patients shift from private-pay to public care, further amplified by providers' contraction of public capacity. For MRI, demand diversion would also increase public congestion, but the expansion of public capacity would offset much of this effect, leaving relatively small net losses even among patients most likely to use private-pay care. The mechanisms isolated in Figure 3 therefore combine to reduce patient surplus across the entire distribution of private-pay propensity, including among patients unlikely to use private-pay care.

8 Conclusion

We study how private-pay care affects welfare in a healthcare system with universal public coverage. Our estimates for diagnostic imaging show that private-pay care generates substantial benefits by allowing patients to sort according to their willingness to pay for shorter waits and reducing congestion in publicly funded care. Disallowing private-pay care would reduce patient surplus even among patients unlikely to use it. These benefits coexist with substantial inefficiencies in providers' pricing and capacity decisions. The socially optimal provision scheme would lower private prices and expand private capacity, attracting patients away from congested public queues. Capacity controls alone would instead achieve considerably smaller welfare gains because providers' prices would continue to limit this diversion.

Our results also suggest substantial gains from changing how publicly funded capacity is purchased. Financing capacity at production cost would require substantially less spending than purchasing slots at existing reimbursement rates to achieve the same welfare. This finding motivates further research on procurement mechanisms that account for providers' heterogeneous costs and incentives to invest. Providers' opportunities to serve private-pay patients would be central to this design problem, since they affect the terms on which providers would supply publicly funded capacity.

References

Agarwal, N., I. Ashlagi, M. A. Rees, P. Somaini, and D. Waldinger (2021): "Equilibrium Allocations Under Alternative Waitlist Designs: Evidence From Deceased Donor Kidneys," *Econometrica*, 89, 37–76.

Agarwal, N., C. Hodgson, and P. Somaini (2025): "Choices and Outcomes in Assignment Mechanisms: The Allocation of Deceased Donor Kidneys," *Econometrica*, 93, 395–438.

Akbarpour, M., P. Dworczak, and S. D. Kominers (2024): "Redistributive Allocation Mechanisms," *Journal of Political Economy*, 132, 1831–1875.

Almagro, M., F. Barbieri, J. C. Castillo, N. G. Hickok, and T. Salz (2026): "Optimal Urban Transportation Policy: Evidence from Chicago," *Mimeo*.

Atal, J. P., J. I. Cuesta, F. González, and C. Otero (2024): "The Economics of the Public Option: Evidence from Local Pharmaceutical Markets," *American Economic Review*, 114, 615–644.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Berry, S. T. (1994): "Estimating Discrete-Choice Models of Product Differentiation," *RAND Journal of Economics*, 25, 242–262.

Berry, S. T., and P. A. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797.

——— (2024): "Nonparametric Identification of Differentiated Products Demand Using Micro Data," *Econometrica*, 92, 1135–1162.

Besley, T., and S. Coate (1991): "Public Provision of Private Goods and the Redistribution of Income," *American Economic Review*, 81, 979–984.

Biglaiser, G., and C.-t. A. Ma (2007): "Moonlighting: Public Service and Private Practice," *RAND Journal of Economics*, 38, 1113–1133.

Blackorby, C., and D. Donaldson (1988): "Cash Versus Kind, Self-Selection, and Efficient Transfers," *American Economic Review*, 78, 691–700.

Bloch, F., and D. Cantala (2017): "Dynamic Assignment of Objects to Queuing Agents," *American Economic Journal: Microeconomics*, 9, 88–122.

Brancaccio, G., M. Kalouptsidi, and T. Papageorgiou (2025): "Investment in Infrastructure and Trade: The Case of Ports," *Mimeo*.

Brekke, K. R., and L. Sørgard (2007): "Public Versus Private Health Care in a National Health Service," *Health Economics*, 16, 579–601.

Burn, S. (2023): "Mixed Public-Private Provision in Healthcare: Evidence from England."

Canta, C., and M.-L. Leroux (2016): "Public and Private Hospitals, Congestion, and Redistribution," *Journal of Public Economic Theory*, 18, 42–66.

Che, Y.-K., and O. Tercieux (2026): "Optimal Queue Design," *Journal of Political Economy*, 134, 2395–2440.

Condorelli, D. (2013): "Market and Non-Market Mechanisms for the Optimal Allocation of Scarce Resources," *Games and Economic Behavior*, 82, 582–591.

Cook, C., and P. Z. Li (2025): "Value Pricing or Lexus Lanes? The Distributional Effects of Dynamic Tolling," *Mimeo*.

Cuff, K., J. Hurley, S. Mestelman, A. Muller, and R. Nuscheler (2012): "Public and Private Health-Care Financing with Alternate Public Rationing Rules," *Health Economics*, 21, 83–100.

Currie, J., and F. Gahvari (2008): "Transfers in Cash and In-Kind: Theory Meets the Data," *Journal of Economic Literature*, 46, 333–383.

Dinerstein, M., D. Morales, C. Neilson, and S. Otero (2023): "Equilibrium Effects of Public Provision in Education Markets," *Mimeo*.

Dinerstein, M., and T. D. Smith (2021): "Quantifying the Supply Response of Private Schools to Public Policies," *American Economic Review*, 111, 3376–3417.

Eizenberg, A. (2014): "Upstream Innovation and Product Variety in the U.S. Home PC Market," *The Review of Economic Studies*, 81, 1003–1045.

Elliott, J. T., G. V. Houngbonon, M. Ivaldi, and P. T. Scott (2025): "Market Structure, Investment, and Technical Efficiencies in Mobile Telecommunications," *Journal of Political Economy*, 133, 1401–1459.

Epple, D., and R. E. Romano (1998): "Competition Between Private and Public Schools, Vouchers, and Peer-Group Effects," *American Economic Review*, 88, 33–62.

Fabbri, D., and C. Monfardini (2009): "Rationing the Public Provision of Healthcare in the Presence of Private Supplements: Evidence from the Italian NHS," *Journal of Health Economics*, 28, 290–304.

Fan, Y., and C. Yang (2020): "Competition, Product Proliferation, and Welfare: A Study of the US Smartphone Market," *American Economic Journal: Microeconomics*, 12, 99–134.

Fondazione Censis, and Intesa Sanpaolo RBM Salute (2021): "IX Rapporto sulla Sanità Pubblica, Privata e Intermediata,"Technical report, https://www.sanitainformazione.it/wp-content/uploads/2021/02/IX-Rapporto-sulla-Sanita-Pubblica-Privata-e-Intermediata.pdf.

Gahvari, F., and E. Mattos (2007): "Conditional Cash Transfers, Public Provision of Private Goods, and Income Redistribution," *American Economic Review*, 97, 491–502.

Gentzkow, M. (2007): "Valuing New Goods in a Model with Complementarity: Online Newspapers," *American Economic Review*, 97, 713–744.

Goldszmidt, A., L. H. de Frahan, J. A. List, M. Mogstad, I. Muir, and A. Torgovitsky (2026): "Price Elasticities in Logit Demand Models," *Mimeo*.

Gravelle, H., M. Dusheiko, and M. Sutton (2002): "The Demand for Elective Surgery in a Public System: Time and Money Prices in the UK National Health Service," *Journal of Health Economics*, 21, 423–449.

Gravelle, H., and L. Siciliani (2008): "Ramsey Waits: Allocating Public Health Service Resources When There Is Rationing by Waiting," *Journal of Health Economics*, 27, 1143–1154.

Hoel, M., and E. M. Sæther (2003): "Public Health Care with Waiting Time: The Role of Supplementary Private Health Care," *Journal of Health Economics*, 22, 599–616.

Huitfeldt, I., V. Marone, and D. Waldinger (2026): "Designing Dynamic Reassignment Mechanisms: Evidence from GP Allocation," *American Economic Review*, 116, 3036–3075.

Iversen, T. (1997): "The Effect of a Private Sector on the Waiting Time in a National Health Service," *Journal of Health Economics*, 16, 381–396.

Jeon, J., and M. Rysman (2025): "Investment and Usage of the Subsea Internet Cable Network," *Mimeo*.

Kang, Z. Y., and P. Dworczak (2026): "Pareto-Improving Pricing: Why 3 Is Better Than 2," *Mimeo*.

Kang, Z. Y., and M. Watt (2026): "Topping Up and Optimal Redistribution," *Mimeo*.

Leshno, J. D. (2022): "Dynamic Matching in Overloaded Waiting Lists," *American Economic Review*, 112, 3876–3910.

Markkula, T. (2025): "Cutting the Queue to the Dentist: Waiting Time, Public-Private Interaction and Consumer Surplus," *Mimeo*.

Martin, S., N. Rice, R. Jacobs, and P. C. Smith (2007): "The Market for Elective Surgery: Joint Estimation of Supply and Demand," *Journal of Health Economics*, 26, 263–285.

Mendelson, H., and S. Whang (1990): "Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue," *Operations Research*, 38, 870–883.

Miravete, E. J., K. Seim, and J. Thurk (2026): "Elasticity and Curvature of Discrete Choice Demand Models," *Mimeo*.

Moscelli, G., L. Siciliani, N. Gutacker, and H. Gravelle (2016): "Location, Quality and Choice of Hospital: Evidence from England 2002–2013," *Regional Science and Urban Economics*, 60, 112–124.

Naor, P. (1969): "The Regulation of Queue Size by Levying Tolls," *Econometrica*, 37, 15–24.

Nichols, A. L., and R. J. Zeckhauser (1982): "Targeting Transfers Through Restrictions on Recipients," *American Economic Review*, 72, 372–377.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2015): "Moment Inequalities and Their Application," *Econometrica*, 83, 315–334.

Riganti, A., L. Siciliani, and C. V. Fiorio (2017): "The Effect of Waiting Times on Demand and Supply for Elective Surgery: Evidence from Italy," *Health Economics*, 26, 92–105.

Russo, A. (2024): "Waiting or Paying for Healthcare: Evidence from the Veterans Health Administration," *Mimeo*.

Siciliani, L., and R. Verzulli (2009): "Waiting Times and Socioeconomic Status Among Elderly Europeans: Evidence from SHARE," *Health Economics*, 18, 1295–1306.

Sullivan, M. R. (2026): "Identifying Demand with Network Externalities," *Mimeo*.

Waldfogel, J. (2003): "Preference Externalities: An Empirical Study of Who Benefits Whom in Differentiated-Product Markets," *The RAND Journal of Economics*, 34, 557–568.

Waldinger, D. (2021): "Targeting In-Kind Transfers through Market Design: A Revealed Preference Analysis of Public Housing Allocation," *American Economic Review*, 111, 2660–2696.

Wollmann, T. G. (2018): "Trucks without Bailouts: Equilibrium Product Characteristics for Commercial Vehicles," *American Economic Review*, 108, 1364–1406.

Yang, F., P. Dworczak, and M. Akbarpour (2025): "Comparison of Screening Devices," *Mimeo*.

Appendix contents

A Data and patient sample construction **44**

A.1 Administrative data 44

A.2 Scraped and auxiliary data 46

A.3 Simulated patient sample 49

B Additional descriptive evidence **49**

B.1 Summary statistics 49

B.2 Capacity responses to realized wait times 51

B.3 Private wait times across scans within a modality 51

C Theory **52**

C.1 Queuing model 52

C.2 Production of capacity 54

C.3 Provider optimality conditions 55

C.4 Relation to alternative models of wait times 57

D Estimation **60**

D.1 Queuing model and computation of the stationary distribution 60

D.2 Approximation of the stationary distribution of waits 61

D.3 Private capacity 65

D.4 Computation of the GMM criterion 69

D.5 Patient choice micromoments 72

D.6 Differentiation 74

D.7 Slot costs 75

D.8 Scanner fixed costs 76

D.9 Inference 78

D.10 Imputation of unscraped private prices and wait times 80

E Estimation results **80**

F Counterfactuals **87**

F.1 Setup 87

F.2 Disallowing private-pay care 90

F.3 Efficient mixed systems 91

F.4 Computation 92

F.5 Inference on counterfactuals 98

G Proofs **98**
G.1 Existence of a stationary distribution of wait times 98
G.2 Invertibility of demand 101
G.3 Identification of private capacity 104

H Miscellaneous **106**
H.1 Queue-level arrival objects 106
H.2 Total derivative of arrivals 107
H.3 Second-order approximation of profits 107
H.4 An alternative validation of the queuing model 109

A Data and patient sample construction

↩ Main text: §2.2 Setting and data / Data

↩ See also: §B.1 Descriptives / Summary statistics

This appendix provides additional details on the construction of the analysis sample, the auxiliary datasets used in the empirical analysis, and the simulated patient sample used in estimation.

A.1 Administrative data

↩ Main text: §2.2 Setting and data / Data

↩ See also: §A.3 Data / Simulated patient sample · §D.5 Estimation / Patient choice micromoments

The administrative data contain the universe of referrals issued by the regional healthcare system in 2023 and 2024 and the publicly funded scans linked to those referrals. Each referral records the date of issuance, the prescribed scan, the patient's town of residence, age, and urgency tier. When a publicly funded scan is subsequently performed, we observe its date and the provider that delivered it. All publicly funded scans are observed. The referral identifier allows us to link referrals to subsequent publicly funded care.

Sample construction Our main analysis uses referrals issued in 2024. We construct the analysis sample according to the following steps.

1. *Scan selection* We retain selected common scans belonging to the X-ray, CT, and MRI modalities and exclude scans associated with screening programs, for which the patient decision problem differs from the one studied in the paper. Table A1 reports the full list of scans included in the sample. After excluding mammography, which is primarily associated with a population screening program, the selected scans account for 97% of referrals across the three modalities in 2024.

2. *Patient selection* We restrict attention to adult patients registered as residents of the region we study. This allows us to match towns to the geographic and demographic data used to construct travel times and choice sets. We exclude patients that moved to and currently live in the region we study but are still registered as residents of their previous hometown.

3. *Completion outcome* We classify a referral as *completed* if it generates a publicly funded scan by February 2026 and as *lapsed* otherwise.

4. *Other outcomes* For publicly completed referrals, we define their realized wait time as the number of days between the referral and scan dates. We retain the provider that performed the scan.

5. *Demographics* We map each scan to its modality, regulated co-pay, and provider reimbursement. We merge the patient's town of residence with its income distribution and the travel time to the provider that performed the scan.

Queue flows The main choice analysis uses referrals issued in 2024, but we use referrals from both 2023 and 2024 to construct the history of public queue flows. Consistent with the model, we define a queue as a provider-modality-urgency combination. Since we only observe the chosen provider for patients using publicly funded care, we restrict attention to public queues. We map each publicly completed referral i to the provider $h(i)$ that performed that scan and to the week $t(i)$ in which it was issued. Denoting by $u(i)$ the urgency attached to referral i, we assume referral i arrives to queue $q = (h(i), \text{pub}, u(i))$ in week $t(i)$. The number of arrivals into queue $q = (h, \text{pub}, u)$ in week t is then the number of referrals i with $h(i) = h$, $u(i) = u$, and $t(i) = t$. The throughput of queue $q = (h, \text{pub}, u)$ in week t is instead the observed number of publicly funded procedures delivered by h to patients of urgency u in week t. These weekly arrival and completion counts are used to reconstruct the queue backlog over time and validate the wait model as described in Section 5.1.

Protected-environment outputs The patient-level administrative records are accessible only within a protected data environment. We never accessed the data directly, and only submitted scripts that government officials with the required clearance executed in the protected environment. The scripts reduced the patient-level records to exportable aggregate objects sufficient for estimation and inference.

We export the following objects.

1. *Referral and scan composition* For 2024 referrals, we export the joint distribution of modality, urgency, age group, and province[39] of residence, together with the distribution of individual scans within modality and urgency.

2. *Public provider choices* For publicly completed referrals, we export their allocation across providers by scan and urgency.

3. *Patient-choice moments* By modality, province of residence, referral month, age group, sex, and urgency, we export the share of lapsed referrals and, conditional on public completion, mean realized wait and travel time. We also export interactions of these outcomes with the income distribution of the patient's town, using town-level income-bin probabilities.

[39] A "province" is one of 12 geographical subregions, each comprising several municipalities. They are typically clustered around large cities and their population goes from roughly 180,000 to over 3 million inhabitants (median: 570,000; standard deviation: 800,000).

4. *Moment covariance* We export referral counts, sums of the referral-level quantities underlying the patient-choice moments, and all pairwise sums of products by modality, province of residence, age group, sex, and urgency. These objects allow us to reconstruct the sampling covariance of the moments outside the protected environment.

5. *Queue flows* Using referrals and scans from 2023 and 2024, we export weekly arrivals and completions by provider, modality, and urgency following the procedure described above.

Table A1 Scans included in the analysis

X-ray	CT	MRI
Abdomen	Abdomen*	Abdomen*
Chest	Upper	Upper
Dental arch*	Lower	Lower and pelvis
Esophagus*	Whole	Brain*
Face	Chest*	Breast*
GI tract	Dental arch	Chest*
Upper	Ear*	Face*
Whole	Face*	Heart*
Head and neck soft tissues	Head*	Joints
Hip and pelvis	Kidneys*	Musculoskeletal system*
Lower limb	Lower limb*	Neck*
Lower limb and pelvis	Neck*	Spine*
Patella	Pelvis	
Pituitary fossa	Salivary glands	
Ribs and sternum	Spine*	
Salivary glands	Upper limb*	
Skull and sinuses		
Spine		
Stomach and duodenum		
Trachea		
Upper limb		
Urinary system		
Whole skeleton		

Notes: Table reports the list of scans used in the analysis.
*The scan description combines separate variants performed with and without contrast.
(Click here to return to the data section in the main text.)

A.2 Scraped and auxiliary data

↩ Main text: §2.2 Setting and data / Data

↩ See also: §B.3 Descriptives / Within-modality waits · §D.10 Estimation / Imputation

Scraped data We constructed a dataset with weekly observations of private-pay prices and appointment availability from online booking systems during 2024. We submitted queries on weekends to all accredited providers located in the region of interest that had an online booking system. For each such provider, we tried booking an appointment for all X-ray, CT, and MRI scans listed in the booking system. Figure A1 shows an example of an online booking system.

Each observation in the constructed dataset records provider, scan, posted private-pay price, date of the query, and first available appointment. Private wait time is defined as the number of days between the query and the first available appointment. According to this definition, different scans in the same modality can have different wait times, possibly reflecting heterogeneity in scan duration and scheduling idiosyncrasies. Since our model assumes that wait times are defined at the modality level, whenever such discrepancy arises for some provider-modality-week combination, we use the median observed wait time across scans. Appendix B.3 shows that the majority of scraped wait times are very similar within each week-modality. We manually matched scraped provider names to the providers appearing in the administrative data.

Travel times We construct driving times between every town in the region and every provider location using the OpenStreetMap road network. The same travel-time measure is merged to publicly completed referrals when constructing the observed travel moments as described in Section A.1.

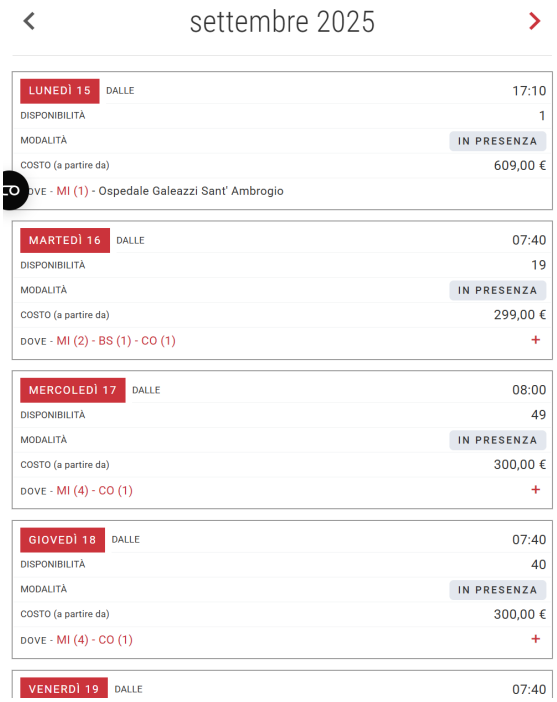
Population and urbanization We obtained resident population counts by municipality, age, and sex from National Statistics Institute (Istat).[40] We aggregate age to the groups used in the empirical model. We also use Istat's municipality-level degree of urbanization, which classifies municipalities as having low, medium, or high urbanization based on population density and settlement patterns.[41] Tiers correspond to rural areas, suburbs, and cities.

Income We obtained municipality-level income information from the Ministry of Economy and Finance statistics on personal income-tax declarations.[42] The source reports numbers of taxpayers and declared income by bin for each municipality. We use these counts to construct the share of taxpayers with declared income below €15,000, between €15,000 and €26,000,

[40] Accessed here on September 7, 2026.
[41] Accessed here on September 7, 2026.
[42] Accessed here on September 7, 2026.

Figure A1 Online booking for a large provider

(a) Query on September 15, 2025



(b) Selecting a slot for September 16, 2025



Notes: Figure reports screenshots from the booking system of a large, multi-facility provider upon submitting a request for an abdomen CT scan on September 15, 2025. Panel (a) shows the available dates. Panel (b) shows the page that opens after selecting September 16 as the desired day. The slots in panel (b) represent appointments at different facilities owned by the same provider. For each facility we can infer the first date in which an appointment at that facility is available. The wait time is then the number of days between the query date and the first availability date.

between €26,000 and €55,000, and above €55,000. We then associate each referral and simulated patient with the income distribution of the corresponding municipality, and use these probabilities when constructing the income interactions described in Section A.1.

Wages We obtained province-level gross hourly earnings from Istat's register of private-sector employment and earnings.[43] Each provider is assigned the hourly wage for workers with a college degree of the province in which it is located.

Procedure codes and regulated payments We map the procedure codes appearing in the administrative and scraped data to a common scan definition and to the X-ray, CT, or

[43] Accessed here on September 7, 2026.

MRI modality using the regional nomenclature for outpatient specialist care. The regional tariff associated with each procedure determines the reimbursement paid to the provider for publicly funded delivery. For non-exempt patients, the public co-pay is the procedure tariff subject to the regional cap of €36 per referral.

A.3 Simulated patient sample

↩ Main text: §5.3 Estimation / Private capacity, patient demand, and scan costs

↩ See also: §A.1 Data / Administrative data · §A.2 Data / Scraped and auxiliary data

We simulate a patient population from the exported 2024 referral distribution and public-use population data. The construction proceeds in three steps.

1. *Referral groups* Let N_g denote the number of referrals observed in the administrative data with combination g of modality, urgency, sex, age group, and province of residence.
2. *Town assignment* We distribute N_g across towns within the corresponding province according to the population distribution of residents with the same age and sex.
3. *Scan assignment* We distribute the resulting mass across individual scans according to their observed frequencies within the corresponding modality and urgency class.

The resulting mass assigned to referral group g, town ℓ, and scan s is

$$M_{g\ell s} = N_g \Pr(\ell \mid \text{province}_g, \text{age}_g, \text{sex}_g) \Pr(s \mid m_g, u_g).$$

This construction preserves the joint distribution of modality, urgency, sex, age group, and province observed among referrals, while using population data to recover within-province geography. Within each referral group, town and scan are assigned from separate conditional distributions.

For estimation, we draw 10,000 simulated patients with replacement with probabilities proportional to $M_{g\ell s}$. Each simulated patient is therefore characterized by a scan $s(i)$, modality $m(i)$, demographic group $g(i)$, and location $\ell(i)$. We combine these characteristics with provider availability, travel times, public co-pays, private-pay prices, and wait times to construct the patient's choice set and the attributes of each available option.

B Additional descriptive evidence

B.1 Summary statistics

↩ Main text: §3 Descriptives

↩ See also: §A.1 Data / Administrative data

Table B1 reports summary statistics of the data we use.

Table B1 Summary statistics

	Overall	Modality		
		X-ray	CT	MRI
Referrals[a]				
Issued	5,615,766	3,134,616	1,208,509	1,272,641
share	1.00	0.56	0.22	0.23
Publicly completed	3,415,411	1,970,273	804,579	640,559
Public completion rate	0.61	0.63	0.67	0.50
Scans	78	30	27	21
Patient demographics[b]				
Male	0.43	0.39	0.50	0.45
Urgency tier				
1	0.04	0.05	0.03	0.01
2	0.12	0.11	0.15	0.13
3	0.21	0.20	0.19	0.28
4	0.63	0.64	0.64	0.58
Age bin				
18–29	0.05	0.05	0.03	0.07
30–39	0.06	0.06	0.04	0.09
40–49	0.11	0.11	0.08	0.15
50–59	0.20	0.20	0.16	0.24
60–69	0.22	0.22	0.24	0.22
70–79	0.22	0.22	0.28	0.17
80+	0.14	0.15	0.18	0.07
Supply[c]				
Providers				
total	279	272	192	172
private	173	167	113	118
scraped	79	73	51	55
Share of public patients				
at private	0.59	0.58	0.53	0.71
at scraped	0.33	0.31	0.33	0.42

Notes: Table reports descriptive statistics of the sample used for the analysis.
[a]This panel reports basic facts about the referrals used in the analysis. A referral is "publicly completed" if it is followed by a subsequent publicly funded scan. Scans within a modality are differentiated by the body part that needs to be scanned. The analysis sample is restricted to referrals issued in 2024 to residents for a subset of common scans. Sample construction details are provided in Appendix A.
[b]This panel reports the demographics of patients as shares conditional on modality. Patients are defined as individuals receiving a referral.
[c]This panel reports providers offering each modality. A provider is classified as offering a modality if it delivered at least 100 publicly funded scans in such modality in 2024.
(Click here to return to the descriptives section in the main text.)

B.2 Capacity responses to realized wait times

↩ Main text: §4 Model

↩ See also: §D.1 Estimation / Queuing model · §C.2 Theory / Production of capacity

The model in Section 4 assumes that providers choose capacity before the short-run path of wait times is realized. This timing rules out a feedback response in which providers observe a high realized wait and subsequently expand capacity. We assess whether such a response is visible in weekly variation within provider-modality cells.

Focusing on lowest-urgency queues, we isolate residual variation in realized waits by estimating

$$w_{hmt} = \alpha_{hm} + \gamma_{mt} + \delta k_{hmt} + u_{hmt}, \tag{B1}$$

where w_{hmt} is the realized public wait, k_{hmt} is contemporaneous capacity, and α_{hm} and γ_{mt} are provider-modality and modality-by-week fixed effects. The residual $\hat{u}_{hmt}$ measures deviations from waits predicted by these controls. For horizons $s \in \{1, 2, 4, 8\}$, we then estimate

$$100 \times \frac{k_{hm\,t+s} - k_{hmt}}{\bar{k}_{hm}} = \beta_s \frac{\hat{u}_{hmt}}{10} + \alpha_{hm} + \gamma_{mt} + \varepsilon_{hm\,t+s}, \tag{B2}$$

where $\bar{k}_{hm}$ is mean provider-modality capacity, so that β_s measures the percentage-point change in relative capacity associated with a ten-day wait residual.

Under a feedback rule in which providers expand capacity after observing unusually high waits, we would expect $\beta_s > 0$, possibly with a delayed response. Instead, Table B2 reports negative coefficients at every horizon, providing no evidence of subsequent expansion in measured public capacity. These estimates are of course not causal, since residual waits may reflect provider-specific conditions that also affect future capacity. For example, scheduled maintenance or staffing shortages may raise current waits and reduce future capacity, even if providers add sessions in response. Alternatively, temporarily expanded capacity may already be reflected in k_{hmt} and subsequently withdrawn as backlogs clear. Providers may also respond by reallocating slots toward private-pay patients, reducing measured public capacity even if total capacity increases. The results are consistent with the model's assumption that capacity is predetermined with respect to short-run wait realizations.

B.3 Private wait times across scans within a modality

↩ Main text: §4.2 Model / Queues §4.3 Model / Supply

↩ See also: §A.2 Data / Scraped and auxiliary data · §C.1 Theory / Queuing model

Our model assigns a common private queue to scans in the same modality at a given provider. To assess the quality of this approximation, we compare first-available appointment waits for

Table B2 Capacity responses to unexpected waits

	Δ future capacity, by horizon			
	1	2	4	8
Unexpected wait (10 days)	-1.20	-2.34	-3.74	-4.80
	(0.12)	(0.22)	(0.38)	(0.67)
Providers	281	282	281	281
provider-modality	646	648	645	645
provider-modality–week	16,011	15,385	14,141	11,626
Provider × modality FE	Yes	Yes	Yes	Yes
Modality × week FE	Yes	Yes	Yes	Yes

Notes: Table reports estimates of the relationship between unexpected public wait times and subsequent changes in public capacity in (B2). The dependent variable is the change in relative capacity between the current week and the indicated horizon. Unexpected wait is the residual from a regression (B1) of realized wait on contemporaneous capacity, provider-modality fixed effects, and modality-by-week fixed effects. Coefficients report the percentage-point change in capacity, relative to average provider-modality capacity, associated with a 10-day increase in unexpected wait. Provider-clustered standard errors are reported in parentheses. The sample covers weeks from January 7 through July 7, 2024.
(Click here to return to the model section in the main text.)

different scans in the same modality, observed at the same provider on the same scraping date. Table B3 shows that waits are identical for 67% of MRI scan pairs, 76% of CT scan pairs, and 83% of X-ray scan pairs. Moreover, 81–92% of scans have scraped waits within one day of the median across scans in the same modality at the same provider and scraping occurrence. These patterns support the common-queue approximation.

C Theory

C.1 Queuing model

↩ Main text: §4.2 Model / Queues · §5.1 Estimation / Queuing model

↩ See also: §D.1 Estimation / Queuing model and approximations

We describe the evolution of queues and generalize the wait-time mapping in Section 4.2 to allow for exogenous fluctuations in weekly capacity.

Patients are assigned to queues according to their chosen provider, access, and modality. Publicly funded patients are additionally triaged by urgency. Publicly funded queues are indexed by (h, pub, m, u) and private-pay queues by (h, pvt, m).

Providers choose mean public and private capacity within each modality before the realization of patient arrivals. Within publicly funded care, the allocation across urgency classes

Table B3 Private wait times across scans within a modality

	MRI	CT	X-ray
Scan pairs with identical waits (%)	67	76	83
Scan pairs with waits differing by at most three days (%)	83	88	91
Scans with waits within one day of the median (%)	81	88	92
Distinct scans	22	27	32
Provider-week observations	1,387	1,584	1,983

Notes: Comparisons use first-available appointment waits for scans in the same modality observed at the same provider on the same scraping occurrence. Each provider-week observation includes at least two distinct scans. The third row reports the percentage of scan observations with waits within one day of the median across scans in the same modality at that provider and week. Scans without an observed wait are excluded.
(Click here to return to the model section in the main text.)

is governed by exogenous weights $\rho_{hmu} \geqslant 0$, with $\sum_u \rho_{hmu} = 1$. The mean weekly capacity allocated to queue q is

$$k_{qt} = \begin{cases} \rho_{hmu} k_{h\,\mathrm{pub}\,mt}, & \text{if } q = (h, \mathrm{pub}, m, u), \\ k_{h\,\mathrm{pvt}\,mt}, & \text{if } q = (h, \mathrm{pvt}, m). \end{cases}$$

The capacity paths $(k_{qt})_t$ may fluctuate over the year because of seasonal conditions, holidays, staffing availability, or other exogenous reasons. We rule out adjustments to these paths in response to realizations of patient arrivals or backlogs.

The definition of throughput and the backlog transition are slightly adapted from Section 4.2 to allow for capacity fluctuations. Throughput in week t at queue q is $y_{qt} = \min\{b_{qt} + a_{qt}, k_{qt}\}$, so that either all patients (queuing and arriving) are treated, or the available slots are exhausted. The backlog evolves according to

$$b_{q,t+1} = b_{qt} + a_{qt} - y_{qt} = \max\{b_{qt} + a_{qt} - k_{qt}, 0\}. \tag{C1}$$

A patient booking an appointment at the beginning of week t faces the backlog b_{qt} inherited from previous weeks. Under first-come, first-served treatment, their first-availability wait is the time required to clear this backlog using the slots available from week t onward. We assume that capacity is distributed uniformly within each week, so that a fraction of a week's duration corresponds to the same fraction of its available slots. The first-availability wait in

week t at queue q is[44]

$$w_{qt} := \inf\left\{x > 0 : \sum_{r=0}^{\lfloor x \rfloor - 1} k_{q,t+r} + \big(x - \lfloor x \rfloor\big) k_{q,t+\lfloor x \rfloor} > b_{qt}\right\}. \tag{C2}$$

The interpolation within each week allows wait times to take noninteger values and avoids discontinuities associated with rounding waits up to whole weeks. When capacity is constant, $k_{qt} = k_q > 0$ for every t, equation (C2) reduces to the formula reported in Section 4.2:

$$w_{qt} = \inf\{x > 0 : x k_q > b_{qt}\} = \frac{b_{qt}}{k_q}.$$

We assume that the future capacity path relevant for determining first availability is known when patients book their appointments. This is consistent with capacity fluctuations not responding to realized congestion.

Let $\mathcal{I}_t$ denote the random set of patients receiving referrals in week t. As in Section 4.1, patient i joins queue $q(i, j)$ when booking at provider-access j, based on their modality and urgency. Conditional on prices $\boldsymbol{p}$ and the vector of waits $\boldsymbol{w}_t$, the probability that patient i joins queue q is

$$\sigma_{iq}(\boldsymbol{p}, \boldsymbol{w}_t) = \sum_{j \in \mathcal{J}_i} \mathbf{1}\{q(i, j) = q\} \sigma_{ij}(\boldsymbol{p}, \boldsymbol{w}_t).$$

Aggregating individual choices gives weekly arrivals

$$a_{qt} = \int_{\mathcal{I}_t} \sigma_{iq}(\boldsymbol{p}, \boldsymbol{w}_t) \, \mathrm{d}i. \tag{C3}$$

Equations (C1), (C2), and (C3) jointly characterize the evolution of the queuing system. Inherited backlogs and predetermined capacity paths determine first-availability waits, which affect patient choices and arrivals, and hence subsequent backlogs.

Appendix G.1 establishes existence of a stationary distribution under constant capacities $k_q > 0$ and iid referral sets $\mathcal{I}_t$. Under these assumptions, the backlog vector $\boldsymbol{b}_t = (b_{qt})_{q \in \mathcal{Q}}$ is a time-homogeneous Markov process. As discussed there, neither assumption is crucial, as dependence of both capacity and referrals on an exogenous Markov process can be allowed.

C.2 Production of capacity

↩ Main text: §4.3 Model / Supply · §5.4 Estimation / Capacity and scanner costs

↩ See also: §C.3 Theory / Provider optimality conditions

Section 4.3 treats $\Phi_h(\boldsymbol{K}_h, \boldsymbol{n}_h)$ as the operating cost of producing capacity conditional on

[44] We use the convention that an empty sum is equal to zero.

inputs. This section derives the properties imposed on Φ_h from standard restrictions on the production technology. Since the argument applies provider by provider, we suppress the provider index h. For each modality m, let $\boldsymbol{x}_m \in \mathbb{R}_+^{L_m}$ denote divisible operating inputs and let $\boldsymbol{z}_m \in \mathbb{R}_{++}^{L_m}$ denote their prices. Technology $\boldsymbol{\Omega} : (\boldsymbol{n}, \boldsymbol{x}) \mapsto \boldsymbol{K}$ maps vectors of scanners and inputs to total capacities in each modality.

Conditional on scanner counts, the provider chooses operating inputs to produce a target capacity vector at minimum cost:

$$\Phi(\boldsymbol{K}, \boldsymbol{n}) \coloneqq \min_{(\boldsymbol{x}_m)_{m \in \mathcal{M}}} \sum_{m \in \mathcal{M}} \boldsymbol{z}_m \cdot \boldsymbol{x}_m \quad \text{sub} \quad \boldsymbol{\Omega}(\boldsymbol{n}, \boldsymbol{x}) \geqslant \boldsymbol{K}.$$

Differentiability of the joint cost function follows from familiar producer theory arguments under standard assumptions (Mas-Colell et al., 1995, Section 5.C).

Costs separate across modalities if production does, i.e., if $\boldsymbol{\Omega}(\boldsymbol{n}, \boldsymbol{x})$ gives capacity in modality m according to $\Omega_m(n_m, \boldsymbol{x}_m)$. In this case, the modality-specific cost functions are

$$\Phi_m(K, n) \coloneqq \min_{\boldsymbol{x} \geqslant \boldsymbol{0}} \left\{ \boldsymbol{z}_m \cdot \boldsymbol{x} : \Omega_m(n, \boldsymbol{x}) \geqslant K \right\}.$$

Given that both the objective and the constraints of the provider's cost-minimization problem are separable across modalities, we also have $\Phi(\boldsymbol{K}, \boldsymbol{n}) = \sum_{m \in \mathcal{M}} \Phi_m(K_m, n_m)$.

Finally, we impose constant returns on the production technology, i.e., that, for every modality m, input bundle $\boldsymbol{x}$, and scalar $t > 0$, $\Omega_m(tn, t\boldsymbol{x}) = t\Omega_m(n, \boldsymbol{x})$. If $\boldsymbol{x}$ can produce K slots using n scanners, then $t\boldsymbol{x}$ can produce tK slots using tn scanners. Conversely, any input bundle feasible for (tK, tn) can be divided by t to obtain a bundle feasible for (K, n). Since expenditure is linear in operating inputs, these two observations imply $\Phi_m(tK, tn) = t\Phi_m(K, n)$. For $n > 0$, define the one-scanner cost function as $\phi_m(q) \coloneqq \Phi_m(q, 1)$. Homogeneity of Φ_m then gives $\Phi_m(K, n) = n\Phi_m(K/n, 1) = n\phi_m(K/n)$.

C.3 Provider optimality conditions

↩ Main text: §4.3 Model / Supply · §5.3 Estimation / Private capacity, patient demand, and scan costs · §5.4 Estimation / Capacity and scanner costs

↩ See also: §D.4 Estimation / Construction of the objective · §D.6 Estimation / Differentiation · §H.3 Miscellaneous / Profit derivatives

We derive the optimality conditions associated with the provider problem in (6). We first consider prices and capacities conditional on scanner counts, then characterize the optimal discrete choice of scanners. To allow for common ownership across providers, we introduce the ownership matrix $\mathcal{H}$, which is defined to be such that $\mathcal{H}_{hh'} = 1$ if providers h and h' share the same owner and zero otherwise.

Private-pay prices Differentiating operating profits with respect to price $p_{h\,\mathrm{pvt}\,s}$ gives the necessary condition

$$0 = a_{h\,\mathrm{pvt}\,s} + \sum_{h'} \mathcal{H}_{hh'} \sum_{s' \in \mathcal{S}} \left[(p_{h'\,\mathrm{pvt}\,s'} - c_{h's'}) \frac{\mathrm{d}a_{h'\,\mathrm{pvt}\,s'}}{\mathrm{d}p_{h\,\mathrm{pvt}\,s}} + (r_{s'} - c_{h's'}) \frac{\mathrm{d}a_{h'\,\mathrm{pub}\,s'}}{\mathrm{d}p_{h\,\mathrm{pvt}\,s}} \right]. \tag{C4}$$

The first term is the additional revenue collected from patients purchasing scan s at the higher price, while the remaining terms capture the resulting changes in private-pay and publicly funded demand. These changes include substitution between access and providers, as well as the feedback from demand to stationary congestion. Because patients requiring different scans within the same modality share queues, a change in the price of one scan can also affect demand for other scans through its impact on wait times. The derivatives in (C4) include both direct choice responses and responses mediated by the stationary distribution of waits, as formalized in Appendix H.2.

Capacity choices Fix scanner counts $\boldsymbol{n}_h$ and let $K_{hm} = k_{h\,\mathrm{pub}\,m} + k_{h\,\mathrm{pvt}\,m}$ denote total capacity in modality m. The capacity cost function derived in Appendix C.2 satisfies

$$\frac{\partial \Phi_h}{\partial k_{h\,\mathrm{pub}\,m}} = \frac{\partial \Phi_h}{\partial k_{h\,\mathrm{pvt}\,m}} = \phi'_{hm}(K_{hm}/n_{hm}).$$

For either access mode $e \in \{\mathrm{pub}, \mathrm{pvt}\}$, the marginal effect of capacity on operating profits is

$$\sum_{h'} \mathcal{H}_{hh'} \frac{\partial \Pi_{h'}}{\partial k_{hem}} = \sum_{h'} \mathcal{H}_{hh'} \sum_{s \in \mathcal{S}} \left[(p_{h'\,\mathrm{pvt}\,s} - c_{h's}) \frac{\mathrm{d}a_{h'\,\mathrm{pvt}\,s}}{\mathrm{d}k_{hem}} + (r_s - c_{h's}) \frac{\mathrm{d}a_{h'\,\mathrm{pub}\,s}}{\mathrm{d}k_{hem}} \right].$$

Unlike prices, capacity has no direct effect on patient utility, so its effect on demand operates through changes in the stationary distribution of wait times. Additional capacity attracts patients by reducing congestion in the queues to which it is allocated, but may also change arrivals to the provider's other queues as patients substitute between access modes.

Let $\mu_h \geqslant 0$ denote the multiplier on the reimbursement constraint (5). At an interior capacity allocation, the necessary optimality conditions are

$$\sum_{h'} \mathcal{H}_{hh'} \frac{\partial \Pi_{h'}}{\partial k_{h\,\mathrm{pvt}\,m}} = \phi'_{hm}(K_{hm}/n_{hm}), \qquad \sum_{h'} \mathcal{H}_{hh'} \frac{\partial \Pi_{h'}}{\partial k_{h\,\mathrm{pub}\,m}} = \phi'_{hm}(K_{hm}/n_{hm}) + \mu_h \bar{r}_m. \tag{C5}$$

The multiplier satisfies the complementary-slackness condition

$$\mu_h \left(R_h - \sum_m \bar{r}_m k_{h\,\mathrm{pub}\,m} \right) = 0, \qquad \mu_h \geqslant 0.$$

The first condition in (C5) equates the marginal operating profit from an additional private-pay slot to its marginal production cost. The second condition imposes the same equality

for publicly funded capacity, accounting for the shadow cost of the reimbursement budget. Subtracting the two conditions gives

$$\sum_{h'} \mathcal{H}_{hh'} \left[\frac{\partial \Pi_{h'}}{\partial k_{h\,\mathrm{pub}\,m}} - \frac{\partial \Pi_{h'}}{\partial k_{h\,\mathrm{pvt}\,m}} \right] = \mu_h \bar{r}_m.$$

When the reimbursement constraint is slack, $\mu_h = 0$, so the marginal operating profits of public and private capacity are equalized within each modality. When the constraint binds, publicly funded capacity must yield additional operating profits sufficient to compensate for its use of scarce reimbursement budget. Moreover, the difference in marginal returns divided by $\bar{r}_m$ is equalized across modalities with interior capacity allocations. At capacity boundaries, the corresponding equalities are replaced by the usual complementary-slackness inequalities.

Scanner choices The number of scanners is discrete, so its optimality is characterized by inequalities rather than first-order conditions. For fixed scanner counts $\boldsymbol{n}_h$, let $V_h(\boldsymbol{n}_h)$ denote the maximum operating profits net of capacity costs, with prices and capacities reoptimized subject to the reimbursement constraint and other providers' choices held fixed.

$$V_h(\boldsymbol{n}_h) = \max_{\substack{\boldsymbol{p}_h, \boldsymbol{k}_h \\ \sum_m \bar{r}_m k_{h\,\mathrm{pub}\,m} \leqslant R_h}} \left\{ \sum_{h'} \mathcal{H}_{hh'} \Pi_{h'}(\boldsymbol{p}_h, \boldsymbol{p}^*_{-h}, \boldsymbol{k}_h, \boldsymbol{k}^*_{-h}) - \Phi_h(\boldsymbol{k}_{h\,\mathrm{pub}} + \boldsymbol{k}_{h\,\mathrm{pvt}}, \boldsymbol{n}_h) \right\}.$$

The optimal scanner vector $\boldsymbol{n}^*_h$ must satisfy

$$V_h(\boldsymbol{n}^*_h) - V_h(\boldsymbol{n}_h) \geqslant \boldsymbol{\chi}_h \cdot (\boldsymbol{n}^*_h - \boldsymbol{n}_h) \tag{C6}$$

for every feasible alternative scanner vector $\boldsymbol{n}_h$. In particular, let $\boldsymbol{e}_m$ denote a unit vector in modality m. Considering deviations that add or remove one scanner, whenever feasible, gives the necessary inequalities

$$V_h(\boldsymbol{n}^*_h + \boldsymbol{e}_m) - V_h(\boldsymbol{n}^*_h) \leqslant \chi_{hm} \leqslant V_h(\boldsymbol{n}^*_h) - V_h(\boldsymbol{n}^*_h - \boldsymbol{e}_m).$$

The gain from an additional scanner cannot exceed its fixed cost, while the operating-profit loss from removing a scanner cannot be smaller than the fixed cost saved.

C.4 Relation to alternative models of wait times

↩ Main text: §4.2 Model / Queues

↩ See also: §C.1 Theory / Queuing model · §G.1 Proofs / Stationary wait distribution

Textbook queuing models Our backlog process is closely related to the workload processes studied in standard queuing theory (see, e.g., Sections 5 and 9 of Whitt, 2002). The

simplest benchmark is the $M/M/1$ queue, in which customers arrive according to a Poisson process with rate λ, service times are exponentially distributed with rate μ, and one server processes customers in order. If utilization $\rho = \lambda/\mu$ is below one, the stationary mean wait is[45]

$$\mathbb{E}[w] = \frac{\rho}{1-\rho}\frac{1}{\mu}. \tag{C7}$$

More general models allow arbitrary iid interarrival and service-time distributions. A convenient representation starts from cumulative arriving work net of potential service, denoted $x(t)$, and constrains workload not to become negative: $\Gamma(x)(t) = x(t) + \sup_{0\leqslant s\leqslant t}[-x(s)]^+$. The second term simply records service capacity that could not be used because the queue was empty; in queuing terminology, the workload is "reflected" at zero. Closed-form stationary distributions are generally unavailable outside special cases, but heavy-traffic approximations can be obtained by approximating the net-input process with Brownian motion and then applying the same reflection operation (Whitt, 2002, Theorem 5.4.1). For a single-server queue this yields Kingman's approximation

$$\mathbb{E}[w] \approx \frac{\rho}{1-\rho}\frac{c_a^2 + c_s^2}{2}\mathbb{E}[s], \tag{C8}$$

where c_a and c_s measure the variability of interarrival and service times (Kingman, 1961).

Equations (C7) and (C8) treat arrivals as exogenous to the current queue state. Our model instead features current backlogs determining quoted wait times and patients choosing among queues after observing those waits, so that both the mean and variance of arrivals to a given queue depend on the full backlog vector. In particular, a persistently congested queue can satisfy $\mathbb{E}[a_{qt}] = k_q$ while still having a finite stationary wait distribution, because high backlogs reduce subsequent demand. A utilization measure based on unconditional mean arrivals would therefore equal one and make the textbook formulas inapplicable. The literature provides some results that allow arrivals or other queue primitives to vary with congestion when considering a queue in isolation (Miyazawa, 2024, 2025; Li and Pang, 2026). Since our queues interact through patient substitution, applying such approximations separately would discard the cross-queue feedback that generates the state dependence. We therefore compute the stationary distribution of wait times from the joint backlog process.

Queuing models in economics Brancaccio et al. (2025) use the $M/M/K$ structure directly to model congestion at ports. If a ship arrives to find Q ships already at a port with K available berths and mean service time T, their benchmark implies the conditional expected queue wait $\mathbb{E}[w \mid Q] = \mathbf{1}\{Q \geqslant K\}(Q - K + 1)\frac{T}{K}$. Their richer specification allows the service

[45] The formula generalizes to the $M/M/K$ model, where K identical servers operate in parallel, and gives $\mathbb{E}[w] = \frac{C(K,\lambda/\mu)}{K\mu-\lambda}$, with $C(K,\lambda/\mu)$ being the Erlang-C probability that all K servers are occupied.

technology to depend on port inputs and on the number of ships being served. The realized queue therefore determines ships' time at port, but port demand is governed by *expected* congestion rather than by each ship observing its contemporaneous queue-specific wait before choosing where to go. Expectations are of course required to be consistent with equilibrium outcomes, but the stochastic queue affects service outcomes without making arrivals respond directly to the current queue realization.

Markkula (2025) instead uses the steady-state logic of an $M/M/K$ queue to motivate a reduced-form congestion equation rather than imposing the $M/M/K$ formula itself. For public dental practice j in year t, his specification takes the form $\mathbb{E}[w_{jt}] = \psi \frac{q_{jt}}{d_{jt}} + \omega_{jt}$, where q_{jt} is annual demand and d_{jt} is the number of dentists. Demand and mean wait times are jointly determined, so the model retains the feedback from congestion to patient choice. Relative to our model, the principal simplification is that congestion is summarized by one average wait per practice-year and its relation to demand per observed dentist is parameterized directly.

Our model combines state-dependent patient choice with an explicit stochastic queue. A patient receiving a referral in week t observes the current wait quoted by each available alternative, so the realized congestion state affects that week's arrivals and hence subsequent backlogs. Random health shocks then generate a stationary distribution of weekly wait times, and the resulting variation is used directly in demand estimation. Because choice probabilities are nonlinear in wait times, our empirical implementation also accounts for the distribution of waits rather than evaluating demand only at its mean.

The queuing technology additionally provides the link to the capacity object that is central to our supply model. Queue capacity k_q is the latent number of treatment slots available per period rather than an observed provider input. Our stochastic queuing model serves both to describe the wait times entering patient choice and to recover the latent service capacity whose level and allocation we study.

Wait times as capacity-clearing scarcity terms Russo (2024) takes a different approach by fixing VA clinic capacities and solving for wait times that make patient demand exhaust them, while community providers are treated as unconstrained. The capacity-clearing interpretation allows for a simple optimal-assignment interpretation. Denoting aggregate demand at clinic j ad $D_j(\boldsymbol{p}, \boldsymbol{w})$, an equilibrium vector of wait times satisfies

$$D_j(\boldsymbol{p}, \boldsymbol{w}) \leqslant \kappa_j, \qquad w_j \geqslant 0, \qquad w_j\big[\kappa_j - D_j(\boldsymbol{p}, \boldsymbol{w})\big] = 0 \qquad \forall\, j \in \mathcal{J}.$$

Expressing patient i's utility from clinic j in units of wait time as $v_{ij}(\boldsymbol{p}) - w_j$, consider a capacity-constrained assignment that maximizes aggregate wait-metric patient surplus:

$$\max_{\boldsymbol{x}} \quad \int_{\mathcal{I}} \sum_{j \in \mathcal{J}} x_{ij} v_{ij}(\boldsymbol{p}) \, \mathrm{d}i$$

$$\text{sub} \quad \int_{\mathcal{I}} x_{ij}\,\mathrm{d}i \leqslant \kappa_j \quad \forall\, j \in \mathcal{J}, \qquad \sum_{j\in\mathcal{J}} x_{ij} \leqslant 1 \quad \forall\, i \in \mathcal{I}.$$

If λ_j denotes the multiplier on clinic j's capacity constraint, the planner assigns patient i to an option maximizing $v_{ij}(\boldsymbol{p}) - \lambda_j$. Setting $\lambda_j = w_j$ makes this assignment rule identical to decentralized patient choice, while the equilibrium capacity-clearing conditions above give feasibility and complementary slackness. The equilibrium wait vector is therefore a vector of shadow prices supporting the capacity-constrained allocation that maximizes wait-normalized patient surplus.[46] Our model differs because wait times are generated by the stochastic evolution of backlogs: for given prices and capacities, the model generates a stationary distribution of wait times rather than a single wait that clears each provider's capacity.

D Estimation

D.1 Queuing model and computation of the stationary distribution

↩ Main text: §5.1 Estimation / Queuing model · §4.2 Model / Queues

We validate the queuing model of Section 4.2 by reconstructing the evolution of backlogs according to the observed arrivals and throughput of public queues, and testing whether the implied wait times line up with the realized ones. We say that a patient joins queue q in week t if they (i) received their referral in week t, (ii) their scan was completed at the provider and access implied by j, and (iii) their modality and urgency are the ones implied by q. Denote as $\mathcal{I}_{qt}$ the set of patients observed to join queue q in week t. The observed week-t arrivals a_{qt} into queue q are equal to the size of set $\mathcal{I}_{qt}$, while throughput y_{qt} is the number of scans executed in week t. Both are observed for all public queues, and this allows us to operationalize the backlog law of motion. We initialize all queues' backlogs to $b_{q1} = 0$ in the first week of 2023, and update them according to $b_{q,t+1} = \max\{b_{qt} + a_{qt} - y_{qt}, 0\}$. We burn-in the first 20 weeks of 2023 and validate our model on the remaining weeks until the end of 2024. The first-availability wait follows (C2).

Since we observe throughput y_{qt} but not necessarily allocated capacity k_{qt}, we compute wait times implied by (C2) under the assumption that public queues are almost never idle, which implies $y_{qt} = k_{qt}$. Panel A in Table D1 shows that this assumption is particularly fitting for the low urgency tiers, where the majority of patients belongs.

Patient i's realized wait time ω_i is defined as the number of days elapsed between their referral and scan date, and is observed for all patients choosing publicly funded access. Because of deviations from the queue's first-come, first-served protocol (e.g., idiosyncratic elements in

[46] This is an allocation result conditional on prices and capacities and does not eliminate the welfare cost of time spent waiting. Because patients differ in their marginal disutility from wait time, wait-normalized patient surplus is generally different from money-metric patient surplus.

i's daily schedule or holidays), this realized wait can differ from the model-implied wait for the queue selected by i. We assess the predictive performance of our model by comparing implied first-availability wait times to the mean observed wait times, defined as $\bar{w}_{qt} = \frac{1}{|\mathcal{I}_{qt}|} \sum_{i \in \mathcal{I}_{qt}} w_i$ for all public queues q and weeks t. Panel B of Table D1 shows that our model's predictions strongly correlate with realized wait times, especially for lower-urgency, congested queues. Figure D1 illustrates.

Table D1 Queuing model validation

	Urgency tier			
	1	2	3	4
Panel A: Queue activity				
Share of backlogged queue-weeks	0.41	0.92	0.98	0.98
Share of always-backlogged queues	0.05	0.74	0.90	0.91
Queues	656	692	712	736
Queue-weeks	50,059	53,212	54,622	56,504
Total arrivals	194,943	642,634	1,103,015	3,881,119
Panel B: Model fit				
Correlation				
weekly	0.57	0.75	0.83	0.88
3-week moving average	0.72	0.88	0.91	0.91
5-week moving average	0.78	0.92	0.94	0.92
Mean absolute error				
weekly	1.77	3.85	8.48	13.83
3-week moving average	1.32	2.45	5.62	11.21
5-week moving average	1.20	1.94	4.57	10.27

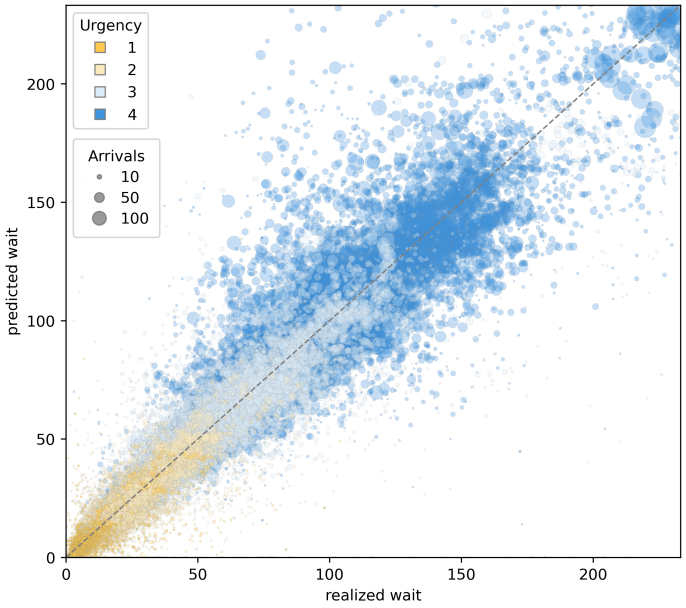
Notes: Panel A reports summaries of queue activity using weeks in the second half of 2023 and the entirety of 2024. According to our model, a queue-week is "backlogged" if there is at least one patient who received a referral in a prior week and is observed treated in that queue at a later week. A queue is "always backlogged" if it is backlogged for at least 95% of weeks. Panel B reports measures of goodness-of-fit of our queuing model. Correlations are weighted by the number of arriving patients in each queue-week. Mean absolute errors are measured in days.
(Click here to return to the estimation section in the main text.)

D.2 Approximation of the stationary distribution of waits

↩ Main text: §5.1 Estimation / Queuing model · §5.3 Estimation / Private capacity, patient demand, and scan costs

This section discusses how we approximate integrals over the stationary distribution of wait times to ease computation. Providers maximize their objective integrated over the stationary

Figure D1 Queuing model validation, MRI



Notes: Figure plots realized mean weekly waits with model-implied ones for MRI publicly funded queues. Dots are queue-week combinations, with colors highlighting their urgency tier and sizes proportional to observed patient arrivals. Model-implied waits are constructed as described in Section D.1.
(Click here to return to the estimation section in the main text.)

distribution of wait times induced by choice variables $(\boldsymbol{p}, \boldsymbol{k})$,[47] and differentiation such distribution is required to evaluate providers' optimality conditions, in Appendices C.3 and H.3. However, obtaining the stationary distribution of wait times by simulation for every candidate $(\boldsymbol{p}, \boldsymbol{k})$ is costly. We therefore use some approximations to reformulate the search for this stationary distribution to the solution of a fixed point equation, which can be differentiated more easily.

We first summarize the stationary distribution of wait times using its first two moments. We assume that the variance of the distribution of wait times at a given queue depends only on the queue's mean wait and capacity through function $\mathbb{V}(w \,|\, \bar{w}, k)$. We then approximate expectations of twice-differentiable functions of wait times as

$$\mathbb{E}_{\boldsymbol{w}\,|\,\boldsymbol{p},\boldsymbol{k}}[g(\boldsymbol{w})] \approx g(\bar{\boldsymbol{w}}) + \frac{1}{2}\sum_q \frac{\partial^2 g(\bar{\boldsymbol{w}})}{\partial w_q^2}\mathbb{V}(w_q \,|\, \bar{w}_q, k_q). \tag{D1}$$

Second, we leverage the observation that, for fixed $(\boldsymbol{p}, \boldsymbol{k})$, stationarity requires the expected backlog in each queue to be unchanged by the backlog transition. Using the backlog law of

[47] For instance, arrivals $a_{js}(\boldsymbol{p}, \boldsymbol{k})$ as functions of choice variables are obtained by integrating arrivals conditional on wait times against the stationary distribution of waits induced by $(\boldsymbol{p}, \boldsymbol{k})$.

motion (C1), this condition is

$$\bar{b}_q = \mathbb{E}_{\boldsymbol{w},\boldsymbol{b}\,|\,\boldsymbol{p},\boldsymbol{k}}\left[\max\left\{b_q + a_q(\boldsymbol{p},\boldsymbol{w}) - k_q, 0\right\}\right]. \tag{D2}$$

The right-hand side of (D2) still depends on the entire stationary distribution of backlogs and wait times through the expectation of the kinked map $x \mapsto \max\{x, 0\}$, where $x = b + a - k$ is net inflow into a queue, i.e., the inherited backlog plus arrivals net of capacity. Direct evaluation would require repeatedly simulating the distribution of net inflow, and hence the stationary distribution of queues, for every candidate value of prices and capacities. We instead replace the kinked map with the smooth approximation:

$$\mathbb{E}[\max\{x, 0\}] \approx \tau \log(1 + \exp\{\mathbb{E}[x]/\tau\}) =: \mathrm{soft}(\mathbb{E}[x]; \tau).$$

In practice, we use queue-specific smoothing scale τ_q, expressed as a function of expected arrivals and queue capacity.

Condition (D2) can then be written as

$$\begin{aligned}
\bar{b}_q &\approx \mathrm{soft}(\bar{b}_q + \mathbb{E}_{\boldsymbol{w}\,|\,\boldsymbol{p},\boldsymbol{k}}[a_q(\boldsymbol{p},\boldsymbol{w})] - k_q; \tau_q) \\
&\approx \mathrm{soft}(\bar{b}_q + a_q(\boldsymbol{p},\bar{\boldsymbol{w}}; \mathbb{V}) - k_q; \tau_q) \\
&= \mathrm{soft}(\bar{b}_q + a_q(\boldsymbol{p},\bar{\boldsymbol{w}}(\bar{\boldsymbol{b}},\boldsymbol{k}); \mathbb{V}) - k_q; \tau_q) \\
&=: \Lambda_q(\bar{\boldsymbol{b}}; \boldsymbol{p}, \boldsymbol{k}, \mathbb{V}, \tau_q),
\end{aligned} \tag{D3}$$

where the second approximation follows from (D1) and the third row comes from the fact that the definition of wait times $w_{qt} = b_{qt}/k_q$ pins down wait times as a function of backlogs and capacity. Stacking conditions (D3) over queues results in a fixed point equation $\bar{\boldsymbol{b}} = \boldsymbol{\Lambda}(\bar{\boldsymbol{b}}; \boldsymbol{p}, \boldsymbol{k}, \mathbb{V}, \boldsymbol{\tau})$, which we solve starting from $\boldsymbol{b} = \boldsymbol{0}$ and Anderson-accelerating a damped version of $\boldsymbol{\Lambda}$. Implicit differentiation then gives

$$\frac{\partial \bar{\boldsymbol{b}}}{\partial \boldsymbol{p}'} = \left[I - \frac{\partial \boldsymbol{\Lambda}}{\partial \bar{\boldsymbol{b}}'}\right]^{-1} \frac{\partial \boldsymbol{\Lambda}}{\partial \boldsymbol{p}'}, \qquad \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{p}'} = \frac{\partial \bar{\boldsymbol{w}}}{\partial \bar{\boldsymbol{b}}'}\frac{\partial \bar{\boldsymbol{b}}}{\partial \boldsymbol{p}'} = \mathrm{diag}(\hat{\boldsymbol{k}})^{-1}\frac{\partial \bar{\boldsymbol{b}}}{\partial \boldsymbol{p}'}.$$

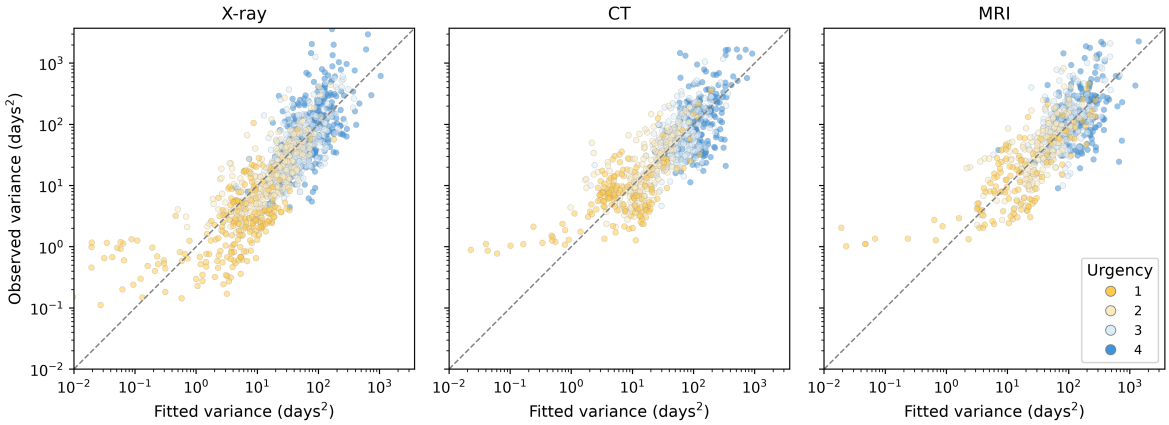
Implementation details are presented in Appendix D.6.

We conclude by elaborating on the estimation of the variance of wait times and the softplus approximation.

Variance of wait times Function $\mathbb{V}$ is interpreted as a primitive of the queuing process and is estimated offline, leveraging observed time-series variation in wait times. For each public queue q, let $\bar{w}_q$ denote its mean wait across weeks and let s_q^2 denote the corresponding sample variance of weekly waits. We estimate $\mathbb{V}(w_q \,|\, \bar{w}_q, k_q)$ by regressing $\log s_q^2$ on a regularized cubic

spline in $\log(1+\bar{w}_q)$, $\log(1+k_q)$, and modality and urgency indicators. At runtime, the fitted function is evaluated at the candidate mean wait and capacity of each queue. Table D2 reports goodness-of-fit measures in log-variance space, Figure D2 illustrates.

Figure D2 Variance model fit



Notes: Figure shows model fit at the queue level by comparing each queue's observed across-week wait variance with the variance predicted by the fitted function $\mathbb{V}(w_q \mid \bar{w}_q, k_q)$. Both axes use a log scale. Dots are public queues. Colors distinguish urgency tiers and panels distinguish imaging modalities. See Table D2 for details.
(Click here to return to the estimation section in the main text.)

Table D2 Variance model fit

	All	X-ray	CT	MRI
Number of queues	2,576	1,105	782	689
Correlation (observed vs. fitted)	0.96	0.92	0.98	0.98
R^2	0.92	0.84	0.95	0.95
Mean absolute error (log)	0.89	1.01	0.78	0.82

Notes: Table reports the goodness-of-fit measures for the variance function estimation.
(Click here to return to the estimation section in the main text.)

Softplus approximation of the expected backlog transition We replace the expectation of the kinked map on the right-hand side of (D2) with the approximation

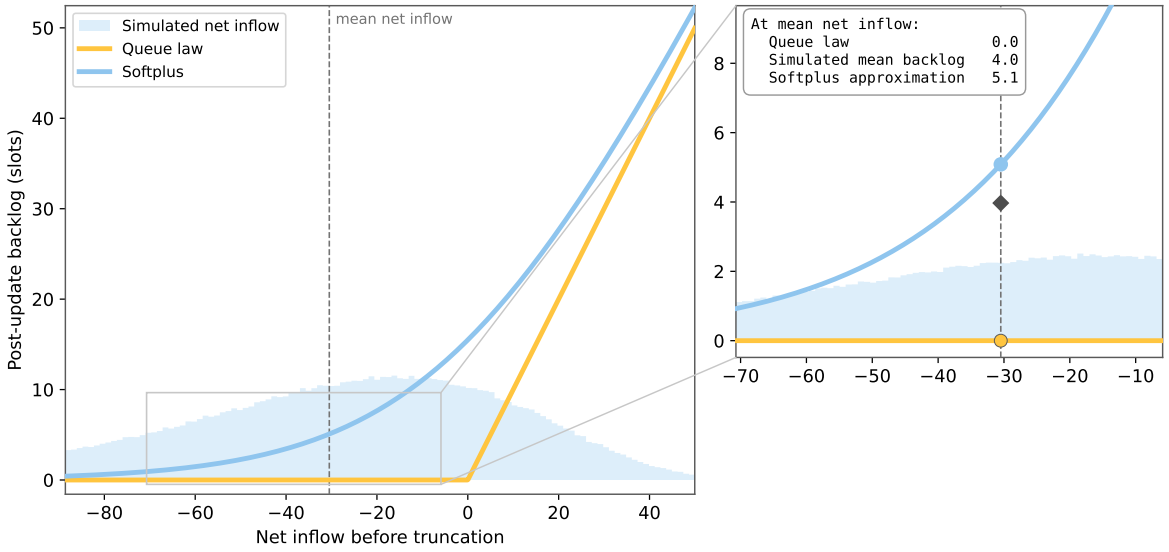
$$\mathbb{E}[\max\{x_q, 0\}] \approx \mathrm{soft}(\mathbb{E}[x_q]; \tau_q).$$

The smoothing scale is queue specific and varies with expected arrivals and capacity. For expected arrivals a_q and capacity k_q, we set $\tau_q = \frac{\sqrt{3}}{\pi}\sqrt{a_q + 0.4^2 k_q^2}$. The scale is recomputed inside every evaluation of the backlog fixed point using the same expected arrivals and capacity

that determine net inflow, and a_q denotes the expected arrivals under (D1).

Figure D3 illustrates the approximation. Away from the kink, the smooth and kinked backlog maps are nearly identical. Near the kink, evaluating the kinked map at mean net inflow can imply zero backlog even when stochastic net inflow generates positive average backlog, while the smooth approximation preserves a positive point-evaluation correction. This is the region in which smoothing is useful for approximating the stochastic backlog update without simulation.

Figure D3 Softplus approximation of the mean expected backlog



Notes: The shaded distribution shows simulated net inflow for a slack X-ray queue evaluated at the estimated model. The left panel compares the kinked backlog map with its smooth approximation over a broad range of net inflows, and the right panel magnifies the neighborhood of mean net inflow.
(Click here to return to the estimation section in the main text.)

D.3 Private capacity

↩ Main text: §5.3 Estimation / Private capacity, patient demand, and scan costs · §5.2 Estimation / Capacity of public queues

↩ See also: §C.1 Theory / Queuing model · §G.3 Proofs / Identification of private capacity

We use lowest-urgency public throughput, which coincides with capacity in persistently congested queues, to recover two sources of within-year capacity variation: (i) a modality-level component common across providers, capturing holidays and seasonal patterns, and (ii) persistent provider-specific deviations from this common pattern. We assume that private capacity follows both components, while its mean level is determined separately by matching observed private waits.

Reconstruction of the private capacity path Denoting the lowest urgency tier by $\underline{u}$, the aggregate weekly publicly funded throughput for lowest-urgency patients is $Y_{\text{pub}\,m\,\underline{u}\,t} = \sum_h y_{h\,\text{pub}\,m\,\underline{u}\,t}$. Bars are used to denote averages over weeks. The (log of the) provider-specific proportional deviation from the modality-wide aggregate pattern is

$$\upsilon_{hmt} = \log\left(\frac{y_{h\,\text{pub}\,m\,\underline{u}\,t} \,/\, \bar{y}_{h\,\text{pub}\,m\,\underline{u}}}{Y_{\text{pub}\,m\,\underline{u}\,t} \,/\, \bar{Y}_{\text{pub}\,m\,\underline{u}}}\right). \tag{D4}$$

The ratio $Y_{\text{pub}\,m\,\underline{u}\,t}/\bar{Y}_{\text{pub}\,m\,\underline{u}}$ captures the common modality-level component, while υ_{hmt} measures each provider's proportional deviation from that pattern. The provider-specific deviation is decomposed as

$$\upsilon_{hmt} = \upsilon^{\text{P}}_{hmt} + \upsilon^{\text{T}}_{hmt}, \quad \upsilon^{\text{P}}_{hmt} = \tau_m \upsilon^{\text{P}}_{hm\,t-1} + \upsilon^{\text{I}}_{hmt}. \tag{D5}$$

Both $\upsilon^{\text{T}}_{hmt}$ and $\upsilon^{\text{I}}_{hmt}$ have mean zero and are serially uncorrelated. We retain the persistent component as the second source of variation in private capacity and filter out the transitory component.

We estimate τ_m and the variance components from the autocovariances of υ_{hmt}. For $\ell = 0, 1, 2$, let $\mathcal{P}_{\ell m}$ contain the valid within-provider observations separated by exactly ℓ calendar weeks. The pooled raw autocovariances are $\hat{\Gamma}_{\ell m} = \frac{1}{|\mathcal{P}_{\ell m}|} \sum_{(h,t)\in\mathcal{P}_{\ell m}} \upsilon_{hmt}\upsilon_{hm\,t-\ell}$, for $\ell = 0, 1, 2$.

The decomposition in (D5) implies $\Gamma_{1m} = \tau_m \operatorname{Var}(\upsilon^{\text{P}}_{hmt})$ and $\Gamma_{2m} = \tau_m^2 \operatorname{Var}(\upsilon^{\text{P}}_{hmt})$, so we estimate

$$\hat{\tau}_m = \frac{\hat{\Gamma}_{2m}}{\hat{\Gamma}_{1m}}, \qquad \hat{\operatorname{Var}}(\upsilon^{\text{P}}_{hmt}) = \frac{\hat{\Gamma}_{1m}}{\hat{\tau}_m}.$$

The remaining variance components follow from $\Gamma_{0m} = \operatorname{Var}(\upsilon^{\text{P}}_{hmt}) + \operatorname{Var}(\upsilon^{\text{T}}_{hmt})$ and stationarity of the AR(1) process:

$$\hat{\operatorname{Var}}(\upsilon^{\text{T}}_{hmt}) = \hat{\Gamma}_{0m} - \hat{\operatorname{Var}}(\upsilon^{\text{P}}_{hmt}), \qquad \hat{\operatorname{Var}}(\upsilon^{\text{I}}_{hmt}) = (1 - \hat{\tau}_m^2)\hat{\operatorname{Var}}(\upsilon^{\text{P}}_{hmt}).$$

Table D3 reports these estimates.

Since only the sum $\upsilon_{hmt} = \upsilon^{\text{P}}_{hmt} + \upsilon^{\text{T}}_{hmt}$ is observed, we use the estimated time-series model to recover $\upsilon^{\text{P}}_{hmt}$ for each provider-modality pair. Conditional on $\hat{\tau}_m$ and the estimated variance components, we recover the persistent component separately for each provider-modality pair using a Kalman filter followed by the fixed-interval smoother of Rauch et al. (1965).

We plug the filtered-and-smoothed estimate $\hat{\upsilon}^{\text{P}}_{hmt}$ into the private-capacity path defined in (D6). Private capacity follows a weekly path informed by seasonal and provider-specific

Table D3 Capacity path parameters

	X-ray	CT	MRI
Persistence (τ_m)	0.716	0.651	0.626
Persistent component (v^{P}) variance	0.145	0.117	0.132
Transitory component (v^{T}) variance	0.089	0.092	0.063
Persistent innovation (v^{I}) variance	0.071	0.068	0.080
Valid observations	3,664	2,581	2,815
Lag-1 pairs	3,534	2,495	2,718
Lag-2 pairs	3,439	2,427	2,647

Notes: Table reports the modality-specific estimates used to decompose the provider-specific capacity deviation v_{hmt} in (D5). Autocovariances are pooled across providers within modality. Lag-1 and lag-2 pairs use observations separated by exactly one and two calendar weeks, respectively.
(Click here to return to the discussion of private capacity in the main text.)

persistent patterns, scaled by mean private capacity $k_{h\,\mathrm{pvt}\,m}$:

$$k_{h\,\mathrm{pvt}\,mt} = k_{h\,\mathrm{pvt}\,m}\,\frac{Y_{\mathrm{pub}\,m\,\underline{u}\,t}}{\bar{Y}_{\mathrm{pub}\,m\,\underline{u}}}\exp(\hat{v}^{\mathrm{P}}_{hmt})\Bigg/\left(\frac{1}{T}\sum_{\tau=1}^{T}\frac{Y_{\mathrm{pub}\,m\,\underline{u}\,\tau}}{\bar{Y}_{\mathrm{pub}\,m\,\underline{u}}}\exp(\hat{v}^{\mathrm{P}}_{hm\tau})\right). \tag{D6}$$

The aggregate throughput ratio captures fluctuations common across providers, while $\exp(\hat{v}^{\mathrm{P}}_{hmt})$ captures persistent provider-specific deviations. The denominator normalizes the weekly profile to have mean one, so $k_{h\,\mathrm{pvt}\,m}$ is mean private capacity.

Figure D4a plots the normalized paths of $Y_{\mathrm{pub}\,m\,\underline{u}\,t}\,/\,\bar{Y}_{\mathrm{pub}\,m\,\underline{u}}$ and $y_{h\,\mathrm{pub}\,m\,\underline{u}\,t}\,/\,\bar{y}_{h\,\mathrm{pub}\,m\,\underline{u}}$ for a representative provider. Figure D4b compares the observed deviation v_{hmt} with this smoothed persistent component $\hat{v}^{\mathrm{P}}_{hmt}$ for the same representative provider, where the deviation v_{hmt} is constructed following (D4).

Inversion The relative weekly capacity path is determined by public throughput, while its unknown mean is recovered by matching model-implied and scraped private waits. For each private queue $q = (h, \mathrm{pvt}, m)$ and candidate mean capacity k_q, weekly capacity follows (D6). Starting from $b_{q1} = 0$ in the first week of 2023, we recursively compute

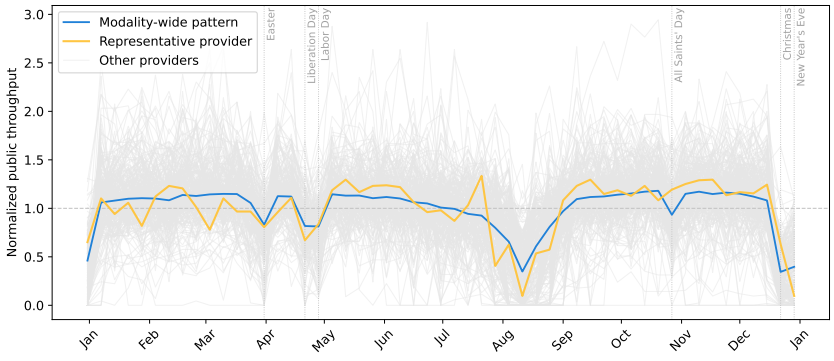
$$w_{qt}(k_q) = \inf\left\{x > 0 : \sum_{r=0}^{\lfloor x\rfloor-1} k_{q\,t+r}(k_q) + \big(x - \lfloor x\rfloor\big)k_{q\,t+\lfloor x\rfloor}(k_q) > b_{qt}(k_q)\right\},$$
$$b_{q\,t+1}(k_q) = \max\big\{b_{qt}(k_q) + a_{qt}(w_{qt}(k_q)) - k_{qt}(k_q), 0\big\}.$$

The weekly referral mass is the observed modality-level total.

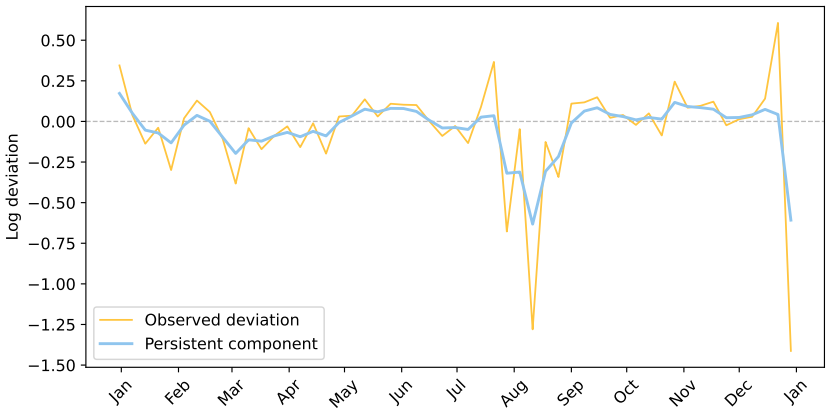
Only queue q's wait varies in this recursion; prices and all other waits remain at their observed week-specific values. We use a script font ($\mathcal{w}$) for observed waits. The logit structure of

Figure D4 Reconstruction of private capacity paths

(a) Capacity paths



(b) Provider-specific decomposition



Notes: Panel (a) plots the normalized weekly capacity paths for X-rays used in the private-capacity inversion. The plot highlights the common variation given by the modality-level component $Y_{\text{pub}\,m\,\underline{u}\,t}/\bar{Y}_{\text{pub}\,m\,\underline{u}}$ and the realized path $y_{h\,\text{pub}\,m\,\underline{u}\,t}/\bar{y}_{h\,\text{pub}\,m\,\underline{u}}$ for a representative provider. The former is the ratio between weekly and mean aggregate throughput (which coincides with capacity under the "persistent congestion" assumption), where the aggregation is over providers and the mean over weeks. The latter is the ratio between a specific provider's weekly and mean throughput. Capacity falls broadly during the summer and around major holidays, overlaid in the plot. The paths of other providers are plotted in the background. Panel (b) plots the observed provider-specific deviation v_{hmt} and its estimated persistent component $\hat{v}^{\text{P}}_{hmt}$ for the representative provider shown in panel (a). Estimation of the persistent component follows the filtering-and-smoothing procedure described in Section D.3 applied to the time series of provider-specific deviation.
(Click here to return to the discussion of private capacity in the main text.)

choice probabilities then allows us to update queue-q choice probabilities without reevaluating the complete choice problem. If $\sigma_{iqt}(w_{qt})$ denotes patient i's probability of choosing queue q at its observed wait w_{qt}, then at counterfactual wait w,

$$\sigma_{iqt}(w) = \frac{\sigma_{iqt}(w_{qt}) \exp\left\{-\beta^{\text{wait}}_{g(i)}(w - w_{qt})\right\}}{1 - \sigma_{iqt}(w_{qt}) + \sigma_{iqt}(w_{qt}) \exp\left\{-\beta^{\text{wait}}_{g(i)}(w - w_{qt})\right\}}.$$

Weekly arrivals $a_{qt}(w)$ are obtained by aggregating these probabilities over the reconstructed weekly referral mass.

The private-wait target is constructed from the 2024 scraped first-availability waits as discussed in Appendix A. A Monday slot available on Saturday is recorded as a 2-day wait, but it is effectively indicative of a slot being available on the first business day from the query date. We therefore add a "lead" ℓ to account for the fact that scraping occurred during the weekend but slots are only available during weekdays. To read the positive waits observed on Saturdays as essentially no congestion, we replace the observed wait w_{qt} by $\max\{w_{qt} - \ell, \epsilon\}$ before averaging, where $\ell \in \{0, 1, 2\}$ based on the scraping day-of-week and $\epsilon > 0$ ensures that the target wait stays away from the zero boundary. Let $\mathcal{T}_{2024}$ denote the weeks of 2024 and let $\bar{w}_q$ denote the resulting mean congestion wait.

For $\bar{w}_q > 0$, we pick mean private capacity $\hat{k}_q$ to solve

$$M_q^{\mathrm{W}}(k_q) := \frac{1}{|\mathcal{T}_{2024}|} \sum_{t \in \mathcal{T}_{2024}} w_{qt}(k_q) - \bar{w}_q = 0.$$

Figure D5 illustrates the identifying variation for a representative queue. Over the relevant range of capacities, the map M_q^{W} is decreasing: higher mean capacity shifts the implied wait distribution toward shorter waits, and the recovered capacity $\hat{k}_q$ is the value at which the wait residual crosses zero. Results on the continuity and monotonicity of M_q^{W} are presented in Section G.3. These root-finding operations are separable across queues conditional on the candidate demand parameters and reconstructed capacity paths. We exploit this embarrassingly parallel structure by stacking queues and candidate capacity values as array dimensions and evaluating their realized paths in a single JAX-compiled operation, with `jax.lax.scan` carrying each queue's backlog forward over weeks. Each evaluation of $M_q^{\mathrm{W}}(k_q)$ is therefore batched across all queues, so the one-dimensional root solves proceed in parallel.

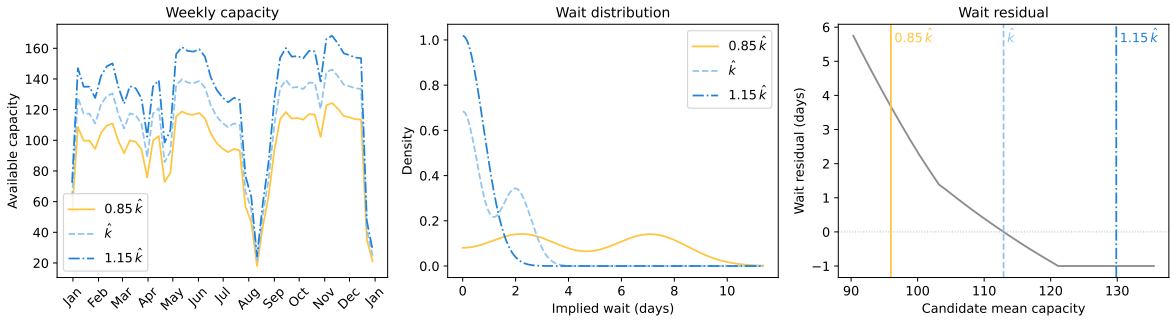
D.4 Computation of the GMM criterion

↩ Main text: §5.3 Estimation / Private capacity, patient demand, and scan costs

↩ See also: §C.3 Theory / Provider optimality conditions · §D.2 Estimation / Stationary wait approximation · §D.5 Estimation / Patient choice micromoments · §D.6 Estimation / Differentiation

The nonlinear optimizer searches over the preference parameters $\boldsymbol{\beta}$. For every candidate $\boldsymbol{\beta}$, we recover provider quality $\boldsymbol{\delta}(\boldsymbol{\beta})$ from the public-share inversion, recover private capacity $\hat{\boldsymbol{k}}_{\mathrm{pvt}}(\boldsymbol{\beta})$ from the private-wait inversion, and solve for the stationary queue environment implied by the resulting demand and capacity vectors. The scan-cost parameters $\boldsymbol{\gamma}$ are concentrated out conditional on these recovered quantities. Algorithm 1 describes the algorithm. The remainder of this section discusses the key points.

Figure D5 Identification of private capacity



Notes: Figure illustrates the private-capacity inversion for a representative queue by plotting capacity paths, implied wait distributions, and wait residuals under three capacity values. The chosen values are $0.85\hat{k}$, $\hat{k}$, and $1.15\hat{k}$, where $\hat{k}$ is the estimated capacity for the representative queue. The left panel plots the weekly capacity paths implied by the three mean capacity guesses, constructed following (D6) and using the provider's estimated deviations and the queue's modality aggregate trends. The right panel plots the wait residual M_q^{W} against candidate mean capacity, with the colored points corresponding to the three capacity values shown in the other panels. The wait residual is decreasing over the relevant range and crosses zero at $\hat{k}$.
(Click here to return to the discussion of private capacity in the main text.)

Demand inversion For each candidate $\boldsymbol{\beta}$, we recover provider qualities $\boldsymbol{\delta}(\boldsymbol{\beta})$ by matching predicted and observed annual shares of publicly funded alternatives within each modality. Denoting the observed shares by $\mathcal{s}_{h\,\mathrm{pub}\,m}$, we solve $\sigma_{h\,\mathrm{pub}\,m}(\boldsymbol{p}, \boldsymbol{W}; \boldsymbol{\beta}, \boldsymbol{\delta}(\boldsymbol{\beta})) = \mathcal{s}_{h\,\mathrm{pub}\,m}$. We use the standard contraction map of Berry et al. (1995), starting from an initial quality vector $\boldsymbol{\delta}^{(0)}$ and updating

$$\delta_{hm}^{(r+1)} = \delta_{hm}^{(r)} + \log \mathcal{s}_{h\,\mathrm{pub}\,m} - \log \sigma_{h\,\mathrm{pub}\,m}(\boldsymbol{p}, \boldsymbol{W}; \boldsymbol{\beta}, \boldsymbol{\delta}^{(r)}).$$

Proposition 1 establishes that the recovered vector is unique whenever a solution exists.

Pricing feedback The pricing first-order conditions derived in (C4) take the matrix form (H5), with derivative matrices $\boldsymbol{\Delta}_e$ defined in Appendix H.3. For access $e \in \{\mathrm{pub}, \mathrm{pvt}\}$, we decompose these matrices into direct (D) and indirect-through-wait (I) components:

$$\boldsymbol{\Delta}_e = \boldsymbol{\Delta}_e^{\mathrm{D}} + \boldsymbol{\Delta}_e^{\mathrm{I}}, \qquad \boldsymbol{\Delta}_e^{\mathrm{D}} = -\mathcal{H} \odot \left(\frac{\partial \boldsymbol{a}_e}{\partial \boldsymbol{p}'} \right)', \qquad \boldsymbol{\Delta}_e^{\mathrm{I}} = -\mathcal{H} \odot \left(\frac{\partial \boldsymbol{a}_e}{\partial \bar{\boldsymbol{w}}'} \frac{\mathrm{d}\bar{\boldsymbol{w}}}{\mathrm{d}\boldsymbol{p}'} \right)'.$$

The first component holds stationary waits fixed. The second contains the change in arrivals generated by the response of stationary waits to prices. Both components incorporate the second-order approximation of expected arrivals through (D1).

Constructing $\boldsymbol{\Delta}_e^{\mathrm{I}}$ explicitly is computationally expensive. The matrix $\mathrm{d}\bar{\boldsymbol{w}}/\mathrm{d}\boldsymbol{p}'$ has one column for every pricing direction, and each column requires solving the linearization of the stationary queue fixed point. The standard adjoint approach to implicit differentiation moves

from the observation that, in some circumstances, the object of interest is the product of a full derivative matrix and some vector, rather than the matrix itself. In our case, only products of $\boldsymbol{\Delta}_e^{\mathrm{I}}$ with the public and private margins matter for pricing first-order conditions. These products can be computed using transpose linear solves and Jacobian products without constructing the underlying derivative matrices.

For a candidate cost vector $\boldsymbol{c}$, collect the wait-response terms into

$$\mathcal{F}(\boldsymbol{c}) \coloneqq \boldsymbol{\Delta}_{\mathrm{pvt}}^{\mathrm{I}}(\boldsymbol{p}-\boldsymbol{c}) + \boldsymbol{\Delta}_{\mathrm{pub}}^{\mathrm{I}}(\boldsymbol{r}-\boldsymbol{c})$$

and rewrite (H5) as

$$0 = \boldsymbol{a}_{\mathrm{pvt}} - \boldsymbol{\Delta}_{\mathrm{pvt}}^{\mathrm{D}}(\boldsymbol{p}-\boldsymbol{c}) - \boldsymbol{\Delta}_{\mathrm{pub}}^{\mathrm{D}}(\boldsymbol{r}-\boldsymbol{c}) - \mathcal{F}(\boldsymbol{c}).$$

Conditional on the value of $\mathcal{F}(\boldsymbol{c})$, this system can be inverted for marginal costs. However, evaluating $\mathcal{F}(\boldsymbol{c})$ requires the same marginal costs through the public and private margins.

Let $\mathcal{C}(\tilde{\boldsymbol{c}};\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{k})$ be the value of $\boldsymbol{c}$ that solves $0 = \boldsymbol{a}_{\mathrm{pvt}} - \boldsymbol{\Delta}_{\mathrm{pvt}}^{\mathrm{D}}(\boldsymbol{p}-\boldsymbol{c}) - \boldsymbol{\Delta}_{\mathrm{pub}}^{\mathrm{D}}(\boldsymbol{r}-\boldsymbol{c}) - \mathcal{F}(\tilde{\boldsymbol{c}})$. At each candidate $\boldsymbol{\beta}$, we solve $\boldsymbol{c} = \mathcal{C}(\boldsymbol{c};\boldsymbol{\beta},\boldsymbol{\delta}(\boldsymbol{\beta}),\boldsymbol{k}(\boldsymbol{\beta}))$ using Anderson acceleration, with safeguards that keep trial costs positive. The solution is denoted by $\hat{\boldsymbol{c}}(\boldsymbol{\beta})$.

Criterion Let X^{c} stack the regressors in (3) over the provider-scan cells entering the pricing moments, and let Z^{c} stack the corresponding instruments. The pricing moments are

$$M^{\mathrm{P}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{N^{\mathrm{P}}} Z^{\mathrm{c}\prime}\big[\log \hat{\boldsymbol{c}}(\boldsymbol{\beta}) - X^{\mathrm{c}}\boldsymbol{\gamma}\big].$$

Conditional on $\boldsymbol{\beta}$, the pricing moments are affine in $\boldsymbol{\gamma}$, so the scan-cost coefficients can be concentrated out exactly. We choose $\boldsymbol{\gamma}^*(\boldsymbol{\beta})$ to minimize the weighted quadratic form in the pricing moments. Its first-order condition is

$$X^{\mathrm{c}\prime} Z^{\mathrm{c}} W^{\mathrm{P}} Z^{\mathrm{c}\prime} X^{\mathrm{c}} \boldsymbol{\gamma}^*(\boldsymbol{\beta}) = X^{\mathrm{c}\prime} Z^{\mathrm{c}} W^{\mathrm{P}} Z^{\mathrm{c}\prime} \log \hat{\boldsymbol{c}}(\boldsymbol{\beta}).$$

The matrix multiplying $\boldsymbol{\gamma}^*(\boldsymbol{\beta})$ does not depend on $\boldsymbol{\beta}$ and is factored once before optimization. We solve this linear system rather than forming its inverse.[48]

Let $M^{\mathrm{D}}(\boldsymbol{\beta})$ denote the patient choice moments. The concentrated criterion is

$$Q(\boldsymbol{\beta}) = M^{\mathrm{D}}(\boldsymbol{\beta})' W^{\mathrm{D}} M^{\mathrm{D}}(\boldsymbol{\beta}) + M^{\mathrm{P}}(\boldsymbol{\beta},\boldsymbol{\gamma}^*(\boldsymbol{\beta}))' W^{\mathrm{P}} M^{\mathrm{P}}(\boldsymbol{\beta},\boldsymbol{\gamma}^*(\boldsymbol{\beta})).$$

[48] The implementation bounds predicted log marginal costs to prevent numerical overflow. The affine representation requires these bounds to be inactive on all observations entering the pricing moments. We verify this condition at the conditional solution and reject candidates for which it fails.

Algorithm 1 Joint estimator

Require: Fixed weights $W^{\mathrm{D}}, W^{\mathrm{P}}$, public capacities $\hat{\boldsymbol{k}}_{\mathrm{pub}}$, starting qualities $\boldsymbol{\delta}^{(0)}$, and positive starting costs $\boldsymbol{c}^{(0)}$.

Cache regression solve: $C_\gamma \leftarrow X^{\mathrm{c}\prime} Z^{\mathrm{c}} W^{\mathrm{P}} Z^{\mathrm{c}\prime} X^{\mathrm{c}}$ and factor C_γ.
procedure JointCriterion($\boldsymbol{\beta}$)
  Initial provider qualities: $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta}^{(0)}$.
  repeat
    Public-share inversion: $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} + \log \boldsymbol{s}_{\mathrm{pub}} - \log \boldsymbol{\sigma}_{\mathrm{pub}}(\boldsymbol{p}, \boldsymbol{W}; \boldsymbol{\beta}, \boldsymbol{\delta})$.
  until $\| \log \boldsymbol{\sigma}_{\mathrm{pub}} - \log \boldsymbol{s}_{\mathrm{pub}} \|_\infty \leqslant \varepsilon_\delta$
  Patient-choice moments: Evaluate $M^{\mathrm{D}}(\boldsymbol{\beta})$.
  for all scraped private queues (h, m) **do**
    Private-wait inversion: Solve (6) for $k_{h\,\mathrm{pvt}\,m}$.
  Combine capacities: $\boldsymbol{k} \leftarrow (\hat{\boldsymbol{k}}_{\mathrm{pub}}, \boldsymbol{k}_{\mathrm{pvt}})$.
  Stationary queues: Solve (D3) for $\bar{\boldsymbol{b}}$.
  Queue linearization: $A \leftarrow I - \partial \boldsymbol{\Lambda} / \partial \bar{\boldsymbol{b}}'$.
  Direct pricing terms: Evaluate $\boldsymbol{a}_{\mathrm{pvt}}, \boldsymbol{\Delta}^{\mathrm{D}}_{\mathrm{pvt}}, \boldsymbol{\Delta}^{\mathrm{D}}_{\mathrm{pub}}$.
  Initial scan costs: $\boldsymbol{c} \leftarrow \boldsymbol{c}^{(0)}$.
  loop
    for all providers h entering the pricing conditions **do**
      Queue adjoint: Solve $A' \boldsymbol{\lambda}_h = \partial \tilde{\Pi}^{\mathcal{H}}_h / \partial \bar{\boldsymbol{b}}$.
      Wait feedback: $\mathcal{F}_{hs}(\boldsymbol{c}) \leftarrow -(\partial \boldsymbol{\Lambda} / \partial p_{hs})' \boldsymbol{\lambda}_h$ for each priced scan s.
    Direct cost solve: Solve $(\boldsymbol{\Delta}^{\mathrm{D}}_{\mathrm{pvt}} + \boldsymbol{\Delta}^{\mathrm{D}}_{\mathrm{pub}}) \tilde{\boldsymbol{c}} = \boldsymbol{\Delta}^{\mathrm{D}}_{\mathrm{pvt}} \boldsymbol{p} + \boldsymbol{\Delta}^{\mathrm{D}}_{\mathrm{pub}} \boldsymbol{r} + \mathcal{F}(\boldsymbol{c}) - \boldsymbol{a}_{\mathrm{pvt}}$.
    if $\| \tilde{\boldsymbol{c}} - \boldsymbol{c} \|_\infty \leqslant \varepsilon_c$ **then**
      Cost convergence: **break**.
    Accelerate cost iteration: $\boldsymbol{c} \leftarrow \mathrm{Anderson}(\boldsymbol{c}, \tilde{\boldsymbol{c}})$, enforcing $\boldsymbol{c} > \boldsymbol{0}$.
  Concentrate coefficients: Solve $C_\gamma \boldsymbol{\gamma} = X^{\mathrm{c}\prime} Z^{\mathrm{c}} W^{\mathrm{P}} Z^{\mathrm{c}\prime} \log \boldsymbol{c}$ using cached factors.
  Pricing moments: $M^{\mathrm{P}} \leftarrow (N^{\mathrm{P}})^{-1} Z^{\mathrm{c}\prime} [\log \boldsymbol{c} - X^{\mathrm{c}} \boldsymbol{\gamma}]$.
  Joint criterion: $Q \leftarrow M^{\mathrm{D}\prime} W^{\mathrm{D}} M^{\mathrm{D}} + M^{\mathrm{P}\prime} W^{\mathrm{P}} M^{\mathrm{P}}$.
  Implicit criterion gradient: $\boldsymbol{g}_\beta \leftarrow \mathrm{d}Q / \mathrm{d}\boldsymbol{\beta}$ using Appendix D.6.
  return $(Q, \boldsymbol{g}_\beta, \boldsymbol{\delta}, \boldsymbol{k}_{\mathrm{pvt}}, \boldsymbol{c}, \boldsymbol{\gamma})$.
Preference search: Minimize $Q(\boldsymbol{\beta})$ using JointCriterion($\boldsymbol{\beta}$) and its returned gradient.
Joint estimates: **return** $\hat{\boldsymbol{\beta}}$ and the corresponding $(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{k}}_{\mathrm{pvt}}, \hat{\boldsymbol{c}}, \hat{\boldsymbol{\gamma}})$.

Notes: $\boldsymbol{\sigma}_{\mathrm{pub}}$ and $\boldsymbol{s}_{\mathrm{pub}}$ stack predicted and observed public shares. Share inversion and patient-choice moments use the realized prices and waits. The private-wait inversion simulates each queue's own wait response as described above. The stationary queue environment is used for pricing, with $\tilde{\Pi}^{\mathcal{H}}_h$ denoting the operating profits internalized under common ownership, as defined in Appendix H.3.

D.5 Patient choice micromoments

↩ Main text: §5.3 Estimation / Private capacity, patient demand, and scan costs

This section provides the complete list of patient choice moments used in estimation. The micromoments are based on three outcomes: realized wait time among patients completing publicly funded care, travel time among the same patients, and an indicator that a referral is

lapsed.[49] For each outcome, the model counterpart is constructed using the same conditioning variables and weighting as the corresponding empirical moment.

Table D4 summarizes the moments entering M^{D}. Each row represents a separate collection of moments indexed by all values of the conditioning variables shown in the second column.

Table D4 Patient choice micromoments

Outcome	Conditioning variables	Sample
Wait		
Mean wait	Modality × age bin × month	Public completions
Mean wait	Modality × urgency × month	Public completions
Mean wait	Modality × sex × month	Public completions
Mean wait	Modality × province × month	Public completions
Distance		
Mean travel time	Modality × age bin	Public completions
Mean travel time	Modality × urgency	Public completions
Mean travel time	Modality × sex	Public completions
Mean travel time	Modality × province	Public completions
Lapsed		
Lapsed share	Modality × age bin	All referrals
Lapsed share	Modality × urgency	All referrals
Lapsed share	Modality × sex	All referrals
Lapsed share	Modality × province	All referrals
Lapsed share	Modality × province × urgency	All referrals
Lapsed share	Modality × province × age bin	All referrals
Income		
Mean wait	Modality × income bin	Public completions
Mean travel time	Modality × income bin	Public completions
Lapsed share	Modality × income bin	All referrals

Notes: Table reports the list of micromoments used in estimation. A "province" is one of 12 geographical subregions, each comprising several municipalities. They are typically clustered around larger cities and their population goes from roughly 180,000 to over 3 million inhabitants (median: 570,000; standard deviation: 800,000). Age bins are: 18–29, 30–39, 40–49, 50–59, 60–69, 70–79, over 80. Income bins coincide with income tax thresholds and are (in thousands of euros): 0–15, 15–26, 26–55, over 55.
(Click here to return to the estimation section in the main text.)

The empirical conditional mean for an outcome $x_i \in \{w_i, d_i\}$ and a conditioning set C is

[49] Referrals expire after 6 months, and recall that we classify a referral issued in 2024 as "lapsed" if it did not lead to a publicly funded treatment by February 2026. See Section 3.

computed as

$$\bar{x}_C = \frac{\sum_i \mathbf{1}\{i \in C, e(j_i) = \text{pub}, x_i \text{ observed}\} x_i}{\sum_i \mathbf{1}\{i \in C, e(j_i) = \text{pub}, x_i \text{ observed}\}}.$$

The lapsed-referral moments replace the outcome in the numerator with $\mathbf{1}\{i \text{ is lapsed}\}$ and use all patients in the conditioning cell as the denominator.

Individual income is not observed in the administrative data, but a town-level distribution over income bins is available from income tax records. Let π_{iq} denote the share of taxpayers in patient i's town belonging to income group q. For an outcome $x_i \in \{w_i, d_i\}$, the income-specific public mean for modality m and income group q is

$$\bar{x}_{mq} = \frac{\sum_i \mathbf{1}\{m(i) = m, e(j_i) = \text{pub}, x_i \text{ observed}\} \pi_{iq} x_i}{\sum_i \mathbf{1}\{m(i) = m, e(j_i) = \text{pub}, x_i \text{ observed}\} \pi_{iq}}.$$

The corresponding income-specific lapsed-referral share is

$$\bar{L}_{mq} = \frac{\sum_i \mathbf{1}\{m(i) = m, i \text{ is lapsed}\} \pi_{iq}}{\sum_i \mathbf{1}\{m(i) = m\} \pi_{iq}}.$$

The same income weights are used when constructing the model counterparts.

D.6 Differentiation

↩ Main text: §5.3 Estimation / Private capacity, patient demand, and scan costs

↩ See also: §D.4 Estimation / Computation of the GMM criterion · §D.2 Estimation / Stationary wait approximation

We compute the gradient of $Q(\boldsymbol{\beta})$ accounting for the dependence of provider quality, private capacity, stationary queues, and recovered marginal costs on the preference parameters. Each of these quantities solves an inner problem, so we obtain its derivative implicitly from the corresponding defining equations rather than differentiating through the numerical iterations. The derivatives of provider quality and private capacity follow from the public-share and private-wait inversions, respectively. The derivative of recovered marginal costs follows from the pricing fixed point described in Appendix D.4.

For stationary queues, write $\bar{\boldsymbol{b}} = \boldsymbol{\Lambda}(\bar{\boldsymbol{b}}; \boldsymbol{p}, \boldsymbol{k}, \boldsymbol{\beta}, \boldsymbol{\delta})$ and define $A := I - \partial\boldsymbol{\Lambda}/\partial\bar{\boldsymbol{b}}'$. For any argument $\boldsymbol{x}$ of the fixed-point map, holding the remaining arguments fixed, implicit differentiation gives

$$A \frac{\mathrm{d}\bar{\boldsymbol{b}}}{\mathrm{d}\boldsymbol{x}'} = \frac{\partial\boldsymbol{\Lambda}}{\partial\boldsymbol{x}'}. \tag{D7}$$

In evaluating the pricing conditions, (D7) gives the response of stationary waits to prices. We use transpose linear solves to compute the products of these derivatives with provider profit

margins without constructing the full derivative matrices. In differentiating $Q(\boldsymbol{\beta})$, the same relation accounts for the changes in stationary queues induced directly by preferences and indirectly through recovered provider quality and private capacity.

We implement these calculations in JAX (Bradbury et al., 2018), using automatic differentiation for Jacobian products and the remaining patient choice and moment calculations. Combining these derivatives with the implicit response of recovered marginal costs gives the gradient of the concentrated criterion.

D.7 Slot costs

↩ Main text: §4.3 Model / Supply · §5.4 Estimation / Capacity and scanner costs

↩ See also: §C.2 Theory / Production of capacity · §C.3 Theory / Provider optimality conditions · §H.3 Miscellaneous / Profit derivatives

We take patient preferences, scan marginal costs, and capacities recovered in the preceding steps as given. Optimality conditions are imposed only on privately owned providers. The slot-cost estimation sample is restricted to scraped provider-modalities. Since X-ray installed machines are not observed, we set $n_{h\,\mathrm{XR}} = 1$.

Let $K_{hm} = k_{h\,\mathrm{pub}\,m} + k_{h\,\mathrm{pvt}\,m}$ denote total capacity and n_{hm} the observed scanner count. Using the ownership matrix $\mathcal{H}$ introduced in Appendix C.3, define

$$MR_{hm} = \sum_{h'} \mathcal{H}_{hh'} \frac{\partial \Pi_{h'}}{\partial k_{h\,\mathrm{pvt}\,m}}.$$

Optimality condition (C5) requires $MR_{hm} = \phi'_{hm}(K_{hm}/n_{hm})$.[50] Writing $\boldsymbol{\zeta} = (\boldsymbol{\zeta}'_x, \zeta_{\mathrm{input}})'$, the parameterization in (4) gives the residual

$$\eta_{hm}(\boldsymbol{\zeta}) = \ln MR_{hm} - \boldsymbol{x}'_{hm}\boldsymbol{\zeta}_x - \zeta_{\mathrm{input}} \ln(K_{hm}/n_{hm}).$$

The vector $\boldsymbol{x}_{hm}$ includes fixed effects for modality and providers' ownership group. Because providers choose capacity and scanners in response to slot-cost shocks, capacity per scanner is endogenous. We instrument for $\ln(K_{hm}/n_{hm})$ with predicted private demand evaluated at common reference prices and waits.[51] We predict each provider's private demand using estimated patient preferences and the distribution of referrals over scans, locations, and demographics used in demand estimation. Differences in local demographics and demand shift

[50] Recall that operating profits Π_h deduct scan costs but exclude slot and scanner costs and that the derivative includes the induced changes in stationary waits and patient choices at all providers whose profits are internalized by h.

[51] We set private waits to zero days in all modalities. Public waits are set to 100 days for MRI and CT and 60 days for X-ray. Public prices are set to €36 for MRI and CT and €20.60 for X-ray. Private prices are set to €250 for MRI, €150 for CT, and €50 for X-ray.

the marginal return to additional slots and help predict capacity per scanner.

Collecting this instrument and exogenous characteristics $\boldsymbol{x}_{hm}$ into a vector $\boldsymbol{z}_{hm}^{\text{slot-cost}}$ and stacking these vectors into matrix Z, we estimate the parameters by two-stage least squares:

$$\boldsymbol{g}_N(\zeta) = \frac{1}{N}\sum_{h,m}\eta_{hm}(\zeta)\boldsymbol{z}_{hm}^{\text{slot-cost}}, \qquad \hat{\zeta} = \arg\min_{\zeta}\boldsymbol{g}_N(\zeta)'(Z'Z/N)^{-1}\boldsymbol{g}_N(\zeta).$$

D.8 Scanner fixed costs

↩ Main text: §5.4 Estimation / Capacity and scanner costs

↩ See also: §C.2 Theory / Production of capacity · §C.3 Theory / Provider optimality conditions

Scanner choices bound fixed costs through the requirement that no feasible deviation increase operating profits net of slot and scanner costs. We evaluate a subset of feasible deviations, so the resulting bounds are necessary but not sufficient for optimality.

Deviations A deviation changes the scanner count in one modality while holding scanner counts in other modalities and all other providers' choices at their observed values. For each deviation, we recompute stationary waits and operating profits, including changes in the profits of other providers under common ownership.

For the first type of deviation, we maximize operating profits net of slot costs over private capacity and private-pay prices at each alternative scanner count, holding public capacity fixed. Private-pay prices vary proportionally across scans within the modality.

The second type removes scanners at observed prices while reducing public capacity alone or both public and private capacity in proportion to the scanner count. When only public capacity is reduced, private capacity remains fixed. These partial reductions in public capacity are evaluated when $MR_{hm}^{\text{pub}} \geqslant MR_{hm}$, where $MR_{hm}^{\text{pub}} := \sum_{h'}\mathcal{H}_{hh'}\partial\Pi_{h'}/\partial k_{h\,\text{pub}\,m}$. This is the sign restriction implied by the public-capacity condition and a nonnegative multiplier on the reimbursement cap. We also evaluate removal of all scanners and both public and private capacity for every provider-modality, without imposing this restriction.

For positive scanner counts, we evaluate additions and removals of one or two scanners. The alternative total capacity K must satisfy

$$K \leqslant n\bar{K}_m.$$

The upper limits used to construct deviations assume that scanners can be operated for 16 hours a day and that MRI and CT scans respectively take 30 and 20 minutes.

Bounds To calculate payoffs under these deviations, we integrate the estimated marginal slot-cost function, imposing $\phi_{hm}(0) = 0$. For $\hat{\zeta}_{\text{input}} > -1$, the cost of offering K slots using

$n \geqslant 1$ scanners is

$$\hat{\Phi}_{hm}(K, n) = \frac{MR_{hm}K_{hm}}{1+\hat{\zeta}_{\text{input}}}\left(\frac{K}{K_{hm}}\right)^{1+\hat{\zeta}_{\text{input}}}\left(\frac{n_{hm}}{n}\right)^{\hat{\zeta}_{\text{input}}}.$$

We set $\hat{\Phi}_{hm}(0,0) = 0$.

Let $\mathcal{A}_{hm}(n)$ collect the price and capacity adjustments evaluated at scanner count n. Writing $\Pi_{h',m}$ for operating profits from modality m, define

$$V_{hm}(n) = \max_{(\boldsymbol{p},\boldsymbol{k})\in\mathcal{A}_{hm}(n)}\left\{\sum_{h'}\mathcal{H}_{hh'}\Pi_{h',m}(\boldsymbol{p},\boldsymbol{k}) - \hat{\Phi}_{hm}(k_{h\,\text{pub}\,m} + k_{h\,\text{pvt}\,m}, n)\right\}.$$

Thus $V_{hm}(n)$ deducts scan and slot costs but excludes scanner fixed costs. Slot costs at other providers are unchanged and cancel from the comparisons. At $n = n_{hm}$, the maintained optimality of observed prices and capacities implies that $V_{hm}(n_{hm})$ equals the payoff at the observed choices.

For each evaluated alternative count, scanner optimality requires

$$V_{hm}(n_{hm}) - n_{hm}\chi_{hm} \geqslant V_{hm}(n) - n\chi_{hm}.$$

For additions, rearranging gives a lower bound on χ_{hm}; for removals, it gives an upper bound. Combining the comparisons yields

$$\max_{n>n_{hm}}\frac{V_{hm}(n) - V_{hm}(n_{hm})}{n - n_{hm}} \leqslant \chi_{hm} \leqslant \min_{n<n_{hm}}\frac{V_{hm}(n_{hm}) - V_{hm}(n)}{n_{hm} - n},$$

where the extrema range over evaluated feasible counts. When no count on one side is evaluated, only a one-sided bound is available. An empty interval means that no scanner fixed cost satisfies all the evaluated comparisons. These are outer bounds because additional feasible deviations could tighten them.

At unchanged prices and capacities, scanner counts affect profits only through slot and scanner costs. The one-scanner comparisons then give

$$\chi_{hm} \geqslant \frac{MR_{hm}K_{hm}}{1+\hat{\zeta}_{\text{input}}}\left[1 - \left(\frac{n_{hm}}{n_{hm}+1}\right)^{\hat{\zeta}_{\text{input}}}\right],$$

$$\chi_{hm} \leqslant \frac{MR_{hm}K_{hm}}{1+\hat{\zeta}_{\text{input}}}\left[\left(\frac{n_{hm}}{n_{hm}-1}\right)^{\hat{\zeta}_{\text{input}}} - 1\right].$$

The upper restriction applies when $n_{hm} \geqslant 2$ and the observed total capacity can be offered with one fewer scanner. The other evaluated price and capacity adjustments can tighten both bounds.

D.9 Inference

↩ Main text: §5.3 Estimation / Patient preferences, private capacity, and scan costs · §5.4 Estimation / Capacity and scanner costs

Joint estimator We conduct inference with respect to sampling variation in the patient choice micromoments and the public shares used to recover provider quality. Let $\Psi_n(\boldsymbol{\beta}) \coloneqq \mathrm{d}Q_n(\boldsymbol{\beta})/\mathrm{d}\boldsymbol{\beta}$, let $D \coloneqq \partial\Psi(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}'$, and let Σ denote the asymptotic covariance matrix of $\sqrt{n}\,\Psi_n(\boldsymbol{\beta}_0)$. Under standard regularity conditions for Z-estimation, the estimated covariance matrix of $\hat{\boldsymbol{\beta}}_n$ is

$$\hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_n) = \frac{1}{n}\hat{D}^{-1}\hat{\Sigma}\hat{D}^{-1\prime}.$$

We estimate D by centered finite differences of the criterion gradient, recomputing the inner solutions and concentrating out marginal scan cost parameters at each perturbation.

Let $\boldsymbol{m}_n^{\mathrm{M}}$ collect the patient choice micromoments and $\boldsymbol{m}_n^{\mathrm{S}}$ the public-share moments entering the quality inversion. Their asymptotic covariance matrices are denoted by S^{M} and S^{S}, respectively. Following the notation and method of Conlon and Gortmaker (2025), we express the micromoments as smooth functions $\boldsymbol{m}_n^{\mathrm{M}} = f(\bar{\boldsymbol{v}}_n)$ of averages of patient-level moment parts $\boldsymbol{v}_i$. The delta method gives

$$\hat{S}^{\mathrm{M}} = \frac{\partial f(\bar{\boldsymbol{v}}_n)}{\partial \boldsymbol{v}'}\hat{S}^{\mathrm{V}}\left(\frac{\partial f(\bar{\boldsymbol{v}}_n)}{\partial \boldsymbol{v}'}\right)',$$

where $\hat{S}^{\mathrm{V}}$ is the empirical covariance matrix of the patient-level moment parts. We obtain $\hat{S}^{\mathrm{S}}$ from the multinomial covariance of provider choices within scan-urgency groups, treating patient records as independent draws within each group.

Define $G^a \coloneqq \mathrm{d}\Psi(\boldsymbol{\beta}_0)/\mathrm{d}(\boldsymbol{m}^a)'$ for $a \in \{\mathrm{M}, \mathrm{S}\}$. Assuming that sampling errors in the two sets of moments are uncorrelated, we estimate $\hat{\Sigma} = \hat{G}^{\mathrm{M}}\hat{S}^{\mathrm{M}}\hat{G}^{\mathrm{M}\prime} + \hat{G}^{\mathrm{S}}\hat{S}^{\mathrm{S}}\hat{G}^{\mathrm{S}\prime}$. The derivatives $\hat{G}^{\mathrm{M}}$ and $\hat{G}^{\mathrm{S}}$ account for the responses of provider quality, private capacity, stationary queues, and recovered marginal costs to changes in the empirical moments.

Slot and scanner costs We compute standard errors for $\hat{\boldsymbol{\zeta}}$ using the two-step GMM covariance estimator, accounting for uncertainty in the preceding estimates of patient preferences, and provider qualities and capacities. Let $\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\delta}}', \hat{\boldsymbol{k}}', \hat{\boldsymbol{\zeta}}')'$, and let $\hat{V}_\vartheta$ denote its estimated joint covariance matrix.

For scanner costs, we invert the generalized moment selection tests of Andrews and Soares (2010) separately for each provider-modality, using bootstrap critical values. Let $\mathcal{D}_{hm}$ denote the evaluated feasible deviations, and let n_{hmd} be the scanner count under deviation d. For a

candidate cost χ, define

$$\hat{g}_{hmd}(\chi) = (n_{hmd} - n_{hm})\chi - \left[\Delta\hat{\Pi}_{hmd} - \Delta\hat{\Phi}_{hmd}\right],$$

where $\Delta\hat{\Pi}_{hmd}$ is the estimated change in operating profits, including profits internalized through common ownership, and $\Delta\hat{\Phi}_{hmd}$ is the estimated change in slot costs. Optimality requires $g_{hmd}(\chi_{hm}) \geqslant 0$ for every evaluated deviation. Stacking the estimated comparisons into $\hat{\boldsymbol{g}}_{hm}(\chi)$, we estimate their sampling covariance as

$$\hat{S}_{hm} = \hat{J}_{hm}\hat{V}_{\vartheta}\hat{J}'_{hm}, \qquad \hat{J}_{hm} = \frac{\partial \boldsymbol{g}_{hm}(\chi;\hat{\boldsymbol{\vartheta}})}{\partial \boldsymbol{\vartheta}'}.$$

Let $\hat{s}_{hmd}$ be the square root of the corresponding diagonal entry of $\hat{S}_{hm}$. Using the statistic in equation (3.4) of Andrews and Soares (2010), define

$$t_{hmd}(\chi) = \frac{\hat{g}_{hmd}(\chi)}{\hat{s}_{hmd}}, \qquad T_{hm}(\chi) = \sum_{d\in\mathcal{D}_{hm}} \left[\min\{t_{hmd}(\chi), 0\}\right]^2.$$

For critical-value calculation, their moment-selection rule retains

$$\hat{\mathcal{D}}_{hm}(\chi) = \left\{d \in \mathcal{D}_{hm} : t_{hmd}(\chi) \leqslant \kappa_N\right\}, \qquad \kappa_N = \sqrt{\log N}.$$

We obtain critical values by a parametric bootstrap from the estimated joint asymptotic distribution. For each replication $b = 1, \ldots, B$, we draw $\hat{\boldsymbol{\vartheta}}^{(b)} \sim \text{Normal}\big(\hat{\boldsymbol{\vartheta}}, \hat{V}_{\vartheta}\big)$. At each draw, we reconstruct marginal scan costs and slot-cost schedules and recompute the deviation payoffs. The same joint draw is used for all providers and comparisons. Writing $g_{hmd}(\chi;\hat{\boldsymbol{\vartheta}}^{(b)})$ for the resulting comparison, define

$$T^{(b)}_{hm}(\chi) = \sum_{d\in\hat{\mathcal{D}}_{hm}(\chi)} \left[\min\left\{\frac{g_{hmd}(\chi;\hat{\boldsymbol{\vartheta}}^{(b)}) - \hat{g}_{hmd}(\chi)}{\hat{s}_{hmd}}, 0\right\}\right]^2.$$

The moment-selection set and standard errors are held at their original estimates across replications. We take $\hat{c}_{hm,1-\alpha}(\chi)$ to be the empirical $(1-\alpha)$ quantile of $\{T^{(b)}_{hm}(\chi)\}_{b=1}^{B}$.

Inverting the tests gives

$$\mathcal{C}_{hm,1-\alpha} = \left\{\chi : T_{hm}(\chi) \leqslant \hat{c}_{hm,1-\alpha}(\chi)\right\}.$$

We report the smallest interval containing this confidence set. Under the regularity and bootstrap-consistency conditions of Andrews and Soares (2010), the coverage target is the provider-specific cost χ_{hm}. Testing the individual comparisons underlying $\hat{\chi}^{\text{lower}}_{hm}$ and $\hat{\chi}^{\text{upper}}_{hm}$ accommodates uncertainty about which deviations determine those bounds.

D.10 Imputation of unscraped private prices and wait times

↩ Main text: §5.3 Estimation / Patient preferences, private capacity, and scan costs

Scraped private-pay prices and appointment availability cover only part of patients' choice sets. We impute missing values to retain these alternatives when computing patient choice probabilities. Pricing optimality conditions are imposed only for scans with observed prices, and private capacity is recovered only for queues with scraped waits.

We predict log annual modal prices using ensembles of 500 randomized regression trees. Predictors include scan codes, reimbursement rates, and providers' locations and hospital groups. When a provider has observed prices for other modalities, we also use summaries of these prices as predictors. For waits, we predict $\log(1 + w)$, where w is the weekly median scraped first-availability wait in days. Predictors include modality, provider characteristics, and calendar variables. Predictions are transformed back into euros and days, and missing observations are filled with the corresponding predictions. The completed weekly wait series are then smoothed using a centered five-week moving average. The first part of Table D5 reports median absolute and relative deviations between fitted and observed values on the training sample. Price comparisons include all scans used to train the imputation model, including those outside the estimation sample. For waits, we average observations and predictions over the same scraped weeks within each provider and modality, without smoothing.

We also compare imputed attributes with prices and waits obtained by solving the estimated model. For queues without scraped waits, we infer private capacity from predicted demand at the imputed waits and the median utilization rate among queues with estimated capacity in the same modality. For scans without observed private prices, we assign the median recovered marginal cost for the same scan. We then solve providers' pricing conditions for these prices and recompute equilibrium waits, holding observed private prices and estimated parameters fixed. The second part of Table D5 reports median absolute and relative deviations between imputed and model-implied values. Relative deviations divide each absolute difference by its corresponding imputed value before taking the median. Median relative price deviations range from 9.73% to 15.06%, while median absolute wait deviations range from 0.88 to 1.73 days.

E Estimation results

↩ Main text: §6 Estimation / Results

↩ See also: §D.3 Estimation / Private capacity · §D.7 Estimation / Slot costs · §D.8 Estimation / Scanner fixed costs

This appendix reports additional results related to the estimates summarized in Section 6. Tables E1 and E2 report the estimated levels of publicly funded and private-pay capacity.

Table D5 Imputation of prices and waits

	MRI	CT	X-ray
Fitted vs observed values			
Median price deviation			
absolute (€)	14.70	8.15	2.71
relative (%)	5.47	3.78	4.41
Median wait deviation			
absolute (days)	0.51	0.45	0.42
relative (%)	12.22	11.74	10.59
Model-implied vs imputed values			
Median price deviation			
absolute (€)	31.81	16.58	5.61
relative (%)	15.06	9.73	10.45
Median wait deviation			
absolute (days)	1.73	0.88	1.37
relative (%)	41.03	29.16	37.59

Notes: Fitted values are evaluated on the training sample. Model-implied values are obtained by solving the estimated model, holding observed prices fixed. Prices are compared by provider and scan. Waits are compared by provider and modality. Relative deviations express each absolute difference as a percentage of the corresponding observed or imputed value before taking the median.
(Click here to return to the imputation discussion in the main text.)

Table E3 reports the full set of patient demand coefficients. Figure E1 documents differences in the composition of publicly funded capacity across provider ownership.

Figure E2 relates the estimated marginal return to private capacity and the utilization of private-pay queues to illustrate the role of congestion in determining the value of additional capacity. When a private-pay queue is operated with substantial slack, an additional slot has little effect on expected throughput and therefore generates a small marginal return. As utilization approaches capacity, an additional slot is increasingly likely to translate into additional patients, and marginal returns rise sharply. The pattern is especially pronounced for CT and MRI, for which the profit per additional patient is larger, and provides an intuitive link between the queuing mechanism and the capacity-cost estimates recovered from providers' optimality conditions.

Table E1 Public capacity estimates

	Providers	Capacity						
		mean	10th	25th	50th	75th	90th	total
X-ray								
public	110	169.35	31.54	78.82	156.27	245.85	317.49	18,629
private	172	154.04	29.10	61.21	112.14	196.43	287.41	26,495
CT								
public	85	85.19	10.76	39.83	64.96	101.77	167.41	7,241
private	126	64.59	2.40	15.89	44.86	93.36	144.01	8,138
MRI								
public	69	54.32	0.22	20.85	39.85	81.64	125.14	3,748
private	134	66.95	1.99	18.12	46.80	94.25	145.35	8,971

Notes: Table reports the estimates of total publicly funded capacity in each provider, by modality and provider ownership. Reported capacity values are based on resident completed referrals and correspond to the average number of scans that can be executed in a week. See Section 5.2 for implementation details.
(Click here to return to the results section in the main text.)

Table E2 Private capacity estimates

	mean	10th	25th	50th	75th	90th	total
X-ray	177.05	17.10	48.38	95.50	203.79	342.60	12,924.92
CT	22.43	4.63	7.97	13.33	26.76	37.68	1,143.90
MRI	35.09	4.14	10.21	21.89	37.89	74.90	1,929.89

Notes: Table reports the estimates of total private-pay capacity in each provider, by modality. Reported capacity values are obtained from matching the observed average waits in our scraped data as described in Section 5.3 and Appendix D.3.
(Click here to return to the results section in the main text.)

Table E3 Estimates of patient demand

	X-ray			CT			MRI		
	Price	Wait	Dist.	Price	Wait	Dist.	Price	Wait	Dist.
Income									
$\leqslant$ 15k	-0.066	–	–	-0.023	–	–	-0.016	–	–
15–26k	-0.036	–	–	-0.016	–	–	-0.011	–	–
26–55k	-0.023	–	–	-0.012	–	–	-0.009	–	–
$\geqslant$ 55k	-0.016	–	–	-0.009	–	–	-0.007	–	–
Curvature	0.615	–	–	0.396	–	–	0.345	–	–
Urgency									
Tier 1	–	-0.003	–	–	-0.015	–	–	-0.018	–
Tier 2	–	-0.121	–	–	-0.039	–	–	-0.061	–
Tier 3	–	-0.041	–	–	-0.031	–	–	-0.031	–
Tier 4	–	-0.024	–	–	-0.015	–	–	-0.017	–
Age									
18–29	–	-0.024	-0.003	–	-0.015	-0.005	–	-0.017	-0.005
30–39	–	-0.031	-0.002	–	-0.015	-0.005	–	-0.014	-0.007
40–49	–	-0.026	-0.002	–	-0.008	-0.007	–	-0.012	-0.007
50–59	–	-0.016	-0.003	–	-0.003	-0.009	–	-0.020	-0.002
60–69	–	-0.014	-0.002	–	-0.002	-0.009	–	-0.008	-0.010
70–79	–	-0.013	-0.002	–	-0.001	-0.009	–	-0.010	-0.006
80+	–	-0.018	-0.002	–	-0.003	-0.009	–	-0.012	-0.006
Urban									
Low	–	–	-0.003	–	–	-0.005	–	–	-0.005
Mid	–	–	-0.198	–	–	-0.047	–	–	-0.034
High	–	–	-0.245	–	–	-0.075	–	–	-0.092

Notes: Each entry reports the implied utility coefficient for the indicated demographic group and modality. The price coefficient is parameterized as $\beta_m^{\text{price}} / \text{inc}(i)^{1-\theta}$, where the row labeled "Curvature" reports θ. Income is imputed as €7,500, €20,500, €40,500, and €75,000 for the four income bins, respectively. Dashes indicate that the corresponding demographic characteristic does not affect that coefficient. Prices are measured in euros, waits in days, and distances in minutes. Standard errors in progress, pending export from the protected environment by government personnel.

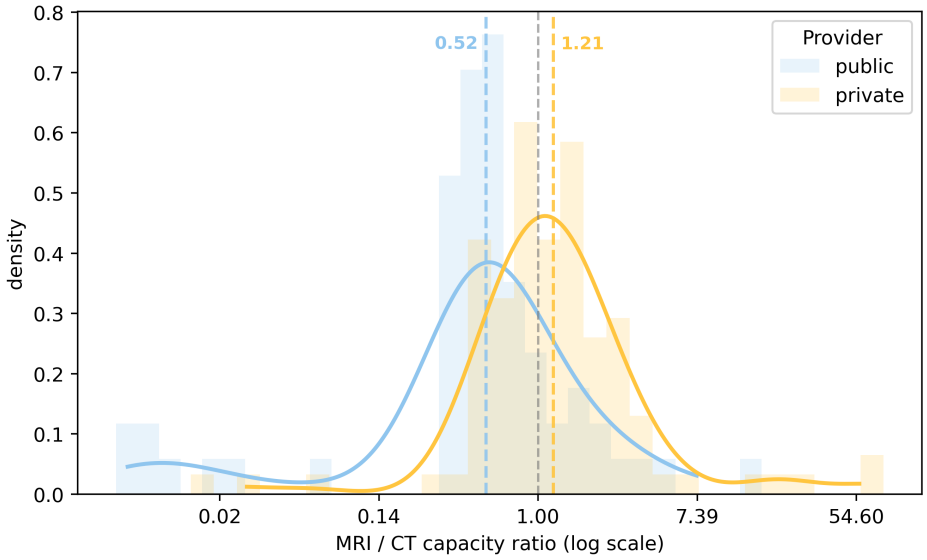
(Click here to return to the results section in the main text.)

Table E4 Estimates of provider cost parameters

	Estimate
Scan marginal costs	
X-ray local wage, $\gamma_{\text{XR,wage}}$	2.82
CT local wage, $\gamma_{\text{CT,wage}}$	7.25
MRI local wage, $\gamma_{\text{MRI,wage}}$	6.10
Slot costs	
Input intensity, ζ_{input}	0.56
Modality intercepts	Yes
Provider-group indicators	Yes

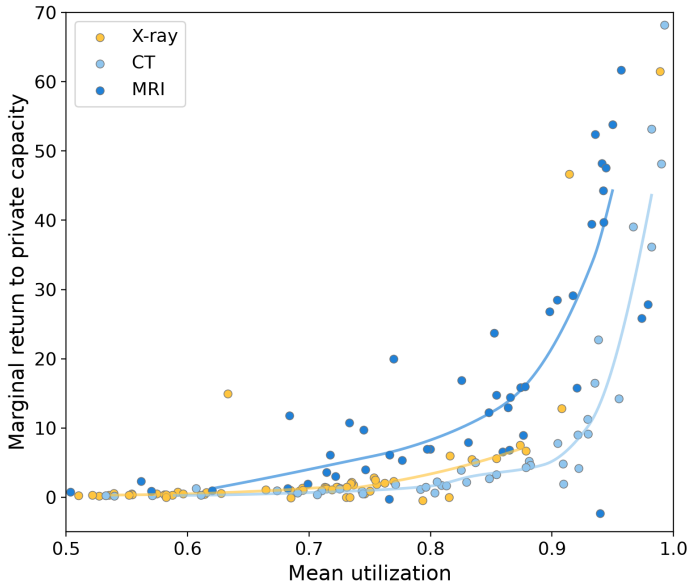
Notes: The first panel reports modality-specific coefficients on log local wages in the scan marginal-cost equation. The second panel reports the slot-cost elasticity estimated by instrumental variables using predicted private demand. The specification includes modality intercepts and provider-group indicators.
(Click here to return to the results section in the main text.)

Figure E1 Publicly funded supply mix



Notes: Figure reports the distribution of the ratio of MRI and CT publicly funded slots across providers, by provider ownership. Dashed vertical lines identify the median of each distribution.
(Click here to return to the results section in the main text.)

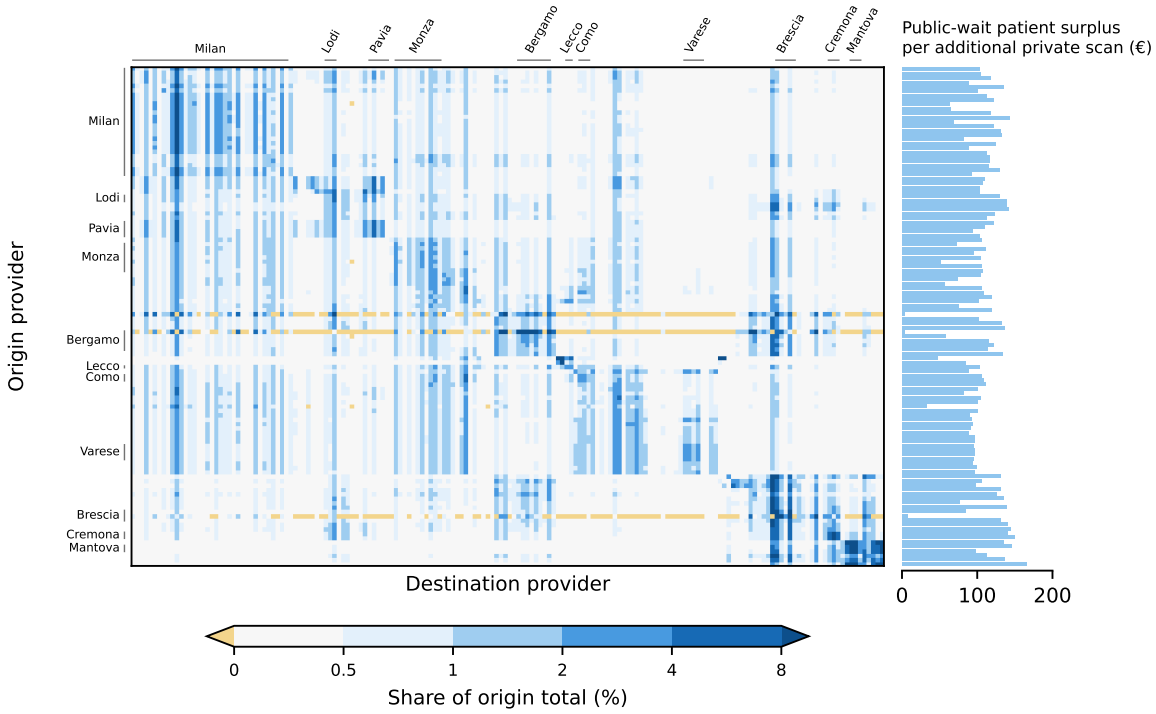
Figure E2 Returns to private capacity and utilization



Notes: Figure pltos the correlation between providers' marginal return to private capacity and the utilization rate of their private queues, by modality. Utilization rates are defined as the ratio between the model-implied expected throughput and the weekly mean capacity.
(Click here to return to the results section in the main text.)

Figure E3 documents the geographic distribution of the congestion spillovers following a 1% reduction in each provider's private-pay MRI prices. This implements the same exercise as Figure 2 for all providers in the data. Table E5 reports the welfare effects of marginal reductions in private-pay prices and expansions of private capacity. Price reductions generate substantially larger welfare gains than capacity expansions, while the vast majority of the patient-surplus benefits from shorter public waits accrue to patients at other providers.

Figure E3 Geographic distribution of congestion spillovers



Notes: Figure shows the geographic distribution of patient surplus gains through shorter public waits following a 1% reduction in all private-pay MRI prices at each "origin" provider. Effects are evaluated at the status quo, holding other providers' prices and all capacities fixed while waiting times adjust endogenously. Each row corresponds to a provider reducing its private-pay prices, and each column to a "destination" provider whose publicly funded patients are affected. There are more columns than rows because publicly owned providers are included among the columns but not among the rows, as they do not set prices. Colors represent the share of the originating provider's total patient-surplus gain accruing to patients queuing at each destination. Blue indicates positive contributions and yellow indicates negative contributions. For each "origin" provider, the bars on the right report the patient surplus gain through reduced public waits per additional private scan induced directly by the price reduction. This normalization accounts for the fact that different locations see different demand responses to price reductions. Providers on both axes are ordered using a Hilbert curve of their geographic coordinates, which approximately preserves proximity. Providers within each metropolitan area are grouped together.
(Click here to return to the results section in the main text.)

Table E5 Welfare effects of private prices and capacity

	X-ray	CT	MRI
Private-pay prices (1% reduction)			
Total welfare change (€/week)	12,965	4,098	10,590
Patient surplus through public waits (€/week)	10,759	1,774	6,092
Share at other providers (%)	92.2	94.8	95.6
Private-pay capacity (1% increase)			
Total welfare change (€/week)	2,954	289	712
Patient surplus through public waits (€/week)	1,602	128	447
Share at other providers (%)	92.2	93.2	95.8

Notes: Table reports welfare effects at the status quo, with wait times adjusting through the stationary queue system. The private-pay price rows reduce all private-pay prices of a modality by 1%, holding capacities fixed. The private-pay capacity rows raise private-pay capacity by 1% at every provider with private-pay capacity, holding public capacity, prices, and scanner counts fixed. *Total welfare* changes include patient surplus, provider operating surplus, public treatment payments, and capacity costs. *Patient surplus through public waits* is the change in patient surplus caused by changes in public waiting times. *Share at other providers* is the part of that change accruing to patients of public queues at providers other than the one whose price or capacity changes. *(Click here to return to the results section in the main text.)*

F Counterfactuals

F.1 Setup

↩ Main text: §7 Counterfactuals

↩ See also: §D.2 Estimation / Stationary wait approximation · §D.6 Estimation / Differentiation

The counterfactuals keep the distribution of referrals, public co-pays, and reimbursement rates fixed at their observed levels. We construct the status quo equilibrium used as the benchmark for our counterfactuals by combining estimated and imputed provider costs, capacities, and prices as follows.

1. *Scan costs* Scraped providers are assigned the scan costs recovered by inverting pricing optimality conditions at observed prices. Unscraped private providers and public providers are assigned imputed scan costs. For each scan, we impute its marginal cost using the median recovered cost among scraped providers.

2. *Capacities* Public capacities are recovered as described in Section 5.2 and are divided across urgency tiers using the status quo proportions. Scraped providers are assigned the private capacities $\hat{\boldsymbol{k}}_{\text{pvt}}$ estimated following the procedure in Section 5.3. Unscraped providers are assigned imputed private capacities. We assign them a private capacity by dividing model-predicted private arrivals at historical wait times by the median

estimated utilization among scraped providers in the same modality, accounting for weekly capacity fluctuations.

3. *Prices* Taking these capacities as given, we obtain prices for all providers (including unscraped ones) by solving the price equilibrium at the estimated provider qualities $\hat{\boldsymbol{\delta}}$ and patient preferences $\hat{\boldsymbol{\beta}}$.

4. *Slot costs* For privately owned scraped and unscraped providers, we recover shocks to marginal slot costs η_{hm} by equating their marginal slot cost to the marginal return to private capacity at the capacities described in point 2 and prices described in point 3. Together with $\hat{\boldsymbol{\zeta}}$, this pins down their slot cost technology via (4). For publicly owned providers, we use the fitted cost schedule with $\eta_{hm} = 0$. We hold these shocks fixed across counterfactuals.

5. *Scanner fixed costs* For scraped providers, we assign scanner fixed costs at the upper bounds obtained in Appendix D.8. For unscraped private providers, we obtain an upper bound using the capacities and slot costs assigned above, requiring that removing scanners while maintaining these capacities would not increase profits. We assign these providers this upper bound. Public providers are assigned the modality-specific median. Denote the scanner fixed costs assigned to provider h in modality m by $\hat{\chi}_{hm}$. For X-ray, we retain the scanner convention used in estimation of setting $n_{h\,\mathrm{XR}} = 1$ and $\chi_{h\,\mathrm{XR}} = 0$ for all providers h. We hold these fixed costs constant across counterfactuals.

The resulting prices and capacities, together with observed scanner counts, define the status quo equilibrium against which we evaluate our counterfactuals.

Let $K_{hm} = k_{h\,\mathrm{pub}\,m} + k_{h\,\mathrm{pvt}\,m}$ denote total capacity at provider h and modality m, and write $\Phi_{hm}(K_{hm}, n_{hm}) = n_{hm}\phi_{hm}(K_{hm}/n_{hm})$. Scanner counts are nonnegative integers, and capacity satisfies $K_{hm} \leqslant \bar{K}_m n_{hm}$, where $\bar{K}_m$ is the maximum number of weekly slots a scanner can sustain.[52]

We use superscript 0 to refer to the status quo. Money-metric patient surplus, net of patient payments, is

$$CS(\boldsymbol{p}, \boldsymbol{k}) = \mathbb{E}\left[\int_{\mathcal{I}_t} \frac{1}{\beta^{\mathrm{price}}_{g(i)}} \ln\left(1 + \sum_{j \in \mathcal{J}_i} \exp\left\{\delta_{jm(i)} - \beta^{\mathrm{price}}_{g(i)} p_{js(i)} - \beta^{\mathrm{wait}}_{g(i)} w_{q(i,j)t} - \beta^{\mathrm{dist}}_{g(i)} d_{\ell(i)h(j)}\right\}\right) \mathrm{d}i \right].$$

The expectation is over weekly referral sets and the stationary distribution of wait times induced by $(\boldsymbol{p}, \boldsymbol{k})$. Dividing by each patient's price coefficient expresses surplus in euros, and integrating over weekly referrals gives euros per week.

[52] We impose upper bounds of 232 weekly slots per MRI scanner and 274 per CT scanner. With an assumption of 30-minute MRI appointments and 20-minute CT appointments, these correspond to 17 and 13 hours per day over a seven-day week.

Total welfare in euros per week is

$$\mathcal{W}(\boldsymbol{p},\boldsymbol{k},\boldsymbol{n}) = \underbrace{CS(\boldsymbol{p},\boldsymbol{k})}_{\text{patient surplus}} + \underbrace{\sum_{h,s}\left[(p_{h\,\text{pvt}\,s}-c_{hs})a_{h\,\text{pvt}\,s}+(r_s-c_{hs})a_{h\,\text{pub}\,s}\right]}_{\text{provider operating profits}} - \underbrace{\sum_{h,m}\Phi_{hm}(K_{hm},n_{hm})}_{\text{slot costs}}$$
$$\underbrace{-\sum_{h,m}\hat{\chi}_{hm}n_{hm}}_{\text{scanner costs}} \underbrace{-\sum_{h,s}(r_s-p_{h\,\text{pub}\,s})a_{h\,\text{pub}\,s}}_{\text{government expenditure}}$$
$$= \underbrace{CS(\boldsymbol{p},\boldsymbol{k})+\sum_{h,e,s}p_{hes}a_{hes}}_{\text{gross patient benefits}} \underbrace{-\sum_{h,e,s}c_{hs}a_{hes}}_{\text{scan costs}} \underbrace{-\sum_{h,m}\Phi_{hm}(K_{hm},n_{hm})}_{\text{slot costs}} \underbrace{-\sum_{h,m}\hat{\chi}_{hm}n_{hm}}_{\text{scanner costs}}.$$

Recall that p_{hes} denotes the amount paid by the patient. For private-pay care, this is the provider's full revenue per scan, whereas for publicly funded care the provider receives reimbursement r_s, financed by the patient's co-pay $p_{h\,\text{pub}\,s}$ and public funds $r_s - p_{h\,\text{pub}\,s}$. Patient payments cancel between patient surplus and provider revenues, while public funding cancels between provider revenues and government expenditure. Welfare therefore equals gross patient benefits minus scan, slot, and scanner costs. This measure assigns equal welfare weight to euros accruing to patients, providers, and taxpayers and assumes no excess cost of raising public funds.

Under *reimbursement financing*, expenditure on modality-m publicly funded care is

$$G_m^R(\boldsymbol{k}) = \sum_h \bar{r}_m k_{h\,\text{pub}\,m}.$$

The status quo modality budget is $B_m^0 = G_m^R(\boldsymbol{k}^0)$. As in the provider problem, this expenditure measure values public slots at the modality's mean reimbursement. Recall that $k_{h\,\text{pub}\,m}$ is a measure of allocated slots to publicly funded care within a provider-modality and some slots can in principle go unused. As discussed in Section 4.3, $k_{h\,\text{pub}\,m}$ coincides with used slots whenever the allocation features persistent congestion, and in that case $G_m^R(\boldsymbol{k})$ coincides with reimbursement paid on *performed* scans, rather than on committed slots.

Under *direct financing* of a public-only system, expenditure instead covers the resources used to deliver care:

$$G_m^D(\boldsymbol{k},\boldsymbol{n}) = \sum_h \sum_{s:m(s)=m} c_{hs}a_{h\,\text{pub}\,s}(\boldsymbol{k}) + \sum_h \Phi_{hm}(k_{h\,\text{pub}\,m},n_{hm}) + \sum_h \hat{\chi}_{hm}n_{hm}.$$

The government finances scanner costs under this regime, including the costs of scanners already installed in the status quo.

F.2 Disallowing private-pay care

↩ Main text: §7.1 Counterfactuals / The value of market provision

↩ See also: §C.3 Theory / Provider optimality conditions · §F.4 Counterfactuals / Computation

All public-only counterfactuals disallow private-pay care and impose $k_{h\,\mathrm{pvt}\,m} = 0$. Patients choose among publicly funded alternatives and foregoing care.

Decentralized public-only equilibrium Private providers choose public capacity and scanner counts to maximize profits:

$$\begin{aligned} \max_{\boldsymbol{k}_{h\,\mathrm{pub}},\boldsymbol{n}_h} \quad & \sum_s (r_s - c_{hs}) a_{h\,\mathrm{pub}\,s} - \sum_m \Phi_{hm}(k_{h\,\mathrm{pub}\,m}, n_{hm}) - \sum_m \hat{\chi}_{hm} n_{hm} \\ \text{sub} \quad & \sum_m \bar{r}_m k_{h\,\mathrm{pub}\,m} \leqslant R_h. \end{aligned}$$

Each provider takes other providers' choices as given and accounts for the effect of its own choices on equilibrium waits. Provider reimbursement caps remain at their status quo levels. Public providers retain their status quo capacities and scanner stocks.

Planner with the current budget The planner chooses public capacities and scanner counts across all providers to maximize welfare subject to each modality's status quo reimbursement budget:

$$\begin{aligned} \max_{\boldsymbol{k}_{\mathrm{pub}},\boldsymbol{n}} \quad & \mathcal{W}(\boldsymbol{k}_{\mathrm{pub}}, \boldsymbol{n}) \\ \text{sub} \quad & G_m^R(\boldsymbol{k}_{\mathrm{pub}}) \leqslant B_m^0, \ \forall m. \end{aligned}$$

The planner can redistribute funding across providers within a modality, but resources cannot be transferred across modalities.

Expanded reimbursement budget For each modality m, we find the smallest reimbursement expenditure that allows the public-only planner to attain its status quo welfare. We allow the planner to incur a marginal cost of public funds $\kappa \geqslant 1$ when raising resources beyond the status quo budget. The planner solves

$$\begin{aligned} \min_{\boldsymbol{k}_{\mathrm{pub}\,m},\boldsymbol{n}_m} \quad & G_m^R(\boldsymbol{k}_{\mathrm{pub}\,m}) \\ \text{sub} \quad & \mathcal{W}_m(\boldsymbol{k}_{\mathrm{pub}\,m}, \boldsymbol{n}_m) \geqslant \mathcal{W}_m^0 \Big(\underbrace{\kappa}_{\text{MCPF}} - 1 \Big) \underbrace{\max\{G_m^R(\boldsymbol{k}_{\mathrm{pub},m}) - B_m^0, 0\}}_{\text{additional money raised}}. \end{aligned}$$

Reimbursement rates remain fixed, so additional expenditure finances more public slots. The status quo budget B_m^0 is taken as given, while funds raised beyond it have marginal cost

$\kappa \geqslant 1$. The additional term captures excess financing costs only on expenditure above B_m^0. The results in Section 7.1 set $\kappa = 1$. The budget increase reported there compares the sum of the solutions to the modality-specific minimization problems with $\sum_m B_m^0$.

Direct investment We solve the same welfare-recovery problem with expenditure evaluated at resource costs:

$$\begin{aligned} \min_{\boldsymbol{k}_{\text{pub}\,m},\, \boldsymbol{n}_m} \quad & G_m^D(\boldsymbol{k}_{\text{pub}\,m}, \boldsymbol{n}_m) \\ \text{sub} \quad & \mathcal{W}_m(\boldsymbol{k}_{\text{pub}\,m}, \boldsymbol{n}_m) \geqslant \mathcal{W}_m^0 + (\kappa - 1)\, \max\{G_m^R(\boldsymbol{k}_{\text{pub},m}) - B_m^0, 0\}. \end{aligned}$$

F.3 Efficient mixed systems

↩ Main text: §7.2 Counterfactuals / Inefficiencies in market provision

↩ See also: §C.3 Theory / Provider optimality conditions · §F.4 Counterfactuals / Computation

The mixed-system planners preserve both access modes. Public expenditure covers reimbursements and the change in slot and scanner costs:

$$G_m^M(\boldsymbol{k}, \boldsymbol{n}) = G_m^R(\boldsymbol{k}) + \sum_h \Phi_{hm} + \sum_h \hat{\chi}_{hm} n_{hm}.$$

Both planners face the status quo public budget separately for each modality.

First best The first-best planner chooses private-pay prices, public and private capacities, and scanner counts:

$$\begin{aligned} \max_{\boldsymbol{p}_{\text{pvt}},\, \boldsymbol{k},\, \boldsymbol{n}} \quad & \mathcal{W}(\boldsymbol{p}, \boldsymbol{k}, \boldsymbol{n}) \\ \text{sub} \quad & G_m^M(\boldsymbol{k}, \boldsymbol{n}) \leqslant B_m^0, \ \forall m. \end{aligned}$$

The planner can change choices at both public and private providers. Provider profits can be driven below their status quo levels and below zero.

Second best The second-best planner chooses capacities and scanner counts, while private providers set private-pay prices to maximize profits. Let $\boldsymbol{p}^*(\boldsymbol{k})$ denote the resulting pricing equilibrium. The planner solves

$$\begin{aligned} \max_{\boldsymbol{k},\, \boldsymbol{n}} \quad & \mathcal{W}(\boldsymbol{p}^*(\boldsymbol{k}), \boldsymbol{k}, \boldsymbol{n}) \\ \text{sub} \quad & G_m^M(\boldsymbol{k}, \boldsymbol{n}) \leqslant B_m^0, \ \forall m. \end{aligned}$$

The planner anticipates providers' price responses to changes in capacity. The difference from the first best measures the welfare cost of leaving private pricing to providers.

F.4 Computation

↩ Main text: §7 Counterfactuals

↩ See also: §F.2 Counterfactuals / Disallowing private-pay care · §F.3 Counterfactuals / Efficient mixed systems · §D.2 Estimation / Stationary wait approximation · §D.6 Estimation / Differentiation

Algorithms 2–7 describe the six counterfactuals. For given prices and capacities, we solve the mean backlog fixed point (D3). We denote the numerical solution initialized at $\boldsymbol{b}_0$ by $\bar{\boldsymbol{b}} = \mathcal{Q}(\boldsymbol{p}, \boldsymbol{k}; \boldsymbol{b}_0)$. Queue solutions and patient choices use the approximation in Appendix D.2.

For an objective $f(\boldsymbol{p}, \boldsymbol{k}, \bar{\boldsymbol{b}}; \boldsymbol{n})$, define $A = I - \partial\boldsymbol{\Lambda}/\partial\bar{\boldsymbol{b}}'$, evaluated at the stationary backlog. Implicit differentiation gives

$$A'\boldsymbol{\lambda}_f = \frac{\partial f}{\partial \bar{\boldsymbol{b}}}, \qquad \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{z}} = \frac{\partial f}{\partial \boldsymbol{z}} + \left(\frac{\partial \boldsymbol{\Lambda}}{\partial \boldsymbol{z}'}\right)' \boldsymbol{\lambda}_f, \quad \boldsymbol{z} \in \{\boldsymbol{p}, \boldsymbol{k}\}. \tag{F1}$$

We solve the transpose linear system for $\boldsymbol{\lambda}_f$ using GMRES. Notationwise, $\partial f/\partial \boldsymbol{z}$ holds the mean backlog fixed, while $\mathrm{d}f/\mathrm{d}\boldsymbol{z}$ includes its response to $\boldsymbol{z}$. In the algorithms, welfare and provider profits are evaluated at stationary queues, and their derivatives include these queue responses. Partial derivatives of welfare with respect to capacity hold prices fixed, as in the main text. We evaluate Jacobian products without constructing the full backlog-response matrices, following Appendix D.6.

For numerical scaling, capacity directions and projections are computed after dividing each k_{hem} by the provider's fixed status quo total capacity K^0_{hm}, then converted back to capacity units. We write Proj for the projection onto the relevant feasible choices. Capacity updates are feasible if the chosen capacity (i) is non-negative, (ii) is below the scanner-dependent physical throughput upper bounds, and (iii) it satisfied the provider's reimbursement-cap constraint (5). Projected-gradient updates use an Armijo line search. Each line search starts at $\alpha_0 > 0$. Each trial allocation recomputes stationary queues and welfare, starting from the preceding accepted backlog. Rejected trials shorten the step to $\rho\alpha$ for $\rho \in (0, 1)$, and failed queue calculations reject the trial.

Scanner choices are handled outside the continuous optimization. For each feasible one-scanner addition or removal, we reoptimize continuous choices and compare the resulting objective with its current value. A private provider reoptimizes its own capacities while holding rivals' choices fixed. A planner reoptimizes all its continuous controls under the relevant fiscal constraints. We accept an improving scanner change and repeat until none is found. X-ray scanner counts retain the convention imposed above.

Algorithm 2 Decentralized public-only equilibrium

Require: Starting capacities and scanner counts $(\boldsymbol{k}, \boldsymbol{n})$, provider caps, and assigned costs.

$\boldsymbol{k}_{\text{pvt}} \leftarrow \mathbf{0}$ and fix public providers' choices at the status quo. — Disallow private care
procedure CapacityEquilibrium$(\boldsymbol{n}, \boldsymbol{k})$
  loop
    $\bar{\boldsymbol{b}} \leftarrow \mathcal{Q}(\boldsymbol{p}, \boldsymbol{k}; \bar{\boldsymbol{b}}_{\text{prev}})$. — Stationary queues
    $\boldsymbol{g}_h \leftarrow \partial \pi_h / \partial \boldsymbol{k}_{h\,\text{pub}}$ using GMRES and (F1). — Profit derivatives
    $\boldsymbol{F}_h^k \leftarrow \boldsymbol{k}_{h\,\text{pub}} - \text{Proj}_h(\boldsymbol{k}_{h\,\text{pub}} + D_h \boldsymbol{g}_h)$. — Capacity conditions
    if $\|\boldsymbol{F}^k\|_D \leqslant \varepsilon$ **then**
      return $\boldsymbol{k}$. — Capacity convergence
    $J_k \leftarrow \mathrm{d}\boldsymbol{F}^k / \mathrm{d}\boldsymbol{k}'$ and solve $J_k \boldsymbol{d} = -\boldsymbol{F}^k$. — Newton direction
    $\alpha \leftarrow \alpha_0$. — Initial step size
    repeat
      $\boldsymbol{k}_{h,\alpha} \leftarrow \text{Proj}_h(\boldsymbol{k}_{h\,\text{pub}} + \alpha \boldsymbol{d}_h)$. — Feasible capacity
      $\bar{\boldsymbol{b}}_\alpha \leftarrow \mathcal{Q}(\boldsymbol{p}, \boldsymbol{k}_\alpha; \bar{\boldsymbol{b}})$ and recompute $\boldsymbol{F}_\alpha^k$. — Trial residual
      if queues fail or $\|\boldsymbol{F}_\alpha^k\|_D \geqslant \|\boldsymbol{F}^k\|_D$ **then**
        $\alpha \leftarrow \rho\alpha$. — Shorten step
    until the trial is accepted
    $(\boldsymbol{k}, \bar{\boldsymbol{b}}) \leftarrow (\boldsymbol{k}_\alpha, \bar{\boldsymbol{b}}_\alpha)$. — Accept capacities
repeat
  $\boldsymbol{k} \leftarrow$ CapacityEquilibrium$(\boldsymbol{n}, \boldsymbol{k})$. — Capacity equilibrium
  for all feasible one-scanner deviations $(h, \boldsymbol{n}')$ **do**
    $\tilde{\boldsymbol{k}}_h \in \arg\max_{\boldsymbol{u}\ \text{feasible}} \pi_h(\boldsymbol{u}, \boldsymbol{k}_{-h}; \boldsymbol{n}')$. — Own capacity response
    if $\pi_h(\tilde{\boldsymbol{k}}_h, \boldsymbol{k}_{-h}; \boldsymbol{n}') > \pi_h(\boldsymbol{k}; \boldsymbol{n})$ **then**
      $(\boldsymbol{n}, \boldsymbol{k}_h) \leftarrow (\boldsymbol{n}', \tilde{\boldsymbol{k}}_h)$ and **break**. — Accept scanner move
until no scanner change is accepted
return $(\boldsymbol{k}, \boldsymbol{n})$. — Return solution

Notes: $\boldsymbol{F}^k$ stacks private providers' capacity residuals. The projection $\text{Proj}_h(\boldsymbol{y})$ minimizes $(\boldsymbol{u}-\boldsymbol{y})' D_h^{-1} (\boldsymbol{u}-\boldsymbol{y})$ over provider h's feasible capacities, where D_h is positive diagonal. The weighted norm is $\|\boldsymbol{F}^k\|_D^2 = \sum_h (\boldsymbol{F}_h^k)' D_h^{-1} \boldsymbol{F}_h^k$. The residual tolerance is ε. The same scaling in the gradient step and projection ensures that a zero residual corresponds to the constrained first-order conditions. Residuals and their derivatives include stationary-wait responses.

Algorithm 3 Public-only planner with the current budget

Require: Starting $(\boldsymbol{k}, \boldsymbol{n})$, expenditure measure G, and budgets $\boldsymbol{B}$.
$\boldsymbol{k}_{\text{pvt}} \leftarrow \boldsymbol{0}$. Disallow private care
procedure PUBLICCAPACITIES($\boldsymbol{n}, \boldsymbol{B}, \boldsymbol{k}$)
  repeat
    $\bar{\boldsymbol{b}} \leftarrow \mathcal{Q}(\boldsymbol{p}, \boldsymbol{k}; \bar{\boldsymbol{b}}_{\text{prev}})$. Stationary queues
    $\boldsymbol{g}_k \leftarrow \partial \mathcal{W} / \partial \boldsymbol{k}$ using GMRES and (F1). Capacity gradient
    Compute the scaled capacity direction $\boldsymbol{d}_k$ from $\boldsymbol{g}_k$ and set $\alpha \leftarrow \alpha_0$. Scaled direction
    repeat
      $\boldsymbol{k}_\alpha \leftarrow \text{Proj}(\boldsymbol{k} + \alpha \boldsymbol{d}_k)$. Feasible capacity
      $\bar{\boldsymbol{b}}_\alpha \leftarrow \mathcal{Q}(\boldsymbol{p}, \boldsymbol{k}_\alpha; \bar{\boldsymbol{b}})$ and evaluate $\mathcal{W}_\alpha$. Trial welfare
      if queues fail, feasibility fails, or the Armijo test fails **then**
        $\alpha \leftarrow \rho \alpha$. Shorten step
    until the trial is accepted
    $(\boldsymbol{k}, \bar{\boldsymbol{b}}) \leftarrow (\boldsymbol{k}_\alpha, \bar{\boldsymbol{b}}_\alpha)$. Accept capacities
  until the constrained capacity conditions converge
  return $\boldsymbol{k}$. Continuous optimum
repeat
  $\boldsymbol{k} \leftarrow$ PUBLICCAPACITIES($\boldsymbol{n}, \boldsymbol{B}, \boldsymbol{k}$). Optimize capacities
  for all feasible one-scanner changes $\boldsymbol{n}'$ **do**
    $\boldsymbol{k}' \leftarrow$ PUBLICCAPACITIES($\boldsymbol{n}', \boldsymbol{B}, \boldsymbol{k}$). Reoptimize candidate
    if $\mathcal{W}(\boldsymbol{k}', \boldsymbol{n}') > \mathcal{W}(\boldsymbol{k}, \boldsymbol{n})$ **then**
      $(\boldsymbol{k}, \boldsymbol{n}) \leftarrow (\boldsymbol{k}', \boldsymbol{n}')$ and **break**. Accept scanner move
until no scanner change is accepted
return $(\boldsymbol{k}, \boldsymbol{n})$. Return solution

Notes: The current-budget planner sets $G = G^R$ and $\boldsymbol{B} = \boldsymbol{B}^0$. Algorithms 4 and 5 run Algorithm 3 separately for each modality, supplying their expenditure measure and trial budget. Each call reoptimizes both capacities and scanner counts and starts from the preceding solution.

Algorithm 4 Expanded reimbursement budget

Require: Welfare targets $\mathcal{W}_m^0$, assigned costs, and budget-search tolerance ε_B.
for all modalities m **do**
  Find $B_L < B_U$ with optimized welfare below $\mathcal{W}_m^0$ at B_L and at least $\mathcal{W}_m^0$ at B_U. Bracket budget
  $(\boldsymbol{k}_U, \boldsymbol{n}_U) \leftarrow$ ALGORITHM 3(m, G^R, B_U). Feasible endpoint
  while $B_U - B_L > \varepsilon_B$ **do**
    Choose $B \in (B_L, B_U)$. Trial budget
    $(\boldsymbol{k}', \boldsymbol{n}') \leftarrow$ ALGORITHM 3(m, G^R, B). Optimize trial budget
    if $\mathcal{W}_m(\boldsymbol{k}', \boldsymbol{n}') \geqslant \mathcal{W}_m^0$ **then**
      $(B_U, \boldsymbol{k}_U, \boldsymbol{n}_U) \leftarrow (B, \boldsymbol{k}', \boldsymbol{n}')$. Lower upper bound
    else
      $B_L \leftarrow B$. Raise lower bound
  $B_m^* \leftarrow B_U$. Required modality budget
return $\sum_m B_m^*$ and $\sum_m B_m^* / \sum_m B_m^0 - 1$. Budget and increase

Algorithm 5 Direct investment to recover status quo welfare

Require: Welfare targets $\mathcal{W}_m^0$, assigned costs, and budget-search tolerance ε_B.

for all modalities m **do**
  Find $B_L < B_U$ with optimized welfare below $\mathcal{W}_m^0$ at B_L and at least $\mathcal{W}_m^0$ at B_U. Bracket resource budget
  $(\boldsymbol{k}_U, \boldsymbol{n}_U) \leftarrow$ Algorithm 3(m, G^D, B_U). Feasible endpoint
  while $B_U - B_L > \varepsilon_B$ **do**
    Choose $B \in (B_L, B_U)$. Trial resource budget
    $(\boldsymbol{k}', \boldsymbol{n}') \leftarrow$ Algorithm 3(m, G^D, B). Optimize trial budget
    if $\mathcal{W}_m(\boldsymbol{k}', \boldsymbol{n}') \geqslant \mathcal{W}_m^0$ **then**
      $(B_U, \boldsymbol{k}_U, \boldsymbol{n}_U) \leftarrow (B, \boldsymbol{k}', \boldsymbol{n}')$. Lower upper bound
    else
      $B_L \leftarrow B$. Raise lower bound
  $B_m^* \leftarrow B_U$. Required resource budget
return $\sum_m B_m^*$. Total resource cost

Algorithm 6 First-best mixed system

Require: Starting prices, capacities, and scanner counts $(\boldsymbol{p}, \boldsymbol{k}, \boldsymbol{n})$ and budgets $\boldsymbol{B}^0$.

procedure MixedControls($\boldsymbol{n}, \boldsymbol{p}, \boldsymbol{k}$)
  repeat
    $\bar{\boldsymbol{b}} \leftarrow \mathcal{Q}(\boldsymbol{p}, \boldsymbol{k}; \bar{\boldsymbol{b}}_{\text{prev}})$. Stationary queues
    $\boldsymbol{g}_p \leftarrow \partial \mathcal{W} / \partial \boldsymbol{p}$ using GMRES and (F1). Price gradient
    $\alpha \leftarrow \alpha_0$. Initial price step
    repeat
      $\boldsymbol{p}_\alpha \leftarrow \text{Proj}(\boldsymbol{p} + \alpha \boldsymbol{g}_p)$. Feasible price trial
      $\bar{\boldsymbol{b}}_\alpha \leftarrow \mathcal{Q}(\boldsymbol{p}_\alpha, \boldsymbol{k}; \bar{\boldsymbol{b}})$ and evaluate $\mathcal{W}_\alpha$. Trial welfare
      if queues fail or the Armijo test fails **then**
        $\alpha \leftarrow \rho \alpha$. Shorten price step
    until the trial is accepted
    $(\boldsymbol{p}, \bar{\boldsymbol{b}}) \leftarrow (\boldsymbol{p}_\alpha, \bar{\boldsymbol{b}}_\alpha)$. Accept prices
    $\boldsymbol{g}_k \leftarrow \partial \mathcal{W} / \partial \boldsymbol{k}$ using GMRES and (F1). Capacity gradient
    Compute the scaled capacity direction $\boldsymbol{d}_k$ from $\boldsymbol{g}_k$ and set $\alpha \leftarrow \alpha_0$. Scaled direction
    repeat
      $\boldsymbol{k}_\alpha \leftarrow \text{Proj}(\boldsymbol{k} + \alpha \boldsymbol{d}_k)$. Feasible capacity
      $\bar{\boldsymbol{b}}_\alpha \leftarrow \mathcal{Q}(\boldsymbol{p}, \boldsymbol{k}_\alpha; \bar{\boldsymbol{b}})$ and evaluate $\mathcal{W}_\alpha$. Trial welfare
      if queues fail, feasibility fails, or the Armijo test fails **then**
        $\alpha \leftarrow \rho \alpha$. Shorten capacity step
    until the trial is accepted
    $(\boldsymbol{k}, \bar{\boldsymbol{b}}) \leftarrow (\boldsymbol{k}_\alpha, \bar{\boldsymbol{b}}_\alpha)$. Accept capacities
  until the constrained price and capacity conditions converge
  return $(\boldsymbol{p}, \boldsymbol{k})$. Continuous optimum
repeat
  $(\boldsymbol{p}, \boldsymbol{k}) \leftarrow$ MixedControls($\boldsymbol{n}, \boldsymbol{p}, \boldsymbol{k}$). Optimize controls
  for all feasible one-scanner changes $\boldsymbol{n}'$ **do**
    $(\boldsymbol{p}', \boldsymbol{k}') \leftarrow$ MixedControls($\boldsymbol{n}', \boldsymbol{p}, \boldsymbol{k}$). Reoptimize candidate
    if $\mathcal{W}(\boldsymbol{p}', \boldsymbol{k}', \boldsymbol{n}') > \mathcal{W}(\boldsymbol{p}, \boldsymbol{k}, \boldsymbol{n})$ **then**
      $(\boldsymbol{p}, \boldsymbol{k}, \boldsymbol{n}) \leftarrow (\boldsymbol{p}', \boldsymbol{k}', \boldsymbol{n}')$ and **break**. Accept scanner move
until no scanner change is accepted
return $(\boldsymbol{p}, \boldsymbol{k}, \boldsymbol{n})$. Return solution

Algorithm 7 Second-best mixed system

Require: Starting $(\boldsymbol{p}, \boldsymbol{k}, \boldsymbol{n})$ and budgets $\boldsymbol{B}^0$.

procedure SecondBestCapacities($\boldsymbol{n}, \boldsymbol{p}, \boldsymbol{k}$)
  loop
    repeat
      $\bar{\boldsymbol{b}} \leftarrow \mathcal{Q}(\boldsymbol{p}, \boldsymbol{k}; \bar{\boldsymbol{b}}_{\text{prev}})$. Stationary queues
      $J \leftarrow \partial \boldsymbol{F}/\partial \boldsymbol{p}'$ and $PJ = LU$. Refresh price factors
      Solve $L\boldsymbol{y} = -P\boldsymbol{F}$ and $U\boldsymbol{d}_p = \boldsymbol{y}$. Newton price direction
      $\boldsymbol{p} \leftarrow \text{Proj}(\boldsymbol{p} + \boldsymbol{d}_p)$. Correct prices
    until the pricing conditions converge
    $J \leftarrow \partial \boldsymbol{F}/\partial \boldsymbol{p}'$ and $PJ = LU$ at the corrected prices. Pricing Jacobian
    $\boldsymbol{g}_p \leftarrow \partial \mathcal{W}/\partial \boldsymbol{p}$ and $\boldsymbol{g}_k \leftarrow \partial \mathcal{W}/\partial \boldsymbol{k}$ using GMRES. Wait-adjusted derivatives
    Solve $U'\boldsymbol{y} = \boldsymbol{g}_p$ and $L'\boldsymbol{u} = \boldsymbol{y}$, then set $\boldsymbol{v} = P'\boldsymbol{u}$. Transpose price solve
    $\boldsymbol{g}_k \leftarrow \boldsymbol{g}_k - H'\boldsymbol{v}$. Allow for price responses
    if the constrained capacity conditions converge **then**
      return $(\boldsymbol{p}, \boldsymbol{k})$. Continuous optimum
    $\boldsymbol{d}_k \leftarrow \text{LBFGS}(\boldsymbol{g}_k)$ and $\alpha \leftarrow \alpha_0$. Capacity direction
    repeat
      $\boldsymbol{k}_\alpha \leftarrow \text{Proj}(\boldsymbol{k} + \alpha \boldsymbol{d}_k)$ and $\boldsymbol{p}_\alpha \leftarrow \boldsymbol{p}$. Feasible capacity trial
      repeat
        $\bar{\boldsymbol{b}}_\alpha \leftarrow \mathcal{Q}(\boldsymbol{p}_\alpha, \boldsymbol{k}_\alpha; \bar{\boldsymbol{b}})$. Trial queues
        Solve $L\boldsymbol{y}_\alpha = -P\boldsymbol{F}(\boldsymbol{p}_\alpha, \boldsymbol{k}_\alpha)$ and $U\boldsymbol{d}_{p,\alpha} = \boldsymbol{y}_\alpha$. Reuse price factors
        $\boldsymbol{p}_\alpha \leftarrow \text{Proj}(\boldsymbol{p}_\alpha + \boldsymbol{d}_{p,\alpha})$. Update trial prices
      until the trial-pricing rule holds or the trial fails
      Recompute stationary queues and $\mathcal{W}(\boldsymbol{p}_\alpha, \boldsymbol{k}_\alpha, \boldsymbol{n})$. Trial welfare
      if an inner solve fails, feasibility fails, or the line-search test fails **then**
        $\alpha \leftarrow \rho\alpha$. Shorten capacity step
    until the trial is accepted
    $(\boldsymbol{p}, \boldsymbol{k}, \bar{\boldsymbol{b}}) \leftarrow (\boldsymbol{p}_\alpha, \boldsymbol{k}_\alpha, \bar{\boldsymbol{b}}_\alpha)$. Accept capacity step
    Update the L-BFGS history when the next corrected gradient is available. Update curvature
repeat
  $(\boldsymbol{p}, \boldsymbol{k}) \leftarrow$ SecondBestCapacities$(\boldsymbol{n}, \boldsymbol{p}, \boldsymbol{k})$. Optimize capacities
  for all feasible one-scanner changes $\boldsymbol{n}'$ **do**
    $(\boldsymbol{p}', \boldsymbol{k}') \leftarrow$ SecondBestCapacities$(\boldsymbol{n}', \boldsymbol{p}, \boldsymbol{k})$. Reoptimize candidate
    if $\mathcal{W}(\boldsymbol{p}', \boldsymbol{k}', \boldsymbol{n}') > \mathcal{W}(\boldsymbol{p}, \boldsymbol{k}, \boldsymbol{n})$ **then**
      $(\boldsymbol{p}, \boldsymbol{k}, \boldsymbol{n}) \leftarrow (\boldsymbol{p}', \boldsymbol{k}', \boldsymbol{n}')$ and **break**. Accept scanner move
until no scanner change is accepted
return $(\boldsymbol{p}, \boldsymbol{k}, \boldsymbol{n})$. Return solution

Notes: the pricing optimality conditions after substituting in stationary queues are denoted by $\boldsymbol{F}(\boldsymbol{p}, \boldsymbol{k}) = \boldsymbol{0}$. Define $J = \partial \boldsymbol{F}/\partial \boldsymbol{p}'$ and $H = \partial \boldsymbol{F}/\partial \boldsymbol{k}'$, both including stationary wait responses. The planner's capacity derivative is

$$J'\boldsymbol{v} = \frac{\partial \mathcal{W}}{\partial \boldsymbol{p}}, \qquad \frac{\mathrm{d}\mathcal{W}(\boldsymbol{p}^*(\boldsymbol{k}), \boldsymbol{k}, \boldsymbol{n})}{\mathrm{d}\boldsymbol{k}} = \frac{\partial \mathcal{W}}{\partial \boldsymbol{k}} - H'\boldsymbol{v}.$$

The LU factorization $PJ = LU$ supplies both the transpose solve for $\boldsymbol{v}$ and the price corrections during backtracking. Capacity directions are computed by L-BFGS in the same normalized coordinates described above. The trial-pricing and second-best line-search stopping rules remain those of the numerical implementation.

F.5 Inference on counterfactuals

↩ Main text: §7 Counterfactuals

↩ See also: §D.9 Estimation / Inference · §D.8 Estimation / Scanner fixed costs

We quantify sampling uncertainty in counterfactual outcomes using the parametric bootstrap described in Appendix D.9. For each replication $b = 1, \ldots, B$, we draw $\hat{\boldsymbol{\vartheta}}^{(b)} \sim \text{Normal}(\hat{\boldsymbol{\vartheta}}, \hat{V}_{\vartheta})$, where $\boldsymbol{\vartheta}$ collects patient preferences, provider qualities, capacities, and slot-cost coefficients. At each draw, we reconstruct marginal scan costs and slot-cost schedules and repeat the complete capacity-imputation and cost-assignment procedures described above. This preserves dependence across providers' assigned costs and between those costs and the parameters determining operating profits and slot costs. For each replication, we reconstruct the status quo and solve each counterfactual using the same parameter draw and assigned cost vector. For each reported statistic, we report the empirical $\alpha/2$ and $1-\alpha/2$ quantiles across replications.

G Proofs

G.1 Existence of a stationary distribution of wait times

↩ Main text: §4.2 Model / Queues · §5.1 Estimation / Queuing model

↩ See also: §C.1 Theory / Queuing model · §D.2 Estimation / Stationary wait approximation

We prove the following result, stated in the main text as well. Without loss, we restrict attention to queues with strictly positive capacity: a zero-capacity queue is interpreted as a provider not offering that queue's modality and hence not entering any patient's choice set. We also maintain that patient sets $\mathcal{I}_t$ are iid and that each queue's capacity is fixed across periods. These assumptions are made for convenience of exposition: generalizing the proof's argument trades cumbersome notation for little insight. We comment on this at the end of this section.

Proposition G1 (Stationary wait times)
If $\beta^{\text{wait}}_{g(i)} > 0$ almost surely, the backlog and wait-time processes admit a stationary distribution.

Proof. Let $\mathcal{I}$ denote the population from which the random sets $\mathcal{I}_t$ are drawn and let $M := \int_{\mathcal{I}} 1 \, \mathrm{d}i < \infty$. Fix prices and capacities, with $k_q > 0$ for every $q \in \mathcal{Q}$. From the constant-capacity wait mapping in Section 4.2, the wait associated with backlog b_q is $w_q(b_q) = \frac{b_q}{k_q}$.

Step 1: arrivals to a queue vanish as its own backlog diverges.

Fix q. For a patient i who can join queue q, let $\bar{v}_{iq}$ collect the deterministic utility of the corresponding alternative excluding its wait-time component. Since the outside option has utility normalized to zero before the logit shock, the logit choice probability satisfies

$$\sigma_{iq}(\boldsymbol{p}, \boldsymbol{w}) \leqslant \min\left\{1, \exp\left(\bar{v}_{iq} - \beta^{\text{wait}}_{g(i)} w_q\right)\right\}.$$

Arrivals into queue q then satisfy

$$a_{qt} = \int_{\mathcal{I}_t} \sigma_{iq}(\boldsymbol{p}, \boldsymbol{w}_t)\,\mathrm{d}i \leqslant \int_{\mathcal{I}_t} \min\left\{1, \exp\left(\bar{v}_{iq} - \beta^{\text{wait}}_{g(i)} w_{qt}\right)\right\}\mathrm{d}i. \tag{G1}$$

Since $\beta^{\text{wait}}_{g(i)} > 0$ almost surely, the integrand on the right-hand side of (G1) converges to zero for almost every i as $w_{qt} \to \infty$. Moreover, since $\mathcal{I}_t \subseteq \mathcal{I}$,

$$a_{qt} \leqslant \int_{\mathcal{I}} \min\left\{1, \exp\left(\bar{v}_{iq} - \beta^{\text{wait}}_{g(i)} w_{qt}\right)\right\}\mathrm{d}i.$$

The integrand is bounded by one and $\mathcal{I}$ has finite mass, so the dominated convergence theorem implies that the right-hand side converges to zero as $w_{qt} \to \infty$. Since $w_{qt} = b_{qt}/k_q$, there exists $\bar{b}_q < \infty$ such that

$$b_{qt} \geqslant \bar{b}_q \implies a_{qt} \leqslant \frac{k_q}{2} \tag{G2}$$

for every realization of $\mathcal{I}_t$ and every value of the other queues' backlogs. It is easy to confirm that we can take $\bar{b}_q$ to satisfy $\bar{b}_q \geqslant k_q$.

Step 2: the backlog law of motion admits an invariant set.

The backlog law is

$$b_{q\,t+1} = \max\{b_{qt} + a_{qt} - k_q, 0\}. \tag{G3}$$

Hence, (G2) and (G3) give

$$b_{qt} \geqslant \bar{b}_q \implies b_{q\,t+1} \leqslant b_{qt} - \frac{k_q}{2}. \tag{G4}$$

Since we additionally have $a_{qt} \leqslant M$ for every q and t, it follows that the chain never leaves the following set

$$\mathcal{B} := \bigtimes_{q \in \mathcal{Q}} [0, \bar{b}_q + M].$$

Indeed, if $b_{qt} < \bar{b}_q$, then $b_{q\,t+1} \leqslant b_{qt} + a_{qt} \leqslant \bar{b}_q + M$; if $b_{qt} \geqslant \bar{b}_q$, (G4) implies $b_{q\,t+1} \leqslant b_{qt}$. Since $\mathcal{Q}$ is finite, $\mathcal{B}$ is a compact subset of $\mathbb{R}^{|\mathcal{Q}|}$. We can therefore restrict the backlog process to the compact state space $\mathcal{B}$.

Step 3: the Markov transition kernel on $\mathcal{B}$ admits an invariant probability measure.

It remains to establish that the backlog process restricted to $\mathcal{B}$ admits an invariant probability measure. Let λ denote the distribution of the iid random sets $\mathcal{I}_t$, and let $K(\boldsymbol{b}, I)$ denote the next-period backlog vector obtained from current backlog $\boldsymbol{b}$ and realization I of the health

shock through (G1) and (G3). The transition kernel of the backlog process is therefore

$$\nu(\boldsymbol{b}, A) := \Pr(K(\boldsymbol{b}, \mathcal{I}_t) \in A) = \int \mathbf{1}\{K(\boldsymbol{b}, I) \in A\}\lambda(\mathrm{d}I),$$

where A is a generic Borel-measurable subset of $\mathcal{B}$. Since $\mathcal{I}_t$ is independent of the history of the process and $\boldsymbol{b}_{t+1} = K(\boldsymbol{b}_t, \mathcal{I}_t)$, the conditional distribution of $\boldsymbol{b}_{t+1}$ given the history depends on the history only through $\boldsymbol{b}_t$, making $(\boldsymbol{b}_t)_t$ Markov. Since the $\mathcal{I}_t$ are identically distributed, the kernel ν does not depend on t, so the chain is time homogeneous.

For each realization I, the map $\boldsymbol{b} \mapsto K(\boldsymbol{b}, I)$ is continuous because $\boldsymbol{b} \mapsto \boldsymbol{w}(\boldsymbol{b})$ is continuous, logit choice probabilities are continuous in wait times, continuity passes through the integral in (G1) by dominated convergence, and the maximum operator in (G3) is continuous. Hence, dominated convergence implies that

$$\nu f(\boldsymbol{b}) := \int_{\mathcal{B}} f(\boldsymbol{b}')\nu(\boldsymbol{b}, \mathrm{d}\boldsymbol{b}') = \int f(K(\boldsymbol{b}, I))\lambda(\mathrm{d}I)$$

is continuous in $\boldsymbol{b}$ for every bounded continuous function $f : \mathcal{B} \to \mathbb{R}$. We conclude that ν is a Feller kernel, that is, $\nu f(\boldsymbol{b}) := \int_{\mathcal{B}} f(\boldsymbol{b}')\nu(\boldsymbol{b}, \mathrm{d}\boldsymbol{b}')$ is continuous in $\boldsymbol{b}$ for every bounded continuous function $f : \mathcal{B} \to \mathbb{R}$ on $\mathcal{B}$ (Meyn and Tweedie, 1993, Section 6.1.1).

For a probability measure ρ on $\mathcal{B}$, write $(\rho\nu)(A) := \int_{\mathcal{B}} \nu(\boldsymbol{b}, A)\rho(\mathrm{d}\boldsymbol{b})$ for its distribution after one transition. Define the t-period transition kernels recursively by

$$\nu^0(\boldsymbol{b}, \cdot) := \delta_{\boldsymbol{b}}, \qquad \nu^{t+1}(\boldsymbol{b}, \cdot) := \nu^t(\boldsymbol{b}, \cdot)\nu,$$

where $\delta_{\boldsymbol{b}}$ denotes the Dirac measure concentrated on $\boldsymbol{b}$. Thus $\nu^t(\boldsymbol{b}, \cdot)$ is the distribution of $\boldsymbol{b}_t$ conditional on $\boldsymbol{b}_0 = \boldsymbol{b}$.

Choose any $\boldsymbol{b}_0 \in \mathcal{B}$, let $\mu_t := \nu^t(\boldsymbol{b}_0, \cdot)$, and define the Cesàro averages $\bar{\mu}_T := \frac{1}{T}\sum_{t=0}^{T-1} \mu_t$. Since $\mathcal{B}$ is compact, the set of probability measures on $\mathcal{B}$ is sequentially compact under the topology of weak convergence (Aliprantis and Border, 2006, Theorems 3.28 and 15.11). Hence there is a subsequence $(\bar{\mu}_{T_r})_r$ of $(\bar{\mu}_T)_T$ converging weakly to a probability measure π. By construction, $\mu_t\nu = \mu_{t+1}$, so

$$\bar{\mu}_T\nu - \bar{\mu}_T = \frac{1}{T}(\mu_T - \mu_0). \tag{G5}$$

Therefore, for every bounded continuous function f,

$$\int_{\mathcal{B}} f \,\mathrm{d}(\bar{\mu}_T\nu) - \int_{\mathcal{B}} f \,\mathrm{d}\bar{\mu}_T = \frac{1}{T}\left[\int_{\mathcal{B}} f \,\mathrm{d}\mu_T - \int_{\mathcal{B}} f \,\mathrm{d}\mu_0\right] \to 0.$$

Since ν is Feller, νf is continuous. Moreover, $|\nu f(\boldsymbol{b})| \leqslant \int_{\mathcal{B}} |f(\boldsymbol{b}')|\nu(\boldsymbol{b}, \mathrm{d}\boldsymbol{b}') \leqslant \|f\|_\infty$, so νf is

bounded. Weak convergence of $\bar{\mu}_{T_r}$ to π therefore implies

$$\int_{\mathcal{B}} f \, \mathrm{d}(\bar{\mu}_{T_r}\nu) = \int_{\mathcal{B}} \nu f \, \mathrm{d}\bar{\mu}_{T_r} \to \int_{\mathcal{B}} \nu f \, \mathrm{d}\pi = \int_{\mathcal{B}} f \, \mathrm{d}(\pi\nu),$$
$$\int_{\mathcal{B}} f \, \mathrm{d}\bar{\mu}_{T_r} \to \int_{\mathcal{B}} f \, \mathrm{d}\pi.$$

Combining these limits with (G5) yields

$$\int_{\mathcal{B}} f \, \mathrm{d}(\pi\nu) = \int_{\mathcal{B}} f \, \mathrm{d}\pi$$

for every bounded continuous f. By the uniqueness part of the Riesz representation theorem (Aliprantis and Border, 2006, Corollary 14.15), this implies $\pi\nu = \pi$, so π is a stationary distribution of the backlog process. Since $\boldsymbol{w}(\boldsymbol{b}) = (b_q/k_q)_{q\in\mathcal{Q}}$ is a measurable function of $\boldsymbol{b}$, the pushforward of π under $\boldsymbol{w}$ is a stationary distribution of the wait times process. ■

The assumptions of iid health shocks and time-invariant capacities are made for expositional convenience. Seasonality can be accommodated by augmenting the state. For illustration, let $z_t \in \mathcal{Z}$ be a finite-state Markov process that collects seasonal conditions, define capacities to be $k_q(z_t)$, and let the conditional distribution of $\mathcal{I}_t$ be $\lambda(\cdot \mid z_t)$. The relevant state is then $(\boldsymbol{b}_t, z_t)$, whose transition law is time homogeneous even though the conditional distributions of capacities and patient arrivals vary over time. The proof's argument extends provided that capacities are uniformly bounded away from zero and the conditions ensuring that arrivals vanish as own backlog diverges hold uniformly over $z \in \mathcal{Z}$. Deterministic seasonality is nested by letting z_t be the calendar state, with a deterministic transition to the next seasonal state.

G.2 Invertibility of demand

↩ Main text: §4.1 Model / Patient choice · §5.3 Estimation / Private capacity, patient demand, and scan costs

↩ See also: §D.4 Estimation / Construction of the objective · §D.6 Estimation / Differentiation

We prove Proposition 1 in a more general demand model. Assume the indirect utility of consumer i from product $j \in \mathcal{J}$ is $v_{ij} = g(\delta_j, \mu_{ij}, \varepsilon_{ij})$, with the outside option satisfying $\delta_0 = \mu_{i0} = 0$. Demand for product j is defined as $\sigma_j(\boldsymbol{\delta}) = \Pr_{\boldsymbol{\mu},\boldsymbol{\varepsilon}}(g(\delta_j, \mu_{ij}, \varepsilon_{ij}) \geqslant \max_k g(\delta_k, \mu_{ik}, \varepsilon_{ik}))$. The set of products $\mathcal{J}$ is finite.

Assumption G1 (Dimensionality reduction)
There is a projection[53] *map* $h : \mathcal{J} \to \mathcal{J}$ *such that* $\delta_j = \delta_{h(j)}$ *for all* j.

[53] Map h is a projection if $h(h(j)) = h(j)$ for all $j \in \mathcal{J}$.

Define the image of $\mathcal{J}$ through h as $\mathcal{H} \coloneqq h(\mathcal{J})$ and let the $\mathcal{H}$-restricted demand vector be $\sigma_{\mathcal{H}}(\boldsymbol{\delta}) \coloneqq (\sigma_j(\boldsymbol{\delta}))_{j \in \mathcal{H}}$. We allow $\delta_j \in \mathbb{R}$, but Assumption G1 implies that possible vectors $\boldsymbol{\delta}$ live in a space isomorphic to $\mathbb{R}^{|\mathcal{H}|}$ instead of $\mathbb{R}^{|\mathcal{J}|}$.[54] With $\sigma_{\mathcal{H}}$ mapping to *a subset* of the $|\mathcal{H}|$-dimensional simplex, our goal is to impose reasonable assumptions to prove injectivity of $\sigma_{\mathcal{H}}$ (i.e., that $\boldsymbol{\delta} \neq \boldsymbol{\delta}'$ implies $\sigma_{\mathcal{H}}(\boldsymbol{\delta}) \neq \sigma_{\mathcal{H}}(\boldsymbol{\delta}')$).

Assumption G2 (Linearity)
Indirect utility satisfies $g(\delta_j, \mu_{ij}, \varepsilon_{ij}) = \delta_j + \mu_{ij} + \varepsilon_{ij}$.

Assumption G3 (Nondegenerate shocks)
Shocks $\boldsymbol{\varepsilon}_i$ *have an everywhere-positive density on* $\mathbb{R}^{|\mathcal{J}|+1}$ *conditional on almost every* $\boldsymbol{\mu}_i$.

Assumptions G2–G3 are standard in the applied literature. When h is the identity map, results from Berry (1994) apply to ensure injectivity of $\sigma_{\mathcal{H}} = \sigma$. The next result shows that injectivity is preserved by any map h.

Proposition G2 (Injectivity of restricted demand)
Under Assumptions G1–G3, $\sigma_{\mathcal{H}}$ *is injective.*

Proof. Take $\boldsymbol{\delta} \neq \boldsymbol{\delta}'$ and define $\Delta_j \coloneqq \delta_j - \delta'_j$ for $j \in \mathcal{J}$ and $\Delta_0 \coloneqq 0$. Without loss, assume $\max_{j \in \mathcal{H}} \Delta_j > 0$, and let $j^* \in \mathcal{H}$ attain this maximum.

Assumption G1 implies $\Delta_j = \Delta_{h(j)}$, so $\Delta_{j^*} \geqslant \Delta_k$ for every $k \in \mathcal{J}$, while $\Delta_{j^*} > \Delta_0 = 0$.

Define $v_{ij}(\boldsymbol{\delta}) \coloneqq g(\delta_j, \mu_{ij}, \varepsilon_{ij})$. Assumption G2 implies that, for all $k \in \{0\} \cup \mathcal{J}$,

$$v_{ij^*}(\boldsymbol{\delta}) - v_{ik}(\boldsymbol{\delta}) = v_{ij^*}(\boldsymbol{\delta}') - v_{ik}(\boldsymbol{\delta}') + \Delta_{j^*} - \Delta_k.$$

Since $\Delta_{j^*} - \Delta_k \geqslant 0$, whenever j^* is chosen under $\boldsymbol{\delta}'$, it is also chosen under $\boldsymbol{\delta}$, and therefore

$$\sigma_{j^*}(\boldsymbol{\delta}) \geqslant \sigma_{j^*}(\boldsymbol{\delta}').$$

To establish strict inequality, note that, at any $\boldsymbol{\mu}_i$ for which Assumption G3 holds, there is a nonempty open set of realizations of $\boldsymbol{\varepsilon}_i$ for which the outside option is preferred to j^* under $\boldsymbol{\delta}'$, j^* is preferred to the outside option under $\boldsymbol{\delta}$, and j^* is preferred to every other inside alternative under both vectors. Assumption G3 assigns positive probability to this set, making the inequality strict.

Formally, consider the set of shock realizations for which the outside option is uniquely optimal under $\boldsymbol{\delta}'$ while j^* is uniquely optimal under $\boldsymbol{\delta}$. We show that this set contains a nonempty open subset of $\mathbb{R}^{|\mathcal{J}|+1}$.

[54] Since map h is a projection, the spanned space has dimension exactly equal $|\mathcal{H}|$. Absent such restriction, the space of allowed $\boldsymbol{\delta}$ vectors would be isomorphic to $\mathbb{R}^L$, possibly with $L < |\mathcal{H}|$. As an illustration, assume $\mathcal{J} = \{a, b, c\}$ and let h be the map $a \mapsto b$, $b \mapsto c$, $c \mapsto a$. Such map is not a projection, $h(\mathcal{J}) = \mathcal{J}$, and the space of feasible $\boldsymbol{\delta}$ is isomorphic to $\mathbb{R}$.

Set $\bar{\varepsilon}_{ij^*} = 0$ and $\bar{\varepsilon}_{i0} = \delta'_{j^*} + \mu_{ij^*} + \frac{\Delta_{j^*}}{2}$. For every $k \in \mathcal{J} \setminus \{j^*\}$, there exists $\bar{\varepsilon}_{ik}$ such that

$$\delta'_k + \mu_{ik} + \bar{\varepsilon}_{ik} < \bar{\varepsilon}_{i0} \quad \text{and} \quad \delta_k + \mu_{ik} + \bar{\varepsilon}_{ik} < \delta_{j^*} + \mu_{ij^*} + \bar{\varepsilon}_{ij^*}$$

because there are finitely many alternatives and $\bar{\varepsilon}_{ik}$ is unrestricted below. At the resulting shock vector $\bar{\boldsymbol{\varepsilon}}_i$, under $\boldsymbol{\delta}'$ we have

$$\bar{\varepsilon}_{i0} - \left(\delta'_{j^*} + \mu_{ij^*} + \bar{\varepsilon}_{ij^*}\right) = \frac{\Delta_{j^\star}}{2} > 0,$$

and the construction of $\bar{\varepsilon}_{ik}$ makes the outside option strictly preferable to every other inside alternative. Hence the outside option is uniquely optimal under $\boldsymbol{\delta}'$. Under $\boldsymbol{\delta}$,

$$\delta_{j^*} + \mu_{ij^*} + \bar{\varepsilon}_{ij^*} - \bar{\varepsilon}_{i0} = \frac{\Delta_{j^\star}}{2} > 0,$$

and the construction of $\bar{\varepsilon}_{ik}$ makes j^* strictly preferable to every other inside alternative. Hence j^* is uniquely optimal under $\boldsymbol{\delta}$.

All these inequalities are strict and indirect utilities are continuous in $\boldsymbol{\varepsilon}_i$, so they continue to hold on some open neighborhood of $\bar{\boldsymbol{\varepsilon}}_i$. Assumption G3 assigns strictly positive conditional probability to this neighborhood. Therefore, for almost every $\boldsymbol{\mu}_i$, there is strictly positive probability that a consumer switches from the outside option under $\boldsymbol{\delta}'$ to j^* under $\boldsymbol{\delta}$. Integrating over $\boldsymbol{\mu}_i$ gives

$$\sigma_{j^*}(\boldsymbol{\delta}) > \sigma_{j^*}(\boldsymbol{\delta}'),$$

so that $\sigma_{\mathcal{H}}(\boldsymbol{\delta}) \neq \sigma_{\mathcal{H}}(\boldsymbol{\delta})$ and injectivity follows. ■

This immediately implies that there is at most one vector $\boldsymbol{\delta}$ satisfying $\sigma_{\mathcal{H}}(\boldsymbol{\delta}) = \boldsymbol{s}_{\mathcal{H}}$, where $\boldsymbol{s}_{\mathcal{H}}$ is a vector of observed shares of "reference" products $j \in \mathcal{H}$. Compared to Berry (1994), one loses existence guarantees when h is not the identity. This is illustrated by a simple example. Assume $\mathcal{J} = \{1, 2\}$, $h(1) = h(2) = 1$, $\mu_{ij} = 0$ almost surely for all j, and that ε_{ij} are logit shocks independent across i and j. Since the logit shocks imply that the two products have the same market share, the equation $\sigma_1(\boldsymbol{\delta}) = s_1$ has no solution whenever $s_1 \geqslant \frac{1}{2}$.

The general lesson is that the restrictions imposed by map h may prevent the model from spanning the whole simplex, and this can lead to non-existence of vectors $\boldsymbol{\delta}$ rationalizing observed shares. In practice, demand inversion can therefore fail at some candidate demand parameters. It might be tempting to try to identify the set of demand parameters that are potentially consistent with the observed restricted vectors of market shares and restrict the search to such set. While this characterization might be obtained analytically for low-dimensional, stylized parameterizations of the model, we could not find a convenient procedure to obtain

the desired set in reasonable specifications. We rarely experienced failures of inversion.

G.3 Identification of private capacity

↩ Main text: §4.2 Model / Queues · §5.3 Estimation / Private capacity, patient demand, and scan costs

↩ See also: §C.1 Theory / Queuing model · §D.3 Estimation / Private capacity

We conclude by discussing the roots of M_q^{W}. The notation follows that of Sections 5.3 and D.3. Fix a private queue and the remaining parameters and realized histories used in the inversion as described in Section D.3. Define

$$g_t := \frac{Y_{\mathrm{pub}\,m\,\underline{u}\,t}}{\bar{Y}_{\mathrm{pub}\,m\,\underline{u}}} \exp\left(\hat{v}^{\mathrm{P}}_{hmt}\right) \Bigg/ \left(\frac{1}{T}\sum_{\tau=1}^{T} \frac{Y_{\mathrm{pub}\,m\,\underline{u}\,\tau}}{\bar{Y}_{\mathrm{pub}\,m\,\underline{u}}} \exp\left(\hat{v}^{\mathrm{P}}_{hm\tau}\right)\right),$$

so that (D6) can be written as $k_{h\,\mathrm{pvt}\,mt} = k_{h\,\mathrm{pvt}\,m} g_t$ and $\frac{1}{T}\sum_{t=1}^{T} g_t = 1$. Suppressing the queue indices and writing the candidate mean capacity as $k > 0$, weekly capacity is therefore kg_t. For elapsed time $x \geqslant 0$, define cumulative relative capacity from week t as

$$G_t(x) := \sum_{r=0}^{\lfloor x \rfloor - 1} g_{t+r} + \big(x - \lfloor x \rfloor\big) g_{t+\lfloor x \rfloor}.$$

Equation (**??**) then gives

$$w_t(k) = \inf\left\{x > 0 : kG_t(x) > b_t(k)\right\}.$$

We say that a relative capacity path is *regular* if $g_t > 0$ for every t, $\sup_t g_t < \infty$, and $\sum_{s=t}^{\infty} g_s = \infty$ for every t.

Lemma 1 (Continuity)
If the relative capacity path is regular, the wait moment M^{W} is continuous.

Proof. Regularity implies that G_t is continuous, strictly increasing, and unbounded, so it admits a continuous inverse such that $w_t(k) = G_t^{-1}(b_t(k)/k)$. Write the backlog recursion as $b_{t+1}(k) = \max\left\{b_t(k) + a_t(w_t(k)) - kg_t, 0\right\}$ and note that logit choice probabilities are continuous in the queue's wait and bounded, while the realized weekly referral mass is finite, so arrivals $a_t(w_t(k))$ are continuous in k. The initial backlog $b_0(k)$ is fixed, hence continuous. The recursion then implies that $b_t(k)$ is continuous for all t. Continuity of G_t^{-1} gives continuity of $w_t(k)$. Their finite average in $M^{\mathrm{W}}(k)$ is therefore also continuous, yielding the result. ∎

We impose an additional, mild regularity condition on the distribution of referrals. We show existence under the initialization of the backlog process we used in the simulation, i.e., setting the backlog (and hence the wait) of the first week of 2023 as 0.

Assumption G4 (Local patient support)
All weeks have positive referral mass in all modalities: $\int_{\mathcal{I}_t} \mathbf{1}\{m(i) = m\}\,\mathrm{d}i > 0$ *for all* m *and* t. *Providers have positive local patient mass:* $\int_{\mathcal{I}_t} \mathbf{1}\{m(i) = m,\, d_{\ell(i)h} \leqslant \bar{d}\}\,\mathrm{d}i > 0$ *for all* h, m, t.

Lemma 2 (Existence)
If the relative capacity path is regular and Assumption G4 holds, $M^{\mathrm{W}}(k) = 0$ *has a solution.*

Proof. The realized-path initialization implies $b_{t_0}(k) = w_{t_0}(k) = 0$ for every $k > 0$ for the first week in 2023. Assumption G4 implies that a positive mass of the modality-m referral population has h in its choice set. Logit choice implies that arrivals in t_0 equal some $a_0 > 0$ that does not depend on k.

Let $t_1 = \min \mathcal{T}_{2024}$ be the first week of 2024. Since 2023 arrivals are nonempty, the backlog at the beginning of 2024 is bounded below by a_0 net of the cumulative 2023 capacity $k \sum_{s=t_0}^{t_1-1} g_s$: $b_{t_1}(k) \geqslant a_0 - k \sum_{s=t_0}^{t_1-1} g_s$. Because $(g_t)_t$ is regular, cumulative 2023 capacity is finite, so $b_{t_1}(k) \geqslant a_0/2$ for all sufficiently small k. Regularity further gives $\bar{g} := \sup_t g_t < \infty$, so $G_{t_1}(x) \leqslant \bar{g}x$ and $w_{t_1}(k) \geqslant \frac{a_0}{2k\bar{g}}$ for all k. It then follows that $\lim_{k\to 0^+} w_{t_1}(k) = \infty$, and therefore $M^{\mathrm{W}}(k) > 0$ for all sufficiently small k.

For the opposite sign, let A_t denote the total realized referral mass in week t. Over the finite realized history through the end of 2024, each A_t is finite and regularity gives $g_t > 0$, so $k^\star := \max_t \frac{A_t}{g_t} < \infty$. For $k > k^\star$, weekly capacity kg_t exceeds the total referral mass and therefore exceeds arrivals to q in every week. Starting from $b_1 = 0$, the backlog remains zero throughout the realized path, so $w_t(k) = 0$ for every $t \in \mathcal{T}_{2024}$ and $M^{\mathrm{W}}(k) = -\bar{\omega} < 0$.

Continuity from Lemma 1 and the intermediate value theorem establish existence. ∎

The initialization $b_1 = 0$ is a boundary condition of the realized-path simulation used to construct M^{W}, rather than an assumption on the structural backlog process. Lemma 2 therefore establishes existence for the moment restriction as implemented in estimation. The simulation starts in 2023, while the wait restriction is constructed from 2024 observations, so the initialization is separated from the target period by a one-year burn-in. The result can be extended to any fixed finite initial backlog.

Uniqueness requires a strict monotonicity assumption that need not hold in general. As mentioned in Section 5.3, an increase in capacity directly reduces wait but, by making the queue more attractive, also increases arrivals. Monotonicity requires the latter feedback to be sufficiently weak relative to the former. In practice, we found monotonicity to hold on the economically relevant range of capacities, with multiple roots arising mostly when the optimizer explores parameter guesses under which both arrivals to a queue and its candidate capacity are close to 0.

H Miscellaneous

H.1 Queue-level arrival objects

↩ Main text: §4.3 Model / Supply

↩ See also: §C.1 Theory / Queuing model · §H.3 Miscellaneous / Profit derivatives

This section collects the queue-level arrival objects used in Sections 4.3 and 5.3. Partition the set of queues as $\mathcal{Q} = \mathcal{Q}_{\text{pub}} \cup \mathcal{Q}_{\text{pvt}}$, where a publicly funded queue $q \in \mathcal{Q}_{\text{pub}}$ is indexed by provider, modality, and urgency, while a private-pay queue $q \in \mathcal{Q}_{\text{pvt}}$ is indexed by provider and modality. For each queue q, let $h(q)$, $e(q)$, and $m(q)$ denote its provider, access, and modality, and let $u(q)$ additionally denote urgency when $q \in \mathcal{Q}_{\text{pub}}$. Recall that ρ_{hmu} denoted the exogenous share of publicly funded slots provider h allocates to urgency u in modality m. In what follows, for provider-access pair $he = j \in \mathcal{J}$, we use notation a_{hes} and a_{js} interchangeably.

The queue-level capacities and their derivatives as a function of capacity choices are

$$k_q(\boldsymbol{k}) = \begin{cases} \rho_{hmu} k_{h\,\text{pub}\,m}, & \text{if } q = (h, \text{pub}, m, u), \\ k_{h\,\text{pvt}\,m}, & \text{if } q = (h, \text{pvt}, m), \end{cases}$$

$$\frac{\partial k_q}{\partial k_{hem}} = \begin{cases} \rho_{hmu}, & \text{if } e = \text{pub and } q = (h, \text{pub}, m, u), \\ 1, & \text{if } e = \text{pvt and } q = (h, \text{pvt}, m), \\ 0, & \text{otherwise.} \end{cases}$$

For patient i and alternative j, the queue map introduced in Section 4.1 is

$$q(i,j) = \begin{cases} (h(j), \text{pub}, m(i), u(i)), & \text{if } e(j) = \text{pub}, \\ (h(j), \text{pvt}, m(i)), & \text{if } e(j) = \text{pvt}. \end{cases}$$

This map returns the queue patient i would join were they to choose alternative j.

For a realized set of patients $\mathcal{I}_t$, scan-specific arrivals to alternative j and queue q are

$$a_{jst}(\boldsymbol{p}, \boldsymbol{w}_t) := \int_{\mathcal{I}_t} \mathbf{1}\{s(i) = s\} \sigma_{ij}(\boldsymbol{p}, \boldsymbol{w}_t)\, \mathrm{d}i,$$
$$a_{qst}(\boldsymbol{p}, \boldsymbol{w}_t) := \int_{\mathcal{I}_t} \mathbf{1}\{s(i) = s\} \sum_{j \in \mathcal{J}_i} \mathbf{1}\{q(i,j) = q\} \sigma_{ij}(\boldsymbol{p}, \boldsymbol{w}_t)\, \mathrm{d}i,$$

Total queue arrivals are

$$a_{qt}(\boldsymbol{p}, \boldsymbol{w}_t) = \sum_{s \in \mathcal{S}} a_{qst}(\boldsymbol{p}, \boldsymbol{w}_t).$$

Taking expectations over the referral-generating process conditional on waits $\boldsymbol{w}$, define

$$a_{js}(\boldsymbol{p}, \boldsymbol{w}) := \mathbb{E}[a_{jst}(\boldsymbol{p}, \boldsymbol{w})],$$
$$a_{q}(\boldsymbol{p}, \boldsymbol{w}) := \mathbb{E}[a_{qt}(\boldsymbol{p}, \boldsymbol{w})].$$

Expected arrivals under the stationary queue environment induced by $(\boldsymbol{p}, \boldsymbol{k})$ are

$$a_{js}(\boldsymbol{p}, \boldsymbol{k}) = \int a_{js}(\boldsymbol{p}, \boldsymbol{w}) \, \mathrm{d}F_{\boldsymbol{w}|\boldsymbol{p},\boldsymbol{k}}(\boldsymbol{w}),$$
$$a_{q}(\boldsymbol{p}, \boldsymbol{k}) = \int a_{q}(\boldsymbol{p}, \boldsymbol{w}) \, \mathrm{d}F_{\boldsymbol{w}|\boldsymbol{p},\boldsymbol{k}}(\boldsymbol{w}).$$

H.2 Total derivative of arrivals

↩ Main text: §4.3 Model / Supply

↩ See also: §H.1 Miscellaneous / Queue-level arrivals · §D.6 Estimation / Differentiation

Recall that expected arrivals as functions of choice variables are obtained by integrating over the stationary distribution of wait times:

$$a_{js}(\boldsymbol{p}, \boldsymbol{k}) = \int a_{js}(\boldsymbol{p}, \boldsymbol{w}) \, \mathrm{d}F_{\boldsymbol{w}|\boldsymbol{p},\boldsymbol{k}}(\boldsymbol{w}).$$

Let $\mathbf{e}_{hs'}$ denote a perturbation of private-pay price $p_{hs'}$. The total price derivative of stationary arrivals is

$$\frac{\mathrm{d}a_{js}(\boldsymbol{p}, \boldsymbol{k})}{\mathrm{d}p_{hs'}} = \int \frac{\partial a_{js}(\boldsymbol{p}, \boldsymbol{w})}{\partial p_{hs'}} \, \mathrm{d}F_{\boldsymbol{w}|\boldsymbol{p},\boldsymbol{k}}(\boldsymbol{w}) + \lim_{\tau \to 0} \frac{1}{\tau} \left[\int a_{js}(\boldsymbol{p}, \boldsymbol{w}) \, \mathrm{d}F_{\boldsymbol{w}|\boldsymbol{p}+\tau\mathbf{e}_{hs'},\boldsymbol{k}}(\boldsymbol{w}) - \int a_{js}(\boldsymbol{p}, \boldsymbol{w}) \, \mathrm{d}F_{\boldsymbol{w}|\boldsymbol{p},\boldsymbol{k}}(\boldsymbol{w}) \right].$$

The first term is the direct effect of price on patient choice holding waits fixed, while the second captures the induced change in the stationary distribution of waits. A similar expression applies to capacity derivatives, with the difference that there is no direct effect on patient choice. This formulation is quite general as it puts no restrictions on the stationary distribution of wait times (besides existence). In particular, it does not require it to have a density, which would generally be false as atoms at zero can arise.

H.3 Second-order approximation of profits

↩ Main text: §4.3 Model / Supply · §5.3 Estimation / Private capacity, patient demand, and scan costs

↩ See also: §D.2 Estimation / Stationary wait approximation · §C.3 Theory / Provider optimality conditions

This section gives the complete derivatives of the second-order approximation in equation (D1). We follow the same notation and make use of the same objects obtained in Section H.1. Applying equation (D1) to the conditional arrival function gives

$$\tilde{a}_{js}(\boldsymbol{p}, \bar{\boldsymbol{w}}, \boldsymbol{k}) = a_{js}(\boldsymbol{p}, \bar{\boldsymbol{w}}) + \frac{1}{2} \sum_{q \in \mathcal{Q}} \frac{\partial^2 a_{js}(\boldsymbol{p}, \bar{\boldsymbol{w}})}{\partial w_q^2} \mathbb{V}\big(w_q \,|\, \bar{w}_q, k_q\big), \tag{H1}$$

with the queue-specific version $\tilde{a}_q$ obtained analogously.

Arrivals derivatives Consider a private-pay price $p_{hs'}$. Because prices do not directly change queue capacities, differentiating (H1) gives

$$\begin{aligned} \frac{\mathrm{d}\tilde{a}_{js}}{\mathrm{d}p_{hs'}} = \frac{\partial a_{js}}{\partial p_{hs'}} + \frac{1}{2} \sum_q \frac{\partial^3 a_{js}}{\partial w_q^2 \partial p_{hs'}} \mathbb{V}\big(w_q \,|\, \bar{w}_q, k_q\big) \\ + \sum_{q'} \left[\frac{\partial a_{js}}{\partial w_{q'}} + \frac{1}{2} \sum_q \frac{\partial^3 a_{js}}{\partial w_q^2 \partial w_{q'}} \mathbb{V}\big(w_q \,|\, \bar{w}_q, k_q\big) + \frac{1}{2} \frac{\partial^2 a_{js}}{\partial w_{q'}^2} \frac{\partial \mathbb{V}\big(w_{q'} \,|\, \bar{w}_{q'}, k_{q'}\big)}{\partial \bar{w}_{q'}} \right] \frac{\mathrm{d}\bar{w}_{q'}}{\mathrm{d}p_{hs'}}. \end{aligned} \tag{H2}$$

All derivatives of a_{js} in (H2) are evaluated at $(\boldsymbol{p}, \bar{\boldsymbol{w}})$, and the derivatives of stationary mean waits are the objects described in Section D.6.

Consider a capacity choice k_{hem}. Conditional patient choice depends on capacity only through wait times, so holding mean waits fixed the only direct effect of capacity in (H1) operates through the variance functions:

$$\begin{aligned} \frac{\mathrm{d}\tilde{a}_{js}}{\mathrm{d}k_{hem}} = \frac{1}{2} \sum_q \frac{\partial^2 a_{js}}{\partial w_q^2} \frac{\partial \mathbb{V}\big(w_q \,|\, \bar{w}_q, k_q\big)}{\partial k_q} \frac{\partial k_q}{\partial k_{hem}} \\ + \sum_{q'} \left[\frac{\partial a_{js}}{\partial w_{q'}} + \frac{1}{2} \sum_q \frac{\partial^3 a_{js}}{\partial w_q^2 \partial w_{q'}} \mathbb{V}\big(w_q \,|\, \bar{w}_q, k_q\big) + \frac{1}{2} \frac{\partial^2 a_{js}}{\partial w_{q'}^2} \frac{\partial \mathbb{V}\big(w_{q'} \,|\, \bar{w}_{q'}, k_{q'}\big)}{\partial \bar{w}_{q'}} \right] \frac{\mathrm{d}\bar{w}_{q'}}{\mathrm{d}k_{hem}}. \end{aligned} \tag{H3}$$

Equation (H3) makes explicit that capacity affects expected arrivals through two channels under the approximation: changes in stationary mean waits and changes in the variance of waits around those means.

Profit derivatives Using the approximated stationary arrivals, provider h's operating profits are

$$\tilde{\Pi}_h = \sum_{s \in \mathcal{S}} \Big[(p_{hs} - c_{hs}) \tilde{a}_{h\,\mathrm{pvt}\,s} + (r_s - c_{hs}) \tilde{a}_{h\,\mathrm{pub}\,s} \Big].$$

To allow for common ownership as in Section 5.3, let $\mathcal{H}$ be the ownership matrix, with entry $\mathcal{H}_{hh'}$ equal to 1 if h and h' share the same owner and to 0 otherwise. The operating profits internalized by provider h are $\tilde{\Pi}_h^{\mathcal{H}} := \sum_{h'} \mathcal{H}_{hh'} \tilde{\Pi}_{h'}$.

For private-pay price $p_{hs'}$,

$$\frac{\mathrm{d}\tilde{\Pi}_h^{\mathcal{H}}}{\mathrm{d}p_{hs'}} = \tilde{a}_{h\,\mathrm{pvt}\,s'} + \sum_{h'} \mathcal{H}_{hh'} \sum_s \left[(p_{h's} - c_{h's}) \frac{\mathrm{d}\tilde{a}_{h'\,\mathrm{pvt}\,s}}{\mathrm{d}p_{hs'}} + (r_s - c_{h's}) \frac{\mathrm{d}\tilde{a}_{h'\,\mathrm{pub}\,s}}{\mathrm{d}p_{hs'}} \right], \tag{H4}$$

where the arrival derivatives are given by equation (H2).

Stacking these derivatives across private-pay prices for access $e \in \{\mathrm{pub}, \mathrm{pvt}\}$ gives

$$\boldsymbol{\Delta}_e = -\mathcal{H} \odot \left(\frac{\mathrm{d}\tilde{\boldsymbol{a}}_e}{\mathrm{d}\boldsymbol{p}'} \right)',$$

and therefore the pricing conditions in Section 5.3 can be written as

$$\mathbf{0} = \tilde{\boldsymbol{a}}_{\mathrm{pvt}} - \boldsymbol{\Delta}_{\mathrm{pvt}}(\boldsymbol{p} - \boldsymbol{c}) - \boldsymbol{\Delta}_{\mathrm{pub}}(\boldsymbol{r} - \boldsymbol{c}). \tag{H5}$$

For capacity choice k_{hem}, prices and reimbursements do not change directly, so

$$\frac{\mathrm{d}\tilde{\Pi}_h^{\mathcal{H}}}{\mathrm{d}k_{hem}} = \sum_{h'} \mathcal{H}_{hh'} \sum_s \left[(p_{h's} - c_{h's}) \frac{\mathrm{d}\tilde{a}_{h'\,\mathrm{pvt}\,s}}{\mathrm{d}k_{hem}} + (r_s - c_{h's}) \frac{\mathrm{d}\tilde{a}_{h'\,\mathrm{pub}\,s}}{\mathrm{d}k_{hem}} \right], \tag{H6}$$

where the arrival derivatives are given by equation (H3).

H.4 An alternative validation of the queuing model

↩ Main text: §5.1 Estimation / Queuing model · §5.2 Estimation / Capacity of public queues

↩ See also: §C.1 Theory / Queuing model · §D.1 Estimation / Queuing model validation

The estimation of public capacity described in Section 5.2 exploits stationarity together with the observation that throughput coincides with available capacity when a queue is backlogged. This section provides a complementary validation exercise that requires substantially weaker assumptions: that treatment is first come, first served and that there is a law of large numbers for the process of available slots.

Fix a public queue q and let $(k_{qt})_{t \geqslant 1}$ denote its process of available weekly slots. Let $w_{qt}(b)$ denote the corresponding first-availability wait in queue-week qt as a function of backlog b, constructed using the within-week interpolation in equation (**??**). This formula relies on first come, first served treatment; our assumption on the queuing protocol is only needed to construct this measure of wait times.

Assumption H1 (Capacity process)
For all queues q and starting weeks t, there is $k_q \in (0, \infty)$ such that $\frac{1}{n} \sum_{r=0}^{n-1} k_{q,t+r} \to k_q$ almost surely.

Assumption H1 places only a long-run restriction on available capacity. It requires neither

stationarity nor independence of $(k_{qt})_{t\geqslant 1}$, but only that its time average converges almost surely to a finite, positive limit.

Proposition H1 (Wait-backlog asymptotics)
Under Assumption H1, for all queues q and weeks t,

$$\lim_{b\to\infty} \frac{w_{qt}(b)}{b} = \frac{1}{k_q} \qquad \textit{almost surely.}$$

Proof. Fix a queue q and starting week t, and consider the probability-one event on which Assumption H1 holds. Let $S_n := \sum_{r=0}^{n-1} k_{q,t+r}$, so that $S_n/n \to k_q$. Let $S(x)$ be the linearly interpolated cumulative capacity appearing in equation (**??**), and set $n = \lfloor x \rfloor$. Since $S_n \leqslant S(x) \leqslant S_{n+1}$,

$$\frac{n}{x}\frac{S_n}{n} \leqslant \frac{S(x)}{x} \leqslant \frac{n+1}{x}\frac{S_{n+1}}{n+1}.$$

As $x \to \infty$, both outer expressions converge to k_q by Assumption H1, so $S(x)/x \to k_q$ by the squeeze theorem. Hence, for every $\varepsilon \in (0, k_q)$ and all sufficiently large x, $(k_q - \varepsilon)x < S(x) < (k_q+\varepsilon)x$. Since $w_{qt}(b) = \inf\{x > 0 : S(x) > b\}$, these inequalities imply that $b/(k_q+\varepsilon) \leqslant w_{qt}(b) \leqslant b/(k_q - \varepsilon) + 1$ for sufficiently large b. Dividing by b, taking $b \to \infty$, and then letting $\varepsilon \downarrow 0$ gives the result. ∎

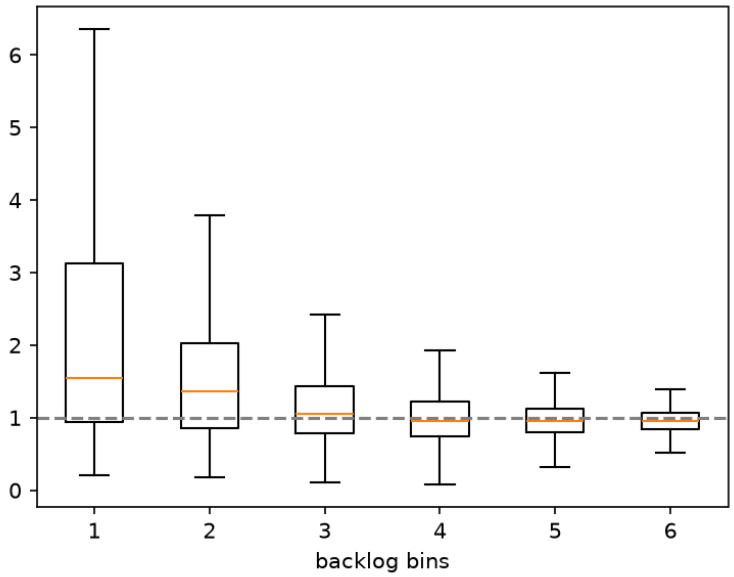
Proposition H1 gives a simple implication of the queuing model that survives after stripping away the stationary-distribution assumptions used elsewhere in the estimation. The proposition implies that, even without our maintained assumptions, sufficiently long queues must clear at the average rate at which slots become available. Equivalently, sufficiently long queues are cleared at the long-run average service rate. In the fixed-capacity version of the model, the relationship is exact because $w_{qt} = b_{qt}/k_q$. The result shows that the same relationship holds asymptotically when available capacity fluctuates across weeks.

We reconstruct b_{qt} from observed arrivals and throughput as described in Section 5.1. Let $\bar{w}_{qt}$ denote the mean realized wait in days among patients joining queue q in week t, and define the corresponding wait in weeks as $\tilde{w}_{qt} := \bar{w}_{qt}/7$. To validate the throughput-based capacity estimates used in the main analysis, define

$$D_{qt} := \hat{k}_q \frac{\tilde{w}_{qt}}{b_{qt}},$$

where $\hat{k}_q$ is the public-capacity estimate obtained as described in Section 5.2. If $\hat{k}_q$ recovers the relevant long-run service capacity, Proposition H1 predicts that $D_{qt} \to 1$ as $b_{qt} \to \infty$. Figure H1 shows that this prediction holds in the data.

Figure H1 Wait-backlog diagnostic for public MRI queues



Notes: The figure illustrates the statistic $D_{qt} = \hat{k}_q \tilde{w}_{qt}/b_{qt}$ for public MRI queues. Queues are binned based on their mean realized backlog, imputed using the queue recursion described in Section 5.1.

References

Aliprantis, C. D., and K. C. Border (2006): *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Berlin, Heidelberg: Springer, 3rd edition.

Andrews, D. W. K., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Berry, S. T. (1994): "Estimating Discrete-Choice Models of Product Differentiation," *RAND Journal of Economics*, 25, 242–262.

Bradbury, J., R. Frostig, P. Hawkins et al. (2018): "JAX: Composable Transformations of Python+NumPy Programs."

Brancaccio, G., M. Kalouptsidi, and T. Papageorgiou (2025): "Investment in Infrastructure and Trade: The Case of Ports," *Mimeo*.

Conlon, C., and J. Gortmaker (2025): "Incorporating Micro Data into Differentiated Products Demand Estimation with PyBLP," *Journal of Econometrics*, 105926.

Kingman, J. F. C. (1961): "The Single Server Queue in Heavy Traffic," *Mathematical Proceedings of the Cambridge Philosophical Society*, 57, 902–904.

Li, B., and G. Pang (2026): "Single-Server Queues with State-Dependent Hawkes Arrivals," *Mathematics of Operations Research*.

Markkula, T. (2025): "Cutting the Queue to the Dentist: Waiting Time, Public-Private Interaction and Consumer Surplus," *Mimeo*.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995): *Microeconomic Theory*, New York: Oxford University Press.

Meyn, S. P., and R. L. Tweedie (1993): *Markov Chains and Stochastic Stability*, Communications and Control Engineering, London: Springer.

Miyazawa, M. (2024): "The Stationary Distributions of State-Dependent Diffusions Reflected at One and Two Sides," *Mimeo*.

——— (2025): "Diffusion Approximation of the Stationary Distribution of a Two-Level Single Server Queue," *Advances in Applied Probability*, 57, 1167–1205.

Rauch, H. E., F. Tung, and C. T. Striebel (1965): "Maximum Likelihood Estimates of Linear Dynamic Systems," *AIAA Journal*, 3, 1445–1450.

Russo, A. (2024): "Waiting or Paying for Healthcare: Evidence from the Veterans Health Administration," *Mimeo*.

Whitt, W. (2002): *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*, New York: Springer.