

Working Paper in Economics No. 864

Rationing under Excess Demand: Evidence on Providers' Adjustment Margins

**Roxanne Kovacs, Timothy Abuya, Timothy Powell-
Jackson**

Department of Economics, October 2026

ISSN 1403-2473 (Print)
ISSN 1403-2465 (Online)



UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

Rationing under Excess Demand: Evidence on Providers' Adjustment Margins

Roxanne Kovacs* Timothy Abuya[†] Timothy Powell-Jackson[‡]

Abstract

Excess demand is pervasive in public service delivery, but its consequences depend on how providers allocate scarce capacity. We sent standardised patients – trained auditors posing as real patients – to 198 primary-care facilities in Nairobi. Across 1,326 pre-scheduled visits, we exploit within-facility variation in workload. We find that when queue length exceeds the typical daily workload per provider, waiting time is 21 minutes longer (0.4 SD). We find no reduction in consultation effort and no statistically significant deterioration in care quality or access. This suggests short-run congestion is borne primarily through users' time and need not reduce service quality.

JEL Codes: I11, I18, D24, H51

Keywords: Excess demand; Non-price rationing; Queuing; Congestion; healthcare

*Department of Economics, University of Gothenburg, Vasagatan 1, 411 24 Göteborg, Sweden. Email: roxanne.kovacs@economics.gu.se.

[†]Population Council Kenya, Nairobi, Kenya.

[‡]Department of Global Health and Development, London School of Hygiene and Tropical Medicine, London, UK.

1 Introduction

Excess demand is a common feature of public service delivery. Many essential services are deliberately provided or regulated in ways that prevent prices from clearing the market (Nichols et al. 1971, Besley & Coate 1991, Taylor et al. 2003), so that scarcity must be resolved through non-price rationing. Where users queue for services rather than being rationed ex ante, frontline providers can respond to short-run increases in demand by letting users wait, reducing the attention devoted to each user, or restricting access. Which margin adjusts determines whether congestion costs appear as lost time, poorer service quality, or unmet demand. We provide novel evidence on the margins through which providers absorb short-run excess demand, using walk-in primary healthcare as an empirical setting. Although we study healthcare, the underlying allocation problem plausibly arises in a range of other queue-based public services, including welfare and employment offices, civil registration agencies, and licensing offices.

We use rich micro-level data from standardised patients (SPs) – a healthcare analogue of audit studies – attending a representative sample of primary care facilities in Nairobi, Kenya. SPs are trained to pose as real patients, present identical clinical cases to providers, and give a detailed report of what occurred during the consultation. The SP approach allows us to observe provider behaviour under routine conditions while holding case mix fixed. SPs conduct 1,326 consultations across 198 facilities, following a predetermined schedule – thereby avoiding selection into queues based on congestion. We compare visits within facilities (and, in some specifications, within healthcare workers), controlling for clinical case and visit timing. Identification requires that, conditional on these fixed effects, differences in workload observed across SP visits are not systematically associated with other factors that affect provider behaviour.

We measure workload using the number of patients waiting to see a provider when the SP arrives, and operationalise excess demand by benchmarking queue length against the facility’s typical daily workload per provider. Thirteen percent of consultations occur during periods of excess demand, when an average of 20 patients are waiting, compared with 3 otherwise. Our sample covers all public primary care facilities and 40% of private-for-profit facilities serving Nairobi County’s population of approximately five million. This is a high-workload and low-quality setting: providers see an average of 19 patients per day (30 in public facilities), and only 36% of SPs receive guideline-indicated care, consistent with evidence from other LMICs (Das et al. 2008).

When queue length at SP arrival exceeds the facility’s usual daily workload per provider, waiting time is 21 minutes (0.4 SD) longer. However, provider effort does not fall, with estimates sufficiently precise to rule out even small reductions in consultation duration and checklist adherence. We find no statistically significant deterioration in clinical or interpersonal quality. We also find no evidence of task-shifting, and no SP was refused care or asked to return. Results hold at high levels of workload, across sectors, among providers with sufficient clinical knowledge, and in specifications with healthcare worker fixed effects.

Although we cannot definitively distinguish between mechanisms, we consider two explanations for why providers appear to rely on rationing through waiting. Under *downward rigidity in quality*, care is already near a minimum quality floor; under *fixed consultation routines*, providers deliver similar care regardless of workload. Distinguishing between these mechanisms is relevant for policy, as reductions in workload would improve service quality under downward rigidity but have limited effects under fixed routines. We examine provider behaviour at low workload – where the mechanisms generate opposing predictions – and find patterns more consistent with fixed consultation routines.

Our findings have important implications for policies aimed at reducing congestion in public service delivery. Costly expansions in capacity are often motivated by concerns that long queues compromise service quality, leading to poorer decisions or worse outcomes for users (Dixit 2002, Prendergast 2007, Finan, F 2017, Hemrajani & Jr 2024, Danziger et al. 2011, Liu & Mahler 2026). The results suggest that reducing short-run congestion need not produce improvements in service quality. While we study healthcare, the underlying allocation problem arises more broadly in queue-based settings in which citizens wait for access to a frontline provider, including welfare and employment offices, civil registration agencies, licensing authorities, and visa and consular offices. These frontline service environments remain common in high-income countries and constitute a primary mode of public service delivery in many LMICs (Pepinsky et al. 2017, Finan, F 2017).

This paper makes two primary contributions to the literature. First, we contribute to the literature on excess demand and non-price rationing (Barzel 1974, Deaton 1981, Besley & Coate 1991) by providing micro-level evidence on how frontline providers absorb short-run excess demand across multiple margins. This allows us to identify whether the burden of congestion falls on users’ time or on the care they receive. We show that congestion substantially increases waiting time while consultation effort remains unchanged, with no statistically significant deterioration in service quality, evidence of task-shifting or access restrictions. Results suggest that in our setting, short-run congestion is borne through users’

time, even though alternative forms of rationing are institutionally feasible.

Second, this paper contributes to research on excess demand in healthcare specifically. Existing work largely focuses on waiting lists or queueing in emergency departments (Propper 1995, Cullis & Jones 2000, Hoe 2022, Richards-Shubik et al. 2022, Woodworth 2020, Martin & Smith 1999, Martins & Filipe 2020, Turner et al. 2020, Strobel 2024), with only limited evidence from primary care. Prior studies typically instrument primary care physician’s absences and examine narrow behavioural outcomes – such as antibiotic prescribing, referrals, or virtual consultations – finding small or non-significant effects (Shurtz et al. 2022, Pilvar & Watt 2025). Guided by our conceptual framework, we provide the first broad analysis across the full set of providers’ plausible adjustment margins. We also add to the limited causal evidence on how provider workload affects healthcare quality in LMIC primary care (Mæstad et al. 2010, Kovacs & Lagarde 2022), where poor quality of care is often viewed as being the result of high workload. The few existing studies focus on rural, low-workload settings with substantial slack capacity and examine intensive-margin quality outcomes (provider effort and clinical quality), leaving open the possibility that null results reflect limited demand. We study a high-workload urban environment and examine how providers adjust along multiple margins. This suggests that slack capacity alone is unlikely to explain prior null findings and instead points toward a potential behavioural mechanism, i.e. relatively fixed consultation routines.

The remainder of this paper proceeds as follows. Section 2 provides a brief conceptual framework. Section 3 describes the study setting and design. Section 4 describes the data and empirical strategy. Section 5 presents our main findings and section 6 provides a discussion and concluding remarks.

2 Conceptual framework

Excess demand is a common feature of public service delivery. In queue-based service environments in particular, short-run fluctuations in demand often generate periods of high provider workload. However, the existence of excess demand alone is not sufficient to characterize its welfare consequences. These depend on how scarce capacity is allocated: different adjustment margins impose different types of costs on users and therefore imply different potential gains from relaxing capacity constraints. We focus on short-run fluctuations in

demand relative to typical capacity, where prices are fixed¹, and distinguish three margins that affect users in different ways: rationing through waiting, rationing on quality along the intensive margin, and rationing on quality along the extensive margin.

First, providers may respond to excess demand via *rationing through waiting*, allowing waiting times to increase while leaving the content of service provision unchanged. In this case, congestion imposes costs primarily through users' time. Second, providers may ration quality along the *intensive margin*, maintaining access while reducing the intensity or quality of service provided. In our setting, this may involve shorter consultations, fewer history questions or physical examinations, or poorer interpersonal care. In this case, congestion imposes costs through reductions in the quality of service received. Third, providers may ration quality along the *extensive margin* by restricting access to services or reallocating users across providers. This may involve redirecting users to return at another time, referring them elsewhere, assigning them to less-skilled personnel, or refusing service altogether. In this case, congestion affects whether, when, or from whom users receive services.

The potential gains from reducing short-run congestion depend on the margin through which excess demand is absorbed. When excess demand is absorbed through waiting, reducing congestion primarily reduces the time users spend in queues. When providers instead ration along the intensive margin, reducing congestion may improve the quality of service received, while under extensive-margin rationing it may improve access to services. This distinction matters because, although we do not rank these outcomes in welfare terms, policies aimed at reducing congestion are often motivated by concerns that high workload compromises service quality. Yet if providers do not ration quality when demand is high, reducing short-run congestion need not produce immediate improvements in service quality. Our empirical analysis exploits within-facility variation in workload across SP visits and examines adjustments along each margin. Table 1 summarises the predicted responses and provides a roadmap for the empirical analysis.

¹In our setting there is no dynamic pricing and although informal payments could in principle allow patients to bypass queues, no SP was asked for such payments.

Table 1: Empirical tests and predicted responses under excess demand

Adjustment margin	Empirical test	Predicted response
Rationing through waiting	Waiting time	↑
Quality rationing (intensive)	Consultation duration	↓
	Checklist adherence	↓
	Clinical quality	↓
	Interpersonal quality	↓
Quality rationing (extensive)	Provider skill level	↓
	Access	↓

Institutional determinants of adjustment margins: Which margin absorbs excess demand depends on the institutional constraints and incentives faced by providers. For example, providers may be less willing to let queues grow when waiting times are monitored or when they stand to benefit financially from increasing throughput, as under fee-for-service payment. Quality rationing may be more feasible when service quality is difficult to observe and providers face limited accountability for reducing it, while extensive-margin adjustment may depend on the discretion providers have over access and assignment, including whether they can defer or redirect users.

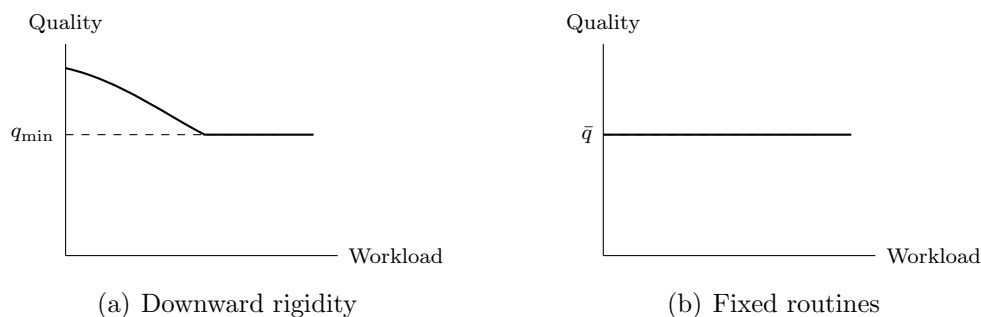
Many non-urgent queue-based public services are organised in ways that, in principle, leave scope for all of these adjustment margins. This includes welfare and employment offices, civil registration agencies, licensing authorities, visa and consular services, as well as the setting we study – walk-in primary care. Waiting is a routine feature of accessing such services, particularly in LMICs, where long queues at public offices are widespread. Public-sector workers are also typically salaried, limiting direct incentives to increase throughput (Pepinsky et al. 2017). In addition, service quality is often difficult to observe in the public sector and, particularly in LMICs, provider accountability is limited (Finan, F 2017). Frontline workers in many of these settings also retain discretion over access, including asking users to return or redirecting them elsewhere. Hence, the margin through which excess demand is absorbed is ultimately an empirical question.

Mechanisms: As we document later in the paper, rationing through waiting appears to be the dominant adjustment margin to short-run congestion in our setting. Whilst we cannot definitively identify the underlying mechanism, we consider two explanations for the absence of quality rationing that have different implications for policy. Under *downward rigidity in quality*, providers are already close to a minimum level of care, leaving little room to reduce

quality further as workload rises. In this case, reducing workload should allow quality to increase. Under *fixed consultation routines*, providers instead deliver a similar package of care regardless of workload, implying that reducing workload alone would have limited effects on quality. Distinguishing between these explanations is therefore important for understanding whether policies that expand capacity can improve service quality.

While the two mechanisms may imply similar average levels of quality, they differ in whether and how quality responds to changes in workload, as illustrated in Figure 1. Both mechanisms generate similar predictions at high workload – namely, no quality rationing along the intensive or extensive margin. However, they diverge in their implications at very low workload, which we examine empirically. Under downward rigidity, we would expect effort or quality to be higher when congestion is absent, whereas under fixed routines, we would expect behaviour to vary little with workload.

Figure 1: Alternative mechanisms



3 Study design

3.1 Study setting

This study was conducted in Nairobi County, Kenya. We focus on primary healthcare facilities, which serve as the main point of contact for patients seeking care for non-urgent conditions. Consultations are typically walk-in rather than appointment-based, as is common in primary care facilities in LMICs. When a patient arrives at a facility, they usually pay a consultation fee at reception and join the queue outside the office of a specific provider. Providers then generally move between their consultation room and the queuing area to call in the next patient – allowing them to observe the length of the queue. There is no

formal symptom-based triage. Patients are typically assigned to providers by service type, with midwives providing antenatal and family planning care and nurses often specialising in paediatric or adult care. Most adult curative consultations are conducted by clinical officers and nurses. As in most low and middle-income countries, provider discretion is high and accountability mechanisms are weak (Björkman & Svensson 2009): malpractice litigation is extremely rare, professional regulatory bodies seldom sanction providers, and patients generally lack the legal access or resources required to pursue formal complaints (Chilumpha et al. 2023).

Data were collected from 250 small to medium-sized primary care facilities (levels 2 and 3), comparable to UK GP practices. The sample includes all public primary care facilities and a random sample of 40% of private-for-profit facilities serving the approximately five million residents of Nairobi County. Facilities were drawn from the Kenya Master Health Facility Registry (KMHFR) and screened through brief site visits. We excluded facilities that did not provide general outpatient primary care or were not operational at the time of sampling.² Ethical approval for the study was obtained from the Amref Ethics and Scientific Review Committee (AMREF-ESRC P1563/2023), the National Commission for Science, Technology and Innovation (NACOSTI/P/24/33778), and the Nairobi City County Health Research Ethics Committee (NCCG/HWN/REC/534).

We combine three sources of data. First, unannounced SP visits measure provider behaviour across adjustment margins and queue length at arrival. Second, a facility survey provides staffing and outpatient volumes, which we use to construct facility-specific measures of typical workload and identify periods of excess demand. Third, a healthcare worker survey measures staff characteristics and clinical knowledge and assesses SP detection.

3.2 Standardised patients

SPs are healthy individuals trained to portray a specific medical condition in a consistent and clinically accurate manner. SPs visited healthcare providers undercover (posing as real patients), presenting a pre-specified case that closely followed a standardised script. Shortly after each consultation, SPs completed a structured debrief recording waiting time, queue length at arrival, consultation duration, history taking, physical examinations, diagnostic tests, medications, fees, and bedside manner. SPs are widely regarded as the gold standard

²Exclusions included dental practices, laboratories, specialist clinics, hospital-based clinics, facilities serving narrowly defined populations, and facilities in inaccessible or restricted locations.

for measuring the clinical quality of outpatient care (Peabody et al. 2000, Wiseman et al. 2019). Because the underlying condition is fixed and known to the researcher, SP data allow direct assessment of whether providers offered guideline-indicated care – which is difficult to achieve with observational data. Moreover, because SPs are indistinguishable from regular patients and visits are unannounced, they provide an unbiased measure of provider performance in routine clinical practice.

SPs presented cases of stable angina, asthma, tuberculosis, upper respiratory tract infection, and family planning. All cases have been used in previous SP studies in Kenya and other LMICs (Das et al. 2016, Daniels et al. 2017, Kwan 2020, Kovacs et al. 2022, Kovacs & Lagarde 2022). Case presentations were deliberately unambiguous and fully consistent with textbook descriptions of the relevant conditions – see Appendix A for details. The study design scheduled eight SP visits per facility: one visit each for asthma, tuberculosis, upper respiratory tract infection, and family planning, and four visits presenting the stable angina case (two male and two female SPs reporting identical symptoms). A small number of scheduled visits (36 visits in 21 facilities) did not occur because facilities reported that they did not provide the relevant service (primarily family planning consultations) or because facilities were not open on the day the SP visit was scheduled.

Our identification strategy relies on three key features of the study design. First, SPs present pre-specified clinical cases following standardised scripts, ensuring that the underlying condition and case presentation were identical across visits. This removes variation in case mix that might otherwise be correlated with provider behaviour. Second, providers were unaware that SP visits would occur – as we obtained ethical approval with a waiver of informed consent. Providers therefore had no opportunity to anticipate visits or adjust their behaviour in advance. Third, SP visits followed a strict pre-defined schedule fixed prior to fieldwork and based solely on logistical considerations. Unlike normal patients, SPs therefore neither selected whether to seek care nor altered the timing of visits in response to workload. The schedule was designed to minimise travel distances for fieldworker teams, ensure that each facility was visited by all SPs, and limit facilities to at most one visit per day. Roughly half of visits occurred between 6:00 and 10:30, consistent with typical patient behaviour in the study setting, where patients often arrive before noon.

SPs were recruited from the general population in Nairobi³ and we exclude individuals with

³We advertised the position through newspaper advertisements and postings in community spaces such as churches and community centres. All SPs had completed primary education or some secondary education – similar to middle and lower-middle-class Kenyans.

prior research participation or ties to the healthcare sector⁴. SPs completed two weeks of intensive training to ensure consistent case portrayal and reliable recall of consultation details. We selected 20 SPs for the angina case (10 male and 10 female) and eight SPs each for the remaining conditions. During fieldwork, SPs met with their supervisor each morning and travelled independently to facilities. SPs consulted whichever provider was on duty – as normal patients would. Shortly after each visit, SPs were debriefed using a structured questionnaire and submitted all prescriptions, medications or referrals, which were coded by medically trained enumerators. SPs also reported provider names and physical descriptions – which were used to identify providers for the purpose of linking data⁵.

Data collection occurred in two waves: May–June 2024 (wave 1) and February–March 2025 (wave 2). In wave 1, six SPs visits took place in each of the 250 sampled facilities (asthma, tuberculosis, upper respiratory tract infection, family planning, female angina patient, male angina patient). In wave 2, two further SP visits (female angina patient, male angina patient) were conducted, as part of endline data collection for a training intervention. We include wave 2 observations from the 125 baseline facilities that received no intervention. In total, SPs completed 1,676 consultations, of which 57% (718) were for stable angina.

3.3 Facility survey

Between the two waves of SP data collection (August–September 2024), we conducted a facility survey to measure staffing and outpatient consultation volumes, which we use to construct facility-specific measures of typical daily workload per provider. Enumerators compiled a roster of providers conducting outpatient consultations, recording their position and full-time or part-time status. They also verified outpatient consultation registers to obtain the total number of consultations conducted at each facility in March, April, May, and June 2024. Together, these data provide the benchmark against which we measure short-run excess demand at the time of each SP visit. The survey also collected information on facility characteristics, including infrastructure, medical equipment, services provided, opening hours, and fees. Facilities without accessible or complete consultation registers were excluded from the analysis, leaving 198 facilities with complete staffing and consultation-volume data.

⁴Including having a healthcare provider in the immediate family, prior employment in the medical sector, or having had strong positive or negative experiences with healthcare providers.

⁵If providers did not introduce themselves during the consultations, names were obtained via name tags, signatures on prescriptions or referrals letters or via facility visits taking place after all SP visits were completed.

3.4 Provider survey

In parallel with the facility survey, we collected data from healthcare providers who had seen SPs. The survey included detailed information on provider characteristics as well as clinical vignettes – short descriptions of patient cases commonly used to measure provider knowledge – identical to the SP cases (see Appendix B). Vignettes were presented as short paragraphs, summarising all relevant information, and providers were asked how they would diagnose each case, fully aware their knowledge was being tested.

Providers also participated in a detection survey in which they were informed that SPs had visited their facility in recent months and asked whether they believed they had consulted such patients. Providers were asked to report gender and symptoms and whether their suspicions arose during the consultation (e.g. because of SP behaviour or the symptoms reported) or after the consultation (e.g. because SPs did not return for follow-ups or did not pay for drugs). Detection rates were low (20 SPs, or 1% of visits), and these consultations are excluded from the analysis.

4 Data and empirical strategy

4.1 Measuring key adjustment margins

Guided by the conceptual framework, we examine how providers respond to excess demand along three distinct adjustment margins: rationing through waiting, quality rationing along the intensive-margin and quality rationing along the extensive-margin.

To measure rationing through waiting, we use waiting time, defined as the number of minutes between an SP’s arrival at the facility and entry into the consultation room.

We measure intensive-margin quality rationing using three sets of outcomes. First, provider effort is measured by consultation duration and checklist adherence, defined as the share of relevant history questions and physical examinations completed (Appendix A). Second, clinical quality is measured by whether the SP received guideline-indicated care (Appendix Table A1). Third, interpersonal quality is measured using two SP-reported indices capturing bedside manner (e.g. ensuring privacy, limiting distractions and explaining diagnoses and treatments) and subjective experience (agreement with statements such as “The provider

was unkind”, “The provider was attentive”; see Appendix Table A3) – see Appendix Tables A2 and A3.

To capture quality rationing along the extensive margin, we examine task-shifting (cadre of the provider seen) and access restrictions (whether SPs were refused care or asked to return on another day; neither occurred).

4.2 Measuring workload and excess demand

Realised workload is measured via queue length, defined as the number of patients already waiting at the facility when the SP arrived. When SPs arrived at the facility, they generally paid for the consultation and joined a queue outside the office of a given provider. We refer to the number of patients already waiting in this queue as Q_{ipf} for consultation i with provider p in facility f .

Our preferred measures of workload benchmark queue length to facility-specific workload. Following Kovacs & Lagarde (2022), we construct a measure of typical daily workload per provider, WL_f , as the average number of consultations in facility f over a four-month period divided by the number of full-time staff attached to the facility⁶, assuming 20 working days per month and an equal distribution of consultations across providers, i.e. $WL_f \equiv Average_consult_f / (20 \times Staff_f)$.

We define a binary measure of excess demand as

$$ExcessDemand_{ipf}^B = 1\{Q_{ipf} > WL_f\},$$

which equals one when the number of patients waiting at arrival exceeds the facility-specific typical daily workload per provider. We also define a continuous measure,

$$ExcessDemand_{ipf}^C = \frac{Q_{ipf}}{WL_f},$$

which captures queue length relative to typical workload and is standardised to have mean zero and unit variance. We also use quintiles and deciles of this measure to allow for non-

⁶The number of full-time staff attached to a facility is highly similar to the number of staff that facility managers estimate to be present on a normal day. Based on the staffing roster, there are 3.2 full-time providers working in study facilities and facility managers estimate that there are 3.3 staff present on a normal day. For 72% of facilities, the discrepancies between the roster and managers’ estimates is less than two providers.

linear responses.

4.3 Identification

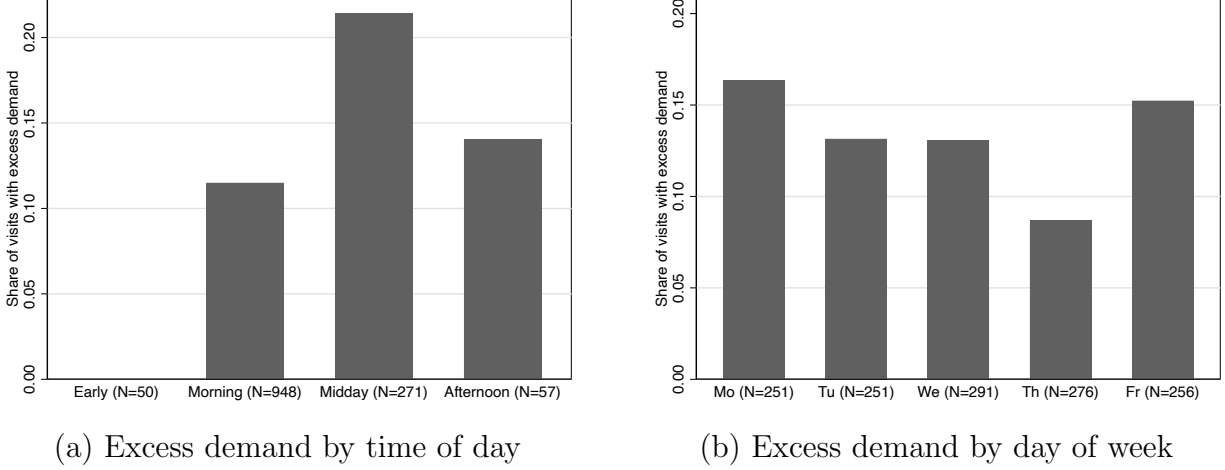
Identification exploits within-facility variation in workload across SP visits. Eight SP visits were scheduled per facility at different days and times according to a pre-defined timetable fixed prior to fieldwork and independent of daily conditions at the facility. Because workload varies systematically with time of day and day of week (Figure 2), all specifications control for these predictable fluctuations in demand. The SP design addresses two key identification challenges. First, SPs neither condition their decision to seek care on workload nor alter visit timing when queues are long, eliminating endogenous selection into queues. Second, as SPs present standardised clinical cases, case mix is fixed across visits, and providers were unaware visits were taking place.

Identification requires that, conditional on facility fixed effects, clinical case, and visit timing, variation in workload across SP visits is not systematically associated with other determinants of provider behaviour. As a robustness check, we estimate specifications with provider fixed effects, comparing visits to the same healthcare worker at different levels of workload. The average provider saw three SPs, with between one and eight visits per provider.

Our approach relates to a large literature studying excess demand in healthcare settings, particularly in emergency departments (Hoe 2022, Francetic et al. 2024, Turner et al. 2020, Martins & Filipe 2020). These studies typically exploit observational variation in demand, using detailed temporal controls to account for systematic fluctuations in workload. Our SP design additionally holds case mix fixed and eliminates patient selection into queues.

There is substantial within-facility variation in exposure to excess demand. Among facilities that experience excess demand, 90% are observed both with and without excess demand; on average, 45% of SP visits to these facilities occur during periods of excess demand. Facilities observed with and without excess demand are similar in infrastructure, equipment, and distance to referral facilities, although excess demand is more common in larger and public facilities (see Appendix Table D1). Providers are similar in skills, clinical knowledge, training, and psychological wellbeing; those exposed to excess demand report seeing more patients per day and earning lower salaries – as is the case in the public sector (see Appendix Table D2).

Figure 2: Excess demand by timing of visits



Notes: Share of visits with excess demand by time of day and day of week. Excess demand is defined as $Q_{ipf} > W_f$, where Q_{ipf} is the number of patients waiting when the SP arrived and W_f is the facility-specific average daily workload per provider. Time-of-day bins are defined as early morning (07:00–08:59), morning (09:00–11:59), midday (12:00–13:59), and afternoon (14:00–17:59). Sunday is excluded.

4.4 Empirical approach

We estimate the effect of excess demand on each outcome using linear regression models that exploit within-facility variation in excess demand at the time of the SP visit. For consultation i with provider p in facility f , we estimate:

$$Y_{ipf} = \beta W_{ipf} + \gamma' X_{ipf} + \alpha_f + \varepsilon_{ipf}, \quad (1)$$

where Y_{ipf} captures rationing through waiting or quality rationing along the intensive or extensive margin, and W_{ipf} measures workload. α_f denotes facility fixed effects, so identification comes from variation in workload across SP visits within the same facility. X_{ipf} includes pre-specified controls for SP case, day of week, and time of day. In additional specifications, we replace facility fixed effects with provider fixed effects, exploiting variation in workload across visits to the same healthcare worker.

4.5 Descriptives

The analysis draws on 1,326 standardised patient visits across 198 facilities – excluding visits in facilities for which we do not have sufficient information to calculate typical workload WL_f , as well as SP visits detected during the consultations. Overall, 13% of SP visits occurred during periods of excess demand (i.e. $ExcessDemand_{ipf}^B = 1$), where there were on average 20 patients waiting to be seen, compared to 3 patients otherwise.

The average number of patients seen per provider per day (WL_f) is 19, with higher workloads in public facilities (30 patients) compared to private-for-profit facilities (13 patients). This is comparable to public primary-care physicians in London (who see around 20 patients) (Pilvar & Watt 2025).

Appendix Table C1 describes the study sample and shows facility and provider characteristics. As shown in Table 2, SP waited on average 38 minutes and consultations lasted 11 minutes on average – slightly above the traditional NHS benchmark of 10 minutes per GP consultation. Yet, providers completed only 40% of recommended checklist items (history questions and physical examinations) and guideline-indicated care was provided to only 36% of SPs. Moreover, there is a substantial gap between knowledge and practice, as 62% of providers correctly managed a vignette identical to the SP case. Overall, for 61% of SP cases that were incorrectly managed providers demonstrated sufficient knowledge in the vignette.

Interpersonal quality was high on both measures. Providers completed on average 80% of procedural behaviours consistent with good bedside manner (ensuring privacy, limiting distractions, explaining procedures and diagnoses) and received favourable subjective evaluations of bedside manner, with scores concentrated toward the more positive end of the distribution (with 1 being the best possible score).

Table 2: Summary statistics on provider behaviour

	Observations	Mean	Std. dev.	Min	Max
Waiting time (min.)	1,326	37.93	51.92	1	344
Consultation duration(min.)	1,326	10.73	8.91	1	121
Checklist adherence (prop.)	1,326	0.40	0.19	0	0.93
Correct case management	1,326	0.36	0.48	0	1
Correct diagnosis in vignette	892	0.62	0.48	0	1
Bedside manner (behaviour)	1,326	0.79	0.16	0.17	1
Bedside manner (subjective)	1,326	1.97	0.70	1	5

Note: Bedside manner (behaviour) captures the proportion of behaviours adopted that are in-line with good bedside manner: ensuring privacy, limiting distractions and explaining treatments, diagnoses and examinations. Bedside manner (subjective) refers to a composite score of patients’ subjective experience, where 1 is the best possible score and 5 is the worst possible score (this index is standardised to a mean of zero and an SD of one in the analysis). The correct diagnosis in vignette refers to whether providers correctly diagnosed a hypothetical patient corresponding to the SP case – available for the subset of providers who completed the provider survey.

5 Results

5.1 Rationing through waiting

We first examine rationing through waiting. Column (1) of Table 3 uses queue length as a measure of realised demand. Each additional patient in the queue increases waiting time by 1.7 minutes. Columns (2) and (3) show our preferred measures of excess demand relative to typical workload. Waiting times increase by 21 minutes (0.4 SD) when demand exceeds the typical daily workload per provider, while a one standard deviation increase in our continuous measure of excess demand increases waiting time by around 9 minutes. Notably, these increases are not mechanically implied as providers could adjust along other margins, for example by shortening consultations. Effects are similar across public and private facilities (Appendix Table E1). Overall, the results show that waiting is an important margin of adjustment to short-run excess demand.

Table 3: Effects on rationing through waiting

	(1)	(2)	(3)
Realised demand (queue length)	1.653*** (0.307)		
Above typical workload (binary)		20.961*** (6.497)	
Relative to typical workload (continuous)			9.209*** (3.022)
Observations	1,326	1,326	1,321
R-squared	0.639	0.607	0.608

Note: The dependent variable is waiting time, measured in minutes between arrival at the facility and entry into the consultation room. Column (1) reports results using realised demand, measured by queue length (Q_{ipf}), defined as the number of patients waiting when the SP arrived. Columns (2) and (3) report results using measures of excess demand relative to facility-specific typical workload. *Above typical workload (binary)* is an indicator equal to one if queue length exceeds the facility-specific average daily workload per provider ($1\{Q_{ipf} > WL_f\}$). *Relative to typical workload (continuous)* scales queue length by this benchmark (Q_{ipf}/WL_f) and is standardised to have mean zero and unit variance. All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.2 Quality rationing along the intensive-margin

We next examine whether providers ration quality along the intensive margin by reducing effort or service quality. Panel A of Table 4 uses queue length to measure realised demand. Across consultation duration, checklist adherence, clinical quality, and interpersonal quality, the estimates are small, precisely estimated, and provide no evidence that provider performance deteriorates as queues lengthen.

Panel B shows our preferred measures of excess demand relative to typical workload. We focus on the binary measure, which directly captures periods when queue length at arrival exceeds the facility’s typical daily workload per provider, although results using the continuous measure are highly similar. We find no strong evidence of quality rationing. The estimates for provider effort are particularly precise: consultation duration increases by 1.34 minutes (95% CI $[-0.78; 3.46]$), while checklist adherence increases by 1.9 percentage points (95% CI $[-1.4; 5.2]$ pp), ruling out meaningful reductions in either measure. Correct case management increases by 2.4 percentage points (95% CI $[-8.8; 13.6]$ pp), relative to a base-line of 36% – leaving room for both a meaningful decline and a sizeable improvement in clinical quality. While interpersonal quality is *a priori* a margin that may adjust readily

to higher workload, estimates for both measures are close to zero: the null is precise for procedural bedside manner (95% CI $[-0.037; 0.037]$) and less so for subjective experience (95% CI $[-0.251; 0.097]$).

A key concern is that queue length partly reflects providers' endogenous responses to workload, as providers may accelerate consultations when queues form. However, we precisely estimate effects on duration and find no decline with excess demand. In addition, for prior increases in service speed to explain this null finding, SPs would need to systematically arrive after providers had already reduced congestion through faster consultations – unlikely given the predetermined timing of visits. Results are also highly similar – although less precise – with provider fixed effects, suggesting that the null is not driven by persistent differences in consultation speed across providers (Appendix Table G2). Moreover, our preferred measures capture congestion relative to facility-specific typical workload rather than queue length alone, reducing concerns that faster providers mechanically generate shorter queues. Finally, waiting times rise sharply during periods of excess demand (Table 3), indicating that any increase in consultation speed is insufficient to offset congestion.

The absence of intensive-margin adjustment is robust across sectors, despite differences in workload and incentives between public and private facilities (Appendix Table F1). It is also not explained by limited provider knowledge: restricting the sample to providers who correctly diagnose a vignette matching the SP case yields similar null effects on effort and quality (Appendix Table F2). Finally, we find no evidence of input rationing, with fees and drug prescriptions unchanged under excess demand (Appendix Table G1).⁷ Taken together, these results provide little evidence that providers respond to high workload by quality rationing along the intensive margin.

⁷We also note that none of our SPs were asked for informal payments by healthcare providers.

Table 4: Effects on intensive-margin rationing

	(1)	(2)	(3)	(4)	(5)
	Duration	Checklist	Correct	Behaviours	Experience
Panel A: Realised demand					
Queue length	0.026	0.001	0.001	0.001	0.001
	(0.021)	(0.001)	(0.002)	(0.001)	(0.003)
Observations	1,326	1,326	1,326	1,326	1,326
R-squared	0.370	0.585	0.374	0.348	0.542
Panel B: Excess demand relative to typical workload					
Above typical workload (binary)	1.341	0.019	0.024	-0.000	-0.077
	(1.083)	(0.017)	(0.057)	(0.019)	(0.089)
Observations	1,326	1,326	1,326	1,326	1,326
R-squared	0.370	0.585	0.374	0.347	0.542
Relative to typical workload (continuous)	0.555*	0.009	0.014	-0.001	0.024
	(0.321)	(0.006)	(0.019)	(0.006)	(0.047)
Observations	1,321	1,321	1,321	1,321	1,321
R-squared	0.370	0.585	0.373	0.343	0.542

Note: Panel A reports results for realised demand, measured by queue length (Q_{ipf}), defined as the number of patients waiting when the SP arrived. Panel B reports results for excess demand relative to facility-specific typical workload. *Above typical workload (binary)* is an indicator equal to one if queue length exceeds the facility-specific average daily workload per provider ($1Q_{ipf} > WL_f$). *Relative to typical workload (continuous)* scales queue length by this benchmark (Q_{ipf}/WL_f) and is standardised to have mean zero and unit variance. Columns (1)–(2) measure provider effort: consultation duration (minutes) and checklist adherence (the proportion of relevant history questions asked and physical examinations performed). Columns (3)–(5) capture different dimensions of quality of care. Correct is a binary indicator for clinically correct case management (whether SPs received clinically indicated advice, referrals, or treatments). Behaviours captures the proportion of procedural actions consistent with good bedside manner (ensuring privacy, limiting distractions, explaining the diagnosis, treatment, or examinations). Experience is a standardised index capturing the SP’s subjective assessment of the consultation based on agreement with nine statements (e.g. “The provider was attentive” or “The provider was unkind”). All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. The sample includes 196 facilities. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.3 Quality rationing along the extensive-margin

We next examine whether excess demand induces extensive-margin rationing through task-shifting – shifting patients away from doctors and clinical officers toward nurses – or through restrictions on access, such as redirecting or refusing service.

Table 5 reports results for the subsample of visits (67%) for which provider cadre can be identified. Panel A shows no systematic link between realised demand and provider assignment. Panel B turns to our preferred measures of excess demand relative to typical workload. Again, we find no evidence of task-shifting. Under the binary specification, the estimated effects on the probability of seeing a skilled or high-skilled provider are 1.2 percent-

age points (95% CI $[-5.9; 8.3]$) and 1.8 percentage points (95% CI $[-10.0; 13.6]$), respectively. The estimated effect on the probability of seeing a nurse is similarly close to zero (95% CI $[-11.6; 10.4]$). Results using the continuous measure are highly similar. Nor do we find evidence of restrictions on access: no SP was refused care, redirected, or asked to return later. This is not driven by the urgency of the cases, as none were medical emergencies and several could plausibly have been deferred (in particular family planning and upper respiratory tract infection). Taken together, we find no evidence suggesting that providers respond to excess demand through extensive-margin rationing.

Table 5: Effects on extensive-margin rationing

	(1) Skilled	(2) High-skilled	(3) Nurse	(4) Doctor	(5) Clinical officer
Panel A: Realised demand					
Queue length	-0.000 (0.002)	0.003 (0.002)	-0.004*** (0.001)	-0.000 (0.000)	0.004 (0.002)
Observations	892	892	892	892	892
R-squared	0.608	0.666	0.673	0.805	0.698
Panel B: Excess demand relative to typical workload					
Above typical workload (binary)	0.012 (0.036)	0.018 (0.060)	-0.006 (0.056)	0.015 (0.018)	0.003 (0.062)
Observations	892	892	892	892	892
R-squared	0.608	0.664	0.670	0.805	0.696
Relative to typical workload (continuous)	0.001 (0.009)	0.020 (0.014)	-0.020* (0.012)	0.001 (0.003)	0.020 (0.014)
Observations	890	890	890	890	890
R-squared	0.605	0.660	0.666	0.805	0.694

Note: Panel A reports results for realised demand, measured by queue length (Q_{ipf}), defined as the number of patients waiting when the SP arrived. Panel B reports results for excess demand relative to facility-specific typical workload. *Above typical workload (binary)* is an indicator equal to one if queue length exceeds the facility-specific average daily workload per provider ($1Q_{ipf} > WL_f$). *Relative to typical workload (continuous)* scales queue length by this benchmark (Q_{ipf}/WL_f) and is standardised to have mean zero and unit variance. Outcomes are indicator variables capturing the type of provider seen by the SP. Skilled provider equals one if the SP was seen by a nurse, clinical officer, or doctor. High-skilled equals one if the SP was seen by a clinical officer or doctor. The remaining columns indicate whether the SP was seen by a nurse, doctor, or clinical officer, respectively. All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. The analysis uses data from providers who saw SPs and completed the provider survey (187 facilities). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.4 Robustness

We conduct a number of robustness checks. First, we find no evidence of intensive or extensive-margin rationing even at the highest levels of workload. Appendix H.1 present

results using quintiles and deciles of excess demand, which are in line with our main analysis. Second, measurement error in the number of patients waiting may be more severe in larger facilities, where multiple providers may draw from a common patient queue. Appendix H.2 therefore restricts the analysis to small (level 2) facilities, for which workload is measured more precisely and results are highly similar to the full sample.

5.5 Mechanisms: Why is quality not rationed under excess demand

The absence of quality rationing may reflect either downward rigidity (limited scope to reduce quality below a minimum) or fixed consultation routines that vary little with workload. The two explanations are difficult to disentangle at high workload but have different implications at low workload: downward rigidity would imply higher effort or quality, whereas fixed routines would imply little change.

Table 6 reports estimates when no patients were waiting and when workload was in the lowest decile of the excess-demand distribution. We find no evidence of higher effort or quality at low workload: estimates are close to zero across consultation duration, clinical quality, and interpersonal quality, while checklist adherence is slightly lower. We similarly find no extensive-margin adjustment (Appendix I). While these patterns do not identify the underlying mechanism, they are more consistent with relatively fixed consultation routines than with downward rigidity.

Table 6: Rationing along the intensive margin at very low workload

	(1)	(2)	(3)	(4)	(5)
	Duration	Checklist	Correct	Behaviours	Experience
No patients waiting	-0.476	-0.023**	0.013	-0.011	0.073
	(0.630)	(0.011)	(0.035)	(0.013)	(0.060)
Observations	1,326	1,326	1,326	1,326	1,326
R-squared	0.370	0.587	0.374	0.348	0.543
Lowest excess demand decile	-0.504	-0.022**	0.014	-0.012	0.065
	(0.632)	(0.011)	(0.035)	(0.014)	(0.059)
Observations	1,326	1,326	1,326	1,326	1,326
R-squared	0.370	0.587	0.374	0.348	0.543

Note: This table examines provider behaviour at very low levels of workload. Panel A uses an indicator for zero queue length at arrival, defined as $1\{Q_{ipf} = 0\}$. Panel B uses an indicator for whether excess demand falls in the lowest decile of its standardised distribution, $1\{Q_{ipf}/WL_f \leq P_{10}\}$. Outcomes capture provider effort and quality of care: consultation duration (minutes), effort (the proportion of relevant history questions asked and physical examinations performed), correct case management (binary), and two standardised indices of interpersonal quality (procedural behaviours and subjective experience). All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

6 Conclusion

Excess demand is pervasive in public service delivery, but its welfare consequences depend on how providers respond to it. This paper uses rich micro-level data to examine how frontline providers allocate scarce capacity when demand exceeds providers' typical daily workload. Using evidence from walk-in primary healthcare facilities in Nairobi, we study whether excess demand is absorbed through waiting or quality rationing. We find that when queue length at arrival exceeds the typical daily workload per provider, waiting time is approximately 21 minutes longer (0.4 SD). We find no statistically significant reductions in consultation duration, checklist adherence, correct management, service quality, provider assignment or access. We also find no evidence for quality rationing at the highest levels of workload, among providers with sufficient clinical knowledge, and in specifications including healthcare worker fixed effects. Finally, we consider potential mechanisms for the absence of quality rationing and find evidence that is broadly consistent with providers following fixed consultation routines that vary little with workload.

Our findings contribute to the economic literature on congestion and non-price rationing in public service delivery. While canonical models often treat waiting time as an equilibrium outcome (Barzel 1974, Deaton 1981, Besley & Coate 1991), we provide micro-level evidence on how excess demand is absorbed at the point of service when frontline providers have

discretion over how scarce capacity is allocated. Results also speak to a common concern motivating policies to reduce congestion: that long queues and high provider workloads compromise the quality of public services. In our setting, however, short-run excess demand is borne primarily through users' time rather than deterioration in the services they receive. Reducing short-run congestion therefore need not produce improvements in service quality.

Funding: This research was funded by a Wellcome Research Fellowship in Humanities and Social Science 219744/Z/19/Z, an Adlerbertska Foundations grant (Erikka Adlers stipendium för forskning inom hållbarhet i ett helhetsperspektiv) and Carl Bennet.

Declarations of interest: None.

Disclosure Statement: The author declares that they have no relevant or material financial interests that relate to the research described in this paper.

References

- Barzel, Y. (1974), ‘Discrimination by Waiting Time in Merit Goods’, *Journal of Law and Economics* .
URL: <https://www.jstor.org/stable/1813429?seq=1>
- Besley, T. & Coate, S. (1991), ‘Public Provision of Private Goods and the Redistribution of Income’, *The American Economic Review* **81**(4), 979–984.
URL: <https://www.jstor.org/stable/2006658>
- Björkman, M. & Svensson, J. (2009), ‘Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda *’, *Quarterly Journal of Economics* **124**(2), 735–769.
URL: <https://academic.oup.com/qje/article-lookup/doi/10.1162/qjec.2009.124.2.735>
- Chilumpha, M., Chatha, G., Umar, E., McKee, M., Scott, K., Hutchinson, E. & Balabanova, D. (2023), ‘We stay silent and keep it in our hearts’: a qualitative study of failure of complaints mechanisms in Malawi’s health system’, *Health Policy and Planning* **38**(Supplement.2), ii14–ii24.
URL: <https://doi.org/10.1093/heapol/czad043>
- Cullis & Jones (2000), Chapter 23 Waiting lists and medical treatment: Analysis and policies, in ‘Handbook of Health Economics’, Vol. 1, Elsevier, pp. 1201–1249.
URL: <https://www.sciencedirect.com/science/chapter/handbook/abs/pii/S1574006400800360>
- Daniels, B., Dolinger, A., Bedoya, G., Rogo, K., Goicoechea, A., Coarasa, J., Wafula, F., Mwaura, N., Kimeu, R. & Das, J. (2017), ‘Use of standardised patients to assess quality of healthcare in Nairobi, Kenya: a pilot, cross-sectional study with international comparisons’, *BMJ Global Health* **2**(2), e000333.
URL: <http://gh.bmj.com/lookup/doi/10.1136/bmjgh-2017-000333>
- Danziger, S., Levav, J., Avnaim-Pesso, L. & Kahneman, D. (2011), ‘Extraneous factors in judicial decisions’, *Proceedings of the National Academy of Sciences of the United States of America* **108**(17), 6889–6892.
URL: <https://www.jstor.org/stable/41242099>
- Das, Holla, A., Mohpal, A. & Muralidharan, K. (2016), ‘Quality and Accountability in

- Healthcare Delivery: Audit-Study Evidence from Primary Care in India', *American Economic Review* **106**(12), 3765–3799.
- Das, J., Hammer, J. & Leonard, K. (2008), 'The Quality of Medical Advice in Low-Income Countries', *The Journal of Economic Perspectives* **22**(2), 93–114.
URL: <https://www.jstor.org/stable/27648243>
- Deaton, A. S. (1981), Theoretical and empirical approaches to consumer demand under rationing, in A. Deaton, ed., 'Essays in the Theory and Measurement of Consumer Behaviour: In Honour of Sir Richard Stone', Cambridge University Press, Cambridge, pp. 55–72.
URL: <https://www.cambridge.org/core/books/essays-in-the-theory-and-measurement-of-consumer-behaviour/theoretical-and-empirical-approaches-to-consumer-demand-under-rationing/EB0D3549CD60E60A4F37157EC929A774>
- Dixit, A. (2002), 'Incentives and Organizations in the Public Sector: An Interpretative Review', *Journal of Human Resources* **37**(4), 696–727.
URL: <https://EconPapers.repec.org/RePEc:uwp:jhriss:v:37:y:2002:i:4:p:696-727>
- Finan, F (2017), The Personnel Economics of the Developing State, in 'Handbook of Economic Field Experiments', Vol. 2, North-Holland, pp. 467–514.
URL: <https://www.sciencedirect.com/science/article/pii/S2214658X16300034>
- Francetic, I., Meacock, R. & Sutton, M. (2024), 'Free-for-all: Does crowding impact outcomes because hospital emergency departments do not prioritise effectively?', *Journal of Health Economics* **95**, 102881.
URL: <https://www.sciencedirect.com/science/article/pii/S0167629624000262>
- Hemrajani, R. & Jr, T. H. (2024), 'The Effects of Decision Fatigue on Judicial Behavior: A Study of Arkansas Traffic Court Outcomes', *Journal of Law and Courts* **12**(2), 435–443.
URL: <https://www.cambridge.org/core/journals/journal-of-law-and-courts/article/abs/effects-of-decision-fatigue-on-judicial-behavior-a-study-of-arkansas-traffic-court-outcomes/8B7EB8735C10F7730FB402D6F2E80D70>
- Hoe, T. P. (2022), 'Does Hospital Crowding Matter? Evidence from Trauma and Orthopedics in England', *American Economic Journal: Economic Policy* **14**(2), 231–262.
URL: <https://www.aeaweb.org/articles?id=10.1257/pol.20180672>
- Kovacs, R. J., Lagarde, M. & Cairns, J. (2022), 'Can patients improve the quality of care they receive? Experimental evidence from Senegal', *World Development* **150**, 105740.
- Kovacs, R. & Lagarde, M. (2022), 'Does high workload reduce the quality of healthcare? Evidence from rural Senegal', *Journal of Health Economics* **82**, 102600.
URL: <https://www.sciencedirect.com/science/article/pii/S0167629622000200>
- Kwan, A. (2020), 'Can we improve quality of care in private health sectors? Evidence from a randomized field experiment in Kenya', *UC Berkeley Electronic Theses and Dissertations* p. 67.
URL: <https://escholarship.org/uc/item/98384265>
- Liu, Z. & Mahler, A. C. (2026), 'The impact of working conditions on productivity: Evidence from the U.S. public defense system', *Journal of Public Economics* **257**, 105630.

- URL: <https://www.sciencedirect.com/science/article/pii/S0047272726000666>
- Martin, S. & Smith, P. C. (1999), ‘Rationing by waiting lists: an empirical investigation’, *Journal of Public Economics* **71**(1), 141–164.
- URL: <https://www.sciencedirect.com/science/article/pii/S004727279800067X>
- Martins, B. & Filipe, L. (2020), ‘Doctors’ response to queues: Evidence from a Portuguese emergency department’, *Health Economics* **29**(2), 123–137. .eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/hec.3957>.
- URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.3957>
- Mæstad, O., Torsvik, G. & Aakvik, A. (2010), ‘Overworked? On the relationship between workload and health worker performance’, *Journal of Health Economics* **29**(5), 686–698.
- URL: <http://www.ncbi.nlm.nih.gov/pubmed/20633940>
- Nichols, D., Smolensky, E. & Tideman, T. N. (1971), ‘Discrimination by Waiting Time in Merit Goods’, *American Economic Review* .
- Peabody, J. W., Luck, J., Glassman, P., Dresselhaus, T. R. & Lee, M. (2000), ‘Comparison of vignettes, standardized patients, and chart abstraction: A prospective validation of 3 methods of measuring quality.’, *Jama* **283**(13), 1715–1722.
- Pepinsky, T. B., Pierskalla, J. H. & Sacks, A. (2017), ‘Bureaucracy and Service Delivery’, *Annual Review of Political Science* **20**, 249–268.
- URL: <https://www.annualreviews.org/content/journals/10.1146/annurev-polisci-051215-022705>
- Pilvar, H. & Watt, T. (2025), ‘The effect of workload on primary care doctors on referral rates and prescription patterns: evidence from English NHS’, *The European Journal of Health Economics* **26**(5), 817–837.
- URL: <https://doi.org/10.1007/s10198-024-01742-7>
- Prendergast, C. (2007), ‘The Motivation and Bias of Bureaucrats’, *American Economic Review* **97**(1), 180–196.
- URL: <https://www.aeaweb.org/articles?id=10.1257/aer.97.1.180>
- Propper, C. (1995), ‘The Disutility of Time Spent on the United Kingdom’s National Health Service Waiting Lists’, *The Journal of Human Resources* **30**(4), 677.
- URL: <https://www.jstor.org/stable/146227?origin=crossref>
- Richards-Shubik, S., Roberts, M. S. & Donohue, J. M. (2022), ‘Measuring quality effects in equilibrium’, *Journal of Health Economics* **83**, 102616.
- URL: <https://www.sciencedirect.com/science/article/pii/S0167629622000364>
- Shurtz, I., Eizenberg, A., Alkalay, A. & Lahad, A. (2022), ‘Physician workload and treatment choice: the case of primary care’, *The RAND Journal of Economics* **53**(4), 763–791. .eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1756-2171.12425>.
- URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1756-2171.12425>
- Strobel, S. (2024), ‘Who responds to longer wait times? The effects of predicted emergency wait times on the health and volume of patients who present for care’, *Journal of Health Economics* **96**, 102898.
- Taylor, G. A., Tsui, K. K. K. & Zhu, L. (2003), ‘Lottery or waiting-line auction?’, *Journal of Public Economics* **87**(5), 1313–1334.
- URL: <https://www.sciencedirect.com/science/article/pii/S0047272701001967>

- Turner, A. J., Anselmi, L., Lau, Y.-S. & Sutton, M. (2020), ‘The effects of unexpected changes in demand on the performance of emergency departments’, *Health Economics* **29**(12), 1744–1763. [eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hec.4167](https://onlinelibrary.wiley.com/doi/pdf/10.1002/hec.4167).
URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.4167>
- Wiseman, V., Lagarde, M., Kovacs, R., Wulandari, L. P. L., Powell-Jackson, T., King, J., Goodman, C., Hanson, K., Miller, R., Xu, D., Liverani, M., Yeung, S., Hompashe, D., Khan, M., Burger, R., Christian, C. S. & Blaauw, D. (2019), ‘Using unannounced standardised patients to obtain data on quality of care in low-income and middle-income countries: key challenges and opportunities’, *BMJ Global Health* **4**(5), e001908.
URL: <http://gh.bmj.com/lookup/doi/10.1136/bmjgh-2019-001908>
- Woodworth, L. (2020), ‘Swamped: Emergency Department Crowding and Patient Mortality’, *Journal of Health Economics* **70**, 102279.

7 Online Appendix

A Standardised patients

A.1 Clinical quality of care

Table A1: Guideline-indicated care

Case	Correct	Harmful
Stable angina	ECG (performed or ordered); antiplatelet agents; statins; beta-blockers; long-acting dihydropyridine calcium channel blockers; or referral to a higher-level facility	NSAIDs
Tuberculosis	No TB treatment initiated; referral; or TB diagnostic testing (sputum smear microscopy [BAAR], chest X-ray, GeneXpert, or Mantoux test)	Systemic corticosteroids
Asthma	Oral corticosteroids or bronchodilators	NSAIDs
URTI	No antibiotics	Systemic corticosteroids
Family planning	No combined hormonal contraceptive methods	—

Notes: NSAIDs denote non-steroidal anti-inflammatory drugs. Classifications reflect guideline-consistent management for Kenyan primary care.

A.2 Interpersonal quality of care

Table A2: Provider behaviour: procedural measure of bedside manner

Item	
Did the provider use a cell phone at any point during the consultation?	
Was there a TV on in the room?	
Were there other people in the room besides clinic staff?	
Did the provider explain why they were doing specific physical examinations?	
Did the provider explain the diagnosis?	
Did the provider explain why they were recommending specific treatments or course of action?	
<i>Notes:</i> Each item was recorded as a binary response (Yes/No). The analysis uses the proportion of these behaviours that provider adopted. Negative items were reverse coded.	

Table A3: Patient experience: subjective measure of bedside manner

Item	
The provider was taking me seriously.	
The provider was attentive.	
The provider cared about my problems.	
The provider was not listening to what I was saying.	
The provider seemed distracted.	
The provider was unkind.	
The provider made me feel uncomfortable.	
The provider seemed stressed.	
The provider was looking at me while speaking.	
<i>Notes:</i> Items were measured on a 5-point Likert scale ranging from 1 (strongly agree) to 5 (strongly disagree). Negatively worded items were reverse coded. The analysis uses the average across items, standardised to have mean zero and standard deviation one.	

A.3 Stable angina

Case description

Opening statement: I have been having this crushing pain in my chest.

Question	Response (if asked)
Can you describe the pain to me?	Very strong, like crushing, in the middle of my chest (patient indicates the centre of the chest).
Was the pain very strong?	Yes, I had to stop what I was doing and sit down until it stopped.
Where exactly was the pain?	In the middle of my chest (patient presses on the centre of the chest).

Question	Response (if asked)
Did you feel the pain in other places or did it move?	Yes, it moved to the left and towards the shoulder.
Did you lose consciousness?	No.
How long did the pain last?	About five minutes.
What did you do to make it stop?	I sat down and waited.
What were you doing when the pain started?	I was running for a matatu.
Did you feel dizzy?	Yes.
Did you have heartburn?	No.
Do you ride your bike frequently?	No, not very much.
Did you have trouble breathing or feel short of breath?	Yes, the pain cut my breath.
Did the pain get worse when breathing?	No.
Have you had this problem before?	Yes, about three times this month.
What were you doing during a previous episode?	I was walking up the stairs to my flat and had to stop when the pain started.
Was the pain different from previous episodes?	No, it was pretty much the same.
Why did you not go to see a health provider earlier?	I was far from home, it passed afterwards, and I thought it was nothing and that I had eaten and drunk too much.
Did you take any medication for the pain?	No.
Did you feel nauseated or vomit?	No.
What did you eat before the episode?	I ate what I normally eat, rice and meat.
Do you have digestive problems?	No.
Do you have a fever?	No.
Do you have a headache?	No.
Do you smoke?	Yes, almost one pack a day of Sportsman cigarettes.
Do you drink alcohol?	Yes, I drink with my friends at the local club.
What is your diet like?	I like fatty foods (examples from local cuisine).
Do you exercise regularly?	No, not really.
Did anyone in your family have a similar problem?	I do not know.
Is your father alive?	No, he died when he was about 50 years old.
What did he die of?	We do not really know, an infection.
And your mother?	She is well.

Question	Response (if asked)
Do you have a history of cardiac disease or high blood pressure?	I do not know, I do not think so.
Do you have diabetes?	No.
Do you have high cholesterol?	I do not think so.
Do you take any medication?	No.

History questions and physical examinations

Questions

1. Where is the pain located?
2. How long did the pain last?
3. What was the pain like (heavy, dull, sharp)?
4. Did the pain radiate (is the pain “walking” / mobile)?
5. What were you doing when the pain began?
6. Have you had this pain before?
7. Does the pain get more or less if you inhale or exhale?
8. Were you sweating?
9. Do you smoke or have smoked in the past?
10. Are you taking any medication?
11. When did you first experience this pain?
12. How severe was the pain?
13. Does the pain get worse or better when you do certain things?
14. Were you short of breath?
15. Did you feel dizzy?
16. Were you nauseous / did you vomit?
17. Did you have a headache?
18. Did you have a fever?
19. Did you have heartburn?
20. Do you have digestive problems (diarrhoea, constipation)?
21. Have you been diagnosed with high blood pressure?
22. Do you know if you have high cholesterol?
23. Do you have diabetes?
24. Do you drink alcohol?
25. Do you exercise regularly?
26. Do people in your family have similar problems?
27. Your age?

Physical examinations

1. Take temperature with thermometer
2. Take blood pressure
3. Take pulse

4. Take respiratory frequency
5. Use stethoscope at the front of your body (chest)
6. Use stethoscope at the back
7. Measure weight
8. Measure height

A.4 Tuberculosis

Case description

Opening statement: I have a cough that is not getting better.

Question	Response (if asked)
For how long have you been coughing?	Three weeks.
Is the cough dry or productive?	It is productive.
Is there sputum?	Yes.
Do you have trouble breathing?	No.
Is there blood in the sputum?	No.
Describe the sputum to me.	It is a bit yellow and thick.
Do you have chest pain?	Sometimes.
Do you cough all day long?	Yes, I cough a lot.
Is the cough worse in the morning?	Yes, it is worst when I wake up.
Does somebody else in your house have the same cough?	No, I do not think so.
Do you smoke?	I do not smoke.
Did you have fever?	Yes, a bit, I think, at night.
Do you have night sweats?	Yes, my sheets are sometimes wet when I wake up.
Do you have chills?	Yes, sometimes a bit.
Did you lose weight?	Yes, my clothes seem baggy.
How much do you weigh?	I do not know exactly.
How much weight did you lose?	I do not really know.
Did you lose your appetite?	Yes, I do not eat much and I am not really hungry.
Do you drink alcohol?	Not really, only very occasionally for social events.
How do you feel in general?	Very tired and weak.
Do other people in your family cough a lot or have asthma?	No.
Do you have any other symptoms or health problems?	No, I just feel very tired and exhausted.
Do you have a history of tuberculosis?	No.

Question	Response (if asked)
Have you been in contact with anybody who has tuberculosis?	No, I do not think so.
Did you take any medication?	No. I went to the chemist a week ago and they gave me some cough syrup, but it did not help.
Which cough syrup did you take?	I cannot remember which one. I took it three times a day.
Why did you not come to the facility before?	I thought this would pass on its own.
Why are you coming now?	I thought the cough would pass, but it is not.
Do you know your HIV status?	Yes, it is negative.
When was the last time you got tested for HIV?	About six months ago.

History questions and physical examinations

History questions

1. How long have you been coughing?
2. Is the cough productive / is there sputum?
3. Is there blood in the sputum?
4. Do you have chest pain?
5. Do you cough all day?
6. Do you have fever?
7. Do you sweat a lot at night?
8. Did you lose weight?
9. Do you eat normally / do you have an appetite?
10. Do you have trouble breathing?
11. Do people around you have been diagnosed with tuberculosis?
12. Do you smoke or are you exposed to smoke?
13. Do you have chills?
14. How do you feel in general?
15. Do other people in your family cough a lot or have asthma?
16. When were you last tested for HIV / do you know your HIV status?
17. Did you take any medication?

Physical examinations

1. Take temperature with thermometer
2. Measure weight
3. Take blood pressure
4. Take pulse
5. Take respiratory frequency

6. Use stethoscope at the front of your body (chest)
7. Use stethoscope at the back
8. Touch on neck and throat

A.5 Asthma

Case description

Opening statement: In the last couple of months I have had trouble breathing several times. Yesterday again, I could not breathe properly for about ten minutes.

Question	Response (if asked)
What were you doing when it happened?	I was cleaning the house.
More details on the symptoms? / How did it feel?	Suddenly I could not really breathe. It felt like I was suffocating.
How long did it last?	About ten minutes. I had a lot of trouble breathing normally. Afterwards it was better, but I felt quite tired.
Was there a whistling noise?	Yes.
Did you eat anything unusual?	No.
Did this happen before?	Yes, several times.
How many times did this happen?	I am not sure, maybe four or five times.
What were you doing then?	Twice after running for the bus and other times when I had a cold.
How many times does this happen in a month?	At the start about once or twice, now more often.
Did you also cough?	Yes, I also coughed.
Was the cough productive or dry?	It was dry. I was not coughing up anything.
Do you have chest pain?	No.
Did you have a fever?	No.
Did you have any other health problems?	I had a cold and a runny nose for almost two weeks, but now it is fine.
Did you have similar problems in your childhood?	My mother said that when I was younger I sometimes had trouble breathing.
When you were young, did you take anything for this problem?	Not really.

Question	Response (if asked)
Medical problems in your family?	No problems. Except maybe my mother, who often has colds and allergies.
Do you have asthma or other respiratory diseases in your family?	No.
Do you have any allergies?	No.
Do you smoke?	No.
Do you have a throat ache?	No.
Do you have a throat ache or a cold?	I just had a cold for about two weeks.

History questions and physical examinations

Questions

1. What were you doing when this happened? What caused the attack?
2. How long did breathing difficulty last?
3. Is breathing difficulty constant or episodic?
4. What triggers episodes (dust, pollution, bad air)?
5. Had you eaten anything unusual before the episode?
6. Did you take any medication?
7. Did this happen before?
8. Details on previous episodes
9. How often does this happen to you?
10. Did you also cough?
11. Was the cough dry or productive?
12. Did / do you have chest pain?
13. Did / do you have fever?
14. Did you lose weight / appetite?
15. Any other symptoms / problems?
16. Did you have similar problems breathing in your childhood?
17. Do you have cases of asthma or allergies in your family?
18. Do you have allergies?
19. Do you have a throat ache or cold?

Physical examinations

1. Take temperature with thermometer
2. Take blood pressure
3. Take pulse
4. Take respiratory frequency
5. Use stethoscope at the front of your body (chest)
6. Use stethoscope at the back

7. Throat examination
8. Measure weight

A.6 Upper respiratory tract infection (URTI)

Case description

Opening statement: Doctor, I have a flu.

Question	Response (if asked)
What are your symptoms?	I have a headache and a sore throat, and I have been coughing.
How long have you had these symptoms for?	Five days.
Do you have a fever?	No.
Do you cough?	Yes, sometimes.
Did you cough any sputum or mucus?	No.
Do you have a cold, such as a blocked or runny nose?	Yes, I feel stuffy.
How do you feel in general?	Tired and exhausted.
Have you taken any medication?	No.
Do your symptoms get worse at night or change through the day?	No, they are the same during the day and at night.
Is there blood in the sputum?	No.
Do you have chest pain?	No.
Have you lost your appetite?	No.
Do you have trouble breathing?	No.
Did you lose weight?	No.
Do you have night sweats?	No.
Do you smoke?	No.

History questions and physical examinations

Questions

1. How long have you been feeling unwell?
2. What symptoms are you experiencing?
3. Do you cough?
4. Is the cough productive (is there mucus)?
5. Did you take any medication?
6. Do / did you have fever?
7. Do / did you have chest pain?
8. Do you have difficulty breathing?
9. Do you have a cold (blocked or runny nose)?

Physical examinations

1. Take temperature with thermometer
2. Take blood pressure
3. Take pulse
4. Take respiratory frequency
5. Use stethoscope at the front of your body (chest)
6. Use stethoscope at the back
7. Throat examination

A.7 Family planning

Case description

Opening statement: I want to avoid pregnancy and I learned from my cousin that there is a pill that can help.

Question	Response (if asked)
How old are you?	Real age.
How old were you when you got married?	One year before having the first child, based on real age.
How old is your last child? When was your last birth?	About one year ago (13 months).
Are you breastfeeding at the moment?	No. I stopped when my daughter was six months old.
How many children do you have?	I have two children.
How many pregnancies have you had?	Two.
Did you have a miscarriage?	No.
Did you have any stillbirths?	No.
Have you had an abortion?	No.
Which method would you like to use?	My cousin is using the pill and says it is good. I would like to try that.
What pill do you want?	I do not know.
When was your last period?	It finished yesterday.
What is your period like?	My period is regular.
How many days does your period last?	Four or five days.
Since your last period, did you have intercourse?	No. The last time was before my period.
Are you currently using a contraceptive method?	Condoms.
Have you used a contraceptive method before?	Only condoms.
For how many years would you like to avoid pregnancy?	I am not sure, maybe two years.

Question	Response (if asked)
Do you have any health problems?	In general I am fine, but I often have very strong headaches.
Do you frequently have very strong headaches?	Yes.
Do you take medication for these headaches?	I sometimes take paracetamol, but it does not help.
Tell me more about these headaches.	They occur once every two to three weeks. They are very strong, usually on one side. Before they start, I often see zig-zag lines in front of my eyes. The pain is throbbing, and light and noise sometimes make it worse.
What do you do when the headaches start?	I have to lie down. Sometimes they last all day.
When do you usually have these headaches?	It happens often, sometimes before my period and sometimes after.
Since when have you had these headaches?	Since I was a teenager.
Do you smoke cigarettes?	No.
Have you ever had an operation?	No.
Do you have high blood pressure?	No.
Do you have diabetes?	No.
At what age did you get your first period?	Real age.
Do you drink alcohol?	No.
Have you ever been told that you have breast cancer?	No.
Have you ever had a stroke, blood clot, or heart attack?	No.
Do you know your HIV status?	Yes, it is negative.
When were you last tested for HIV?	Five months ago, when I came for medical care.
Do you have a fever?	No.
Do you have a cough?	No.
Is there a history of tuberculosis in your family?	No.
Do you have any other symptoms?	No.
Do you regularly take medication for tuberculosis, seizures, or ritonavir for antiretroviral therapy?	No.
Do you have gall bladder disease, liver disease, or jaundice?	No.
Have you been told that you have high blood pressure?	No.
Have you been told that you have diabetes?	No.
Have you been told that you have a rheumatic disease such as lupus?	No.

Question	Response (if asked)
Do you take any medication?	Only paracetamol for headaches.
Why do you not want more children at the moment?	I just want to take a break for now.
Do you have any medical problems in your family?	My sister was recently diagnosed with breast cancer. Otherwise, no.

History questions and physical examinations

Questions

1. Your age?
2. The age of your last child / date of your last birth?
3. How many pregnancies / children have you had?
4. Do you breastfeed currently?
5. Are you currently using a contraceptive method?
6. Date of your last period?
7. Questions about your period (frequency, number of days, heavy flow, etc.)
8. Any medical problems?
9. Do you frequently have very strong headaches?
10. More details on the headaches (describe pain, location of pain, when do you have headaches?)
11. Do you have a family history of breast cancer?
12. Have you ever had a blood clot?
13. Have you been told that you have high blood pressure?
14. Do you take any medication?
15. Do you smoke cigarettes?
16. Have you ever used a family planning method?
17. Which method would you prefer?
18. For how many years would you like not to have children?

Physical examinations

1. Pregnancy test
2. Measure height
3. Measure weight
4. Take blood pressure
5. Examine eyes
6. Examine breasts
7. Examine abdomen
8. Examine legs
9. Examine thyroid
10. Vaginal exam attempted

B Vignettes

Tuberculosis

A 35-year-old man comes to the facility because he has a cough that is not getting better. He reports that he has been coughing for three weeks. The cough is productive, with yellow sputum and no blood. He does not have chest pain or difficulty breathing but feels weak all the time. He reports fever at night, night sweats, and chills. He has lost weight and has reduced appetite. He does not smoke or drink alcohol and is a normal weight. He took cough syrup, which did not help, and reports no other symptoms.

Stable angina

A 50-year-old man/woman comes to the facility because he/she has chest pain. He/she describes the pain as crushing, in the middle of the chest, and moving towards the left and the shoulder. The pain occurred during physical activity and he/she had to stop what he/she was doing and sit down. He/she felt dizzy but did not vomit. The pain lasted about five minutes and went away with rest. He/she notes that this has happened three times before in similar circumstances, such as running for the bus or walking up stairs. He/she smokes, consumes alcohol frequently, and is overweight. He/she has not taken any medication and has no other symptoms.

Asthma

A 20-year-old woman comes to the facility because of breathing difficulty that has occurred several times. Yesterday, while cleaning the house, she could not breathe properly for about ten minutes. She also had a dry cough. This has happened several times before, after physical activity or when she had a cold, and has started to happen more and more often. She had similar problems as a child. She has no chest pain or fever. She does not smoke or drink alcohol, is a normal weight, has not taken any medication, and has no other symptoms.

Upper respiratory tract infection (URTI)

A 30-year-old man comes to the facility and says, "Doctor, I have a flu". For five days, he has had a sore throat and a dry cough that occurs sometimes. He has no fever and feels tired and exhausted. He has not taken any medication and has no other symptoms.

Family planning

A 27-year-old woman comes to the facility asking for a contraceptive method, noting that her cousin told her that there is a pill that can help. She has two children and has had no miscarriages or abortions. Her youngest child is one year old and she would like to take a two-year break from having children. She is not pregnant and is not breastfeeding. She

suffers from frequent headaches with throbbing pain on one side of the head, sometimes with vision changes, but has no other medical problems. She does not regularly take any medication except paracetamol. She is a normal weight and does not smoke or drink alcohol.

C Descriptives

Table C1: Characteristics of facilities and healthcare workers

	Mean	SD	Min	Max
Panel A: Facilities				
Public facility	0.37	0.49	0	1
Level two facility	0.63	0.49	0	1
Total staff (facility roster)	3.13	3.03	1	24
Distance to referral facility (km)	5.31	5.10	0	30
Equipment index	0.90	0.11	0.52	1.00
Running water	0.98	0.12	0	1
Open on weekends	0.68	0.47	0	1
Provider on normal day (manager estimate)	3.13	2.59	0	15
Consultation rooms	2.34	1.46	0	10
Observations		198		
Panel B: Providers				
Female	0.51	0.50	0	1
Doctor	0.07	0.26	0	1
Nurse	0.28	0.45	0	1
Clinical officer	0.61	0.49	0	1
Skilled provider	0.96	0.21	0	1
Experience in facility (years)	3.13	3.52	0	25
Correct vignette diagnosis	0.60	0.49	0	1
Any training in past year	0.26	0.27	0	1
WHO wellbeing score	18.24	3.94	5	25
Salary below 60,000 KES	0.50	0.50	0	1
Overtime (hours/month)	1.65	5.46	0	48
Reported patients seen per day	34.15	23.87	0	150
Clinical knowledge	0.57	0.20	0	1
Observations		407		

Notes: Panel A reports characteristics of all 198 study facilities. Level two facilities are slightly smaller and less well staffed than larger (level three) facilities. Panel B reports characteristics of providers who completed the provider survey. Overall, 670 providers saw standardised patients; provider-level characteristics are available for 407 of these providers. The equipment index captures the availability of essential medical equipment. Skilled providers are doctors, clinical officers, or nurses. The WHO wellbeing score ranges from 0 to 25, with higher values indicating better psychological wellbeing. Clinical knowledge refers to the proportion of vignettes corresponding to the SP cases (stable angina, asthma, tuberculosis, URTI and family planning) that were correctly diagnosed by providers.

D Workload

Table D1: Characteristics of facilities with and without excess-demand periods

Variable	No excess demand		Excess demand		Difference	
	Mean	SD	Mean	SD	b	p
Public facility	0.25	0.43	0.61	0.49	-0.36	0.00
Total staff (facility roster)	2.30	1.98	4.68	3.94	-2.38	0.00
Level two facility	0.68	0.47	0.52	0.50	0.16	0.03
Working phone	1.09	0.29	1.16	0.37	-0.07	0.17
Running water	0.99	0.09	0.97	0.17	0.02	0.25
Distance to referral facility (km)	5.03	4.71	5.83	5.77	-0.80	0.30
Provider on normal day (manager estimate)	3.03	2.40	3.32	2.92	-0.29	0.46
Consultation rooms	2.32	1.45	2.38	1.49	-0.06	0.79
Equipment index	0.90	0.11	0.90	0.12	0.00	0.84
Observations	129		69		198	

Notes: Facilities with excess demand are those where at least one SP consultation occurred during a period in which queue length exceeded the facility-specific average daily workload per provider ($Q_{ipf} > WL_f$). Level two facilities are smaller and less well staffed than larger level three facilities. We show total full-time staff working in the facility based on the facility roster as well as facility managers' estimate of the number of providers working in the facility on a normal day. The equipment index captures the availability of 27 items of essential equipment.

Table D2: Characteristics of healthcare workers in facilities with and without excess demand

Variable	No excess demand		Excess demand		Difference	
	Mean	SD	Mean	SD	<i>b</i>	<i>p</i>
Patients per day	32.62	22.64	39.18	27.04	-6.56	0.02
Salary below 60,000 KES	0.54	0.50	0.40	0.49	0.14	0.02
Skilled provider	0.95	0.21	0.97	0.18	-0.02	0.49
Doctor	0.08	0.28	0.03	0.18	0.05	0.09
Clinical officer	0.59	0.49	0.67	0.47	-0.08	0.14
Nurse	0.28	0.45	0.26	0.44	0.02	0.77
Experience (years)	2.98	3.25	3.59	4.25	-0.61	0.14
Female	0.49	0.50	0.56	0.50	-0.06	0.27
WHO wellbeing score	18.42	3.99	17.65	3.76	0.77	0.10
Training past year	0.25	0.28	0.27	0.24	-0.02	0.53
Training past five years	0.25	0.28	0.27	0.24	-0.02	0.53
Overtime (hours/month)	1.74	5.50	1.37	5.37	0.36	0.57
Clinical knowledge	0.57	0.50	0.54	0.50	0.03	0.60
Observations	312		95		407	

Notes: Providers exposed to excess demand are those who saw at least one SP case during a consultation in which queue length exceeded the facility-specific average daily workload per provider ($Q_{ipf} > WL_f$). Skilled providers are doctors, clinical officers, or nurses. Experience refers to years working in the facility. The WHO wellbeing score ranges from 0 to 20, with higher values indicating better psychological wellbeing. Clinical knowledge refers to the proportion of vignettes corresponding to the SP cases (stable angina, asthma, tuberculosis, URTI and family planning) that were correctly diagnosed by providers.

E Rationing through waiting

Table E1: Effect of workload on through waiting, by sector

	(1)	(2)
	Public sector	Private sector
	Waiting time	Waiting time
Above typical workload (binary)	24.985*** (9.311)	10.846** (5.452)
Outcome mean	74.721	16.979
Observations	481	845
R-squared	0.546	0.336

Note: Excess demand is a binary indicator equal to one if the provider is observed during a period in which queue length exceeds the facility-specific average daily workload per provider ($Q_{ipf} > WL_f$). The dependent variable is patient waiting time, measured in minutes. Column (1) reports estimates for public-sector facilities and column (2) for private-sector facilities. All specifications include the full set of controls used in the main analysis. Standard errors are shown in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

F Rationing along the intensive margin

F.1 Public and private-sector facilities

Table F1: Effect of workload on provider effort and quality of care by facility ownership

	(1)	(2)	(3)	(4)	(5)
	Duration	Checklist	Correct	Behaviours	Experience
Panel A: Private-sector for-profit facilities					
Above typical workload (binary)	1.771 (2.397)	0.024 (0.029)	0.040 (0.077)	-0.017 (0.031)	-0.128 (0.139)
Observations	845	845	845	845	845
R-squared	0.315	0.631	0.363	0.293	0.600
Panel B: Public facilities					
Above typical workload (binary)	0.957 (0.711)	0.034* (0.019)	-0.010 (0.087)	0.012 (0.023)	-0.093 (0.112)
Observations	481	481	481	481	481
R-squared	0.271	0.446	0.476	0.270	0.447

Note: The table reports estimates separately by facility ownership. To capture workload, we use a binary indicator for excess demand, equal to one for a given SP visit if queue length exceeds the facility-specific average daily workload per provider ($1\{Q_{ipf} > WL_f\}$). Duration is measured in minutes. Checklist captures the proportion of relevant history questions asked and physical examinations performed. Correct is a binary indicator for clinically correct case management. Behaviours and Experience are indices capturing procedural bedside manner and the SP's subjective experience, respectively. All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

F.2 Provider knowledge

Table F2: Effect of workload on provider effort and quality of care: among providers with sufficient knowledge

	(1) Duration	(2) Checklist	(3) Correct	(4) Behaviours	(5) Experience
Above typical workload (binary)	3.186* (1.850)	0.031 (0.031)	-0.020 (0.129)	-0.008 (0.034)	-0.090 (0.157)
Observations	557	557	557	557	557
R-squared	0.582	0.666	0.478	0.493	0.422

Note: The table shows the effect of workload on provider behaviour. The sample is restricted to providers who correctly diagnosed the same case in a vignette, indicating adequate clinical knowledge. Duration is measured in minutes. To capture workload, we use a binary indicator for excess demand, equal to one for a given SP visit if queue length exceeds the facility-specific average daily workload per provider ($1\{Q_{ipf} > WL_f\}$). Checklist captures the proportion of relevant history questions asked and physical examinations performed. Correct is a binary indicator for clinically correct case management during the standardised patient interaction. Behaviours and Experience are indices capturing procedural bedside manner and the SP’s subjective experience, respectively. All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

G Input rationing

Table G1: Effect of workload on fees paid and drug prescriptions (input rationing)

	(1) Fees	(2) Total drugs	(3) Any drugs
Above typical workload (binary)	-0.088 (0.059)	0.040 (0.169)	-0.059 (0.059)
Observations	1,326	1,140	1,140
R-squared	0.679	0.403	0.354

Note: The table reports estimates from regressions examining whether providers adjust input use during periods of excess demand. To capture workload, we use a binary indicator for excess demand, equal to one for a given SP visit if queue length exceeds the facility-specific average daily workload per provider ($1\{Q_{ipf} > WL_f\}$). Column (1) uses total fees charged, standardised to have mean zero and unit variance. Column (2) reports the total number of drugs prescribed, and column (3) indicates whether any drugs were prescribed. Columns (2) and (3) exclude family planning standardised patient cases, for which drug provision is not clinically indicated. All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table G2: Effect of workload on intensive-margin rationing: individual healthcare-worker fixed effects

	(1) Duration	(2) Checklist	(3) Correct	(4) Behaviours	(5) Experience
Above typical workload (binary)	1.931 (2.125)	-0.002 (0.033)	-0.057 (0.102)	-0.041 (0.044)	0.009 (0.171)
Observations	1,302	1,302	1,302	1,302	1,302
R-squared	0.649	0.778	0.653	0.637	0.752

Note: The table reports estimates from regressions of provider effort and quality outcomes on workload, using individual healthcare-worker fixed effects and exploiting within-provider variation across consultations. Healthcare workers saw between 1 and 8 SPs, with 3 consultations per provider on average. Workload is captured by a binary indicator for excess demand, equal to one if queue length exceeds the facility-specific average daily workload per provider ($1\{Q_{ipf} > WL_f\}$). Columns (1)–(2) measure provider effort: consultation duration (minutes) and checklist adherence (the proportion of relevant history questions asked and physical examinations performed). Columns (3)–(5) capture different dimensions of quality of care. Correct is a binary indicator for clinically correct case management (whether SPs received clinically indicated advice, referrals, or treatments). Behaviours captures the proportion of procedural actions consistent with good bedside manner (ensuring privacy, limiting distractions, explaining the diagnosis, treatment, or examinations). Experience is a standardised index capturing the SP’s subjective assessment of the consultation based on agreement with nine statements (e.g. “The provider was attentive” or “The provider was unkind”). All models control for individual healthcare worker, SP case, time of day, and day of the week. Standard errors clustered at the provider level are shown in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

H Robustness

H.1 Non-linear responses to workload

Table H1: Effect of workload on intensive-marging rationing: excess-demand quintiles

	(1) Duration	(2) Checklist	(3) Correct	(4) Behaviours	(5) Experience
2 nd quintile of excess demand	-0.055 (0.997)	0.024 (0.016)	-0.061 (0.061)	0.014 (0.022)	-0.047 (0.127)
3 rd quintile of excess demand	0.480 (0.816)	0.019 (0.014)	0.009 (0.044)	-0.002 (0.015)	-0.115* (0.069)
4 th quintile of excess demand	0.945 (0.816)	0.018 (0.015)	-0.017 (0.044)	0.021 (0.018)	-0.034 (0.078)
5 th quintile of excess demand	0.307 (1.022)	0.028* (0.017)	-0.015 (0.054)	0.019 (0.021)	-0.036 (0.082)
Observations	1,321	1,321	1,321	1,321	1,321
R-squared	0.370	0.586	0.374	0.346	0.543

Note: The table reports estimates from regressions of provider effort and quality outcomes on indicators for quintiles of excess-demand. The first quintile (lowest) is omitted and serves as the reference category. Excess demand (continuous) is measured as Q_{ipf}/WL_f and standardised. Consultation duration is measured in minutes. Checklist captures the proportion of relevant history questions asked and physical examinations performed. Correct is a binary indicator for clinically correct case management. Behaviours and Experience are indices capturing procedural bedside manner and the SP's subjective experience, respectively. All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table H2: Effect of workload on intensive margin rationing: excess-demand deciles

	(1) Duration	(2) Checklist	(3) Correct	(4) Behaviours	(5) Experience
4 th decile of excess demand	-0.075 (0.999)	0.021 (0.016)	-0.064 (0.061)	0.016 (0.022)	-0.054 (0.127)
5 th decile of excess demand	0.526 (0.938)	0.004 (0.016)	-0.013 (0.051)	0.008 (0.019)	-0.153* (0.089)
6 th decile of excess demand	0.367 (1.061)	0.032* (0.017)	0.027 (0.057)	-0.010 (0.016)	-0.087 (0.092)
7 th decile of excess demand	1.448 (1.194)	0.008 (0.018)	-0.006 (0.054)	0.020 (0.021)	-0.022 (0.090)
8 th decile of excess demand	0.410 (0.864)	0.030* (0.018)	-0.026 (0.051)	0.022 (0.021)	-0.044 (0.098)
9 th decile of excess demand	-0.575 (1.108)	0.029 (0.019)	-0.032 (0.060)	0.030 (0.023)	-0.084 (0.087)
10 th decile of excess demand	1.991 (1.281)	0.033 (0.022)	0.024 (0.066)	-0.005 (0.028)	0.069 (0.123)
Observations	1,321	1,321	1,321	1,321	1,321
R-squared	0.373	0.588	0.375	0.347	0.544

Note: The table reports estimates from regressions of provider effort and quality of care on deciles of excess demand. The first decile (lowest) of Q_{ipf}/WL_f is omitted and serves as the reference category. Because the excess demand measure exhibits substantial bunching at zero (no patients waiting), the lower tail of the distribution does not meaningfully split into separate 2nd and 3rd deciles; we therefore report results starting from the 4th decile. Duration is measured in minutes. Checklist captures the proportion of relevant history questions asked and physical examinations performed. Correct is a binary indicator for clinically correct case management. Behaviours and Experience are indices capturing procedural bedside manner and the SP's subjective experience, respectively. All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table H3: Effect of workload on extensive-margin rationing: excess-demand quintiles

	(1) Skilled provider	(2) High-skilled	(3) Nurse	(4) Doctor	(5) Clinical officer
2 nd quintile of excess demand	-0.008 (0.012)	-0.049 (0.050)	0.042 (0.048)	-0.005 (0.036)	-0.044 (0.055)
3 rd quintile of excess demand	-0.021 (0.020)	-0.015 (0.043)	-0.006 (0.037)	0.001 (0.026)	-0.016 (0.046)
4 th quintile of excess demand	-0.012 (0.023)	-0.022 (0.043)	0.010 (0.036)	0.007 (0.024)	-0.029 (0.042)
5 th quintile of excess demand	-0.040* (0.023)	0.024 (0.051)	-0.064 (0.048)	0.018 (0.020)	0.006 (0.050)
Observations	890	890	890	890	890
R-squared	0.610	0.663	0.670	0.806	0.696

Note: The table reports estimates from regressions of extensive-margin rationing outcomes on indicators for quintiles of excess demand. The first quintile (lowest excess demand) is omitted and serves as the reference category. Excess demand (continuous) is measured as Q_{ipf}/WL_f and standardised. Outcomes are indicator variables capturing the type of provider seen by the SP. Skilled provider equals one if the SP was seen by a nurse, clinical officer, or doctor. High-skilled equals one if the SP was seen by a clinical officer or doctor. The remaining columns indicate whether the SP was seen by a nurse, doctor, or clinical officer, respectively. All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. The analysis uses data from providers who saw SPs and completed the provider survey (187 facilities). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table H4: Effect of workload on extensive-margin rationing: excess-demand deciles

	(1)	(2)	(3)	(4)	(5)
	Skilled provider	High-skilled	Nurse	Doctor	Clinical officer
4 th decile of excess demand	-0.009 (0.013)	-0.049 (0.050)	0.040 (0.048)	-0.008 (0.036)	-0.041 (0.056)
5 th decile of excess demand	-0.028 (0.026)	-0.015 (0.054)	-0.013 (0.045)	-0.013 (0.028)	-0.003 (0.053)
6 th decile of excess demand	-0.014 (0.016)	-0.014 (0.049)	0.000 (0.046)	0.013 (0.033)	-0.028 (0.059)
7 th decile of excess demand	-0.009 (0.023)	-0.025 (0.050)	0.016 (0.045)	-0.003 (0.033)	-0.022 (0.049)
8 th decile of excess demand	-0.015 (0.026)	-0.020 (0.053)	0.005 (0.044)	0.020 (0.021)	-0.039 (0.054)
9 th decile of excess demand	-0.052** (0.026)	0.029 (0.051)	-0.081* (0.048)	0.021 (0.021)	0.008 (0.051)
10 th decile of excess demand	-0.011 (0.028)	0.014 (0.065)	-0.025 (0.060)	0.019 (0.020)	-0.005 (0.066)
Observations	890	890	890	890	890
R-squared	0.612	0.663	0.670	0.806	0.696

Note: The table reports estimates from regressions of extensive-margin rationing outcomes on indicators for deciles of excess demand. The first three deciles (lowest excess demand) are omitted and serve as the reference category. Excess demand (continuous) is measured as Q_{ipf}/WL_f and standardised. Outcomes are indicator variables capturing the type of provider seen by the SP. Skilled provider equals one if the SP was seen by a nurse, clinical officer, or doctor. High-skilled equals one if the SP was seen by a clinical officer or doctor. The remaining columns indicate whether the SP was seen by a nurse, doctor, or clinical officer, respectively. All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. The analysis uses data from providers who saw SPs and completed the provider survey (187 facilities). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

H.2 Small facilities

Table H5: Effect of workload on provider effort and intensive margin rationing: Small facilities

	(1) Duration	(2) Checklist	(3) Correct	(4) Behaviours	(5) Experience
Above typical workload (binary)	1.953 (1.964)	0.021 (0.027)	0.100 (0.083)	0.011 (0.022)	-0.099 (0.120)
Observations	831	831	831	831	831
R-squared	0.358	0.569	0.382	0.327	0.564

Note: The table shows the effect of workload on provider behaviour. The sample is restricted to small (Level 2) primary care facilities. To capture workload, we use a binary indicator for excess demand, equal to one for a given SP visit if queue length exceeds the facility-specific average daily workload per provider ($1\{Q_{ipf} > WL_f\}$). Duration is measured in minutes. Checklist captures the proportion of relevant history questions asked and physical examinations performed. Correct is a binary indicator for clinically correct case management. Behaviours and Experience are indices capturing procedural bedside manner and the SP's subjective experience, respectively. All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table H6: Effect of workload on extensive-margin rationing: Small facilities

	(1) Skilled provider	(2) High-skilled	(3) Nurse	(4) Doctor	(5) Clinical officer
Above typical workload (binary)	0.029 (0.034)	0.028 (0.069)	0.002 (0.078)	0.018 (0.030)	0.010 (0.072)
Observations	551	551	551	551	551
R-squared	0.697	0.723	0.719	0.811	0.728

Note: The table shows the effect of workload on extensive-margin rationing. The sample is restricted to small (Level 2) primary care facilities. To capture workload, we use a binary indicator for excess demand, equal to one for a given SP visit if queue length exceeds the facility-specific average daily workload per provider ($1\{Q_{ipf} > WL_f\}$). Outcomes are indicator variables capturing the type of provider seen by the SP. Skilled provider equals one if the SP was seen by a nurse, clinical officer, or doctor. High-skilled equals one if the SP was seen by a clinical officer or doctor. The remaining columns indicate whether the SP was seen by a nurse, doctor, or clinical officer, respectively. All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

I Mechanisms

Table I1: Rationing along the extensive margin at very low workload

	(1)	(2)	(3)	(4)	(5)
	Skilled provider	High-skilled	Nurse	Doctor	Clinical officer
Panel A: No patients waiting					
No patients waiting	0.021 (0.017)	0.011 (0.035)	0.009 (0.030)	-0.007 (0.019)	0.018 (0.036)
Observations	892	892	892	892	892
R-squared	0.606	0.662	0.668	0.805	0.695
Panel B: Lowest decile of excess-demand					
Lowest decile of excess-demand	0.021 (0.017)	0.011 (0.035)	0.009 (0.030)	-0.007 (0.019)	0.018 (0.036)
Observations	892	892	892	892	892
R-squared	0.606	0.662	0.668	0.805	0.695

Note: This table examines whether provider assignment varies at very low levels of workload. Panel A uses an indicator for zero queue length at arrival, defined as $1\{Q_{ipf} = 0\}$. Panel B uses an indicator for whether excess demand falls in the lowest decile of its standardised distribution, $1\{Q_{ipf}/WL_f \leq P_{10}\}$. Outcomes are indicator variables capturing the type of provider seen by the SP. Skilled provider equals one if the SP was seen by a nurse, clinical officer, or doctor. High-skilled equals one if the SP was seen by a clinical officer or doctor. All models include facility fixed effects and control for the SP case, time of day, and day of the week. Standard errors clustered at the facility level are shown in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.