The Price of Pace: Paying Experts by Volume

Amanda Dahlstrand[1] and Shan Huang[2]

[1] *University of Zurich,* [2] *Stockholm School of Economics*

November 22, 2026

Click here for the most recent version

Abstract

Quantity incentives can expand capacity when expert time is scarce, but multitasking theory predicts reduced effort in unmeasured parts of a task. We study individual physicians' staggered switches from fixed hourly wages to quantity incentives within a large Swedish primary care organization. Because patients are assigned to whichever physician is available, we compare similar patients treated by the same physician before and after the switch. We link over 450,000 online consultations to patients' subsequent care across the healthcare system to measure externalities outside the organization. We separate time with the patient, captured well by the organization's internal quality measures, from administrative time, such as for record review, whose implications are captured less well. Quantity incentives shorten consultations from 11.7 to 9.3 minutes on average, primarily through administrative time savings, and raise consultations per physician by 26% without additional hours worked. Internal quality measures are unchanged, and most costly downstream care, including in-person primary care visits, does not increase within seven days; the exception is hospitalizations, rising from 3.8 to 4.9 per 1,000 consultations. Added costs for the public payer are modest and offset by the capacity expansion if extra consultations are valued at public-payer prices, unless additional hospitalizations imply large health losses beyond their monetary costs.

Keywords: quantity incentives, multitasking, time use, primary care physicians, externalities
JEL codes: I11, J22, J33, M52

We thank colleagues and seminar and conference participants at the workshop "Effectiveness of Digital and AI Solutions in Social and Health Care" of the Finnish Institute for Health and Welfare, Aalto University and the University of Turku; the Centre for Health Economic Research (HEFUU) Workshop in Uppsala; RFBerlin; the University of Oslo; and the University of Gothenburg for their insightful comments and feedback. We are grateful to the healthcare company, Kry, for giving us access to their data. All remaining errors are our own. We thankfully acknowledge financial support from Handelsbankens Forskningsstiftelser (Program grant P25-0261). The research project was approved by the Stockholm Ethics Review Board and the Swedish Ethics Review Authority.

1 Introduction

A worker paid per hour earns no more by working faster. A worker paid per client does, as time saved on one client serves another. When expert time is scarce, paying experts by volume can expand capacity without hiring more staff by pricing that time. Yet, such gains may come at a cost: Holmstrom and Milgrom (1991)'s seminal multitasking model shows that rewarding measured dimensions of a task draws effort, attention, or time from unmeasured dimensions. An organization can easily count the patients a physician sees or the students passing the final exams, but not how well the physician reads the record or how much the teacher builds understanding beyond the test. In expert work, these unmeasured dimensions are often consequential, and cutting inputs on them can impose costs on third parties that neither the organization nor the expert internalizes. Quantity incentives are widespread in expert services, yet there is little evidence on how much capacity they create or what costs they impose on third parties.

This tension is particularly pressing in healthcare, where expert shortages are growing as populations age (Association of American Medical Colleges 2024; WHO Regional Office for Europe 2025). Physicians whose pay depends less on their own output see fewer patients (Gaynor and Gertler 1995; Dumont et al. 2008). Access may then be rationed.[1] Quantity incentives could increase access but divert the physician's inputs from unmeasured aspects of care. How to resolve this trade-off is unsettled: in some countries, such as Sweden and Finland, primary care physicians receive fixed wages; in others, such as Germany, France, and Belgium, they are paid by quantity (Kroneman et al. 2013); and in others yet, such as the United States, several arrangements with combinations of salary, volume, and bonus components coexist and vary across practices (Igwe 2026). The same trade-off shapes policy debates over consultation length, where the benefits of more time per patient have to be weighted against reduced access when physician time is scarce.[2]

The contract generating this trade-off is between an organization and its workers. Evaluating it poses three challenges. First, isolating workers' responses to contract changes on comparable tasks is difficult. Healthcare studies, typically using payment reforms, show that financial incentives affect care (Clemens and Gottlieb 2014; Mullen, Frank, and Rosenthal 2010; Échevin and Fortin 2014). However, these reforms alter contracts between payers and healthcare organizations, which may adjust hiring, task assignment, or workers' often unobserved contracts. For example, contracted private ambulance firms cut costs through staffing adjustments, raising mortality (Knutsson and Tyrefors 2022). Even when physicians' own contracts with the organization change, tasks may also adjust as physicians select different types of patients (Alexander 2020). Randomized trials using standardized patients to compare the same doctors under different contracts come closest to isolating

[1] In Sweden, the setting of our study, where individual physicians are almost never paid by quantity, access is rationed by waiting: only 37% of patients report receiving care within two days when ill, the lowest share among comparable countries (Myndigheten för vård- och omsorgsanalys 2021).

[2] For example, The Netherlands has introduced a program supporting longer primary care consultations, and the British Medical Association advocates longer appointments but notes that this may reduce the number of appointments. See the Dutch program (BR/REG-25136) and the British Medical Association's recommendation. Last accessed: 22 October 2026.

incentive effects (Das et al. 2016), but cannot measure actual health outcomes. Observed changes thus reflect organizational adjustments and sorting in addition to workers' responses to incentives. Second, inputs in the multitasking model are rarely observed. Even with time-use data, information within tasks is needed to reveal whether quantity incentives cut time from a task's measured or unmeasured dimensions. Third, consequences may fall outside the contracting organization. For example, a primary care organization may neither pay for nor observe a patient's subsequent hospital care. It can only contract on what it observes well: the quantity of work and its own quality metrics. Studies limited to internal outcomes may therefore miss consequences predicted by the multitasking model. Evaluating volume pay requires holding prices paid to the organization fixed while changing only the worker's contract; comparing the same worker on comparable tasks with and without quantity incentives; and observing time reallocation within a task and costs outside the organization.

We overcome these challenges using more than 450,000 unscheduled video consultations at a large Swedish primary care organization that moved physicians from fixed hourly wages to a quantity-incentives scheme at different times between June 2022 and June 2023. The new scheme kept the hourly wage as a floor and added a piece rate of 20% of the base wage for each additional video consultation per hour above a monthly average of five. First, we isolate physicians' responses to their own contracts; the public payer's contract with the organization remained unchanged. Patients do not know how physicians are paid and are assigned to available physicians who must handle all case types, limiting sorting within narrow time windows. We use an estimator suitable for staggered treatment adoption (Gardner et al. 2026). Supporting our identifying assumptions, we find no differential pre-trends or imbalance in patient characteristics across contracts. Second, we use timestamps to observe how physicians divide their time within and across consultations, separating time spent with the patient from administrative time for that patient (such as record review, documentation, prescribing, and referrals). Third, we link consultations to national registers of prescriptions, specialist, emergency, and hospital care, mortality, and to records of all other primary care for adults in two large regions (about 40% of consultations). This allows us to trace the consequences of physicians' time reallocation outside the organization.

Our conceptual framework builds on models of incentive pay in which an agent chooses a privately costly input (Holmstrom and Milgrom 1991; Baker 1992; Lazear 2000). In our setting, the input is consultation time, divided between patient-facing and administrative work. Time carries an opportunity cost of forgone care for other patients, which corresponds to the cost of effort or attention. Under fixed wages, a physician extends a consultation until the marginal health benefit equals the value of displaced care. Consultation times vary with physicians' relative weights on patient health and the organization's internal quality measures. Quantity incentives add forgone income from completing more consultations to the opportunity cost of time and align the physician's incentives with the organization's, which is paid per consultation. The remaining agency problem is between the physician and the public payer, who bears downstream costs but can verify only a minimum quality standard. Only physicians whose pace under fixed wages is close to the piece-rate threshold adopt quantity incentives and shorten consultations. Because internal quality measures

respond to patient-facing time but not administrative time at the margin, physicians mainly cut administrative time. These cuts can leave measured quality unchanged while worsening patient health, which the organization does not fully measure. Any resulting increase in downstream care costs creates an externality, borne by the public payer and unobserved to the organization.

Within the organization, we find that quantity incentives shorten consultations by 2.4 minutes (21%) from a baseline mean of 11.7 minutes. This reduction is persistent and driven by a 2.1-minute (28%) drop in administrative time compared to a 0.3-minute (7.5%) drop in patient-facing time. The organization's internal quality – measured by patient ratings (to capture satisfaction), along with diagnostic informativeness and counter-guideline antibiotic prescribing (our proxies to capture care process quality) – remains unchanged. Among outcomes not used to evaluate physicians, unscheduled and booked return visits are unaffected, while the share of consultations with a prescription filled from the organization rises by 1.1 percentage points from 51%. From the organization's perspective, the scheme works as intended: physicians work faster, without affecting the quality it measures.

Outside the organization, quantity incentives leave primary care consultations with other organizations, emergency department visits, and specialist care unchanged. Within seven days, urgent care visits rise by 0.07 percentage points (18%) from 0.39%, and hospitalizations by 0.11 percentage points (29%) from 0.38%. The hospitalization increase is not followed by significantly higher one-year mortality. Valuing additional care at the public payer's costs, we estimate a modest increase of 24.9 SEK per consultation (3.3%) in downstream costs for the public payer, almost entirely from additional hospitalizations.

Reductions in administrative time account for nearly 90% of time savings. They amount to about two minutes across the distribution of consultations, regardless of patients' age or prior emergency and hospital care, while time spent with patients changes little. Internal quality measures remain unchanged because they capture administrative work only up to a minimum standard, which physicians exceed under either contract. As our framework predicts, however, these cuts can lower patient health in ways that surface only as rare adverse events elsewhere in the healthcare system. Cost increases follow consultations for acute infections, allergy or asthma symptoms, and requests naming no specific symptom (most often diagnosed as respiratory infections). They do not follow consultations for skin conditions, sexual and reproductive health, or prescription renewals. The most common diagnoses at subsequent hospitalizations are infections that can escalate quickly from common symptoms, such as pneumonia, kidney infections, and peritonsillar abscesses. In such cases, the administrative work being cut – including reviewing patients' pre-consultation symptom reports, comorbidities, or previous prescriptions, and arranging treatment or follow-up care – could otherwise prevent a condition from deteriorating.

Responses to quantity incentives also vary with physicians' initial pace. Voluntary adopters were already faster under fixed wages than non-voluntary adopters, whose average consultation time shows no break at adoption. This pattern follows from the presence of a wage floor, as physicians change their time inputs only if exceeding the piece-rate threshold is worthwhile.[3] Our main estimates

[3]The wage floor allows physicians to keep their pace without losing earnings. For simplicity, the model assumes

identify treatment effects on the treated for voluntary adopters (65% of physicians); given the model implications, the average treatment effects across all physicians are two-thirds this size. Among adopters, initially slower physicians cut consultation time most, to just about 12 minutes (the threshold when ignoring breaks), reducing dispersion in administrative time. Yet downstream costs rise more, if anything, among initially fast physicians cutting administrative time from a low baseline, consistent with diminishing returns. A mandate with a wage floor thus need not speed up all physicians, and further time savings may be costliest among those already working quickly.

Quantity incentives expand capacity without increasing hours worked, adding 39 video consultations per physician per month (26%). These capacity gains draw from existing physicians and work hours, without the need to train new physicians, and are valuable in a system that rations primary care by waiting. In Sweden, fewer than two in five patients are seen within two days when ill, and the national guarantee of a medical assessment within three days is routinely missed (Myndigheten för vård- och omsorgsanalys 2021). Paying by volume improves welfare if the value of additional consultations exceeds health losses and added downstream costs on baseline consultations – those that would have also occurred under fixed wages. We compute the health loss per baseline consultation at which the quantity-incentives scheme breaks even, given our estimates of downstream costs and capacity gains. If additional consultations are worth what the public payer pays for them, quantity incentives increase welfare unless health losses exceed 105 SEK (USD 11) per baseline consultation. If health losses come only from the additional hospitalizations we observe, they must amount to about 95,000 SEK (USD 9,400) per additional hospitalization, above and beyond the public payer's spending on the hospital stay already accounted for, for welfare to fall. At 1,000,000 SEK (USD 99,000) per quality-adjusted life year (QALY), close to the USD 100,000 per QALY that Finkelstein, Hendren, and Luttmer (2019) use to monetize health changes for the general U.S. population, this threshold corresponds to about five weeks in full health.

This paper contributes to research in personnel economics on incentive pay and multitasking. Theory has long recognized multitasking problems in complex tasks (Holmstrom and Milgrom 1991). Much of the empirical evidence for multitasking concerns rewards for measured quality, such as pay-for-performance schemes for teachers (Lavy 2002; Glewwe, Ilias, and Kremer 2010; Muralidharan and Sundararaman 2011). Instead, evidence for quantity incentives comes largely from settings where output is simple and contractible (Lazear 2000; Shearer 2004). We test the implications of multitasking with quantity incentives in complex tasks by directly observing experts' allocation of inputs between rewarded and unmeasured dimensions within tasks and tracing downstream costs from the unmeasured margin. We also resolve the recurring challenge that measured responses to an incentive contract partly reflect sorting (Biasi 2021; Leaver et al. 2021; Brown and Andrabi 2023). By comparing the same physicians under both contracts while they treat patients they cannot choose, we can separate their incentive response from sorting into contracts and tasks. The response we identify – how experts reallocate time within a task when that time acquires a price – also informs contract

that indifferent physicians do not adopt voluntarily. Alternatively, a small adoption cost, such as due to inattention, could produce the same behavior.

design for professions such as lawyers, teachers, or auditors. In these settings, multitasking can likewise impose costs that surface after the task ends and fall outside the contracting organization.

We also contribute to research on externalities from organizational decisions and payment structures in healthcare. An organizational change in one part of the system can shift care and costs elsewhere (Chandra, Gruber, and McKnight 2010; Dahlstrand et al. 2026). We identify a source of these spillovers in the contract between an organization and its workers. Earlier physician-payment studies examine changes in organizational reimbursement (Clemens and Gottlieb 2014; Mullen, Frank, and Rosenthal 2010), contracts that physicians select into and that change which tasks are paid (Dumont et al. 2008; Échevin and Fortin 2014), physicians moving between organizations with different payment schemes (Brekke et al. 2020), or patient sorting across provider contracts (Alexander 2020). Related work on incentive contract design examines performance pay tied to quality measures or clinical inputs, without tracing cost externalities elsewhere in the healthcare system (Mohanan et al. 2021; Rodríguez-Lesmes and Vera-Hernandez 2025). We isolate quantity incentive changes in an individual physician's contract within one organization, holding organizational reimbursement fixed, and price patients' care across the full healthcare system. This links an individual physician's contract design to the welfare consequences of that contract.

Finally, we contribute to research on how workers use their time and what it means for output quantity and quality. Economists have treated time as a scarce input since at least Becker (1965), but only recently have studies measured how workers allocate it across activities or tasks (Bandiera et al. 2020; Coviello, Ichino, and Persico 2015; Dahlstrand et al. 2025; Deming et al. 2026). Our analysis of physician consultations shows why time allocation *within* tasks also matters: Time saved on one patient can be used to treat another, but the consequences for quality depend on what parts physicians shorten, and this choice is shaped by how closely the organization measures each part. In primary care, evidence on the causes and consequences of shorter consultations is largely associational (Reif et al. 2025), or coupled with other organizational changes such as the mode being online or in-person (Dahlstrand, Le Nestour, and Michaels 2024). For dialysis treatment, treating more patients by spending less time on each has been shown to raise infection rates (Grieco and Mcdevitt 2017). In emergency departments, physicians' pace responds to peers (Silver 2021), and as the marginal cost of time rises near shift ends, physicians substitute tests and admissions for time, raising hospital costs (Chan 2018). In such urgent settings, faster care through shorter waiting time targets can reduce mortality (Gruber, Hoe, and Stoye 2023). We study how a price on physician time, set through individual pay incentives rather than institutional targets or schedules, changes time allocation within and across consultations. By measuring both direct capacity gains and downstream costs, we assess the welfare trade-off in physicians' time use in primary care.

The remainder of the paper is organized as follows. Section 2 describes the setting and the quantity-incentives scheme. Section 3 sets up the conceptual framework. Section 4 presents the data. Section 5 discusses the empirical framework. Section 6 reports the estimation results. Section 7 derives the welfare implications of the quantity-incentives scheme. Finally, Section 8 concludes.

2 Setting

2.1 Primary care in Sweden

Sweden's publicly funded healthcare system provides comprehensive universal coverage. The system is largely tax-financed, and patient cost-sharing is low.

Primary care is the main entry point into the health care system – specialist care (except urgent care) requires referral from primary care physicians. Primary care clinicians diagnose and treat common conditions and determine whether patients need downstream care. In Sweden, both private for-profit and public provider organizations deliver primary care under regulated contracts with the regional authorities, with private providers accounting for roughly 40% of primary care centers (The Commonwealth Fund 2020). Provider organizations are reimbursed mostly by capitation for patients registered ('listed') with them and fee-for-service for 'unlisted' patients. Provider organizations can treat both listed and unlisted patients online. Both online and in-person primary care consultations are reimbursed under national health insurance. Online consultations are common in Sweden, accounting for about a quarter of doctor consultations in 2023, and doctors can prescribe medication online (OECD 2025; OECD 2021). The share and scope of online care are similar in many other OECD countries, such as the U.S. (Centers for Medicare and Medicaid Services (CMS) 2025).

Primary care organizations bear little if anything of the cost of the downstream care their patients use: the public payer (the region) pays for specialist, emergency, and hospital care directly.[4]

Virtually all medical records in Sweden are digital. Although they are hosted on different IT systems, a National Patient Overview system for medical records (Nationell patientöversikt) and a National Medication List (Nationell läkemedelslista) for prescription medications collected by each patient exist (Ludvigsson et al. 2025). Information sharing is restricted by privacy law, so that individual doctors can log into and see patients' medical records from other providers when they have a direct need for it to provide care, and when the patient has consented to it, which patients usually do before their consultation. However, the primary care organization as a whole cannot retrieve records about what happens to patients in downstream care after their visit with a physician at the primary care provider. Hence, the organization cannot analyze average post-primary care ED visits or hospitalizations before and after any organizational change.

Physicians in Sweden are not paid directly by the public health insurance and instead predominantly receive *fixed* salaries from the provider organization they work for. Quantity incentives at the individual level are rare for physicians. Similar fixed salaries for physicians are the norm in several other healthcare systems, such as Spain, Italy, Portugal, the U.S. Veterans Affairs system, and hospital physicians in the UK and Norway. In other settings, by contrast, quantity incentives for individual physicians are common, most notably fee-for-service payments made directly to physicians in U.S. Medicare, Germany, Switzerland, and Denmark.

[4] In some regions, the capitation contracts for listed patients include some penalties for organizations whose patients use, e.g., many ED visits, but this amounts to far from the full cost, and primary care provider organizations never face a penalty for downstream care that unlisted patients use.

2.2 The primary care organization

The organization we study is a large private primary care provider operating nationwide (and in several other countries). It employs hundreds of clinicians across Sweden, comparable in scale to large group practices in the U.S. It runs many in-person clinics for registered patients, but also provides nationwide online primary care to any Swedish citizen. Consultations are covered by national health insurance. Online primary care visits carried a copay of 100–250 SEK (about USD 10–25) for adults and were free for minors and the elderly; these payments counted toward the annual copay cap.[5]

To receive online care, patients request consultations through an app after logging in via an electronic identification service (BankID) used for all interactions with banking and public authorities. Before assigning a doctor, the system retrieves basic demographics (gender, age, and region) from civil records, and patients select their main reason for seeking care (*symptom*) from a predefined list. Patients may also provide additional information, such as symptom duration, current medications, and other details. Consultations can be unbooked (drop-ins) or booked, but the vast majority of consultations are drop-ins. A few minutes before a booked consultation, clinicians are blocked from taking on other consultations. Physicians cannot choose whether to take booked consultations.

Importantly, the organization requires that all physicians handle all incoming consultations, with few exceptions such as consultations with infants, which we exclude from the analysis. Physicians cannot select or change their case mix over time, and patients are assigned to physicians without regard to or knowledge of the physician's compensation. There was no public knowledge of the change in the physician compensation system at the organization during the sample period. This became public through news reports only in 2026.

The organization employs physicians as employees or as consultants, which is a choice by the physician and has no impact on their compensation apart from the method in which it is taxed. For simplicity, we refer to compensation under both arrangements as a 'wage'. Physicians can work both online and in person for the primary care organization. However, as the quantity incentives only applied to the online work, we focus in the study on physicians who worked mostly in the online mode for the organization (90% or more of their consultations before the quantity-incentives scheme was rolled out). Almost all of them work part-time for the organization and earn their main income elsewhere, usually at an in-person job, e.g., at a hospital or clinic.

Staffed hours include breaks between consultations. A consultation ends when the physician signs it off. Each consultation consists of patient-facing time in the video meeting and administrative time, which is used to prepare the meeting, review and write the patient's medical records and clinical notes, and write prescriptions, referrals, and care instructions for the patient.

Performance and quality measurement Throughout our study period, before and after the rollout, the organization measures physicians' work mainly on three dimensions, which differ in how continuously they are observed. First, the organization records the number of consultations per

[5] During our sample period, annual out-of-pocket spending for outpatient visits was capped at roughly 1,150 SEK, or about USD 115 (1 SEK = 0.099 USD in 2022, Sveriges Riksbank 2022).

staffed hour for each physician each month, which forms the basis of pay under the quantity-incentives scheme. Second, patients can rate each consultation on a scale from 1 to 5 directly after it ends. The organization collects ratings continuously and provides them to the physician in weekly and monthly reports. About 40% of consultations in our sample have patient ratings. Because patients rate their interaction with the physician, these ratings reflect what happens during the video call and are less sensitive to what the physician does during administrative time. Third, the organization audits a sample of medical records and prescriptions to ensure that documentation is complete and that prescribing follows clinical guidelines. These audits verify that a minimum standard is met, but do not distinguish a physician who barely meets it from one who spends more time thoroughly reviewing a patient's record. Physicians also do not receive regular updates on these measures; they are notified only if quality is consistently below the bar. The organization emphasizes patient ratings and audits as its quality measures in internal and public communication. It also records outcomes such as prescription rates and return visits, but does not use them to evaluate physicians, and it does not observe the care patients later receive from other providers.

Apart from consultations per hour in the quantity-incentives scheme, none of these measures is directly tied to pay. Clinician managers nevertheless consider consultations per hour, patient ratings, audit results, and working hours in performance evaluations, for example in annual wage negotiations. Physicians receive weekly and monthly summaries of their consultations, staffed hours, consultations per staffed hour, average consultation time, and patient ratings. Evaluation and feedback did not change during our study period, so the only change in how a physician's work was measured or rewarded was the introduction of the piece rate.

2.3 Physician compensation at the organization

The baseline scheme. Under the baseline scheme, physicians receive a fixed hourly wage that does not depend on the number of consultations they conduct. Baseline wages vary across physicians with specialization, seniority, recruitment timing, and individual negotiation based on the metrics mentioned above.

The quantity-incentives scheme. Between June 2022 and June 2023, the organization staggered the rollout of an individual volume incentive scheme for physicians working in its digital services. The scheme adds a monthly bonus to the guaranteed hourly wage. Physicians were informed that they would earn the bonus whenever their monthly average exceeded five consultations per staffed hour, with each additional video consultation per hour raising hourly pay by 20% of their base wage.

For video consultations during regular working hours, the scheme is a piece rate above a threshold pace, combined with a guaranteed hourly wage. Let $\bar{W}$ denote a physician's fixed hourly base wage and n her monthly pace, measured as video consultations per staffed hour. Her pay per staffed hour is

$$\bar{W} + p\,(n-5)\,\mathbf{1}\{n \geq 5\}, \tag{1}$$

where $p = \bar{W}/5$ is the piece rate per video consultation above the threshold. At or below five

consultations per hour, the physician receives her base wage. Above the threshold, each additional consultation per staffed hour raises hourly pay by p, or 20% of the base wage (Figure 1a). Both pace and the bonus are calculated over the calendar month.

Equation (1) abstracts from two features of the scheme. First, physicians may also be assigned text-based consultations, each of which counts as half a video consultation when calculating the physician's monthly pace and bonus.. We abstract from them because they on average account for only 2.4% of physicians' monthly consultations during the pre-rollout period (January 2022 – May 2022). Moreover, as we will show in Section 7, the quantity-incentives scheme does not change the number of text-based consultations. Second, overtime and night-time pay multipliers exist. As they apply in the same way to the guaranteed hourly wage and the piece rate, we omit them from the simplified formula. We address potential selection into working hours directly in our main specifications and show in Section 5.3 similar estimates of total consultation time at the physician-month level, suggesting that such selection plays a limited role. Appendix A.1 gives the full payment formula.

Physicians had information about their pace under the baseline scheme, but with no payment tied to it. The new quantity-incentives scheme introduced an explicit financial return to consultation volume, without changing the existing productivity feedback.

Timeline. The quantity-incentives scheme was rolled out in two main stages. First, in June 2022, the organization introduced the quantity-incentives scheme for voluntary participation, which around two-thirds of clinicians eventually ended up taking up. Clinician managers typically communicated the new scheme to their physicians and explained that physicians would never be worse off due to the scheme's wage floor. According to firm representatives, voluntary adopters often included physicians who had previously raised concerns about their compensation. Second, in June 2023, the scheme became mandatory for all physicians, meaning the remaining one-third of physicians were moved onto it. Firm representatives report that physicians generally moved onto the new scheme on the first day of a calendar month, so we treat the first day in the month when a physician was first observed with a quantity incentive as the onset of treatment. Figure 1b shows the resulting distribution of physicians across starting months.

The period we study is largely unaffected by two later changes in the working environment. First, in summer 2024 and thus well after our analysis period ends, the organization introduced non-monetary tournament incentives by allowing better-performing physicians to select future work shifts first. Second, the organization has increasingly adopted AI tools for administrative tasks. However, the only tool that overlaps with our window was the very first test of an AI tool, piloted with a subset of physicians from April 2023 onward. This tool only gave advice on the secondary (less important) diagnosis code. We show that our results are robust to AI tool use by excluding that period in a robustness check. The rollout of further AI tools for administration may have explained why total consultation time kept decreasing in the raw data also after the sample period (Figure 2c).

3 Conceptual framework

This section develops a conceptual framework to guide the empirical analysis. It has three parts. First, physicians produce an output, patient health, which neither the public payer nor the organization observes. As in the multitasking model of Holmstrom and Milgrom (1991), physicians produce the output by allocating a scarce input, their time, between patient-facing and administrative work. Quantity incentives raise the cost of every minute, so physicians shorten both time inputs. Second, the public payer pays the organization per consultation, enforces minimum quality standards for administrative work, and bears downstream costs. The organization can contract with physicians on consultation volume and internal quality, so that the physicians' objective contains what the organization values. Reductions fall mainly on administrative time unless patient health is much more concave in administrative than in patient-facing time. As in Baker (1992), the organization's internal quality measure can diverge from what matters to the public payer: internal quality may be unaffected while patient health falls and downstream costs rise. The setup generates an agency problem between the principal, the public payer, and the agent, the physician (to whom the organization's objectives are passed). Third, the quantity-incentives contract combines a piece rate with a wage guarantee, as in Lazear (2000), so only physicians who gain from exceeding the threshold respond. Physicians differ in how much they value patient health relative to the organization's quality measures and to their own pay, rather than in ability. Their consultation times therefore already differ under fixed wages, and quantity incentives reduce the dispersion among those who respond. Appendix C presents the formal model and derivations.

3.1 Setup

Health production and the organization. Physicians use their time to produce care during consultations. They can input patient-facing time, t_f, or administrative time, t_a. Patient-facing time covers examination and communication during the video call. Administrative time covers consultation-related work performed outside the video call, including reviewing past medical records, reading patient-provided pre-consultation symptom reports, writing new medical records, issuing prescriptions, and arranging referrals and follow-up care.

Let $h(t_f, t_a) = h_f(t_f) + h_a(t_a)$ be the output of a consultation: the monetary value of patient health, where h_f is the health produced during the time spent directly with the patient, and h_a is the health produced during administrative work, such as by a careful review and preparation of the patient's record. Both are increasing with diminishing returns. A consultation benefits the patient only once it receives sufficient time, $h(0, 0) < 0$, so the value of a consultation per minute is highest at a positive length, and physicians do not minimize it.

Neither the public payer nor the organization observes patient health $h(\cdot)$ directly, and thus cannot contract on it. The organization observes and evaluates physicians on (i) the number of consultations per staffed hour, which depends on the time spent per consultation and determines pay under one of the contract types below; and (ii) internal quality q. Internal quality combines

patient ratings, which depend only on patient-facing time, and clinical process quality, which the organization observes only through random audits of medical records and prescriptions. Audits verify that documentation and prescribing meet a minimum standard, so clinical process quality enters internal quality only up to that minimum.

Internal quality $q = q_f + q_a$ has two components. First, patient ratings $q_f(t_f)$ measure patient satisfaction, which rises both with the health produced during the time the physician speaks to the patient and with the patient's experience of the consultation itself, which responds to the time the physician spends with the patient, $q_f = q_f(h_f(t_f), t_f)$.[6] We assume that q_f is strictly increasing and strictly concave in patient-facing time. Second, the organization measures clinical process quality $q_a(t_a)$ only through its audits. This establishes a verifiable minimum quality level, $\underline{q}_a$, that ensures complete records and guideline-concordant prescribing. We assume that q_a increases in t_a up to the administrative time necessary for that minimum quality bar, $\underline{t}_a$, and is flat above it. Internal quality is an incomplete measure of patient health. Patient ratings respond to the health produced during the consultation only as patients perceive it, and not to the health produced by administrative time, which patients do not observe, and the organization's audits verify only a minimum quality standard. We assume the physicians choose $t_a > \underline{t}_a$ under either contract, so that only patient-facing time raises internal quality at the margin. The organization also records outcomes such as overall prescription rates and return visits, but does not use them to evaluate physicians; in the model, they enter downstream costs.

Downstream costs, payments, and minimum quality. Worse patient health leads to more follow-up care, so expected downstream costs $D(h)$ are strictly decreasing in patient health. Any change in patient health is reflected in downstream costs, although D measures the cost of the care that follows, not the value of the health change itself. Downstream costs include the production costs of follow-up care outside the organization, as well as return consultations and prescriptions at the organization, which are not part of its internal quality measures. The public payer bears these costs, and neither the physician nor the organization internalizes them.

The organization has two objectives: securing payment and building a strong reputation. It receives a fixed payment per consultation from the public payer, as long as its records and prescriptions meet the minimum standard, $q_a(t_a) \geq \underline{q}_a$. The public payer reviews the organization's processes to assess compliance with the minimum standard required for payment. The organization's reputation is determined by patient satisfaction and measured directly by q_f. Patient satisfaction matters because dissatisfied patients will seek care from competing healthcare organizations. The public payer does not review patient ratings.

The organization passes both objectives on to the physician: the minimum standard, $q_a \geq \underline{q}_a$, which it needs to meet to receive payment, and the importance of patient satisfaction, q_f, to retain patients. The organization's profits increase with the number of consultations as long as it meets

[6] Patient ratings q_f matter for the organization directly, as they affect its reputation. They do not fully reflect patient health; previous research shows that patient ratings in primary care are not correlated with physicians' effects on patient mortality (Ginja et al. 2025).

the minimum quality standard and receives the fee-for-service reimbursement from the public payer. It contracts with physicians on two observable outcomes: the number of consultations per unit of time worked, and internal quality. Under the quantity-incentives contract, the organization and its physicians are aligned on everything the organization observes, in particular internal quality *as well as* the gains from each additional consultation. Under the fixed-wage contract, the organization does not pass its per-consultation payment to the physician. Physicians additionally value patient health directly, through the weight α below, which the organization cannot observe or reward. Neither the organization nor its physicians bear downstream costs.

Contract type. Over a calendar month, a physician's average time per consultation is $T = t_f + t_a + t_z$, where t_z is break time outside of recorded administrative or patient-facing consultation time, which is taken as given. The physician's pace is $n = \frac{60}{T}$ consultations per staffed hour. At the threshold pace of five consultations per staffed hour, $T = 12$ minutes, including breaks.

The physician receives a guaranteed hourly wage $\bar{W}$ plus a bonus $B_d(n)$ under contract $d \in \{0, 1\}$, where $d = 0$ denotes the fixed-wage contract and $d = 1$ the quantity-incentives contract. The fixed-wage contract pays no bonus, $B_0(n) = 0$. The quantity-incentives contract pays $B_1(n) = p(n-5)\mathbf{1}\{n \geq 5\}$ per staffed hour, where p is the piece rate for each consultation per hour above the threshold ($p = \bar{W}/5$ for video consultations in our setting). The quantity-incentives contract thus combines a guaranteed hourly wage with a piece rate and a threshold, similar to the compensation scheme studied by Lazear (2000).

Away from the threshold, marginal pay per additional consultation is:

$$\text{Marginal pay } B'_d(n) = \mu = \begin{cases} p = \bar{W}/5 & \text{if } d = 1,\ n > 5 \\ 0 & \text{else.} \end{cases}$$

Each consultation also yields a marginal value in monetary terms:

$$\text{Marginal value } v = \alpha\, h(t_f, t_a) + \beta\, q(t_f, t_a),$$

where $\alpha > 0$ weights patient health relative to the physician's own income, reflecting patient-regarding social preferences (Ellis and McGuire 1990; McGuire 2000), and $\beta > 0$ is the weight on internal quality, reflecting that the organization evaluates physicians under either contract. An additional consultation is therefore worth $\mu + v$ to the physician.

The physician's money-metric utility per staffed hour is:

$$U_d = \bar{W} + B_d(n) + nv. \tag{2}$$

Internal quality enters Equation 2 through the per-consultation value v. Patient health not captured by internal quality enters only through the physician's patient-regarding preferences. Neither the physician nor the organization internalizes downstream costs.

First-order conditions. The physician chooses time inputs t_f and t_a to maximize U_d, subject to the minimum quality requirement. More time on a consultation affects utility through two channels.

First, it has the cost of reducing the physician's pace n. Lengthening a consultation reduces the number of consultations generating v and, where applicable, the bonus. The cost of an additional minute, $\lambda \equiv (\mu + v)/T$, is the shadow price of a minute of physician time – the fraction of a new consultation that it displaces, $1/T$, times its marginal value to the physician, $\mu + v$. This shadow price is the same for both time inputs because each lowers the pace by the same amount.

Second, spending more time on a consultation increases a consultation's value v. A minute of patient-facing time raises v through patient health and measured performance. A minute of administrative time raises v only through patient health.

Away from the threshold, the physician optimally chooses time inputs such that the marginal value of a minute of each input is equal to its shadow price.

3.2 Responses to the quantity-incentives contract

Changes in physician inputs. Let T_0 denote time per consultation under the fixed-wage contract. A physician who responds to the quantity-incentives contract chooses a pace above the threshold, raising marginal pay from $\mu = 0$ to $\mu = p$.[7] At the fixed-wage input choices, the quantity-incentives contract raises the shadow price of time, $\lambda = (\mu + v)/T$, by p/T_0. Marginal costs now exceed marginal values, so the physician reduces time inputs until each marginal value again matches the shadow price, subject to the minimum quality requirement.

The physician cuts fewer minutes from the input whose marginal value increases more rapidly as time is reduced. Reducing patient-facing time lowers both internal quality and patient health, whereas reducing administrative time lowers only patient health at the margin. Appendix C.3 derives first-order approximations to these reductions around the fixed-wage input choices. Patient-facing time is reduced by less than administrative time if and only if:

$$-\alpha \frac{\partial^2 h}{\partial t_a^2} < -\alpha \frac{\partial^2 h}{\partial t_f^2} - \beta \frac{\partial^2 q}{\partial t_f^2}. \tag{3}$$

The left-hand side measures the increase in the marginal value of administrative time following a one-minute reduction. The right-hand side measures the corresponding marginal value increase for patient-facing time. As marginal returns to time are diminishing, all second derivatives are negative and every term in the inequality is positive. Condition 3 holds more readily when the weight on internal quality, β, is higher relative to the weight on patient-regarding social preferences, α.

We expect Condition (3) to hold in our setting. As time is reduced, the marginal value of patient-facing time rises through both patient health and patient ratings, whereas the marginal value of administrative time rises only through patient health. The condition therefore holds *unless*

[7] If a physician's pace remains below the threshold, the contract does not affect time inputs; see Section 3.3. If her pace crosses the threshold, time inputs fall by at least the amount needed to reach it. Above the threshold, she faces the same marginal pay $\mu = p$ as a physician who was above it under fixed wages.

the marginal health gain from administrative time rises much faster than the marginal health gain from patient-facing time as time is reduced. For example, reviewing patient records may improve treatment if it uncovers relevant information, such as comorbidities or past prescriptions, but if that information is rare and dispersed, shortening time spent on such administrative work will likely raise the marginal health gain only slightly. Administrative time cannot fall below $\underline{t}_a$, where the minimum standard binds. Above $\underline{t}_a$, audits do not detect the reduction.

Prediction 1: Time inputs.

(a) *Administrative and patient-facing time.* Positive marginal pay reduces administrative time and patient-facing time, decreasing total consultation time. Patient-facing time falls by less than administrative time if the health gain from each additional minute of administrative time rises sufficiently slowly compared to the health gains from additional patient-facing time.

(b) *Internal quality.* Clinical process quality meets the minimum standard under either contract and does not change. Positive marginal pay lowers patient ratings in proportion to the reduction in patient-facing time; as that reduction is small, patient ratings change little. Internal quality does not respond to administrative time at the margin.

Measured and unmeasured outcomes. Neither the physician nor the organization observes care outside the organization. Internal quality $q(\cdot)$ is therefore an incomplete measure of patient health. In particular, it misses rare adverse events outside the organization. Downstream costs reflect any change in patient health. The organization records its own follow-up care but does not use it to evaluate physicians. It does not observe external downstream care, and does not bear the costs of downstream care.

The physician's time inputs t_f and t_a produce three outcomes: (i) the number of consultations per staffed hour, which determines the physician's payoff under the quantity-incentives contract; (ii) internal quality q, measured directly by the organization; and (iii) patient health h, which is partially reflected in q and partially internalized by physicians directly. Both t_f and t_a draw on the same resource, the physician's time. Stronger incentives for a measured outcome, such as the number of consultations, raise the cost of both inputs (Holmstrom and Milgrom 1991). Internal quality rewards additional patient-facing time, but the organization has no quality measure, and no incentive passed from the public payer, to reward administrative time at the margin.

Prediction 2: Multitasking. Positive marginal pay reduces patient health and raises downstream costs. Internal quality does not fully capture the health loss because it does not respond to administrative time at the margin. To first order, patient health falls with reductions in both time inputs, whereas internal quality falls only with reductions in patient-facing time. Downstream costs rise with any loss in patient health.

3.3 Physician heterogeneity

Assume physicians differ in the preference weights α and β. These weights determine n_0, the pace that maximizes consultation value per staffed hour under the fixed-wage contract, and the opportunity cost of working faster. The cost is higher for a physician with a higher weight on patient health, α, if she raises her pace mainly by reducing administrative time.

Adoption. The bonus is paid only above a pace of five consultations per hour. A physician with a lower pace $n_0 < 5$ must reach that threshold, incurring the cost of a faster pace without compensation. Her gain from responding is the bonus earned at her optimal pace above the threshold, net of the cost of a faster pace. The gain is larger the closer n_0 is to the threshold and the lower the physician's cost of a faster pace. A physician who does not gain leaves her time inputs unchanged and, given the guaranteed hourly wage, is equally well off under either contract. We assume she then does not take up the quantity-incentives contract voluntarily.

Dispersion. Every physician who responds to the quantity-incentives contract chooses a pace above the threshold and thus a time per consultation below 12 minutes, including breaks. Physicians initially below the threshold must cut more time the slower their pace under fixed wages. Because fixed wages impose no upper limit on time per consultation, quantity incentives limit dispersion in consultation time among responding physicians. Under Condition (3), physicians mainly reduce administrative time.

Prediction 3: Adoption and dispersion.

(a) *Adoption.* Voluntary adopters tend to have a faster pace under the fixed-wage contract than physicians who adopt only when the quantity-incentives contract becomes mandatory. Only voluntary adopters reduce their time per consultation; physicians who adopt only when the quantity-incentives contract becomes mandatory make no time reduction. The effect for voluntary adopters is then the average treatment effect on the treated (ATT), and the average effect across all physicians is the ATT, multiplied by the share of voluntary adopters.

(b) *Dispersion.* Among voluntary adopters whose pace under the fixed-wage contract was below the threshold, the time reduction is larger the slower their pace, so that dispersion in the time per consultation among adopters is reduced. The reduction in dispersion falls mainly on administrative time.

3.4 Mapping to the empirical analysis

We directly observe physicians' time inputs (total, patient-facing, and administrative consultation time). We measure internal quality by patient satisfaction and, as proxies for care process quality, diagnostic informativeness and counter-guideline antibiotic prescribing within the organization. We value downstream care use at prices paid by the public payer, which are administratively set to

cover the costs of producing care. For follow-up consultations at the organization, we consider the administratively set payment rate during our analysis period, 500 SEK per digital consultation with a doctor.

4 Data

4.1 Sample construction

Data sources. Our data combine internal information from the healthcare firm with administrative records from national Swedish registers. From the firm, we observe all consultations conducted in 2021–2024, all physician contracts including contract updates from October 2022 onwards, and realized salaries under both the quantity incentives and the fixed-wage regime for a subset of physicians. Moreover, the firm supplies us with the monthly list of all physicians working under the quantity-incentives scheme (June 2022–May 2023).[8]

Using national identifiers, we link each consultation to (i) prescriptions, secondary care use, and deaths from the National Board of Health and Welfare (Socialstyrelsen, 2013–2024); (ii) socioeconomic measures – including education, employment, earnings, and civil and migration status – from Statistics Sweden (2013–2024), which we also use to characterize physicians' demographics, earnings, and education; (iii) external primary care for adult patients in two large regions of Sweden, Scania and Stockholm, which together account for about 40.2% of consultations in the analysis sample.

Analysis sample. We follow 266 physicians who were already working primarily online when the rollout began. Physicians enter the sample if they conduct any consultations between January and May 2022, before the start of the rollout, and if fewer than 10% of those consultations are in the organization's non-digital services.

The main analysis sample contains 452,652 unscheduled (drop-in) video consultations with 343,101 patients, which the 266 physicians conducted between January 2022 and May 2023. We construct this sample as follows (see also Appendix Table A1). We exclude consultations with infants and consultations that follow a prior consultation within seven days to restrict the sample to first-contact cases. The analysis window ends before the mandatory adoption date in June 2023 to ensure that physicians on fixed wages are available to compare against physicians on a quantity-incentives contract. Finally, we exclude consultations reporting a rare symptom in the sample (less than 1,000 cases), as well as observations that are singletons in physician fixed effects, date-by-three-hour-window fixed effects, and basic patient controls (self-reported main symptom, gender, age, and region of residence).

Physician-month panels. We also construct a balanced panel at the physician-month level that aggregates physicians' work, containing every month from January 2022 to May 2023 for every

[8]The list is missing for employee physicians in January 2023, for whom we impute the information based on data from the month prior and after.

physician. For each physician-month, we sum minutes worked and count consultations over all of the physician's online meetings, including booked consultations and text-based consultations, which the analysis sample excludes. A month in which a physician has an active contract but conducts no online meetings enters with zero minutes and zero consultations.

Additional samples. We also report descriptives by extending the analysis sample with additional months before January 2022 or after May 2023 for the same set of physicians.

4.2 Variables

Treatment (Q_{jt}). Our treatment variable Q_{jt} indicates whether physician j was compensated under the quantity-incentives scheme in month t. We take information about the contract type from the organization's monthly records. It is therefore available for all physicians on an active contract with the organization.

Fixed effects (δ_j, θ_w) and patient controls (X_i). To identify the causal effect of the quantity-incentives scheme, we include two sets of fixed effects: (i) physician fixed effects δ_j, which account for time-invariant characteristics of the assigned physician; and (ii) date-by-three-hour-window fixed effects θ_w, which account for common variation in patient demand and physician availability. These windows are defined by when patients register their request for a consultation. For example, this allows us to compare consultations that took place between 12 pm and 3 pm on the 15th October 2022. We also include patient controls X_i: For each consultation i, these comprise the patient's self-reported main symptom, gender, age in years, and region of residence. This information matches the case-specific and demographic information which the organization also observes.

Additional patient characteristics. We also observe patient characteristics that are unlikely to be available to the organization: (i) health-risk measures, which capture prior healthcare use (hospital, emergency, urgent, and specialist care) and comorbidities recorded in specialist or hospital care during the three years before the consultation, excluding the final two days; and (ii) socioeconomic characteristics measured in 2021 (indicators for above-median earnings, welfare benefit receipt, education level, employment, civil and migration status, an indicator for whether any socioeconomic information is missing). We use these additional variables in balance tests and robustness checks.

Outcomes within the organization. We distinguish three sets of outcomes recorded within the organization: (i) time inputs which are directly observed (total consultation time, patient-facing time, and administrative time); (ii) internal measures of quality used to evaluate physicians (patient satisfaction and, as proxies for care process quality, diagnostic informativeness and counter-guideline antibiotic prescribing); and (iii) recorded outcomes that are not directly used for evaluation (indicators of any prescribing, any unscheduled return consultation, and any booked return consultations, all within seven days).

Downstream care. To measure downstream healthcare use outside the organization, we construct indicators for the following outcomes within seven days of the consultation date: (i) any prescription issued by another provider; (ii) any primary care consultation with another provider, observed only for patients registered in Scania or Stockholm; (iii) any urgent care use; (iv) any specialist visit; (v) any emergency department (ED) visit; and (vi) any hospitalization. We exclude visits and admissions with diagnoses related to pregnancy or childbirth. (vii) To capture adverse outcomes beyond healthcare use, we also measure mortality within one year using index consultations for which the patient had no other consultation in the preceding year. (viii) Downstream costs capture the total cost of care received within seven days of the consultation, both inside and outside the organization, calculated by counting each type of care at most once and applying the unit costs reported in Appendix Table A6.

Physician labor. At the physician-month level, we construct the following measures: (i) total minutes worked, defined as the total time a physician is staffed during a month (sum of consultation and break minutes), and decomposed into patient-facing, administrative, and break minutes; (ii) the number of consultations, measured separately for video consultations and for text-based consultations; and (iii) consultations per hour worked, calculated as the number of consultations divided by total hours worked within a month. Break minutes measure the time between consecutive meetings within a staffed shift. To measure staffed time, we assume that a shift ends after a gap of at least fifteen minutes or after eight hours staffed.

4.3 Descriptives

Table 1 reports patients' self-reported symptoms and demographics in the analysis sample. Appendix Table A2 reports patients' registered region, health risk, and socioeconomic characteristics. The most common symptoms are skin rash, prescription renewal, and cough.[9] Patients are 29 years old on average, and 60% are female. The table also reports unconditional comparisons between consultations conducted under fixed wages and under quantity incentives. Differences are economically small but sometimes statistically significant, likely because consultations under quantity incentives tend to occur later in the analysis window (Appendix Figure A1). Our empirical strategy compares consultations within date-by-three-hour windows to account for time imbalances.

Table 2 describes physicians in the analysis sample before the quantity-incentives contract rollout. On average, they conduct 4.1 consultations per worked hour during this period and work part-time, with 38 hours per month. Administrative time accounts for about two-thirds of their consultation time. Physicians are 42 years old on average, and 44% are female. Most physicians hold a specialist degree (48%) or are in specialist training (32%), and 11% have a degree from Sweden's highest-ranked medical school. For most physicians, the organization is a secondary employer, with wage income from the organization accounting for about a fifth of their gross wage income. Voluntary adopters of

[9]Some symptoms, such as *Other health inquiries* and *Other care processes*, do not identify specific reasons that patients seek care. For these categories, the most common non-missing diagnosis provided after the consultation is "Acute upper respiratory infection, unspecified".

the quantity-incentives scheme conduct more consultations per hour in the pre-rollout period than those who adopt only when the scheme becomes mandatory (4.3 versus 3.9), are younger, and have higher gross wage income. They otherwise do not differ in gender, training, or hours worked.

Figure 2 shows average consultation time around the onset of quantity incentives. Consultation time falls sharply by about two minutes in the three-hour windows immediately after onset (Panel a) and in the following month (Panel b), then continues to fall. Average consultation time declines from about 12 minutes in early 2022 to 9 minutes in June 2023, when the scheme became mandatory, as part of a broader secular decline over 2021–2024 (Panel c).[10]

Appendix Figure A2 plots average consultation time by onset cohort during the analysis period. Consultation time drops sharply in the month of onset in each larger cohort that adopted voluntarily. For the 94 physicians who adopted when the scheme became mandatory, consultation time does not change discontinuously.

5 Empirical framework

Our main causal effect of interest compares the outcome of a consultation handled by a physician on the quantity-incentives scheme with the (counterfactual) outcome of that consultation, had the same physician worked under the fixed-wage scheme.

We can identify this effect for all physicians who voluntarily adopt the quantity-incentives scheme, that is, we can estimate a treatment effect on the treated (ATT). The ATT is also the effect of interest, since the treatment effect for mandatory adopters is zero given our conceptual framework (see Prediction 3). Below we describe the design, estimation and inference, and assumptions for estimating the ATT.

5.1 Design

Our setting provides two sources of variation that operate at different levels of the data. First, the organization allocates unscheduled (drop-in) consultations to whichever physicians are available at the time, and patients neither observe nor can act on a physician's compensation scheme. Within narrow time windows, it is therefore plausibly as good as random which physician is assigned a consultation and physicians cannot select the type of consultation according to their contract scheme. Second, physicians move onto the quantity-incentives scheme at different points in time. For a given time window, we thus observe both consultations conducted under the quantity-incentives scheme and under the fixed-wage scheme, and for a given physician, we observe consultations she has worked on under the quantity-incentives and under the baseline scheme. We measure outcomes at the consultation level, while the treatment, working under the quantity-incentives scheme, varies at the physician-month level.

[10] From late 2021 to early 2022, Sweden went through a COVID-19 wave and lifted most nationwide restrictions afterward. These common shocks likely affected the composition of cases and drive part of the secular decline. Later, the organization piloted and increasingly adopted AI tools for administrative tasks, and let better-performing physicians select work shifts first.

To estimate the treatment effect on the consultation-level, we include physician and time-window fixed effects. Physician fixed effects account for time-invariant differences between early and late adopters, such as differences in physicians' baseline pace which might be correlated with both the adoption decision and outcomes. Date-by-three-hour-window fixed effects account for common differences across the time windows in which physicians choose to work. The treatment effect is therefore estimated from changes within physicians, accounting for time trends.

For consultation i handled by physician j in date-by-three-hour window w in calendar month t, we estimate:

$$Y_{ijtw} = \gamma Q_{jt} + \delta_j + \theta_w + X_i'\rho + \varepsilon_{ijtw}, \tag{4}$$

where Y_{ijtw} is the consultation outcome; Q_{jt} indicates whether the assigned physician $j = j(i)$ is on the quantity-incentives scheme in calendar month $t = t(w)$; θ_w and δ_j are window and physician fixed effects; and X_i contains patient controls: indicators for reported symptom, gender, age in years, and registered region. Our parameter of interest is γ.

Our analysis sample includes all consultations between January 2022 and May 2023, the last month before all physicians moved onto the quantity-incentives contract, ensuring that there are not-yet-treated physicians in every month.

5.2 Estimation and inference

Because physicians move onto the scheme at different dates, we do not estimate Equation (4) by ordinary least squares (OLS) with two-way fixed effects (TWFE) on the pooled sample. Under staggered treatment adoption, TWFE uses already-treated physicians as part of the comparison group, which can produce negative weights that bias treatment estimates (Goodman-Bacon 2021; Chaisemartin and D'Haultfœuille 2020).

We instead estimate γ in two stages, following Gardner et al. (2026). The first stage uses OLS to estimate the fixed effects and patient controls in the untreated-outcome model, $\delta_j + \theta_w + X_i'\rho$, using only not-yet-treated consultations, that is, consultations that took place before the assigned physician's first month on the quantity-incentives scheme. This allows the physician fixed effects $\widehat{\delta}_j$ to be estimated from treated physicians' own pre-treatment consultations, and the window fixed effects $\widehat{\theta}_w$ to be estimated from not-yet-treated physicians working in a given window. Using estimates of δ_j, θ_w, and ρ, we residualize consultation outcomes: $\widetilde{Y}_{ijtw} = Y_{ijtw} - \widehat{\delta}_j - \widehat{\theta}_w - X_i'\widehat{\rho}$. The second stage then regresses the residualized consultation outcome on Q_{jt} to estimate the treatment effect γ by OLS. Under the identifying assumptions below, $\widehat{\gamma}$ estimates the average treatment effect on treated consultations.

The standard errors are estimated by generalized method of moments (GMM) to account for estimation error in the first stage. Because the identifying variation comes from physicians' adoption of the quantity-incentives scheme over time, we cluster standard errors at the physician level (Abadie et al. 2023).

Physician-month specifications. We also estimate effects on physicians' aggregate work at the physician-month level using a balanced panel as described in Section 4. For physician j in calendar month t, we estimate:

$$\bar{Y}_{jt} = \phi Q_{jt} + \psi_j + \tau_t + \eta_{jt}, \tag{5}$$

where $\bar{Y}_{jt}$ is an aggregated physician-month level outcome; Q_{jt} is the treatment indicator as defined above; ψ_j are physician fixed effects; and τ_t are calendar-month fixed effects. The treatment effect of interest is ϕ.

We estimate ϕ as above. In the first stage, we estimate $\psi_j + \tau_t$ using physician-month data from before each physician's first month on the scheme. In the second stage, we regress the residualized outcome on Q_{jt}. Standard errors are clustered at the physician-level.

Weighting. For the same outcome Y, the main consultation-level specification (Equation 4) and the physician-month specification (Equation 5) may differ in their weighting. For example, in Equation 5, the treatment effect estimated in a physician-month with many consultations receives the same weight as in a physician-month with few consultations. In contrast, treatment effects in Equation 4 are weighted by consultation counts. The treatment effects γ and ϕ will thus not generally be equivalent for the same outcome. For our main results, we treat consultations as the unit of interest and weight by consultation counts.

5.3 Identification

Identification rests on two assumptions corresponding to the two sources of variation. First, a *composition assumption* ensures that there is no selection of patients to physicians working in the same narrow time window based on the physician's quantity-incentives status. Second, an *adoption assumption* ensures that, absent the treatment, a physician's outcomes under the quantity-incentives scheme would have evolved in parallel with those of not-yet-treated physicians on the fixed-wage scheme. We state the assumptions, discuss their relationship, and provide diagnostic evidence.

Assumption 1 (Composition): Quasi-random assignment within time windows. Let $Y_{ijtw}(d)$ denote the potential outcome of consultation i when the assigned physician is compensated under the baseline scheme ($d = 0$) or with quantity incentives ($d = 1$). Let $\mathcal{C}_{jw}(q)$ denote the set of consultations physician j would handle in window w if her contract status in that window were $q \in \{0, 1\}$ (baseline scheme; quantity incentives). We observe the consultations for her realized status Q_{jt}. We assume that the patients a physician handles in a window do not depend on her contract status:

$$\mathbb{E}\left[Y_{ijtw}(0) \mid i \in \mathcal{C}_{jw}(1)\right] = \mathbb{E}\left[Y_{ijtw}(0) \mid i \in \mathcal{C}_{jw}(0)\right]. \tag{6}$$

Under Equation (6), consultations handled by different physicians can differ. Physicians can also differ in characteristics such as their baseline pace, even if baseline pace is correlated with the time of adoption. The composition assumption only rules out that the types of patients assigned to a

physician change systematically once a physician moves onto the quantity-incentives scheme, in ways that are not captured by the narrow time windows. A violation of this assumption would be, for example, if physicians on the quantity-incentives scheme could select into faster consultations with easy patients. Under such a violation, we would attribute physician cream-skimming to an incentive effect of physicians working faster, and potentially underestimate negative effects on patient health if the cream-skimmed patients are healthier.

Assumption 2 (Adoption): Additively separable untreated outcomes. We separately impose the following model for the untreated outcomes of the consultations a physician handles under fixed wages:

$$\mathbb{E}\left[Y_{ijtw}(0) \mid i \in \mathcal{C}_{jw}(0)\right] = \delta_j + \theta_w. \tag{7}$$

Once the physician fixed effects and time-window fixed effects are accounted for, expected untreated outcomes contain no further physician-specific changes associated with adoption.

Equation (7) imposes parallel trends on untreated outcomes. It is violated, for example, if earlier-treated physicians generally decreased their consultation times even without moving onto the quantity-incentives scheme, relative to later-treated control physicians. In the presence of differential trends, we would attribute part of the decline in consultation time to quantity incentives and overstate the treatment effect.

Together, *Assumption 1 (Composition)* and *Assumption 2 (Adoption)* imply $\mathbb{E}[Y_{ijtw}(0) \mid j, w] = \delta_j + \theta_w$ for all observed consultations, including those observed under quantity incentives. While we do not need patient controls X_i for identification, we include them in our main specifications to improve precision. In robustness checks, we show that our estimates are unchanged when we include only the physician and time-window fixed effects.

A differential trend could arise even if *Assumption 1 (Composition)* holds and patients are assigned as good as random within time windows. Similarly, *Assumption 2 (Adoption)* and parallel trends do not restrict selective case assignment in response to the quantity-incentives contract. Below, we test for both assumptions.

Testing for balance. To support the *Assumption 1 (Composition)*, Figure 3 tests balance in patient assignment to a physician under a quantity-incentives or fixed-wage contract. The figure examines balance in the patient controls X_i, as well as additional health-risk and socioeconomic characteristics (see Section 4).

The left panels report coefficients from separate regressions of an indicator for doctor with quantity incentives on each characteristic, controlling for physician and date-by-three-hour-window fixed effects. All estimates are small: no characteristic is associated with a difference of more than 3 percentage points, compared with a mean of 33% in the probability that a doctor is on the quantity-incentives scheme. A joint F-test does not reject the null hypothesis that the coefficients for all characteristics are zero.

The right panels report unconditional normalized differences between consultations under quantity

incentives and those under fixed wages. Even without conditioning on the fixed effects, most normalized differences are below 0.1. For the two characteristics above this threshold, the conditional estimates show no evidence of imbalance after adjusting for fixed effects. The normalized differences are well below the 0.25 threshold for potentially problematic imbalance identified by Imbens and Rubin (2015) and Baker et al. (2026), and generally fall below the stricter 0.1 threshold that, as these authors note, is sometimes used in psychological studies.

To account for minor imbalances in the estimation, we include patient symptoms and demographics as controls in our main specifications. In later robustness checks, we additionally vary the control variables and show that our estimates are unchanged as long as we condition on the physician and time window fixed effects.

Testing for pre-trends. To test *Assumption 2 (Adoption)*, we examine pre-trends using an event study. We use the same two-stage approach as our main specification in Equation (4), but replace Q_{jt} in the second stage with indicators for months relative to the physician's first month on the scheme.[11] This approach keeps our weighting by consultation.

Figure 4 plots the event study estimates for total consultation time. Before treatment onset, all coefficients are close to zero with no pre-trend. Immediately after treatment onset, consultation time drops and remains lower in the following months. Standard errors increase with months since onset because the analysis window includes only early adopters observed many months after onset. Appendix Figure A4 shows that, also for the other main outcomes, the event studies do not indicate pre-trends.

Appendix Figure A3 shows the event study for total consultation time under alternative estimators. Panel (a) estimates the event studies by OLS with TWFE. The TWFE event study coefficient estimates are compressed toward zero compared to our main estimates, indicating that negative weighting bias may be present under the staggered treatment adoption design. Panel (b) compares our main two-stage estimates with those proposed by Callaway and Sant'Anna (2021) or De Chaisemartin and D'Haultfœuille (2026). As these alternative estimators require one observation per physician and month, we use a physician-month panel that averages the consultation outcomes of each physician's consultations in a calendar month. For comparability, we weight our two-stage estimates by physician-months using the inverse of the number of consultations in a month as weights. The three estimators yield very similar results.

6 Results

We first present the effects of quantity incentives on physicians' time inputs and on outcomes recorded by the organization (Prediction 1). We then show the effects on care outside the organization and

[11] Let t_j^* denote physician j's first month on the quantity-incentives scheme and $e = t(w) - t_j^*$ the month relative to onset. The second stage is $\tilde{Y}_{ijtw} = \sum_{e \neq -1} \gamma_e \mathbf{1}\{t(w) - t_j^* = e\} + \nu_{ijtw}$, where the month before onset, $e = -1$, is the reference month. The coefficients for $e < 0$ test for differential trends before adoption, and the coefficients for $e \geq 0$ estimate the effect e months after onset.

on downstream costs (Prediction 2). We also show that the estimates are robust to alternative specifications and samples.

6.1 Within-organization effects

Table 3 reports the effects of quantity incentives on within-organization outcomes. Panel A reports the effects on time inputs. Quantity incentives reduce total consultation time by 2.44 minutes, or 21% of the baseline mean of 11.7 minutes. This reduction is primarily driven by administrative time, which falls by 2.14 minutes (28% of 7.73 minutes), whereas patient-facing time falls only by 0.30 minutes (7.5% of 4.00 minutes).

Panel B reports the effects on patient ratings, measured by top grade and worst grade, and diagnostic informativeness and counter-guideline antibiotic prescribing. These measures proxy the organization's main internal quality measures used to evaluate physicians: patient satisfaction and care process quality. None of the estimates is statistically significant.[12]

Panel C reports outcomes that the organization records but does not use to evaluate physicians. The share of consultations followed by a prescription from the organization within seven days rises by 1.1 percentage points from 51%. The share followed by an unscheduled return consultation rises by 0.26 percentage points from 6.4%, but this increase is not statistically significant. Booked return consultations do not change.

These results are consistent with Prediction 1. First, while physicians reduce total consultation time, they reduce administrative time by much more than patient-facing time. Second, the organization's quality measures change little with a small reduction in patient-facing time, regardless of changes in administrative time.

6.2 Effects outside the organization

Table 4 reports effects of quantity incentives on the care that patients receive outside the organization within seven days of the consultation. Panel A shows that quantity incentives do not shift care to other primary care providers: external primary care consultations and external prescriptions do not change. However, the share of patients with an urgent care visit rises by 0.07 percentage points from 0.39%, or 18.2%.

Panel B further shows that hospitalization rates increase by 0.11 percentage points from 0.38%, or 28.9%. Specialist and emergency department visits do not change. The effect on hospitalizations persists over thirty days (Appendix Table A3). Estimates for avoidable hospitalizations, which timely primary care can prevent, are positive but statistically significant only within 30 days. The increase in hospitalizations is not reflected by higher one-year mortality.

Panel C reports the effect on downstream costs, the costs of providing follow-up care. Downstream costs rise by 24.9 SEK per consultation from 754 SEK, or 3.3%. Additional hospitalizations account

[12] Appendix Table A3 further shows that patient satisfaction in consultations with any grade is unchanged, and that there are no clear quality deteriorations in prescriptions, including no increases in overall antibiotic or opioid prescriptions.

for most of this increase: at 28,488 SEK per hospitalization, an increase of 0.11 percentage points increases the downstream costs by 31 SEK per consultation.

These results are consistent with Prediction 2. The reduction in time inputs leaves the organization's quality measures unchanged but raises downstream costs, almost entirely through care outside the organization, which neither the organization nor the physician observes.

6.3 Robustness

We provide event studies for the other main outcomes reported in Appendix Figure A4. These event studies show a lack of pre-trends for all outcomes. The effects remain persistent over time, but are estimated with lower precision for later relative time periods, when the estimates have to rely on a smaller number of physicians on quantity incentives.

Appendix Table A4 shows the robustness of our results to alternative sets of control variables and estimators. The estimates are unchanged when we (i) include only the physician and date-by-three-hour-window fixed effects; (ii) add patient health-risk and socioeconomic characteristics which the organization does not observe; and (iii) control for the share of other physicians on the quantity-incentives scheme up to one hour prior to the start of a consultation. This specification addresses potential spillovers, as physicians working in the same window draw patients from a common queue. Finally, (iv) we also estimate Equation (4) by OLS with two-way fixed effects. The estimated effects generally become more compressed, as expected if treatment effects are negatively weighted.

Appendix Table A5 shows robustness to alternative sample definitions. The effects are similar when we (i) restrict to each patient's first consultation only; (ii) also include infants; (iii) restrict to consultations up to eight months after onset, beyond which effects are estimated less precisely (Appendix Figure A4); (iv) restrict to consultations from March 2022 onward, which excludes the last months of the COVID-19 pandemic; (v) restrict to consultations up to March 2023, before the organization began piloting an AI tool for administrative tasks; (vi) restrict to physicians who never switch off the scheme.

6.4 Mechanisms

Next, we examine how physicians reduce consultation time and which consultations account for the modest increase in downstream costs. We first analyze shifts in the distribution of consultation time. Then, we compare the effects on consultation time and downstream costs by patient characteristics and presenting complaint (*symptom*). Finally, we examine variation in physician responses to test Prediction 3.

Patient complexity We begin by showing that quantity incentives reduce physicians' use of administrative time broadly, while leaving patient-facing time largely unchanged. Figure 5 compares the distribution of consultation time under fixed wages with the implied distribution under quantity

incentives.[13] The distribution of total consultation time shifts to the left (Panel a), driven almost entirely by a reduction in administrative time (Panel b), while the distribution of patient-facing time remains nearly unchanged (Panel c). The distributions of total and administrative time under fixed wages first-order stochastically dominate those under quantity incentives.

The time reductions are also similar across patient groups. Appendix Figure A7 reports effects by the patient's number of emergency department visits and hospitalizations in the three years before the consultation. Administrative time falls by about two minutes in every group, and we cannot reject equal effects across groups. Similar patterns appear across deciles of downstream costs predicted from the patient's symptom, gender, age, and region (Appendix Figure A5), and across symptom and age groups (Appendix Figure A6). Total and administrative time fall by between two and three minutes across all groups, while patient-facing time changes much less. Thus, physicians reduce their time inputs across consultations with different symptoms and patient characteristics, rather than a subset of 'easy' consultations.

Effects on downstream costs differ by presenting symptom. Appendix Figure A6 shows that costs rise following consultations for acute infections, allergy or asthma symptoms, and 'other symptoms', a category in which the reported reason for seeking care does not identify a specific symptom. We detect no increases following consultations for skin conditions, sexual and reproductive health, or administrative and preventive care. The differences across symptom groups are statistically significant ($p < 0.001$). We find no differences in downstream costs by age (Appendix Figure A5) or prior use of emergency or hospital care (Appendix Figure A7).

The diagnoses during subsequent hospitalizations provide clinical context for these patterns. Appendix Table A8 lists the diagnoses of hospitalizations following consultations in the analysis sample. Most involve acute infections or abdominal conditions, including pneumonia, nephritis, and appendicitis. These conditions can present with common, non-specific symptoms such as a cough or abdominal pain, which can escalate quickly.

Together, these patterns suggest that less thorough assessment or follow-up of a presenting complaint may be one channel behind the small average increase in downstream costs. Reviewing records, reading patients' pre-consultation symptom reports, documenting decisions, issuing prescriptions, and arranging further care require physicians to input administrative time. For a clearly defined task, such as a prescription renewal, cutting administrative time may matter little, but it may be more consequential when symptoms are ambiguous, and conditions can deteriorate quickly. Such acute conditions can arise even in otherwise healthy patients, while patients with established health risks may also receive ongoing care elsewhere. In some consultations at the organization, less time for such work could increase the need for subsequent care outside the organization, with costs borne by the public payer.

Physician heterogeneity Next, we test Prediction 3, which builds on the idea that physicians differ in their preferences and therefore in their time inputs under fixed wages, that is, in their

[13] We estimate these effects using the main specification (4), with indicators $\mathbf{1}\{b \leq Y_i < b+1\}$ for each one-minute bin b as outcomes. The estimates give the effect of quantity incentives on the share of consultations in each bin.

practice style.

Table 2 and Appendix Figure A2 show descriptively that voluntary adopters were closer to the threshold and reduced their consultation time when they moved onto the scheme, whereas non-voluntary adopters did not. Our main estimates identify the effect of quantity incentives on voluntary adopters.

Among adopters, Figure 6 shows that the time reduction is larger for physicians who were slower before they moved onto the quantity-incentives scheme. This heterogeneity is entirely driven by administrative time ($p < 0.001$), whereas patient-facing time falls by about the same small amount in every group ($p = 0.61$). Under quantity incentives, the average time per consultation for all groups is close to or below 12 minutes, such that time inputs become less dispersed. The increase in downstream costs, in contrast, is largest for physicians who already had a high pace and cut administrative time from a low level ($p = 0.002$), consistent with diminishing returns to administrative time in our framework.

Appendix Figure A8 shows the effects by physicians' administrative share of consultation time under fixed wages. Physicians with a high administrative share cut administrative time the most, and physicians with a low administrative share cut patient-facing time somewhat more, so that both time inputs become less dispersed. If different physician preferences generate different optimal time inputs, quantity incentives reduce variation in this practice style.

7 Welfare implications

So far, we have estimated an increase in downstream costs under the quantity-incentives scheme. However, overall welfare effects depend on whether quantity incentives expand the number of consultations that physicians conduct, given that the supply of physicians is constrained, and how patients' health loss is priced. In particular, quantity incentives still raise overall welfare if the value of additional consultations through expanded capacity outweighs the loss in patient health and the increase in downstream costs on baseline consultations. Below, we set up a framework to provide welfare implications.

7.1 Framework

The quantity-incentives contract affects welfare through two channels. First, for the N_0 baseline consultations conducted under the fixed-wage contract, it changes patient health by Δh and downstream costs by ΔD per consultation. Second, it changes the number of consultations that physicians conduct. Let N_d denote total consultations and L_d total staffed hours under contract d over a period, so that $N_d = L_d\, n_d$ with the aggregate pace $n_d = N_d/L_d$. The number of consultations during that period rises by $\Delta N = N_1 - N_0$ if physicians work at a faster pace or staff more hours. Each additional consultation has a net social value V_M. Additional consultations matter because physician supply constraints limit access to care and ration it. Even if different provider organizations compete for patients, they draw on the same limited pool of physicians. Increasing consultations

at unchanged staffed hours, which requires no additional inputs or production costs, therefore expands total capacity, allowing more patients to receive care. Transfers between the public payer, organization, and physician, including the bonus, do not change the resource costs of care. The welfare effect is thus $\Delta\mathcal{W} = N_0(\Delta h - \Delta D) + \Delta N\, V_M$.

Because patient health is not observed in monetary units, we cannot estimate Δh. Instead, we compute the break-even health effect – the change in patient health per baseline consultation that the quantity-incentives contract has to induce for welfare to remain unchanged:

$$\Delta h^{BE} = \Delta D - \frac{\Delta N}{N_0} V_M. \tag{8}$$

If the additional consultations are valuable enough, Δh^{BE} can be negative. The gain from expanded access may then outweigh a worsening of patient health relative to the fixed-wage contract, as long as the loss in patient health is small enough. In other words, the quantity-incentives contract increases welfare if and only if the change in patient health per baseline consultation exceeds the break-even health effect, $\Delta h > \Delta h^{BE}$.

We next take Equation (8) to the data. Section 6 identifies the change in downstream costs ΔD. We aggregate consultations and staffed hours to the physician-month level to estimate the change in the number of consultations ΔN. We do not measure the health effect Δh or the net social value V_M of a marginal consultation. Instead, we report the break-even health effect Δh^{BE} for various values of V_M.

7.2 Capacity

Table 5 reports effects on physician labor at the physician-month level. Quantity incentives increase the number of video consultations per month by 39 from a baseline mean of 152 consultations (26%).[14] They do so by raising consultations per hour worked by 1.08, from a baseline of 4.3 to 5.4 consultations per hour, which is above the bonus threshold. Total minutes worked and break minutes are unchanged, so that the additional consultations come entirely from a faster work pace. Monthly administrative time falls by 207 minutes, while total patient-facing time remains unchanged. These results are driven by video consultations. Text-based consultations are unaffected, and they have a low baseline.

7.3 Back-of-the-envelope calculations

We estimate a downstream cost increase of $\Delta D = 24.9$ SEK per consultation (Table 4, Panel C), and an increase in the number of video consultations by $\Delta N/N_0 = 0.26$ (Table 5, Panel B).

Using these estimates, Figure 7 reports the break-even health effect Δh^{BE} – that is, the monetary change in patient health at which quantity incentives are welfare-neutral – under various assumptions

[14] This capacity increase matches the reduction in consultation time. At unchanged hours, shortening consultations by 2.44 minutes from a baseline of 11.7 minutes raises the number of consultations by $11.7/(11.7 - 2.44) - 1 \approx 26\%$.

on the value of each additional consultation, V_M.[15] At $V_M = 0$, additional consultations have no net social value, and quantity incentives raise welfare only if they raise patient health by at least 25 SEK per baseline consultation to exactly offset the additional downstream costs. At $V_M = 96$ SEK, the value from additional consultations exactly covers the downstream costs, and quantity incentives raise welfare as long as patient health does not fall. Either case, $V_M = 0$ or $V_M = 96$, implies that welfare necessarily falls under the assumption that downstream costs are strictly decreasing in patient health.

We next turn to scenarios based on the public payer's evaluations. At $V_M = 183$ SEK, the average patient fee for an online consultation, the break-even health effect is -23 SEK. At $V_M = 500$ SEK, the public payer's payment per video consultation, the break-even health effect is -105 SEK: welfare rises unless health per baseline consultation falls by more than 105 SEK. Thus, welfare falls only if patient health per baseline consultation falls by more than 21% of the value of an additional consultation (105/500), about the same share by which total consultation time falls (Table 3). At $V_M = 2{,}002$ SEK, a marginal consultation replaces a physical primary care consultation (Appendix Table A6). The break-even health effect is then -494 SEK.

To interpret these break-even thresholds, suppose that the health effect operates only through additional hospitalizations, which rise by 0.11 percentage points per consultation (Table 4). The health loss per additional hospitalization, beyond hospital care costs already included in ΔD, must exceed $105/0.0011 \approx 95{,}000$ SEK (about USD 9,400; three times the care cost) for welfare to fall at $V_M = 500$ SEK. At $V_M = 2{,}002$ SEK, the threshold is $494/0.0011 \approx 449{,}000$ SEK (about USD 44,000; more than fifteen times the care cost).

For comparison, Sweden's National Board of Health and Welfare uses 1,000,000 SEK (about USD 99,000) per quality-adjusted life year (QALY) as an upper-end reference value.[16] Finkelstein, Hendren, and Luttmer (2019) use a similar value of USD 100,000 per QALY to map health gains in the general U.S. population into monetary terms. At this valuation, welfare falls only under quantity incentives if the health loss per additional hospitalization exceeds the equivalent of about five weeks in full health (0.095 QALYs) at $V_M = 500$ SEK, or 23 weeks (0.45 QALYs) at $V_M = 2{,}002$ SEK.

8 Conclusion

When expert time is constrained, paying experts by volume can expand capacity while holding their labor supply fixed. However, quantity incentives can also have negative consequences when experts' work involves unmeasured dimensions. Although quantity incentives are common in expert services, there is little empirical evidence on this trade-off.

This paper identifies the incentive effect on individual experts, separately from the sorting of experts into contracts and tasks, and measures the costs that fall outside the organization. In

[15] Because additional consultations require no more staffed hours, we abstract from production costs in the social value of a marginal consultation, V_M.

[16] Sweden has no fixed cost-effectiveness threshold, but the National Board of Health and Welfare classifies a cost per QALY below 100,000 SEK as low, between 500,000 and 1,000,000 SEK as high, and above 1,000,000 SEK as very high (Socialstyrelsen 2026).

our setting, online primary healthcare in Sweden, quantity incentives shorten consultations by 21%, primarily through administrative time, and raise consultations per physician by 26% without additional hours worked – for the two-thirds of physicians who take up the program before it becomes mandatory. The quality measures the organization observes and uses are unchanged, but hospitalizations and downstream costs, which the organization does not observe, rise. This rise is driven by physicians who were already working at a high pace without quantity incentives. If additional consultations are valued at the public payer's price, the scheme reduces welfare only if the health loss per consultation exceeds 105 SEK *above and beyond* the costs of the additional care. This loss would arise from cutting patient-facing time by 7.5% and administrative time by 28%. If a consultation's health value equals the payer's price of 500 SEK and rose in proportion to its length, a 21% cut would cost about 105 SEK, close to the threshold. The loss is likely smaller: the time removed is mostly administrative work that, for most patients, has no detectable consequences when cut, and the health returns to administrative time diminish, so the minutes cut are among the least valuable.

How to expand access to primary care for an aging population is a key policy question across many countries. When physician time is scarce, policies that lengthen consultations may improve care for patients who are seen but reduce overall capacity. Our findings provide evidence on both sides of this trade-off. Quantity incentives can expand access through shorter consultations, but physicians mostly reduce time that is otherwise spent on patient records and may somewhat increase costs for the payers of downstream care (p-value of 0.08). The policy debates above concern in-person care, where patients are likely sicker than in our online setting, so that shorter consultations may be more costly, but in such settings, access constraints may also be tighter.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?*". In: *The Quarterly Journal of Economics* 138 (1): 1–35. [Link].

Alexander, Diane. 2020. "How Do Doctors Respond to Incentives? Unintended Consequences of Paying Doctors to Reduce Costs". In: *Journal of Political Economy* 128 (11): 4046–4096. [Link].

Association of American Medical Colleges. 2024. *The Complexities of Physician Supply and Demand: Projections From 2021 to 2036*. Report. Washington, DC: Association of American Medical Colleges. [Link].

Baker, Andrew, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna. 2026. "Difference-in-Differences Designs: A Practitioner's Guide". In: *Journal of Economic Literature* 64 (2): 498–557. [Link].

Baker, George P.. 1992. "Incentive Contracts and Performance Measurement". In: *Journal of Political Economy* 100 (3): 598–614. [Link].

Bandiera, Oriana, Andrea Prat, Stephen Hansen, and Raffaella Sadun. 2020. "CEO Behavior and Firm Performance". In: *Journal of Political Economy* 128 (4): 1325–1369. [Link].

Becker, Gary S.. 1965. "A Theory of the Allocation of Time". In: *The Economic Journal* 75 (299): 493–517. [Link].

Biasi, Barbara. 2021. "The labor market for teachers under different pay schemes". In: *American Economic Journal: Economic Policy* 13 (3): 63–102.

Brekke, Kurt R., Tor Helge Holmås, Karin Monstad, and Odd Rune Straume. 2020. "How Does The Type of Remuneration Affect Physician Behavior?: Fixed Salary versus Fee-for-Service". In: *American Journal of Health Economics* 6 (1): 104–138. [Link].

Brown, Christina and Tahir Andrabi. 2023. *Inducing Positive Sorting through Performance Pay: Experimental Evidence from Pakistani Schools*. Tech. rep. Research on Improving Systems of Education (RISE). [Link].

Callaway, Brantly and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with multiple time periods". In: *Journal of Econometrics*. Themed Issue: Treatment Effect 1 225 (2): 200–230. [Link].

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors". In: *The Review of Economics and Statistics* 90 (3): 414–427. [Link].

Centers for Medicare and Medicaid Services (CMS). 2025. *Medicare Telehealth Trends — CMS Data*. [Link].

Cesarini, David, Erik Lindqvist, Robert Östling, and Björn Wallace. 2016. "Wealth, Health, and Child Development: Evidence from Administrative Data on Swedish Lottery Players *". In: *The Quarterly Journal of Economics* 131 (2): 687–738. [Link].

Chaisemartin, Clément de and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects". In: *American Economic Review* 110 (9): 2964–2996. [Link].

Chan, David C.. 2018. "The Efficiency of Slacking off: Evidence From the Emergency Department". In: *Econometrica* 86 (3): 997–1030. [Link].

Chandra, Amitabh, Jonathan Gruber, and Robin McKnight. 2010. "Patient Cost-Sharing and Hospitalization Offsets in the Elderly". In: *American Economic Review* 100 (1): 193–213. [Link].

Clemens, Jeffrey and Joshua D. Gottlieb. 2014. "Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?". In: *American Economic Review* 104 (4): 1320–1349. [Link].

Coviello, Decio, Andrea Ichino, and Nicola Persico. 2015. "The Inefficiency of Worker Time Use". In: *Journal of the European Economic Association* 13 (5): 906–947. [Link].

Dahlstrand, Amanda, Shan Huang, Nestor Le Nestour, and Guy Michaels. 2026. *Organizing Expertise with Externalities: Evidence from Primary Care*. CEPR Discussion Paper 21971. Paris & London: CEPR. [Link].

Dahlstrand, Amanda, Dávid Laszlo, Helena Schweiger, Oriana Bandiera, Andrea Prat, and Raffaella Sadun. 2025. *CEO-Firm Matches and Productivity in 42 Countries*. Working Paper 33324. National Bureau of Economic Research. [Link].

Dahlstrand, Amanda, Nestor Le Nestour, and Guy Michaels. 2024. *Online versus In-Person Services: Effects on Patients and Providers*. Discussion Paper No. 17177. Institute of Labor Economics. [Link].

Das, Jishnu, Alaka Holla, Aakash Mohpal, and Karthik Muralidharan. 2016. "Quality and Accountability in Health Care Delivery: Audit-Study Evidence from Primary Care in India". In: *American Economic Review* 106 (12): 3765–3799. [Link].

De Chaisemartin, Clément and Xavier D'Haultfœuille. 2026. "Difference-in-Differences Estimators of Intertemporal Treatment Effects". In: *Review of Economics and Statistics*, 1–18. [Link].

Deming, David J, Katrine V Løken, Alexander Willén, and Yaling Xu. 2026. *Meetings*. Working Paper 35706. National Bureau of Economic Research. [Link].

Dumont, Etienne, Bernard Fortin, Nicolas Jacquemet, and Bruce Shearer. 2008. "Physicians' multitasking and incentives: Empirical evidence from a natural experiment". In: *Journal of Health Economics* 27 (6): 1436–1450. [Link].

Échevin, Damien and Bernard Fortin. 2014. "Physician payment mechanisms, hospital length of stay and risk of readmission: Evidence from a natural experiment". In: *Journal of Health Economics* 36, 112–124. [Link].

Ellis, Randall P. and Thomas G. McGuire. 1990. "Optimal payment systems for health services". In: *Journal of Health Economics* 9 (4): 375–396. [Link].

Finkelstein, Amy, Nathaniel Hendren, and Erzo F. P. Luttmer. 2019. "The Value of Medicaid: Interpreting Results from the Oregon Health Insurance Experiment". In: *Journal of Political Economy* 127 (6): 2836–2874. [Link].

Gardner, John, Neil Thakral, Linh T Tô, and Luther Yap. 2026. "Two-Stage Differences in Differences".

Gaynor, Martin and Paul Gertler. 1995. "Moral Hazard and Risk Spreading in Partnerships". In: *The RAND Journal of Economics* 26 (4): 591–613. [Link].

Ginja, Rita, Julie Riise, Barton Willage, and Alexander Willén. 2025. "Does Your Doctor Matter?". In: *Journal of Political Economy Microeconomics* 3 (3): 497–538. [Link].

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives". In: *American Economic Journal: Applied Economics* 2 (3): 205–227. [Link].

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing". In: *Journal of Econometrics* 225 (2): 254–277. [Link].

Grieco, Paul L. E. and Ryan C. Mcdevitt. 2017. "Productivity and Quality in Health Care: Evidence from the Dialysis Industry". In: *The Review of Economic Studies* 84 (3 (300)): 1071–1105. [Link].

Gruber, Jonathan, Thomas P. Hoe, and George Stoye. 2023. "Saving Lives by Tying Hands: The Unexpected Effects of Constraining Health Care Providers". In: *The Review of Economics and Statistics* 105 (1): 1–19. [Link].

Holmstrom, Bengt and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design". In: *Journal of Law, Economics, & Organization* 7, 24–52. [Link].

Igwe, Apoorva Rama. 2026. *Physician Compensation Methods: Although Salary-Based Models Dominate, Productivity-Based Models Remain Prominent*. Policy Research Perspectives. Chicago: American Medical Association. [Link].

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. 1st ed. Cambridge University Press. [Link].

Knutsson, Daniel and Björn Tyrefors. 2022. "The Quality and Efficiency of Public and Private Firms: Evidence from Ambulance Services*". In: *The Quarterly Journal of Economics* 137 (4): 2213–2262. [Link].

Kroneman, Madelon, Pascal Meeus, Dionne Sofia Kringos, Wim Groot, and Jouke van der Zee. 2013. "International Developments in Revenues and Incomes of General Practitioners from 2000 to 2010". In: *BMC Health Services Research* 13, 436. [Link].

Lavy, Victor. 2002. "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement". In: *Journal of Political Economy* 110 (6): 1286–1317. [Link].

Lazear, Edward P. 2000. "Performance pay and productivity". In: *American Economic Review* 90 (5): 1346–1361.

Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin. 2021. "Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools". In: *The American Economic Review* 111 (7): 2213–2246. [Link].

Ludvigsson, Jonas F., David Bergman, Catharina Ihre Lundgren, Kristina Sundquist, Jean-Luc af Geijerstam, Anna H. Glenngård, Marie Lindh, Johan Sundström, Johan Kaarme, and Jialu Yao. 2025. "The healthcare system in Sweden". In: *European Journal of Epidemiology* 40 (5): 563–579. [Link].

McGuire, Thomas G.. 2000. "Chapter 9 Physician agency". In: *Handbook of Health Economics*. Vol. 1. Elsevier, pp. 461–536. [Link].

Mohanan, Manoj, Katherine Donato, Grant Miller, Yulya Truskinovsky, and Marcos Vera-Hernández. 2021. "Different Strokes for Different Folks? Experimental Evidence on the Effectiveness of Input and Output Incentive Contracts for Health Care Providers with Varying Skills". In: *American Economic Journal: Applied Economics* 13 (4): 34–69. [Link].

Mullen, Kathleen J., Richard G. Frank, and Meredith B. Rosenthal. 2010. "Can you get what you pay for? Pay-for-performance and the quality of healthcare providers". In: *The RAND Journal of Economics* 41 (1): 64–91. [Link].

Muralidharan, Karthik and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India". In: *Journal of Political Economy* 119 (1): 39–77. [Link].

Myndigheten för vård- och omsorgsanalys. 2021. *Vården ur befolkningens perspektiv 2020: En jämförelse mellan Sverige och tio andra länder*. Rapport 2021:4. Stockholm: Myndigheten för vård- och omsorgsanalys. [Link].

OECD. 2021. *The COVID-19 Pandemic and the Future of Telemedicine*. OECD. [Link].

OECD. 2025. *Health Statistics Data Explorer: Doctor Consultations*. [Link].

Page, Anthea, Sarah Ambrose, John Glover, and Diana Hetzel. 2007. *Atlas of Avoidable Hospitalisations in Australia: Ambulatory Care-Sensitive Conditions*. Adelaide: PHIDU, University of Adelaide.

Reif, Simon, Jan Köhler, Sabrina Schubert, and Benedikt Stelter. 2025. "Contact time in GP Care: Descriptive patterns and a scoping review of the literature". In: *Health Policy* 156, 105315. [Link].

Rodríguez-Lesmes, Paul and Marcos Vera-Hernandez. 2025. *Multitasking, two-part contracts, and bunching: an application to doctors' tasks*. Tech. rep. The IFS. [Link].

Shearer, Bruce. 2004. "Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment". In: *The Review of Economic Studies* 71 (2): 513–534. [Link].

Silver, David. 2021. "Haste or Waste? Peer Pressure and Productivity in the Emergency Department". In: *The Review of Economic Studies* 88 (3): 1385–1417. [Link].

Socialstyrelsen. 2026. *Nationella riktlinjer 2026: Könsdysfori. Bilaga: Fördjupad analys av kostnadseffektivitet*. Tech. rep. 2026-2-10054. Socialstyrelsen. [Link].

Strama. 2019. *Kvalitetsindikatorer för digitala vårdmöten*. Rekommendation. Stockholm: Stramas programråd.

Sveriges Riksbank. 2022. *Exchange rates – SEK to USD (daily rates 2022)*. Sveriges Riksbank statistics. [Link].

The Commonwealth Fund. 2020. *Sweden | International Health Care System Profiles*. [Link].

WHO Regional Office for Europe. 2025. *Health Workforce Migration in the WHO European Region: Country Case Studies from Albania, Armenia, Georgia, Ireland, Malta, Norway, Republic of Moldova, Romania and Tajikistan*. Report WHO/EURO:2025-10839-50611-76542. Copenhagen: World Health Organization, Regional Office for Europe. [Link].

Tables

Table 1. Characteristics of the analysis sample: symptoms and demographics

	Full sample		Doctor with fixed wages		Doctor with quantity incentives		Diff. p-value
Symptoms							
Acute infection (likely bacterial)							
Ear Problem	0.0027	[0.052]	0.0022	[0.047]	0.0038	[0.061]	0.000
Eye Infection	0.048	[0.21]	0.049	[0.22]	0.046	[0.21]	0.000
Sinusitis	0.021	[0.14]	0.022	[0.15]	0.020	[0.14]	0.006
Sore Throat	0.057	[0.23]	0.053	[0.22]	0.064	[0.25]	0.000
Urinary Tract Infection	0.031	[0.17]	0.030	[0.17]	0.033	[0.18]	0.000
Acute infection (likely viral)							
COVID-19	0.0097	[0.098]	0.012	[0.11]	0.0044	[0.066]	0.000
Chickenpox	0.0087	[0.093]	0.0090	[0.094]	0.0081	[0.090]	0.002
Cold Sores	0.013	[0.11]	0.014	[0.12]	0.011	[0.10]	0.000
Cold and Flu	0.021	[0.14]	0.022	[0.15]	0.020	[0.14]	0.000
Cough	0.10	[0.31]	0.100	[0.30]	0.11	[0.32]	0.000
Administrative and preventive care							
Medical Certificate	0.014	[0.12]	0.0091	[0.095]	0.024	[0.15]	0.000
Prescription Renewal	0.15	[0.36]	0.15	[0.36]	0.16	[0.37]	0.000
Return Visit	0.0053	[0.072]	0.0054	[0.073]	0.0051	[0.071]	0.136
Sick Leave	0.0022	[0.047]	0.0014	[0.037]	0.0039	[0.062]	0.000
Allergy and asthma							
Allergy	0.037	[0.19]	0.038	[0.19]	0.035	[0.18]	0.000
Asthma	0.0081	[0.090]	0.0086	[0.092]	0.0070	[0.083]	0.000
General and nonspecific conditions							
Abdominal Pain	0.023	[0.15]	0.023	[0.15]	0.023	[0.15]	0.264
Bites and Stings	0.017	[0.13]	0.018	[0.13]	0.014	[0.12]	0.000
Constipation	0.0026	[0.051]	0.0026	[0.050]	0.0028	[0.053]	0.111
Diarrhea or Vomiting	0.0042	[0.065]	0.0041	[0.064]	0.0043	[0.066]	0.297
Fever	0.0064	[0.080]	0.0058	[0.076]	0.0078	[0.088]	0.000
Headache	0.0062	[0.079]	0.0053	[0.073]	0.0082	[0.090]	0.000
Mental health							
Anxiety	0.0041	[0.064]	0.0036	[0.060]	0.0050	[0.070]	0.000
Depression	0.0039	[0.062]	0.0035	[0.059]	0.0048	[0.069]	0.000
Other Mental Health	0.0025	[0.050]	0.0026	[0.051]	0.0021	[0.046]	0.001
Sleep Disorder	0.011	[0.10]	0.012	[0.11]	0.0087	[0.093]	0.000
Stress	0.0063	[0.079]	0.0045	[0.067]	0.010	[0.100]	0.000
Other symptoms							
Other Care Process	0.042	[0.20]	0.037	[0.19]	0.052	[0.22]	0.000
Other Health Inquiry	0.069	[0.25]	0.078	[0.27]	0.052	[0.22]	0.000
Sexual and reproductive health							
Contraceptives	0.045	[0.21]	0.047	[0.21]	0.039	[0.19]	0.000
Erectile Dysfunction	0.0053	[0.072]	0.0057	[0.075]	0.0044	[0.066]	0.000
Menstruation/Menopause	0.0085	[0.092]	0.0095	[0.097]	0.0066	[0.081]	0.000
Vaginal Conditions	0.0030	[0.054]	0.0021	[0.045]	0.0048	[0.069]	0.000
Skin conditions							
Acne	0.014	[0.12]	0.016	[0.12]	0.010	[0.10]	0.000
Hair Loss	0.0036	[0.060]	0.0040	[0.064]	0.0027	[0.052]	0.000
Nail Problem	0.0042	[0.065]	0.0038	[0.061]	0.0050	[0.071]	0.000
Skin Rash	0.18	[0.39]	0.19	[0.39]	0.17	[0.38]	0.000
Demographics							
Female	0.60	[0.49]	0.60	[0.49]	0.60	[0.49]	0.203
Patient age	29.3	[16.6]	29.2	[16.6]	29.3	[16.7]	0.020
Observations	452,652		304,026		148,626		

Note: This table reports summary statistics for the analysis sample. *Symptoms* reports patients' main symptoms, grouped by clinical category. *Demographics* reports patients' gender and age. Appendix Table A2 reports region, health risk, and socioeconomic characteristics. Standard deviations are in square brackets. Columns two and three split the sample by the physician's quantity-incentives status in the month of the consultation. The final column reports p-values from t-tests comparing the unconditional means.

Table 2. Characteristics of the doctors in the analysis sample

	All doctors		Voluntary onset		Mandatory onset		Difference
	Mean	SD	Mean	SD	Mean	SD	p-value
Work per month (pre-rollout, 2022m1 - 2022m5)							
Consultations per hour worked	4.14	[0.80]	4.29	[0.69]	3.87	[0.92]	0.000
Video consultations	161.1	[169.6]	171.4	[163.9]	142.2	[179.0]	0.193
Consultations, all types	166.3	[177.1]	177.0	[171.7]	146.5	[185.8]	0.190
Text-based consultations	5.18	[11.2]	5.68	[12.7]	4.28	[7.64]	0.265
Minutes patient-facing	700.1	[667.7]	723.7	[653.2]	656.9	[694.9]	0.445
Minutes on administration	1334.3	[1427.6]	1426.0	[1497.4]	1166.5	[1280.8]	0.139
Minutes on break within shifts	272.1	[296.0]	289.6	[298.3]	240.2	[290.6]	0.190
Minutes worked	2306.6	[2310.5]	2439.4	[2356.6]	2063.6	[2215.5]	0.198
Demographics							
Female	0.44	[0.50]	0.43	[0.50]	0.47	[0.50]	0.556
Age in 2022	41.7	[11.6]	40.5	[10.7]	43.9	[12.9]	0.031
Training							
Specialist degree	0.48	[0.50]	0.48	[0.50]	0.48	[0.50]	0.953
In specialist training	0.32	[0.47]	0.34	[0.47]	0.30	[0.46]	0.510
Top medical school degree	0.11	[0.31]	0.093	[0.29]	0.13	[0.34]	0.401
Income in 2021[a]							
Number of employers	1.90	[0.96]	1.91	[0.91]	1.89	[1.05]	0.885
Gross wage income (SEK)	719,668	[372,764]	749,832	[368,872]	665,100	[375,638]	0.087
Wage income from the organization (SEK)	150,402	[298,474]	159,527	[321,420]	133,896	[252,603]	0.488
Doctors	266		172		94		

Note: This table reports summary statistics for physicians in the analysis sample on the physician-level. *Voluntary onset* includes physicians who voluntarily moved on to the quantity-incentives scheme for the first time between June 2022 and May 2023. *Mandatory onset* includes only physicians who moved onto it when the scheme became mandatory in June 2023. *Work per month* reports monthly averages over January–May 2022, before the rollout of the quantity-incentives scheme. All variables are defined in Appendix Table A7. Standard deviations are in square brackets. The final column reports p-values from unequal-variance t-tests comparing means between the two groups.

[a] Observed for 250 of the 266 doctors, 161 of them with voluntary onset.

Table 3. Effect of quantity incentives on within-organization outcomes

Panel A: Time inputs

	Total consultation time (minutes)	Administrative time (minutes)	Patient-facing time (minutes)
Doctor with quantity incentives	-2.44***	-2.14***	-0.30***
	(0.25)	(0.24)	(0.065)
Observations	452,652	452,652	452,652
#Doctors	266	266	266
Baseline mean	11.7	7.73	4.00
Full controls	✓	✓	✓

Panel B: Performance-relevant measures

	Top grade (5/5)	Worst grade (1/5)	Informative diagnosis	Any counter-guideline antibiotic (7 days)
Doctor with quantity incentives	-0.0024	-0.00056	-0.00090	-0.00058
	(0.0042)	(0.00072)	(0.0055)	(0.0010)
Observations	452,652	452,652	452,652	452,652
#Doctors	266	266	266	266
Baseline mean	0.35	0.011	0.73	0.014
Full controls	✓	✓	✓	✓

Panel C: Recorded outcomes (7 days)

	Any prescription by the provider	Any unscheduled return consultation	Any booked return consultation
Doctor with quantity incentives	0.011***	0.0026	0.00055
	(0.0041)	(0.0018)	(0.0018)
Observations	452,652	452,652	452,652
#Doctors	266	266	266
Baseline mean	0.51	0.064	0.044
Full controls	✓	✓	✓

Note: This table presents regression results for our main specification on the effects of quantity incentives on outcomes recorded by the organization. The analysis sample consists of drop-in video consultations (Appendix Table A1). Each column corresponds to a different outcome as defined in Appendix Table A7. The columns report the estimate for *Doctor with quantity incentives*, γ in Equation (4), using the two-stage estimator of Gardner et al. (2026), with the first stage estimated on consultations before each physician's first month on the quantity-incentives scheme. All regressions control for physician fixed effects, date-by-three-hour-window fixed effects, and patient controls X_i as described in the main text (*Full controls*). The baseline mean is the average outcome value for consultations whose physicians are not on the quantity-incentives scheme in that month. Standard errors clustered at the physician level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4. Effect of quantity incentives on downstream care outside the organization

Panel A: Primary care (7 days)

	Any external primary care consultation	Any external prescription	Any urgent care
Doctor with quantity incentives	-0.0029	-0.0023	0.00071*
	(0.0035)	(0.0027)	(0.00040)
Observations	181,994	452,652	452,652
#Doctors	266	266	266
Baseline mean	0.14	0.21	0.0039
Full controls	✓	✓	✓

Panel B: Secondary care (7 days)

	Any specialist visit	Any ED visit	Any hospitalization
Doctor with quantity incentives	-0.00088	-0.00063	0.0011***
	(0.00095)	(0.00083)	(0.00039)
Observations	452,652	452,652	452,652
#Doctors	266	266	266
Baseline mean	0.021	0.018	0.0038
Full controls	✓	✓	✓

Panel C: Mortality and downstream cost

	Death (1 year)	Downstream cost (SEK, 7 days)
Doctor with quantity incentives	0.000090	24.9*
	(0.00018)	(14.3)
Observations	292,255	452,652
#Doctors	266	266
Baseline mean	0.00073	753.5
Full controls	✓	✓

Note: This table presents regression results for our main specification on the effects of quantity incentives on the care patients receive outside the organization. The analysis sample consists of drop-in video consultations (Appendix Table A1). Each column corresponds to a different outcome as defined in Appendix Table A7. The columns report the estimate for *Doctor with quantity incentives*, γ in Equation (4), using the two-stage estimator of Gardner et al. (2026), with the first stage estimated on consultations before each physician's first month on the quantity-incentives scheme. All regressions control for physician fixed effects, date-by-three-hour-window fixed effects, and patient controls X_i as described in the main text (*Full controls*). External primary care is observed for patients registered in Scania or Stockholm. Mortality is measured for consultations at least one year after a patient's previous consultation. The baseline mean is the average outcome value for consultations whose physicians are not on the quantity-incentives scheme in that month. Standard errors clustered at the physician level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5. Effect of quantity incentives on physician labor

Panel A: Time use (per month)

	Total minutes worked	Patient-facing minutes	Administrative minutes	Break minutes
Doctor with quantity incentives	-139.2	61.7	-207.4***	6.60
	(103.8)	(38.6)	(61.1)	(13.2)
Observations	4,522	4,522	4,522	4,522
#Doctors	266	266	266	266
Baseline mean	2131.4	666.4	1215.8	249.2

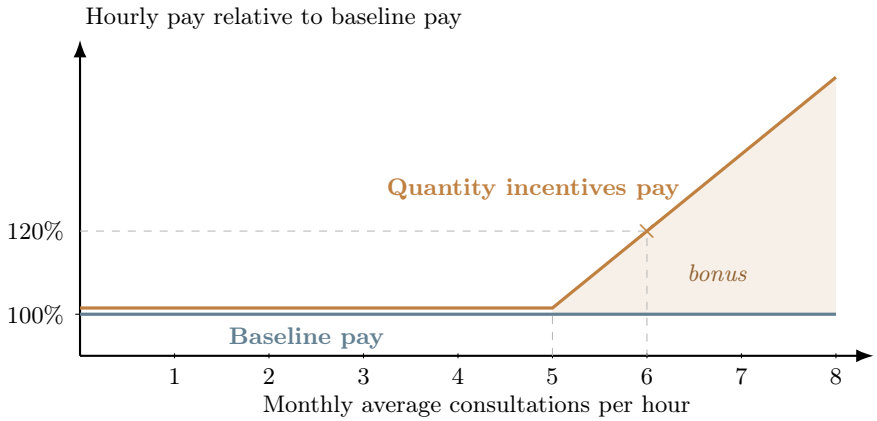
Panel B: Consultations (per month)

	Number of video consultations	Number of consultations	Number of text-based consultations	Consultations per worked hour
Doctor with quantity incentives	39.3***	40.0***	0.70	1.08***
	(10.4)	(11.0)	(1.28)	(0.095)
Observations	4,522	4,522	4,522	3,838
#Doctors	266	266	266	266
Baseline mean	151.7	157.7	6.03	4.28

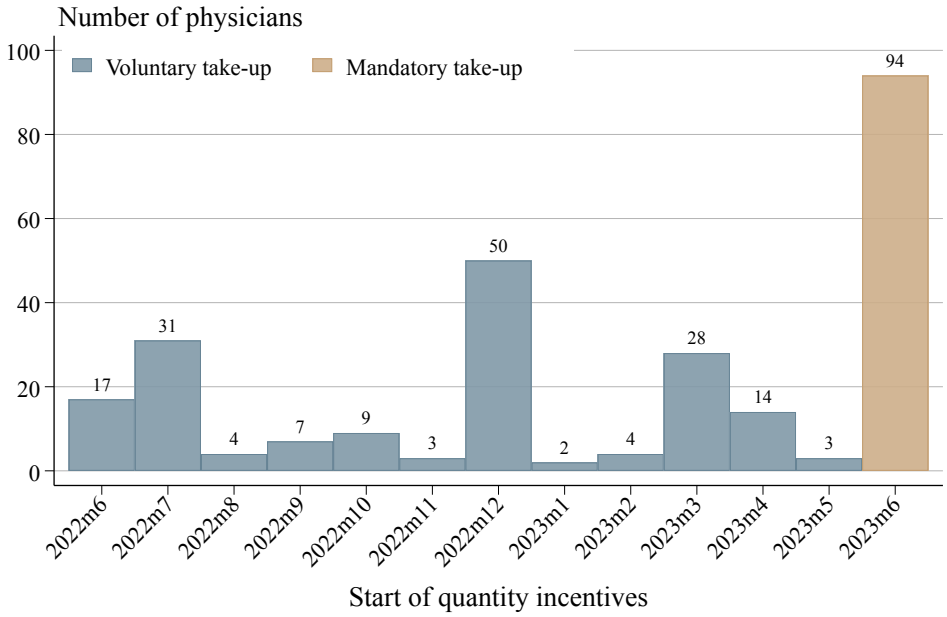
Note: This table presents regression results on the effects of quantity incentives on physicians' labor at the physician-month level. The sample contains every month from January 2022 to May 2023 in which a physician has an active contract with the organization. Months in which a physician has no online consultations enter as zero. Each column corresponds to a different outcome as defined in Appendix Table A7. The columns report the estimate for *Doctor with quantity incentives*, ϕ in Equation (5), using the two-stage estimator of Gardner et al. (2026), with the first stage estimated on physician-months before each physician's first month on the quantity-incentives scheme. All regressions control for physician fixed effects and calendar-month fixed effects. The baseline mean is the average outcome value for physician-months in which the physician is not on the quantity-incentives scheme. Consultations per worked hour is defined only for months with any hours of work. Standard errors clustered at the physician level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figures

Figure 1. The quantity-incentives scheme and its rollout



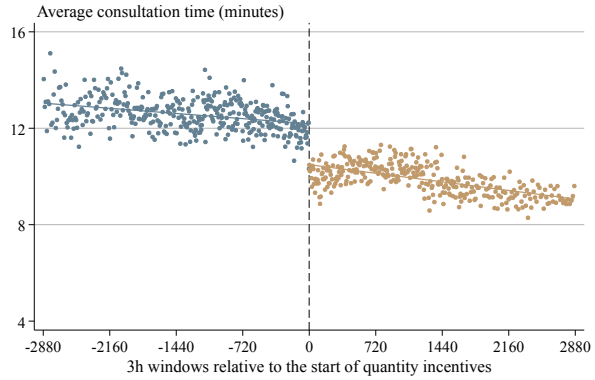
(a) Compensation under the quantity-incentives scheme



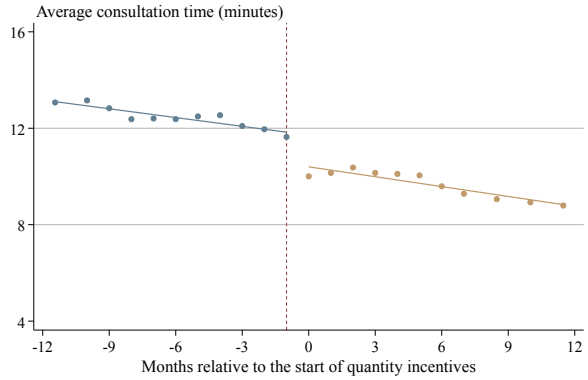
(b) Number of physicians by month of quantity-incentives scheme onset

Notes: Panel (a) shows hourly pay relative to the fixed baseline hourly wage, as a function of the physician's monthly average number of consultations per hour. Below a threshold of five consultations per hour, the physician receives the guaranteed baseline pay. Above the threshold, each additional video consultation per hour, averaged over a month, adds a piece rate (*bonus*) of 20% of the hourly wage. The baseline wage is physician-specific, so the piece rate differs across physicians. Panel (b) shows the number of physicians moving on to the quantity-incentives scheme for the first time in each calendar month. The shading marks the two phases of the rollout: the voluntary take-up period from June 2022 to May 2023; and the mandatory take-up in June 2023.

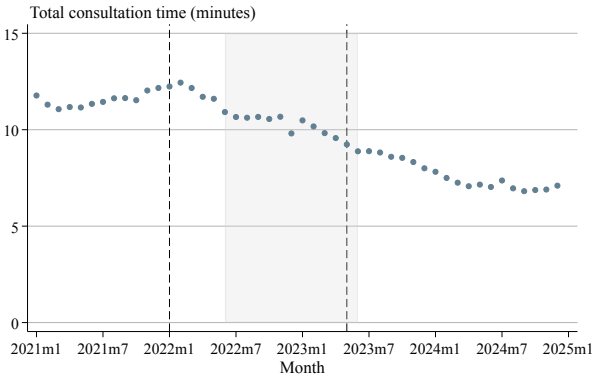
Figure 2. Average consultation time around quantity-incentives onset



(a) Three-hour windows



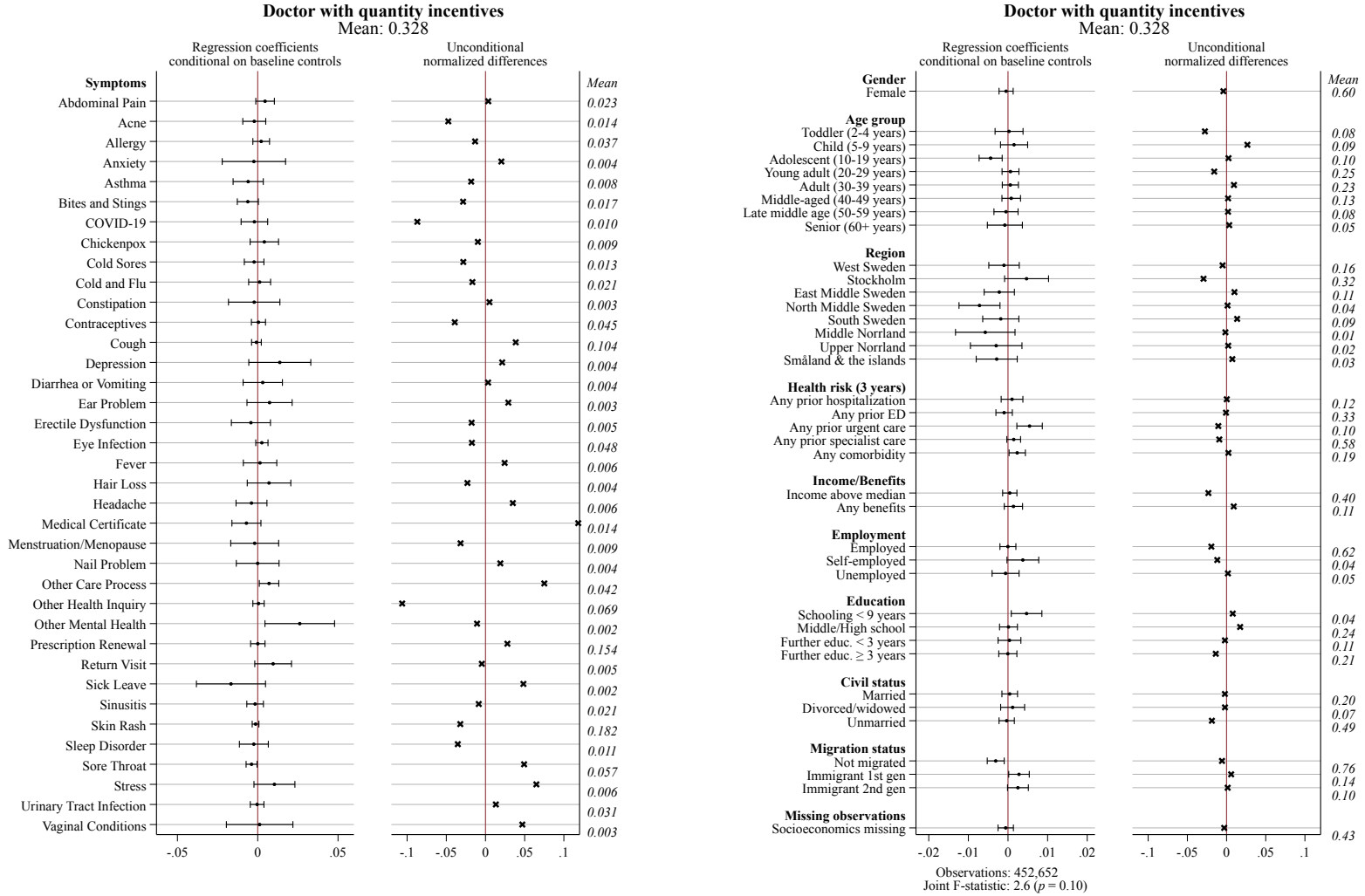
(b) Months



(c) Full data period

Notes: Panels (a) and (b) show binned means of the average total time per consultation (the sum of patient-facing and administrative time in minutes), against time relative to a physician's onset of quantity incentives. Solid lines show linear fits for the periods before onset and from onset on, and the dashed vertical lines mark the window or month before onset. Panel (a) shows one bin per three-hour time window, and Panel (b) shows one bin per month. Panel (c) shows average total consultation time by calendar month over the full data period, January 2021–December 2024. All panels include only consultations with the 266 physicians in our analysis sample, and the restrictions in Appendix Table A1 apply, except that panel (c) does not apply Restriction 7 (Estimation restrictions). The shaded band marks the rollout period, June 2022–June 2023, and the two dashed vertical lines mark the boundaries of the analysis window, January 2022 and May 2023.

Figure 3. Balance test on patient characteristics

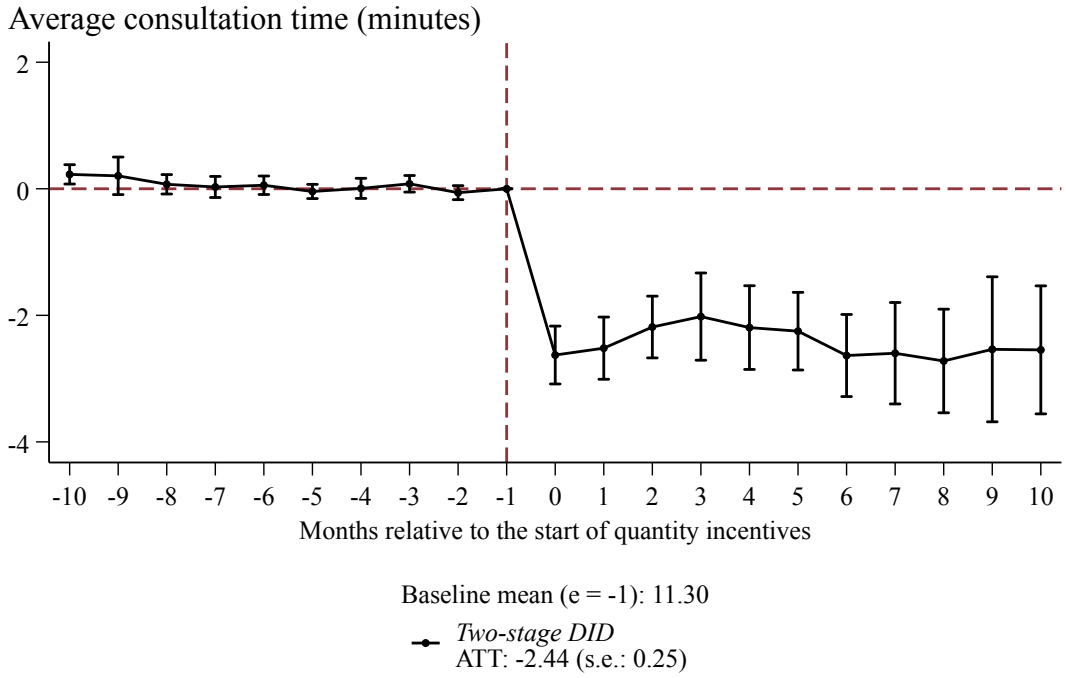


(a) Symptoms

(b) Demographic, health risk, and socioeconomic information

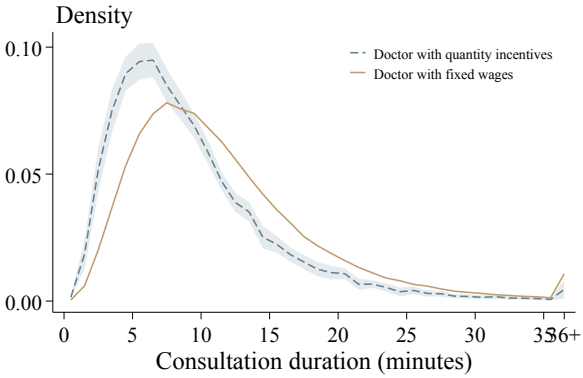
Notes: This figure presents balance tests for the quantity-incentives status of the physician on the consultation. Panel (a) reports results for patients' self-reported main symptoms. Panel (b) reports results for patient demographics, health risk, and socioeconomic characteristics. The left columns report, in each row, the coefficient from a separate regression of quantity-incentives status on the characteristic listed on the vertical axis, using physician and date-by-three-hour-window fixed effects. The F-statistic and p-value test whether all coefficients in the figure are jointly zero, conditional on the same fixed effects, with the p-value computed using a wild cluster bootstrap at the physician level, with Rademacher weights and 1000 replications (Cameron, Gelbach, and Miller 2008). The right columns report the corresponding unconditional normalized differences, where thresholds between 0.1 and 0.25 are commonly used to flag potentially problematic imbalance between groups (Imbens and Rubin 2015; Baker et al. 2026). Numbers on the right show sample means. All characteristics are defined in Appendix Table A7. Horizontal bars show 95% confidence intervals. Standard errors are clustered at the physician level.

Figure 4. Effect of quantity incentives on average consultation time

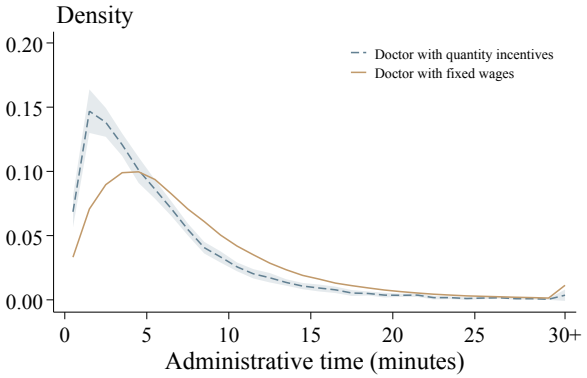


Notes: This figure reports event-study estimates of the effect of quantity incentives on total consultation time, using the two-stage estimator of Gardner et al. (2026) at the consultation level. The first stage estimates, for consultations before each physician's first month on the quantity-incentives scheme, physician and date-by-three-hour-window fixed effects. The second stage regresses total consultation time, residualized of the fixed effects, on indicators for months relative to the first month on the quantity-incentives scheme. The reference month is the month before onset, $e = -1$. Bars show 95% confidence intervals. Standard errors are clustered at the physician level.

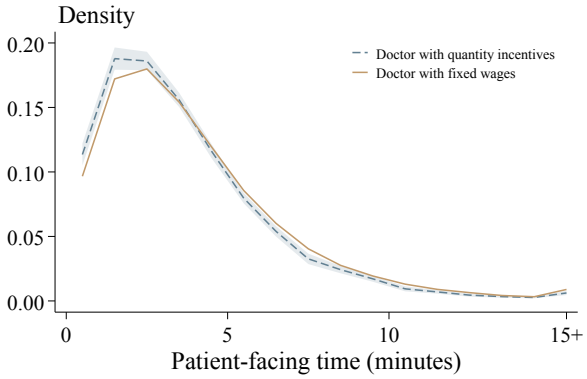
Figure 5. Distribution of consultation time



(a) Total consultation time



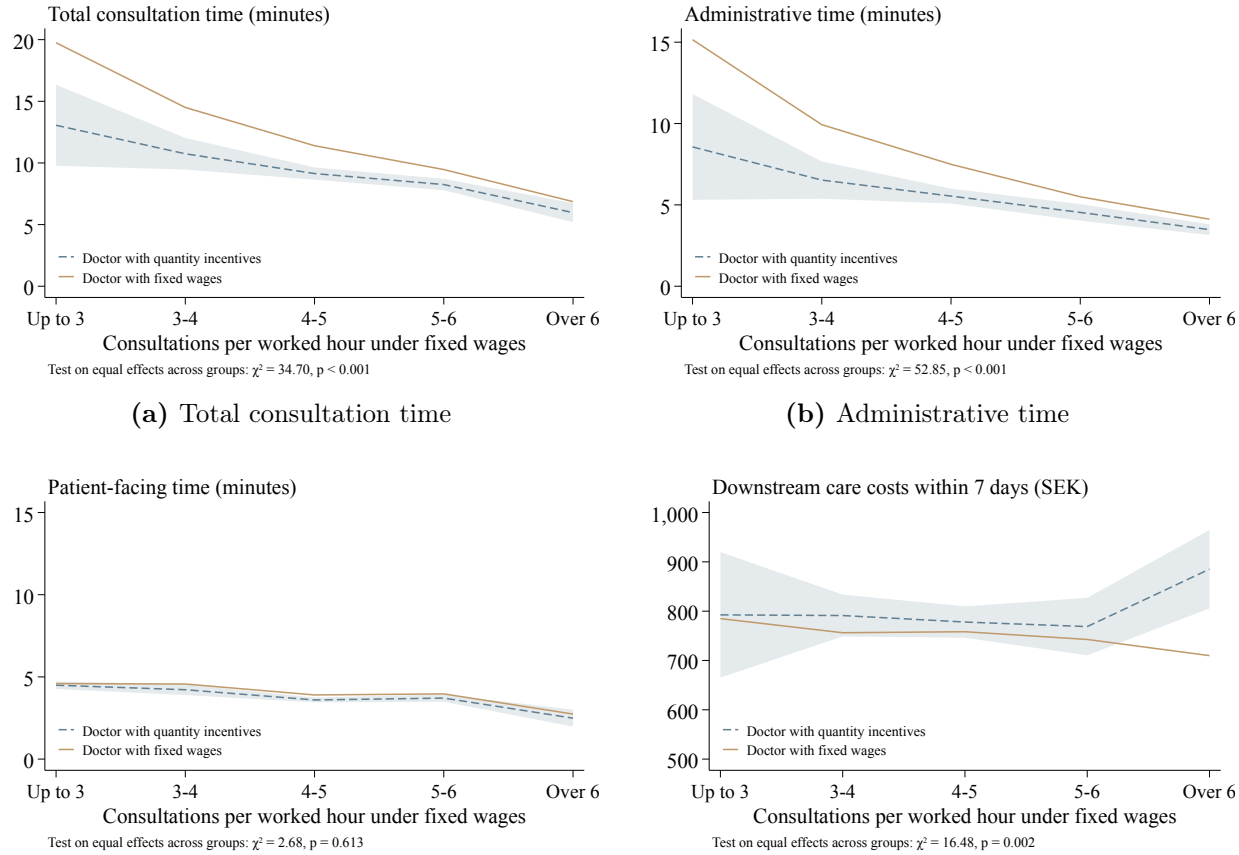
(b) Administrative time



(c) Patient-facing time

Notes: This figure shows consultation time distributions in one-minute bins under fixed wages (solid lines) and quantity incentives (dashed lines). Panels (a), (b), and (c) report total, administrative, and patient-facing time, respectively. Solid lines show the observed share of consultations in each bin under fixed wages. Dashed lines add the estimated effect of quantity incentives on each bin's share to the baseline shares under fixed wages. Each effect is estimated using the main specification (4) and the two-stage estimator of Gardner et al. (2026), with an indicator for consultation time falling in that bin as the outcome. Shaded bands show 95% confidence intervals. Standard errors are clustered at the physician level.

Figure 6. Consultation time under fixed and volume pay, by consultations per hour worked under fixed wages



(a) Total consultation time

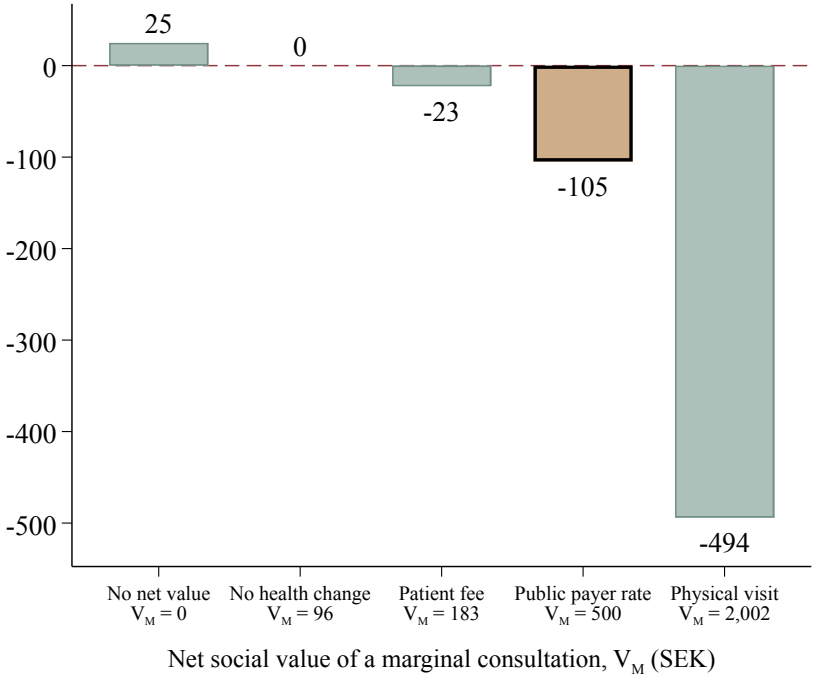
(b) Administrative time

(c) Patient-facing time

(d) Downstream costs

Notes: This figure reports consultation times and downstream costs under fixed wages (solid lines) and quantity incentives (dashed lines) by the physician's consultations per worked hour. Consultations per worked hour is the physician's average number of consultations divided by hours worked during the months before the physician's first month on the quantity-incentives scheme. The groups are up to three consultations per worked hour (2.3% of consultations), more than three and up to four (20.7%), more than four and up to five (58.1%), more than five and up to six (14.5%), and more than six (4.4%). Solid lines show observed means under fixed wages. Dashed lines add each group's estimated effect of quantity incentives to these means. The group effects are estimated using the two-stage estimator of Gardner et al. (2026) with physician and date-by-three-hour-window fixed effects and patient controls X_i in the first stage, and interactions between quantity-incentives status and group indicators in the second stage. The χ^2 statistics test whether these effect estimates are equal across groups. Panels (a), (b), and (c) report total, administrative, and patient-facing time, respectively. Panel (d) shows downstream cost within seven days. Shaded bands show 95% confidence intervals. Standard errors are clustered at the physician level.

Figure 7. Change in patient health value at which quantity incentives are welfare-neutral (SEK per baseline consultation)



Notes: This figure shows the change in patient health value (SEK per baseline consultation) at which switching physicians from fixed wages to quantity incentives is welfare-neutral. Positive values indicate the health gain needed to break even. Negative values indicate the largest health loss consistent with no welfare loss. Welfare increases if the unobserved health change, Δh, exceeds the threshold and decreases if it falls below. The threshold is $\Delta h^{BE} = \Delta D - (\Delta N/N_0)\, V_M$ (Equation 8), where ΔD is the estimated effect on downstream costs per consultation (Table 4, Panel C), $\Delta N/N_0$ is the estimated effect on monthly consultations divided by the baseline mean (Table 5, Panel B), and V_M is the net social value of an additional consultation. *No net value* sets $V_M = 0$. *No health change* sets V_M so that the switch is welfare-neutral without a health change ($\Delta h^{BE} = 0$). *Patient fee* sets $V_M = 183$ SEK, the average patient fee for an online consultation (250 SEK in Stockholm before May 2023, 200 SEK in Scania and Stockholm from May 2023, and 100 SEK in other regions). *Public payer rate* sets $V_M = 500$ SEK. *Physical visit* sets $V_M = 2{,}002$ SEK, the cost of a physical primary care consultation, assuming each additional consultation replaces such a physical visit with the same health benefit (Appendix Table A6).

Appendix

Appendix Tables

Table A1. Restrictions on the main analysis sample

	#Clinicians	#Consultations	
		Total	%Treated
All consultations from 2021-2024	2,458	5,797,962	–
1. Doctors in digital services (during January 2022-May 2022)	285	2,273,854	–
2. Online drop-in consultations between January 2022-May 2023	285	597,435	32.6
3. Patients aged 2 and over	285	580,202	32.7
4. Patient characteristics available	285	547,977	32.6
5. Index consultation (none by the same patient in the prior 7 days)	285	467,511	33.1
6. Symptom with at least 1000 consultations	285	456,320	32.7
7. Estimation restrictions	266	452,652	32.8
Analysis sample	**266**	**452,652**	**32.8**

Notes: This table presents the number of clinicians and consultations remaining after each sample restriction, and the share of consultations whose physician is on the quantity-incentives scheme in that month (*%Treated*). Restriction 1 limits the sample to physicians who worked in the organization's digital services between January and May 2022, before the rollout. Restriction 2 further limits the sample to online drop-in video consultations in the analysis window, January 2022 to May 2023, which spans the staggered rollout from June 2022 to May 2023; *%Treated* is reported from this row onwards. Restrictions 3 and 4 exclude infants (age below two years) and consultations with missing patient characteristics X_i (patient-reported symptom, gender, age, region). Restriction 5 keeps index consultations, that is, consultations with no prior consultation by the same patient in the preceding seven days. Restriction 6 excludes symptoms with fewer than 1,000 consultations. Restriction 7 imposes estimation restrictions by dropping singletons, three-hour time windows without any consultation with a physician not yet on the quantity-incentives scheme, and physicians who do not enter the physician-month estimations. The final row reports the analysis sample.

Table A2. Characteristics of the analysis sample: region, health risk and socioeconomic background

	Full sample		Doctor with fixed wages		Doctor with quantity incentives		Diff. p-value
Region							
West Sweden	0.16	[0.37]	0.16	[0.37]	0.16	[0.37]	0.109
Stockholm	0.32	[0.47]	0.32	[0.47]	0.31	[0.46]	0.000
East Middle Sweden	0.11	[0.31]	0.11	[0.31]	0.11	[0.32]	0.002
North Middle Sweden	0.035	[0.18]	0.035	[0.18]	0.036	[0.19]	0.683
South Sweden	0.089	[0.28]	0.087	[0.28]	0.091	[0.29]	0.000
Middle Norrland	0.013	[0.11]	0.013	[0.11]	0.013	[0.11]	0.651
Upper Norrland	0.016	[0.13]	0.016	[0.13]	0.016	[0.13]	0.497
Småland & the islands	0.031	[0.17]	0.031	[0.17]	0.032	[0.18]	0.020
Health risk factors							
Any prior hospitalization (3 years)	0.12	[0.32]	0.12	[0.32]	0.12	[0.32]	0.927
Any prior ED (3 years)	0.33	[0.47]	0.33	[0.47]	0.33	[0.47]	0.819
Any prior urgent care (3 years)	0.097	[0.30]	0.098	[0.30]	0.095	[0.29]	0.001
Any prior specialist care (3 years)	0.58	[0.49]	0.58	[0.49]	0.57	[0.49]	0.004
Any comorbidity (3 years)	0.19	[0.39]	0.19	[0.39]	0.19	[0.39]	0.450
Socioeconomic characteristics							
Income above median	0.40	[0.49]	0.40	[0.49]	0.39	[0.49]	0.000
Any benefits	0.11	[0.31]	0.11	[0.31]	0.11	[0.31]	0.004
Employed	0.62	[0.49]	0.62	[0.49]	0.61	[0.49]	0.000
Self-employed	0.043	[0.20]	0.044	[0.20]	0.041	[0.20]	0.000
Unemployed	0.053	[0.22]	0.053	[0.22]	0.053	[0.22]	0.573
Schooling $<$ 9 years	0.043	[0.20]	0.043	[0.20]	0.044	[0.21]	0.012
Middle/High school	0.24	[0.43]	0.23	[0.42]	0.24	[0.43]	0.000
Further educ. $<$ 3 years	0.11	[0.32]	0.11	[0.32]	0.11	[0.32]	0.513
Further educ. $\geq$ 3 years	0.21	[0.41]	0.21	[0.41]	0.21	[0.41]	0.000
Married	0.20	[0.40]	0.20	[0.40]	0.20	[0.40]	0.501
Divorced/widowed	0.072	[0.26]	0.072	[0.26]	0.072	[0.26]	0.526
Unmarried	0.49	[0.50]	0.49	[0.50]	0.48	[0.50]	0.000
Not migrated	0.76	[0.43]	0.76	[0.43]	0.75	[0.43]	0.067
Immigrant 1st gen	0.14	[0.35]	0.14	[0.35]	0.14	[0.35]	0.063
Immigrant 2nd gen	0.10	[0.30]	0.10	[0.30]	0.10	[0.30]	0.649
Socioeconomics missing	0.43	[0.50]	0.43	[0.50]	0.43	[0.49]	0.342
Observations	452,652		304,026		148,626		

Note: This table reports summary statistics for the analysis sample. *Region* reports the patient's region of residence. *Health risk factors* report any hospitalization, emergency department visit, urgent care use, specialist care use, and any comorbidities diagnosed in the three years before the consultation, excluding the day of the consultation and the day before. *Socioeconomic characteristics* report earnings above the sample median of 275,100 SEK among patients aged 20 or older, benefit receipt, employment status, education, civil status, and migration background. Table 1 reports symptoms and demographics. All characteristics are defined in Table A7. Standard deviations are in square brackets. Columns two and three split the sample by the physician's quantity-incentives status in the month of the consultation. The final column reports p-values from t-tests comparing the unconditional means.

Table A3. Effect of quantity incentives on additional outcomes

Panel A: Patient satisfaction

	Grade, if any
Doctor with quantity incentives	0.0066
	(0.0086)
Observations	186,694
#Doctors	266
Baseline mean	4.74
Full controls	✓

Panel B: Prescriptions

	Any new external prescription (7 days)	Any new provider prescription (7 days)	Any mental health drug prescription (7 days)	Any opioid prescription (7 days)	Any cough medication prescription (7 days)
Doctor with quantity incentives	-0.0051**	0.0076**	0.0021*	0.000034	0.0024
	(0.0022)	(0.0037)	(0.0012)	(0.000054)	(0.0018)
Observations	452,652	452,652	452,652	452,652	452,652
#Doctors	266	266	266	266	266
Baseline mean	0.14	0.39	0.051	0.000089	0.092
Full controls	✓	✓	✓	✓	✓

Panel C: Antibiotics

	Any antibiotic (7 days)	Any broad-spectrum antibiotic (7 days)	Any rare antibiotic (7 days)
Doctor with quantity incentives	0.00066	0.000049	-0.0013*
	(0.0018)	(0.00031)	(0.00076)
Observations	452,652	452,652	452,652
#Doctors	266	266	266
Baseline mean	0.097	0.0029	0.019
Full controls	✓	✓	✓

Panel D: Secondary care

	Any specialist visit (30 days)	Any ED visit (30 days)	Any hospitalization (30 days)	Any avoidable hospitalization (7 days)	Any avoidable hospitalization (30 days)
Doctor with quantity incentives	0.00041	-0.00079	0.0019***	0.00020	0.00045**
	(0.0016)	(0.0012)	(0.00054)	(0.00017)	(0.00022)
Observations	425,051	425,051	425,051	452,652	425,051
#Doctors	266	266	266	266	266
Baseline mean	0.071	0.032	0.0079	0.00057	0.0011
Full controls	✓	✓	✓	✓	✓

Panel E: Other follow-up consultations

	Any psychologist consultation (7 days)	Any in-person consultation (7 days)	Any external primary or community care (7 days)	Any psychologist consultation (30 days)	Any in-person consultation (30 days)
Doctor with quantity incentives	0.00071	0.00055	-0.0011	0.0012*	0.000055
	(0.00050)	(0.00062)	(0.0039)	(0.00068)	(0.00097)
Observations	452,652	452,652	181,994	425,051	425,051
#Doctors	266	266	266	266	266
Baseline mean	0.0061	0.0096	0.18	0.010	0.021
Full controls	✓	✓	✓	✓	✓

Note: This table presents regression results for our main specification on the effects of quantity incentives on additional outcomes. Each column corresponds to a different outcome as defined in Appendix Table A7. The columns report the estimate for *Doctor with quantity incentives*, γ in Equation (4). The patient's grade in Panel A is observed only for consultations that patients rate. For outcomes measured within thirty days, we only consider consultations that take place more than thirty days after the patient's previous consultation. External primary care is observed only for patients registered in Scania or Stockholm. The baseline mean is the average outcome value for consultations whose physicians are not on the quantity-incentives scheme in that month. Standard errors clustered at the physician level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4. Effect of quantity incentives, robustness checks with alternative estimations

Care at the provider

	Total consultation time (minutes)	Administrative time (minutes)	Patient-facing time (minutes)	Top grade	Worst grade	Informative diagnosis	Any counter-guideline antibiotic (7 days)	Any prescription (7 days)	Any drop-in follow-up (7 days)	Any booked follow-up (7 days)
Panel A: Main specification										
Doctor with quantity incentives	-2.44***	-2.14***	-0.30***	-0.0024	-0.00056	-0.00090	-0.00058	0.011***	0.0026	0.00055
	(0.25)	(0.24)	(0.065)	(0.0042)	(0.00072)	(0.0055)	(0.0010)	(0.0041)	(0.0018)	(0.0018)
Panel B: Time and physician fixed effects only										
Doctor with quantity incentives	-2.44***	-2.14***	-0.30***	-0.0017	-0.00051	-0.0010	-0.00097	0.013***	0.0025	0.0013
	(0.25)	(0.24)	(0.065)	(0.0042)	(0.00073)	(0.0055)	(0.0011)	(0.0047)	(0.0018)	(0.0019)
Panel C: Time and physician fixed effects + patient controls + additional patient characteristics										
Doctor with quantity incentives	-2.44***	-2.14***	-0.30***	-0.0019	-0.00057	-0.00076	-0.00057	0.011***	0.0025	0.00047
	(0.25)	(0.24)	(0.065)	(0.0042)	(0.00072)	(0.0055)	(0.0010)	(0.0041)	(0.0018)	(0.0018)
Panel D: Time and physician fixed effects + patient controls + additional patient characteristics + share of other doctors with quantity incentives										
Doctor with quantity incentives	-2.46***	-2.15***	-0.31***	-0.0015	-0.00051	-0.00018	-0.00033	0.012***	0.0027	0.00032
	(0.25)	(0.24)	(0.066)	(0.0043)	(0.00074)	(0.0055)	(0.0010)	(0.0042)	(0.0018)	(0.0017)
Panel E: TWFE OLS										
Doctor with quantity incentives	-2.16***	-1.88***	-0.28***	-0.0033	-0.00067	-0.0036	-0.0013	0.0067**	0.0022	0.00031
	(0.21)	(0.19)	(0.055)	(0.0036)	(0.00062)	(0.0049)	(0.00097)	(0.0034)	(0.0014)	(0.0014)
Observations	452,652	452,652	452,652	452,652	452,652	452,652	452,652	452,652	452,652	452,652
Doctors	266	266	266	266	266	266	266	266	266	266
Baseline mean	11.7	7.73	4.00	0.35	0.011	0.73	0.014	0.51	0.064	0.044

Downstream care

	Any external primary care (7 days)	Any external prescription (7 days)	Any urgent care (7 days)	Any specialist care (7 days)	Any ED (7 days)	Any hospitalization (7 days)	Death (1 year)	Downstream cost (SEK, 7 days)
Panel A: Main specification								
Doctor with quantity incentives	-0.0029	-0.0023	0.00071*	-0.00088	-0.00063	0.0011***	0.000090	24.9*
	(0.0035)	(0.0027)	(0.00040)	(0.00095)	(0.00083)	(0.00039)	(0.00018)	(14.3)
Panel B: Time and physician fixed effects only								
Doctor with quantity incentives	-0.0018	-0.0020	0.00069*	-0.00067	-0.00068	0.0011***	0.000094	25.8*
	(0.0036)	(0.0028)	(0.00041)	(0.00097)	(0.00083)	(0.00039)	(0.00019)	(14.1)
Panel C: Time and physician fixed effects + patient controls + additional patient characteristics								
Doctor with quantity incentives	-0.0028	-0.0028	0.00064	-0.0010	-0.00066	0.0011***	0.000094	23.0
	(0.0035)	(0.0027)	(0.00040)	(0.00093)	(0.00083)	(0.00039)	(0.00018)	(14.3)
Panel D: Time and physician fixed effects + patient controls + additional patient characteristics + share of other doctors with quantity incentives								
Doctor with quantity incentives	-0.0039	-0.0032	0.00056	-0.00095	-0.00050	0.0010**	0.000044	22.1
	(0.0035)	(0.0028)	(0.00041)	(0.00094)	(0.00084)	(0.00040)	(0.00019)	(14.7)
Panel E: TWFE OLS								
Doctor with quantity incentives	0.000032	0.0010	0.00046	-0.00043	-0.00020	0.00078**	-0.000044	22.7*
	(0.0031)	(0.0022)	(0.00037)	(0.00073)	(0.00073)	(0.00036)	(0.00015)	(13.3)
Observations	181,917	452,652	452,652	452,652	452,652	452,652	292,223	452,652
Doctors	266	266	266	266	266	266	266	266
Baseline mean	0.14	0.21	0.0039	0.021	0.018	0.0038	0.00073	753.5

Note: This table presents regression results on the effects of quantity incentives under alternative specifications for the main analysis sample. Each column corresponds to a different outcome as defined in Appendix Table A7. The upper block reports the same outcomes as Table 3 and the lower block the same as Table 4. Panels A to D use the two-stage estimator of Gardner et al. (2026). Panel A repeats the main specification, with physician fixed effects, date-by-three-hour-window fixed effects, and patient controls X_i. Panel B drops the patient controls X_i. Panel C adds the additional patient characteristics on health risk and socioeconomic background. Panel D further adds the share of the other physicians who started a consultation in the preceding hour and were on the quantity-incentives scheme. Panel E estimates the main specification by ordinary least squares with two-way fixed effects. The baseline mean is the average outcome value for consultations whose physicians are not on the quantity-incentives scheme in that month. Standard errors clustered at the physician level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5. Effect of quantity incentives, robustness checks on alternative samples

Care at the provider

	Total consultation time (minutes)	Administrative time (minutes)	Patient-facing time (minutes)	Top grade	Worst grade	Informative diagnosis	Any counter-guideline antibiotic (7 days)	Any prescription (7 days)	Any drop-in follow-up (7 days)	Any booked follow-up (7 days)
Panel A: First consultation per patient only										
Doctor with quantity incentives	-2.47***	-2.17***	-0.30***	0.00079	-0.00085	-0.0029	-0.000094	0.013***	0.0024	-0.0011
	(0.25)	(0.24)	(0.064)	(0.0047)	(0.00085)	(0.0058)	(0.0012)	(0.0044)	(0.0019)	(0.0018)
Observations	343,022	343,022	343,022	343,022	343,022	343,022	343,022	343,022	343,022	343,022
Doctors	266	266	266	266	266	266	266	266	266	266
Baseline mean	11.8	7.79	4.04	0.35	0.011	0.74	0.014	0.50	0.059	0.043
Panel B: All ages (infant restriction removed)										
Doctor with quantity incentives	-2.43***	-2.12***	-0.30***	-0.0022	-0.00044	-0.0017	-0.00060	0.011***	0.0027	0.00039
	(0.25)	(0.24)	(0.064)	(0.0042)	(0.00071)	(0.0054)	(0.0010)	(0.0041)	(0.0018)	(0.0017)
Observations	469,369	469,369	469,369	469,369	469,369	469,369	469,369	469,369	469,369	469,369
Doctors	285	285	285	285	285	285	285	285	285	285
Baseline mean	11.8	7.74	4.01	0.35	0.011	0.73	0.013	0.50	0.064	0.045
Panel C: Dropping data after relative month +8										
Doctor with quantity incentives	-2.42***	-2.10***	-0.32***	-0.0029	-0.00073	-0.0012	-0.00080	0.011***	0.0022	0.00041
	(0.24)	(0.22)	(0.063)	(0.0041)	(0.00072)	(0.0052)	(0.0010)	(0.0040)	(0.0017)	(0.0016)
Observations	437,949	437,949	437,949	437,949	437,949	437,949	437,949	437,949	437,949	437,949
Doctors	266	266	266	266	266	266	266	266	266	266
Baseline mean	11.7	7.73	4.00	0.35	0.011	0.73	0.014	0.51	0.064	0.044
Panel D: Restricted to post-Covid period (consultations from March 2022)										
Doctor with quantity incentives	-2.53***	-2.22***	-0.31***	-0.0036	-0.00063	-0.00031	-0.00018	0.010**	0.0011	-0.00024
	(0.27)	(0.25)	(0.068)	(0.0040)	(0.00082)	(0.0055)	(0.0010)	(0.0043)	(0.0018)	(0.0018)
Observations	406,833	406,833	406,833	406,833	406,833	406,833	406,833	406,833	406,833	406,833
Doctors	266	266	266	266	266	266	266	266	266	266
Baseline mean	11.6	7.61	4.01	0.36	0.012	0.74	0.014	0.51	0.063	0.044
Panel E: Restricted to consultations until March 2023 (administrative AI rollout)										
Doctor with quantity incentives	-2.58***	-2.26***	-0.33***	-0.0021	-0.0012	-0.0042	-0.00035	0.011***	0.0033*	-0.00016
	(0.27)	(0.26)	(0.070)	(0.0047)	(0.00077)	(0.0057)	(0.0012)	(0.0043)	(0.0019)	(0.0018)
Observations	401,351	401,351	401,351	401,351	401,351	401,351	401,351	401,351	401,351	401,351
Doctors	266	266	266	266	266	266	266	266	266	266
Baseline mean	11.8	7.77	4.00	0.35	0.011	0.73	0.014	0.51	0.064	0.045
Panel F: Excluding clinicians who ever switch off quantity incentives										
Doctor with quantity incentives	-2.49***	-2.18***	-0.31***	-0.0030	-0.00049	0.00080	-0.00027	0.010**	0.0030	0.00084
	(0.26)	(0.24)	(0.067)	(0.0043)	(0.00074)	(0.0055)	(0.0011)	(0.0041)	(0.0019)	(0.0018)
Observations	427,984	427,984	427,984	427,984	427,984	427,984	427,984	427,984	427,984	427,984
Doctors	248	248	248	248	248	248	248	248	248	248
Baseline mean	11.7	7.72	3.99	0.35	0.011	0.74	0.014	0.51	0.063	0.044

Downstream care

	Any external primary care (7 days)	Any external prescription (7 days)	Any urgent care (7 days)	Any specialist care (7 days)	Any ED (7 days)	Any hospitalization (7 days)	Death (1 year)	Downstream cost (SEK, 7 days)
Panel A: First consultation per patient only								
Doctor with quantity incentives	-0.0073*	-0.0047	0.00066	-0.0013	-0.0012	0.0014***	0.000095	27.7*
	(0.0041)	(0.0029)	(0.00043)	(0.0010)	(0.00096)	(0.00044)	(0.00019)	(16.0)
Observations	133,345	343,022	343,022	343,022	343,022	343,022	273,546	343,022
Doctors	266	266	266	266	266	266	266	266
Baseline mean	0.15	0.20	0.0039	0.020	0.018	0.0038	0.00072	745.6
Panel B: All ages (infant restriction removed)								
Doctor with quantity incentives	-0.0028	-0.0024	0.00084**	-0.00097	-0.00059	0.0011***	0.000077	26.8*
	(0.0035)	(0.0027)	(0.00043)	(0.00094)	(0.00084)	(0.00040)	(0.00018)	(14.6)
Observations	183,182	469,369	469,369	469,369	469,369	469,369	303,037	469,369
Doctors	283	285	285	285	285	285	285	285
Baseline mean	0.14	0.20	0.0048	0.021	0.019	0.0041	0.00070	764.9
Panel C: Dropping data after relative month +8								
Doctor with quantity incentives	-0.0025	-0.0018	0.00066*	-0.00089	-0.00061	0.00086**	0.000080	19.6
	(0.0035)	(0.0026)	(0.00038)	(0.00094)	(0.00083)	(0.00037)	(0.00017)	(14.1)
Observations	176,506	437,949	437,949	437,949	437,949	437,949	282,676	437,949
Doctors	266	266	266	266	266	266	266	266
Baseline mean	0.14	0.21	0.0039	0.021	0.018	0.0038	0.00073	753.5
Panel D: Restricted to post-Covid period (consultations from March 2022)								
Doctor with quantity incentives	0.00075	-0.0035	0.00043	-0.00061	-0.00081	0.00081*	0.00012	15.0
	(0.0036)	(0.0031)	(0.00041)	(0.0010)	(0.00090)	(0.00042)	(0.00021)	(15.6)
Observations	163,208	406,833	406,833	406,833	406,833	406,833	264,995	406,833
Doctors	266	266	266	266	266	266	266	266
Baseline mean	0.14	0.21	0.0039	0.021	0.018	0.0037	0.00075	752.4
Panel E: Restricted to consultations until March 2023 (administrative AI rollout)								
Doctor with quantity incentives	-0.00078	-0.0018	0.00058	-0.0011	-0.00042	0.00071*	0.00011	15.6
	(0.0039)	(0.0029)	(0.00040)	(0.0010)	(0.00085)	(0.00039)	(0.00019)	(14.7)
Observations	162,558	401,351	401,351	401,351	401,351	401,351	258,623	401,351
Doctors	266	266	266	266	266	266	266	266
Baseline mean	0.14	0.20	0.0039	0.021	0.018	0.0038	0.00073	750.3
Panel F: Excluding clinicians who ever switch off quantity incentives								
Doctor with quantity incentives	-0.0039	-0.0015	0.00063	-0.00083	-0.00083	0.0010**	0.000026	22.4
	(0.0036)	(0.0027)	(0.00040)	(0.00098)	(0.00086)	(0.00040)	(0.00019)	(14.7)
Observations	172,337	427,984	427,984	427,984	427,984	427,984	276,011	427,984
Doctors	248	248	248	248	248	248	248	248
Baseline mean	0.14	0.20	0.0039	0.021	0.018	0.0039	0.00072	750.8

Note: This table presents regression results for our main specification on the effects of quantity incentives on alternative samples. Each column corresponds to a different outcome as defined in Appendix Table A7. The upper block reports the same outcomes as Table 3 and the lower block the same as Table 4. Each panel changes one restriction of the main analysis sample while keeping all others fixed. Panel A keeps only each patient's first consultation in the analysis sample. Panel B includes consultations with infants. Panel C drops consultations more than eight months after the physician's first month on the quantity-incentives scheme. Panel D starts the analysis window in March 2022 to exclude the Covid-19 months. Panel E ends the analysis window in March 2023, before the organization began piloting an AI tool for administrative tasks. Panel F drops the 18 physicians who ever switch off the quantity-incentives scheme. The baseline mean is the average outcome value for consultations whose physicians are not on the quantity-incentives scheme in that month. Standard errors clustered at the physician level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6. Costs of healthcare services covered by Sweden's public healthcare system

Service	Cost in SEK	Year	Source	Link
Return consultation	500	2022–2023	The recommended payment from the public payer to healthcare providers for digital healthcare services is 500 SEK. We use this payment rate for all return consultations at the provider.	Sveriges Kommuner och Regioner (SKR)[1] Last access: 19 May 2026
Prescription	371.23	2022	Average total cost of a prescription including patient co-pays. Average computed over all prescriptions to patients with a Kry visit in 2022, excluding infants (patients below age 2).	Sveriges Kommuner och Regioner (SKR)[1] Last access: 19 May 2026
External physical primary care consultation	2,002	2019–2020	Average total cost of a physical doctor visit for 2019–2020 (Region Östergötland estimate).	Myndigheten för vård- och omsorgsanalys[2] Last access: 19 May 2026
Specialist visit	3,930.68	2019	Median cost of a doctor's visit across various specialties in Northern Sweden: internal medicine (excluding cardiology), rheumatology, pulmonology, infectious diseases, dermatology, surgery, oncology, urology, orthopedics, ophthalmology, otolaryngology, pediatrics, gynecology, obstetrics, rehabilitation, psychiatry, and child psychiatry.	Own calculation, based on information from Norra sjukvårdsregionförbundet (NRF)[3] Last access: 19 May 2026
Emergency department visit	3,991.5	2019–2020	Average costs in Southern Sweden, $(3,963 \text{ SEK} + 4,020 \text{ SEK}) \times 0.5 = 3,991.5$ SEK.	Södra Sjukvårdsregionen[4] Last access: 5 Nov 2024
Urgent care center visit	2,002	2019–2020 (information from 2024)	Information from 2024 states that "an appointment with a doctor at a vårdcentralen (healthcare centre) or a visit to the närakuten (urgent care centre) costs SEK 2,093." We thus assume that an urgent care visit has the same costs as a primary doctor care visit and map this to the 2019–2020 cost levels, when a doctor visit cost 2,002 SEK.	Region Stockholm (1177)[5] Last access: 5 Nov 2024
Hospitalization	28,487.5	2020	Hospitalization costs consist of four components: the one-time admission fee (Intagning, vårdavdelning), admission costs including physician consultations (Intagning, läkarinsats), daily medical care costs including physician consultations (Läkarinsats per vårddag), and daily nursing care costs (Omvårdnadsdag). Given an average length of stay of three days, all costs except the initial admission-related expenses are multiplied by three. Because care intensity differs across clinical specialties, the final cost is defined as the median total cost for a three-day hospitalization across acute care and internal medicine, pediatric surgery and neonatal care, pediatrics, endocrinology, dermatology, ophthalmology, hematology, oncology, cardiology, pulmonology, infectious diseases, gastroenterology, surgery, obstetrics and gynecology, neurosurgery, geriatrics, neurology, rehabilitation medicine, nephrology, rheumatology, orthopedics, hand surgery, plastic surgery, and ear, nose, and throat.	Own calculation, based on information from Södra Sjukvårdsregionen[6] Last access: 5 Nov 2024

[1] Full source: https://skr.se/download/18.546bb11919c226dc53431aab/1770189157692/Digital-utomlansvard-Rekommendation-SKR-2026.pdf
[2] Full source: https://www.vardanalys.se/rapporter/besok-via-natet/
[3] Full source: https://www.norrasjukvardsregionforbundet.se/halso-och-sjukvard/avtal-och-priser/arkiv/
[4] Full source: https://sodrasjukvardsregionen.se/verksamhet/avtal-priser/regionala-priser-och-ersattningar-foregaende-ar/
[5] Full source: https://www.1177.se/en/Stockholm/other-languages/other-languages/soka-vard/hitta-ratt-vard-nara-dig—andra-sprak—stockholms-lan/
[6] Full source: https://sodrasjukvardsregionen.se/verksamhet/regionala-priser-och-ersattningar/regionala-priser-och-ersattningar-foregaende-ar/

Table A7. Description of key variables

Variable	Description
Treatment	
Doctor with quantity incentives (Q_{jt})	Indicates whether physician j is compensated under the quantity-incentives scheme in calendar month t, taken from the provider's monthly list of physicians on the scheme (June 2022–May 2023). The list is missing for employee physicians in January 2023, for whom we impute the status from the month prior and after. Single-month gaps between months on the scheme are filled. Treatment onset is the first month on the scheme. (Source: Provider)
Fixed effects	
Physician fixed effects	Set of indicator variables for the physician handling the consultation. (Source: Provider)
Time window fixed effects	Set of indicator variables for the date-by-three-hour window in which the consultation starts. Windows are 0–3 am, 3–6 am, 6–9 am, 9 am–12 pm, 12–3 pm, 3–6 pm, 6–9 pm, and 9 pm–12 am of a given calendar date, such as April 1, 2022, 12–3 am. (Source: Provider)
Patient controls (X_i)	
Age	Indicators for the patient's age at the end of the calendar year of each consultation, with a separate indicator for each age from 2 to 69 and a single indicator for ages 70 and older. Consultations with infants (below age 2) are excluded. (Source: SCB)
Female	Indicates whether the patient is female. (Source: Provider)
Patient's region	Set of indicator variables for the patient's area of residence, based on the registered municipality of residence in 2021. We use Sweden's eight national areas (riksområden): West Sweden, Stockholm, East Middle Sweden, North Middle Sweden, South Sweden, Middle Norrland, Upper Norrland, and Småland and the islands, plus one indicator for a missing region. (Source: SCB)
Symptom category	Set of indicator variables for the main symptom the patient reports when requesting the consultation. Symptom categories with fewer than 1,000 consultations in the analysis window are excluded from the sample. The full list of symptoms is shown in Table 1. (Source: Provider)
Outcomes: Time inputs	
Total consultation time (minutes)	Total time the physician records on the consultation, in minutes. Values above 180 minutes are set to missing. Equals the sum of patient-facing and administrative time. (Source: Provider)
Patient-facing time (minutes)	Time in the video call with the patient, in minutes. Values above 120 minutes are set to missing. (Source: Provider)
Administrative time (minutes)	Total consultation time minus patient-facing time, in minutes. Covers work on the consultation outside the video call, such as documentation, prescriptions, and referrals. Consultations with non-positive administrative time are excluded from the sample. (Source: Provider)
Outcomes: Performance-relevant measures	
Top grade (5/5)	Indicates whether the patient rates the consultation with the top grade, 5 on a scale from 1 to 5. Coded 0 if the patient does not rate the consultation. (Source: Provider)
Worst grade (1/5)	Indicates whether the patient rates the consultation with the worst grade, 1 on a scale from 1 to 5. Coded 0 if the patient does not rate the consultation. (Source: Provider)
Grade, if any	The patient's rating of the consultation on a scale from 1 to 5. Missing if the patient does not rate the consultation. (Source: Provider)
Informative diagnosis	Indicates whether the main (first-listed) ICD-10 diagnosis the physician records at the consultation belongs to a chapter other than R (symptoms and signs, e.g., cough, rash) and Z (factors influencing health status and contact with health services). Coded 0 if the diagnosis is missing, invalid, or U99. (Source: Provider)
Outcomes: Recorded outcomes at the provider	
Any prescription by the provider (7 days)	Indicates whether the patient fills any prescription written by the provider within 7 days of the consultation, whether or not the drug is new to the patient. Prescriptions are attributed to the provider by the prescriber's workplace identifier. (Source: NBHW)
Any unscheduled return consultation (7 days)	Indicates whether the patient has a later unscheduled (drop-in) video consultation with the provider within 7 days, with any clinician and for any symptom. (Source: Provider)
Any booked return consultation (7 days)	Indicates whether the patient has a later booked (not drop-in) video consultation with the provider within 7 days, with any clinician and for any symptom. (Source: Provider)
Any psychologist consultation (7 days)	Indicates whether the patient has a later video consultation with a psychologist at the provider within 7 days. (Source: Provider)
Any in-person consultation (7 days)	Indicates whether the patient has an in-person consultation at one of the provider's physical clinics within 7 days. (Source: Provider)

Table A7. Description of key variables (Continued)

Variable	Description
Outcomes: Downstream care outside the provider	
Any external primary care consultation (7 days)	Indicates whether the patient has a primary care consultation with a provider other than the online provider within 7 days. Defined only for patients residing in Scania or Stockholm and observed only in these regions' records. In Scania, we count new and follow-up in-person visits with a doctor or nurse at a primary care clinic, excluding maternal and child health care; in Stockholm, we count any contact at a primary care clinic. (Source: Region Scania, Region Stockholm)
Any external primary or community care (7 days)	As *Any external primary care consultation*, but counting a contact with any type of clinician at any type of clinic in the regional records, e.g., including care at maternity or child clinics. Defined only for patients residing in Scania or Stockholm. (Source: Region Scania, Region Stockholm)
Any external prescription (7 days)	Indicates whether the patient fills any prescription written by a provider other than the online provider within 7 days, whether or not the drug is new to the patient. (Source: NBHW)
Any new external (provider) prescription (7 days)	Indicates whether the patient fills a new prescription written by a provider other than the online provider (by the online provider) within 7 days. A prescription is "new" if no drug with the same ATC code at the chemical-subgroup level (5 characters, e.g., J01CA) was dispensed to the patient in the 6 months before the consultation. (Source: NBHW)
Any urgent care (7 days)	Indicates any outpatient visit at an urgent care unit (local emergency unit, *närakut*, or other acute unit) within 7 days of the consultation. Excludes visits with any ICD-10 diagnosis in chapter O (pregnancy and childbirth). (Source: NBHW)
Any specialist visit (7 days)	Indicates any non-acute specialist outpatient visit within 7 days of the consultation, excluding emergency department and urgent care visits as well as phone and letter contacts. Excludes visits with any ICD-10 diagnosis in chapter O. (Source: NBHW)
Any ED visit (7 days)	Indicates any emergency department visit within 7 days of the consultation, counting visits to hospital acute units with one or more somatic specialties and to psychiatric acute units. Excludes visits with any ICD-10 diagnosis in chapter O. (Source: NBHW)
Any hospitalization (7 days)	Indicates any inpatient admission within 7 days of the consultation. Excludes admissions with any ICD-10 diagnosis in chapter O. (Source: NBHW)
Any avoidable hospitalization (7 days)	Indicates any inpatient admission within 7 days of the consultation with a diagnosis for an ambulatory care sensitive condition, for which timely primary care can prevent admission (e.g., diabetes, dehydration, epilepsy, otitis, hypertension and heart failure, upper respiratory infections, chronic lung disease, dental conditions, ulcers, pyelonephritis), with separate lists for adults and children, following Page et al. (2007). Excludes admissions with any ICD-10 diagnosis in chapter O. (Source: NBHW)
Death (1 year)	Indicates whether the patient dies within 365 days of the consultation. (Source: NBHW, Cause of Death Register)
Outcomes: Types of prescriptions	
Any mental health drug prescription (7 days)	Indicates whether the patient fills a prescription by the provider for a psycholeptic (ATC N05) or an antidepressant (ATC N06A) within 7 days, following Cesarini et al. (2016). (Source: NBHW)
Any opioid prescription (7 days)	Indicates whether the patient fills a prescription by the provider for an opioid (ATC N02A) within 7 days. (Source: NBHW)
Any cough medication prescription (7 days)	Indicates whether the patient fills a prescription by the provider for a cough and cold preparation (ATC R05) within 7 days. (Source: NBHW)
Any counter-guideline antibiotic prescription (7 days)	Indicates whether the patient fills an antibiotic (ATC J01) prescribed by the provider within 7 days of a consultation whose main diagnosis is one for which the Swedish quality indicators for digital consultations (Strama 2019) advise against antibiotic treatment at a digital visit: acute tonsillitis, sinusitis, bronchitis, and upper respiratory infection; pneumonia; acute otitis media; acne; impetigo; abscess, furuncle, and cellulitis; other skin infections; erysipelas; and a missing diagnosis. For acute cystitis and Lyme disease, only antibiotics outside the recommended groups count (J01CA and J01XE for cystitis, J01CE for Lyme disease). Coded 0 for all other consultations. (Source: NBHW, Provider)
Any antibiotic (7 days)	Indicates whether the patient fills a prescription for an antibacterial for systemic use (ATC J01) from any provider within 7 days. (Source: NBHW)
Any broad-spectrum antibiotic (7 days)	Indicates whether the patient fills a prescription for a broad-spectrum antibiotic from any provider within 7 days: penicillins with beta-lactamase inhibitors (J01CR), second- and third-generation cephalosporins (J01DC, J01DD), fluoroquinolones (J01MA), or macrolides other than erythromycin (J01FA, excluding J01FA01). (Source: NBHW)
Any rare antibiotic (7 days)	Indicates whether the patient fills a prescription from any provider within 7 days for an antibiotic (J01) other than the four that each account for at least 5% of antibiotic fills before the rollout: phenoxymethylpenicillin (J01CE02), pivmecillinam (J01CA08), nitrofurantoin (J01XE01), and flucloxacillin (J01CF05). (Source: NBHW)

Table A7. Description of key variables (Continued)

Variable	Description
Outcomes: Cost	
Downstream cost (SEK, 7 days)	Cost of the care the patient uses within 7 days of the consultation, inside and outside the provider, in SEK. Sums the cost of up to one occurrence of each of: a return video consultation at the provider, a prescription fill, an external primary care consultation, an urgent care visit, a specialist visit, an ED visit, and a hospitalization, each priced as in Table A6. The consultation itself is not included. For patients outside Scania and Stockholm, an external primary care consultation is inferred from any external prescription fill within 7 days. (Source: Provider, NBHW, Region Scania, Region Stockholm)
Outcomes: Physician labor (physician-month level)	
Total minutes worked	Sum of consultation time and break minutes in the month, over all of the physician's online meetings. (Source: Provider)
Patient-facing minutes	Sum of patient-facing time over all of the physician's online meetings in the month. (Source: Provider)
Administrative minutes	Sum of administrative time over all of the physician's online meetings in the month. (Source: Provider)
Break minutes	Sum of the gaps between consecutive meetings within a shift. A shift ends when the gap to the next meeting is 15 minutes or longer, or after 8 hours. (Source: Provider)
Number of consultations, all types	Number of online consultations the physician conducts in the month, across all booking types and modes. (Source: Provider)
Number of video consultations	Number of video consultations the physician conducts in the month, booked or unbooked. (Source: Provider)
Number of text-based consultations	Number of text-based consultations the physician conducts in the month, booked or unbooked. (Source: Provider)
Consultations per worked hour	Number of consultations across all booking types and modes, divided by total hours worked in the month. Text-based consultations are weighted by one-half. Missing in months without work. (Source: Provider)
Additional controls: Health risk	
Any prior hospitalization	Indicates any inpatient admission within the 3 years before the consultation, excluding the day of the consultation and the day before. Excludes admissions with any ICD-10 diagnosis in chapter O. (Source: NBHW)
Any prior ED visit	Indicates any emergency department visit (hospital or psychiatric acute unit) within the 3 years before the consultation, excluding the day of the consultation and the day before. Excludes visits with any ICD-10 diagnosis in chapter O. (Source: NBHW)
Any prior urgent care	Indicates any urgent care visit (*närakut* or other acute unit) within the 3 years before the consultation, excluding the day of the consultation and the day before. Excludes visits with any ICD-10 diagnosis in chapter O. (Source: NBHW)
Any prior specialist care	Indicates any non-acute specialist outpatient visit within the 3 years before the consultation, excluding the day of the consultation and the day before, and excluding phone and letter contacts. Excludes visits with any ICD-10 diagnosis in chapter O. (Source: NBHW)
Any comorbidity	Indicates an Elixhauser comorbidity index of 1 or higher, based on all diagnoses recorded in inpatient and specialist outpatient care within the 3 years before the consultation, excluding the day of the consultation and the day before. The Elixhauser comorbidities are adapted from 4-digit to 3-digit ICD-10 codes to match the reporting in the NBHW registers. (Source: NBHW)

Table A7. Description of key variables (Continued)

Variable	Description
Additional controls: Socioeconomic characteristics	
Income above median	Indicates whether the patient's annual earned income (wages and business income) in 2021 is above the median across all consultations with patients aged 20 or older, before sample restrictions. Defined for patients aged 20 or older. (Source: SCB)
Any benefits	Indicates receipt of at least one social insurance or housing benefit in 2021: youth disability compensation, sickness or disability compensation, sickness benefits, rehabilitation benefits, or housing allowance. Defined for patients aged 20 or older. (Source: SCB)
Employment status	Set of indicator variables for the patient's occupational position in November 2021: employed; self-employed; no occupational position. Defined for patients aged 20 to 67. (Source: SCB)
Education level	Set of indicator variables for the highest completed education in 2021: compulsory schooling (9 years or less); upper secondary school; post-secondary education of less than 3 years; post-secondary education of 3 years or more. Defined for patients aged 25 or older. (Source: SCB)
Civil status	Set of indicator variables for the patient's civil status in 2021: married or registered partner; divorced or widowed; unmarried. Defined for patients aged 18 or older. (Source: SCB)
Migration status	Set of indicator variables for the patient's migration background: born in Sweden with at least one Swedish-born parent; first-generation immigrant (born outside Sweden); second-generation immigrant (born in Sweden to two foreign-born parents). (Source: SCB)
Socioeconomics missing	Indicates that any of the socioeconomic variables (income, benefits, employment status, education level, civil status, or migration status) is missing. (Source: SCB)
Variables used in robustness checks	
Consultations per worked hour under fixed wages	The physician's number of consultations divided by hours worked, over the months before her first month on the quantity-incentives scheme. Used to group physicians into up to 3, 3–4, 4–5, 5–6, and more than 6 consultations per worked hour. (Source: Provider)
Administrative time share under fixed wages	The physician's total administrative time divided by her total consultation time, over the consultations before her first month on the quantity-incentives scheme. Used to group physicians into quintiles. (Source: Provider)
Prior healthcare use	The number of the patient's emergency department visits and hospitalizations in the 3 years before the consultation, excluding the day of the consultation and the day before, and excluding visits and admissions with any ICD-10 diagnosis in chapter O. Used to group consultations into 0, 1, 2, 3, and 4 or more prior events. (Source: NBHW)
Predicted downstream cost	The part of downstream cost (SEK, 7 days) explained by the patient controls X_i in a regression that also includes physician and time window fixed effects, estimated on consultations before each physician's first month on the quantity-incentives scheme. Used to group consultations into deciles. (Source: Provider, NBHW, Region Scania, Region Stockholm)
Share of other doctors with quantity incentives	Among consultations started by other physicians in the 60 minutes before the consultation starts, the share conducted by physicians on the quantity-incentives scheme in that month. (Source: Provider)
Physician characteristics	
Work per month before the rollout	The physician's monthly average over January to May 2022, before any physician was compensated under quantity incentives, counting all of the physician's online meetings with complete key variables. (Source: Provider)
Demographics	The physician's sex and age in 2022. (Source: SCB)
Training	Indicators for holding a specialist degree, being in specialist training, and holding a degree from a top medical school. Specialty is from the provider. The physician's medical school is taken from the graduation municipality, the earliest observed for the physician in 2016–2021. (Source: Provider, SCB)
Income in 2021	The number of employers filing an employer declaration for the physician, gross wage income, and wage income from the provider. The provider is identified as the employer most frequent among physicians' three largest workplaces; where the provider is not among a physician's three largest workplaces, income from the provider is imputed from the remaining wage income. (Source: SCB)

Notes: Windows of k days count events from the calendar date of the consultation (day 0) through day k; events on the day of the consultation are included. Prescriptions are dated by the dispensing date. Outcomes measured over 30 days (Appendix Table A3) are defined as their 7-day counterparts with a 30-day window. Outcomes over 30 days (1 year) are measured on consultations that take place more than 30 (365) days after the patient's previous such consultation. NBHW: National Board of Health and Welfare (Socialstyrelsen). SCB: Statistics Sweden.

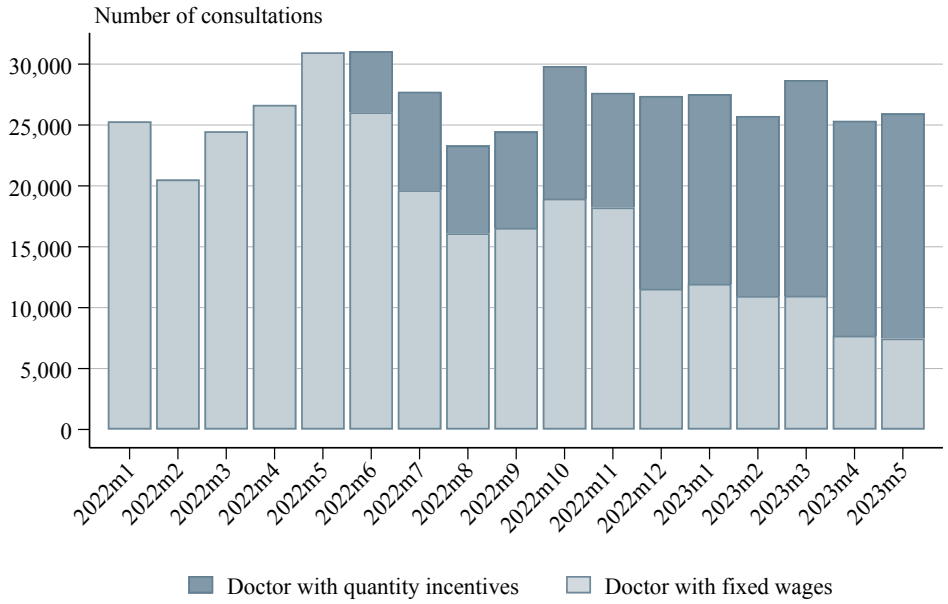
Table A8. Diagnoses set in the ED, hospital, and as cause of death

ICD code	Diagnosis	# Cases	# Patients
Panel A. ED visits within 7 days (8,104 consultations, 7,855 patients)			
R104X	Stomach ache	665	644
R074	Chest pain, unspecified	252	248
J069	Acute upper respiratory infection, unspecified	167	167
R519	Headache	166	165
R509	Fever, unspecified	160	160
R104	Other and unspecified abdominal pain	156	156
B349	Viral infection, unspecified	146	146
R060	Dyspnoea	144	142
F419	Anxiety disorder, unspecified	123	123
J189	Pneumonia, unspecified	107	107
J039	Acute tonsillitis, unspecified	104	103
T784	Allergy, unspecified	96	95
N390	Urinary tract infection, site not specified	94	94
U071	COVID-19, virus identified	87	87
A099	Gastroenteritis and colitis of unspecified origin	80	80
R219	Rash non specified	80	80
L509	Urticaria, unspecified	78	78
R429	Dizziness and vertigo	76	76
N109	Acute tubulo-interstitial nephritis	72	72
H109	Conjunctivitis, unspecified	69	69
Panel B. Hospitalizations within 7 days (1,829 consultations, 1,785 patients)			
K358	Acute appendicitis, other and unspecified	44	44
R104X	Stomach ache	42	42
N109	Acute tubulo-interstitial nephritis	40	40
K353	Acute appendicitis with localized peritonitis	36	36
A469	Erysipelas	31	31
J159	Bacterial pneumonia, unspecified	31	31
J189	Pneumonia, unspecified	31	31
J369	Peritonsillar abscess	29	29
J039	Acute tonsillitis, unspecified	26	26
K610	Anal abscess	24	24
R104	Other and unspecified abdominal pain	24	21
F419	Anxiety disorder, unspecified	21	20
F603	Emotionally unstable personality disorder	21	15
U071	COVID-19, virus identified	20	20
K800	Calculus of gallbladder with acute cholecystitis	19	19
K567	Ileus, unspecified	17	16
F100	Mental and behavioural disorders due to use of alcohol: Acute intoxication	16	16
R429	Dizziness and vertigo	15	15
N201	Calculus of ureter	14	14
J101	Influenza with other respiratory manifestations, seasonal influenza virus identified	13	13
Panel C. Deaths within 1 year (214 consultations, 214 patients)			
X70	Intentional self-harm by hanging, strangulation and suffocation	15	15
C349	Malignant neoplasm: Bronchus or lung, unspecified	14	14
R99	Cause of death not defined	12	12

Note: This table presents the 20 most frequent ICD-10 diagnoses recorded in secondary care that follow a consultation in the analysis sample: emergency department visits within 7 days (Panel A), hospitalizations within 7 days (Panel B), and cause of death within 1 year (Panel C). Panel C is restricted to consultations that take place more than one year after the patient's previous consultation. Diagnoses with fewer than ten cases are suppressed.

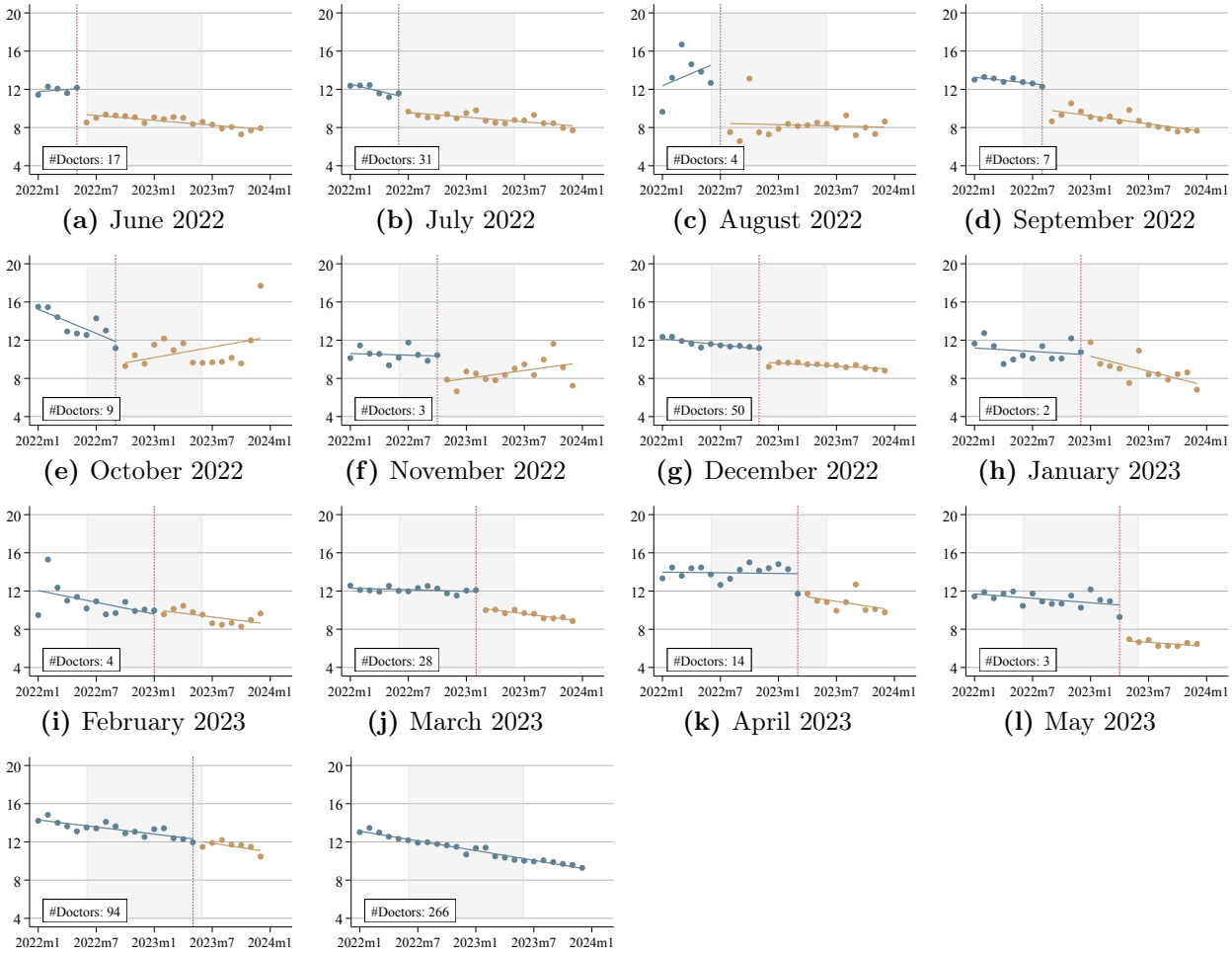
Appendix Figures

Figure A1. Consultations per month by treatment status



Notes: This figure shows the monthly number of consultations in the analysis sample from January 2022 through May 2023. The lighter bars show the number of consultations conducted by physicians paid under the fixed-wage scheme. The darker bars show the number of consultations conducted by physicians paid under the quantity-incentives scheme.

Figure A2. Average consultation time by onset month



(a) June 2022 **(b)** July 2022 **(c)** August 2022 **(d)** September 2022

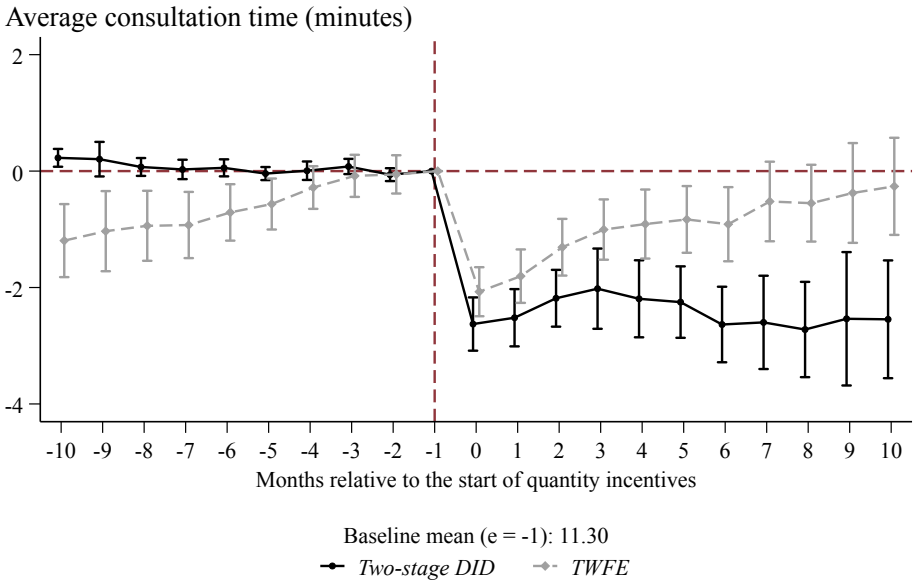
(e) October 2022 **(f)** November 2022 **(g)** December 2022 **(h)** January 2023

(i) February 2023 **(j)** March 2023 **(k)** April 2023 **(l)** May 2023

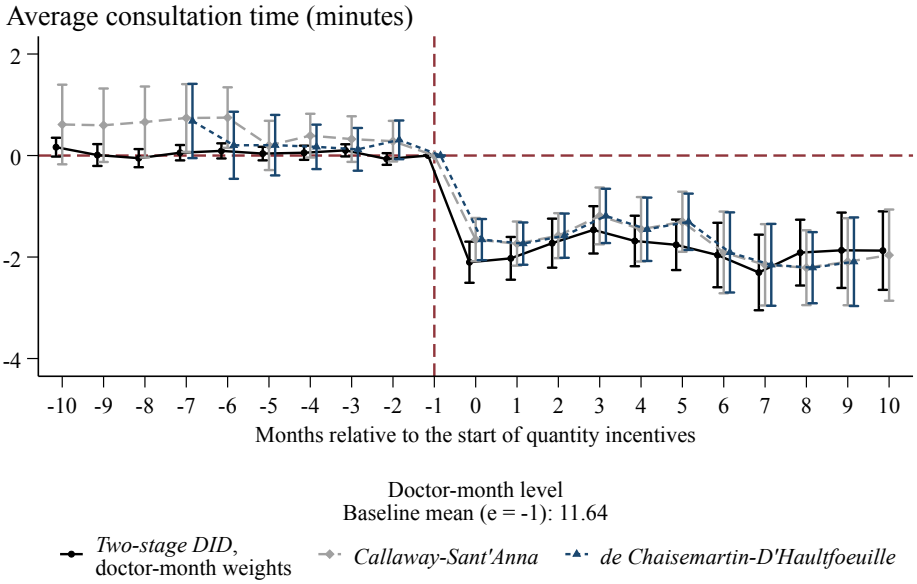
(m) June 2023 (mandatory) **(n)** All physicians

Notes: Each cohort panel shows binned means of the average total time per consultation (the sum of patient-facing plus administrative time in minutes), against calendar month, for physicians who first adopt the quantity-incentives scheme in the month indicated in the panel label. Panel (m) shows physicians who moved to the quantity-incentives scheme when it became mandatory. Panel (n) pools all physicians. Linear trends are fitted separately before adoption and from the adoption month onward. The dashed vertical line marks the month before adoption; the shaded area marks the rollout period, June 2022–June 2023. Each panel reports its physician count.

Figure A3. Alternative estimators



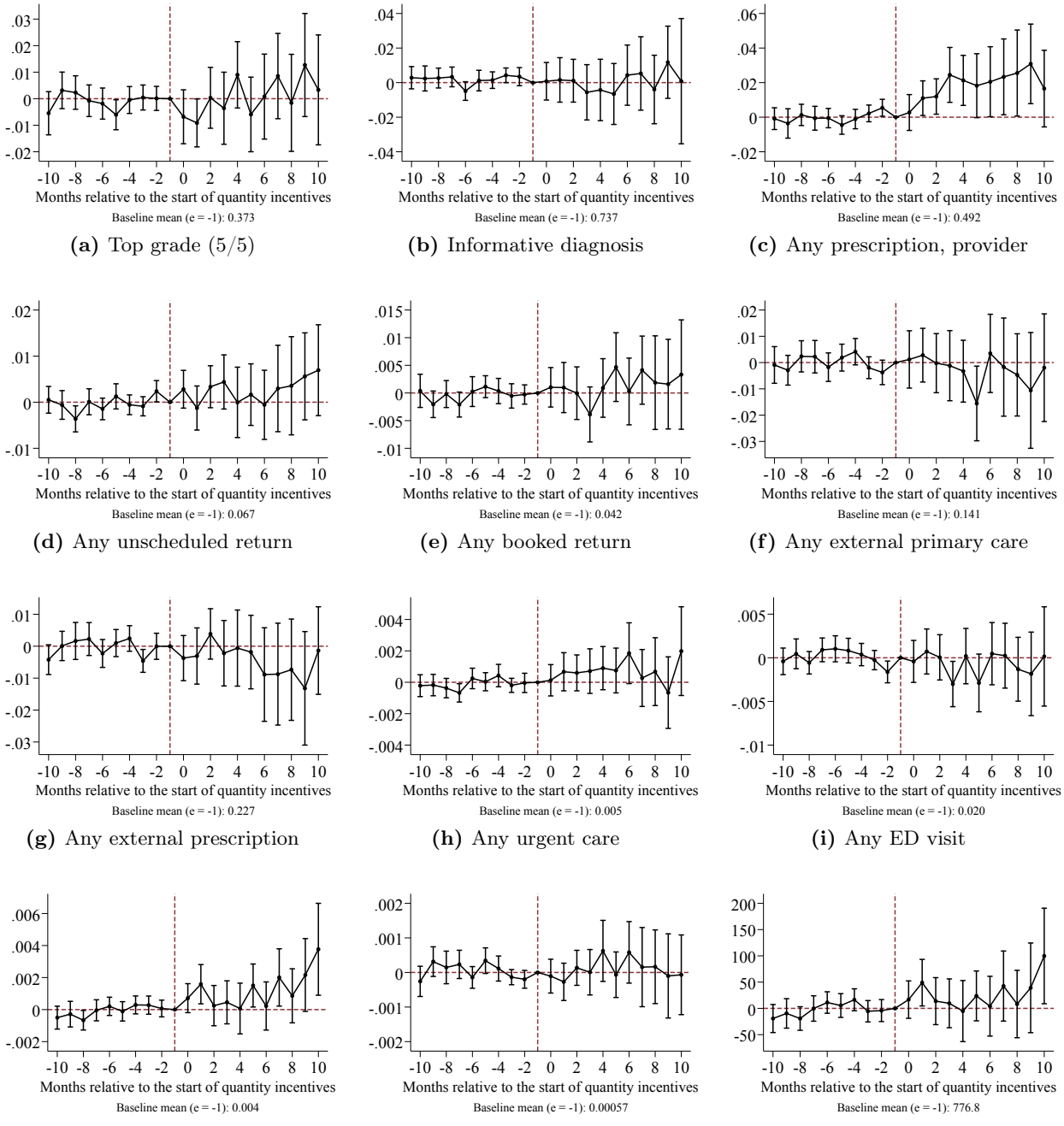
(a) Consultation-level estimates



(b) Estimates with doctor-month weights

Notes: This figure reports event-study estimates of the effect of quantity incentives on total consultation time under alternative estimators. Panel (a) is at the consultation level. The Gardner et al. (2026) two-stage estimator (Two-stage DID) is the main specification of Figure 4. Two-way fixed effects (TWFE) estimates Equation (4) with indicators for months relative to onset by OLS. Panel (b) shows estimates with physician-month weights. Two-stage DID estimates Equation (4) with each consultation weighted by the inverse of its physician's number of consultations in that month. Callaway and Sant'Anna (2021) and De Chaisemartin and D'Haultfœuille (2026) estimators use the physician-month panel with the physician's average total consultation time as the outcome. The baseline mean is the average outcome value in the month before onset. Bars show 95% confidence intervals. Standard errors are clustered at the physician level.

Figure A4. Event studies for the main outcomes



(a) Top grade (5/5) **(b)** Informative diagnosis **(c)** Any prescription, provider

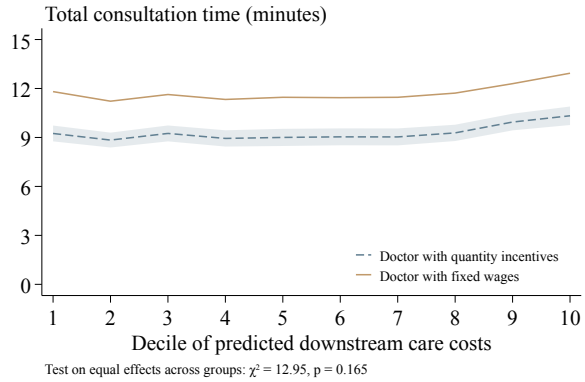
(d) Any unscheduled return **(e)** Any booked return **(f)** Any external primary care

(g) Any external prescription **(h)** Any urgent care **(i)** Any ED visit

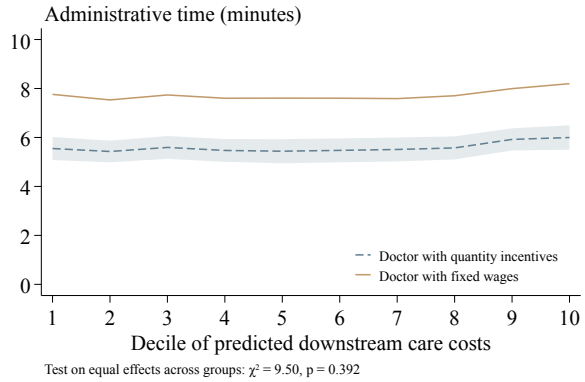
(j) Any hospitalization **(k)** Death (1 year) **(l)** Downstream cost (SEK)

Notes: This figure reports consultation-level event-study estimates of the effects of quantity incentives on within-organization outcomes and downstream care, using the two-stage estimator of Gardner et al. (2026). Outcomes are defined in Appendix Table A7. The first stage estimates physician and date-by-three-hour-window fixed effects using each physician's pre-adoption consultations. The second stage regresses the residualized outcome on indicators for months relative to the first month of quantity incentives. The reference month is the month before first adoption ($e = -1$); baseline means are measured in this month. External primary care is observed for patients registered in Scania or Stockholm. Mortality is measured for consultations at least one year after a patient's previous consultation. Bars show 95% confidence intervals; standard errors are clustered at the physician level.

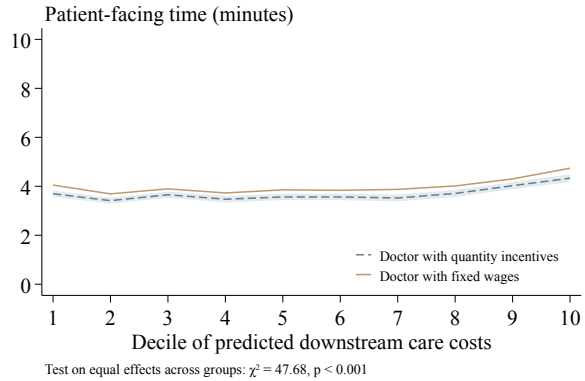
Figure A5. Consultation time under fixed and volume pay, by predicted baseline downstream costs



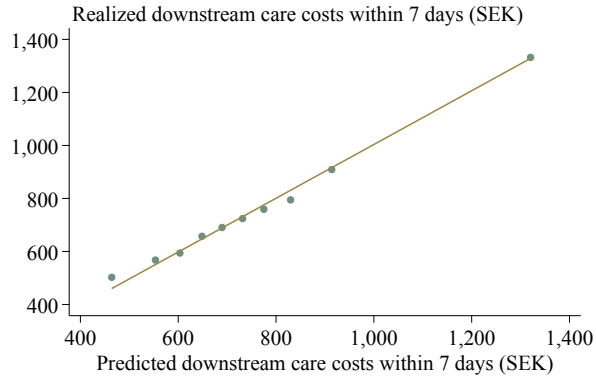
(a) Total consultation time



(b) Administrative time



(c) Patient-facing time

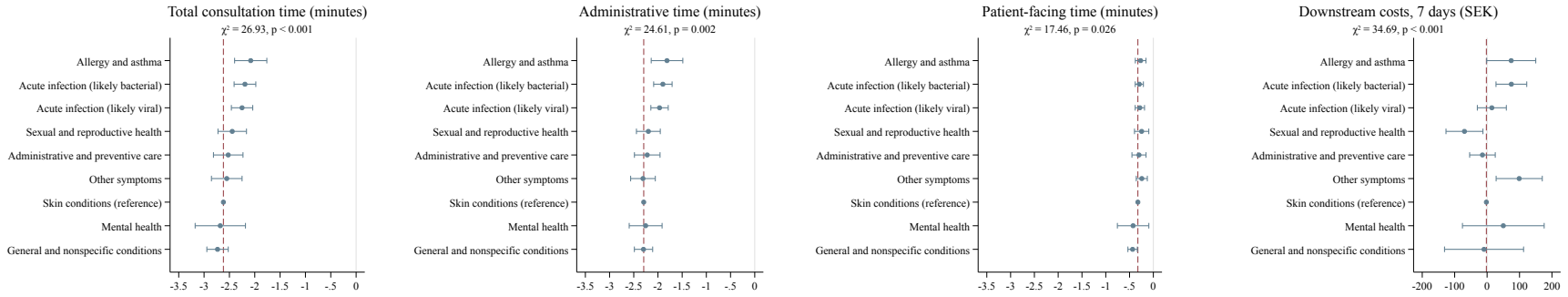


(d) Cost prediction

Notes: This figure reports consultation times under fixed wages (solid lines) and quantity incentives (dashed lines) by decile of patients' predicted downstream cost. Solid lines show observed means under fixed wages. Dashed lines add each decile's estimated effect of quantity incentives to these means. Predicted downstream cost is the part of downstream cost within seven days explained by the patient controls X_i in a regression with physician and date-by-three-hour-window fixed effects, estimated on consultations before each physician's first month on the quantity-incentives scheme. The decile effects are estimated using the two-stage estimator of Gardner et al. (2026) with the same fixed effects and patient controls X_i in the first stage, and interactions between quantity-incentives status and decile indicators in the second stage. The χ^2 statistics test whether the estimated effects are equal across deciles. Panels (a), (b), and (c) report total, administrative, and patient-facing time, respectively. Panel (d) plots realized against predicted downstream cost by decile, with a linear fit through the decile means. Shaded bands show 95% confidence intervals. Standard errors are clustered at the physician level.

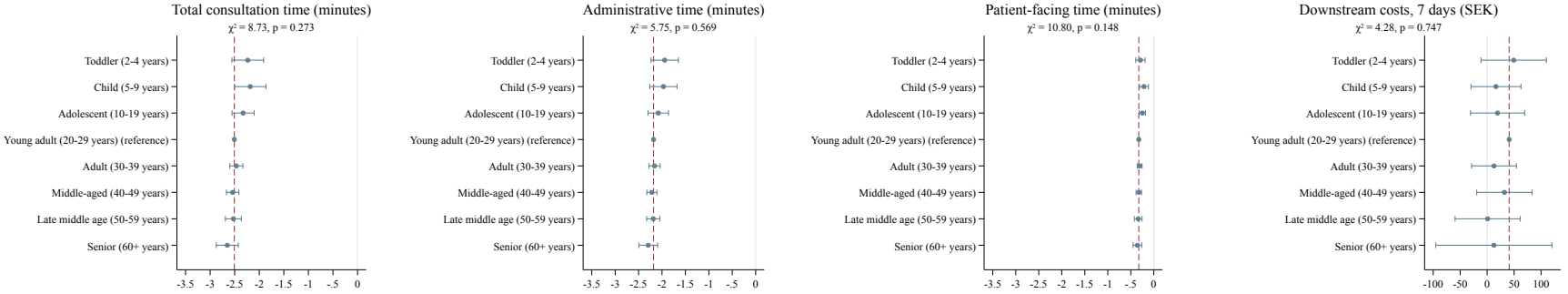
Figure A6. Heterogeneity in the effect of quantity incentives

Panel A. By symptom group



(A1) Total consultation time **(A2)** Administrative time **(A3)** Patient-facing time **(A4)** Downstream cost (7d)

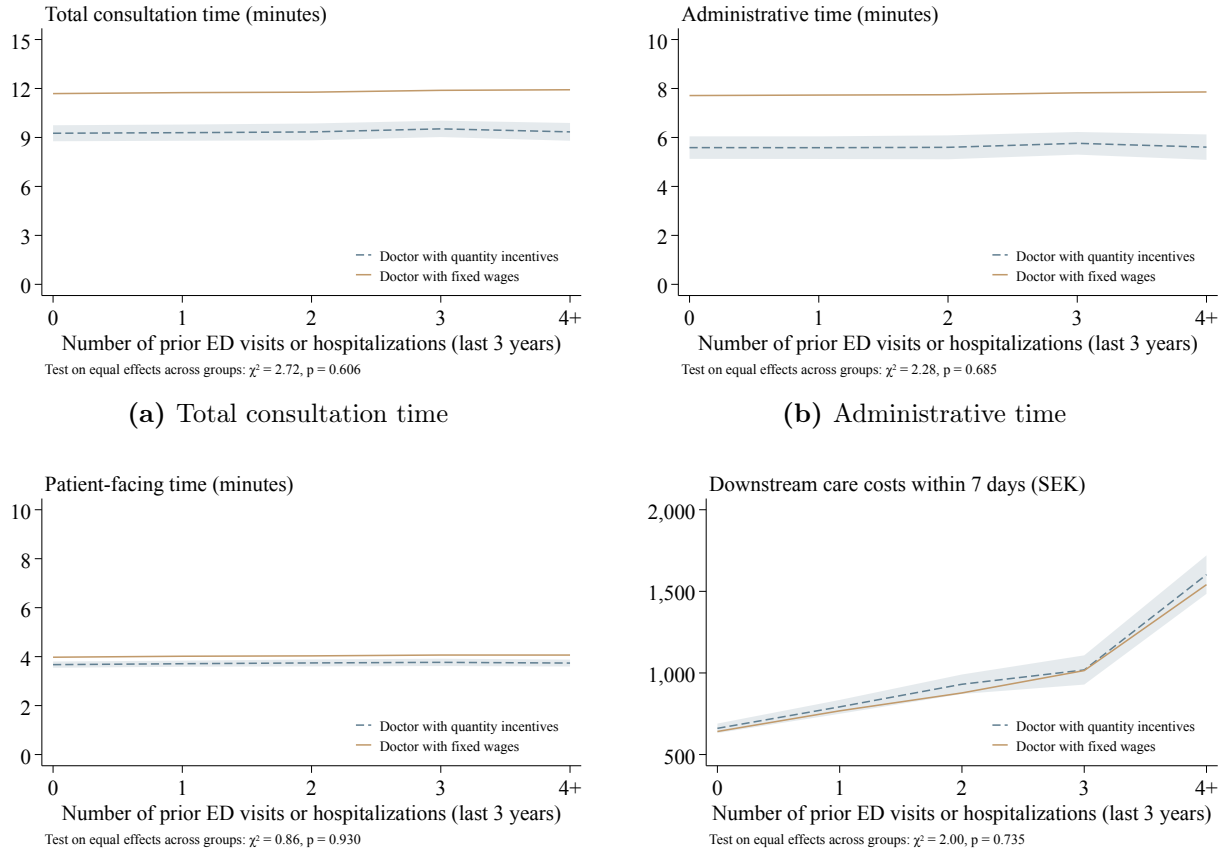
Panel B. By patient age



(B1) Total consultation time **(B2)** Administrative time **(B3)** Patient-facing time **(B4)** Downstream cost (7d)

Notes: This figure reports the effect of quantity incentives on each outcome by the clinical group of the patient's reported symptom (Panel A) and by patient age group (Panel B). The group effects are estimated using the two-stage estimator of Gardner et al. (2026) with physician and date-by-three-hour-window fixed effects and patient controls X_i in the first stage, and interactions between quantity-incentives status and group indicators in the second stage. The dashed vertical line marks the treatment effect estimate on the reference group. Bars show 95% confidence intervals for the difference between each group's effect and the reference group's effect. The χ^2 statistics test whether these effect estimates are equal across groups. Standard errors are clustered at the physician level.

Figure A7. Consultation time under fixed and volume pay, by prior healthcare use



(a) Total consultation time

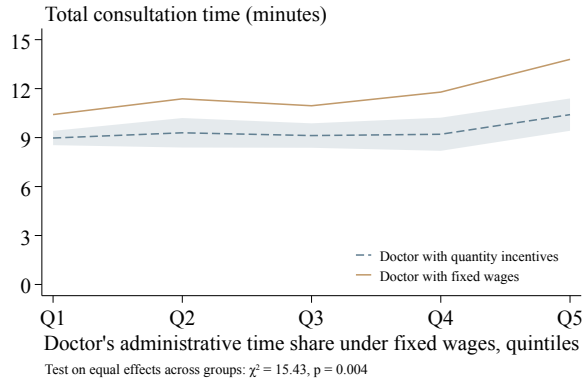
(b) Administrative time

(c) Patient-facing time

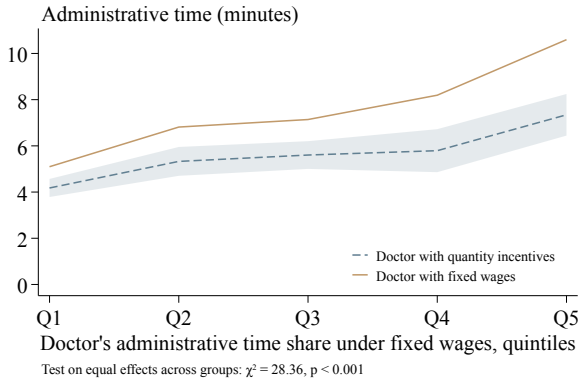
(d) Downstream costs

Notes: This figure reports consultation times and downstream costs under fixed wages (solid lines) and quantity incentives (dashed lines) by patients' prior healthcare use. Prior healthcare use is the number of emergency department visits and hospitalizations in the three years before the consultation. The groups are zero (64.2% of consultations), one (17.6%), two (8.0%), three (3.9%), and four or more (6.3%) events. Solid lines show observed means under fixed wages. Dashed lines add each group's estimated effect of quantity incentives to these means. The group effects are estimated using the two-stage estimator of Gardner et al. (2026) with physician and date-by-three-hour-window fixed effects and patient controls X_i in the first stage, and interactions between quantity-incentives status and group indicators in the second stage. The χ^2 statistics test whether the estimated effects are equal across groups. Panels (a), (b), and (c) report total, administrative, and patient-facing time, respectively. Panel (d) shows downstream costs from healthcare use within seven days, calculated using Appendix Table A6. Shaded bands show 95% confidence intervals. Standard errors are clustered at the physician level.

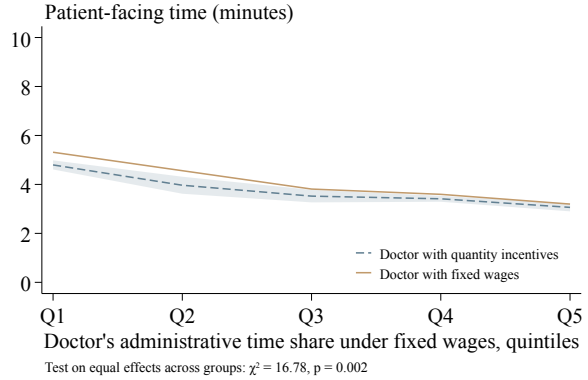
Figure A8. Consultation time under fixed and volume pay, by doctor's administrative time use with fixed wages



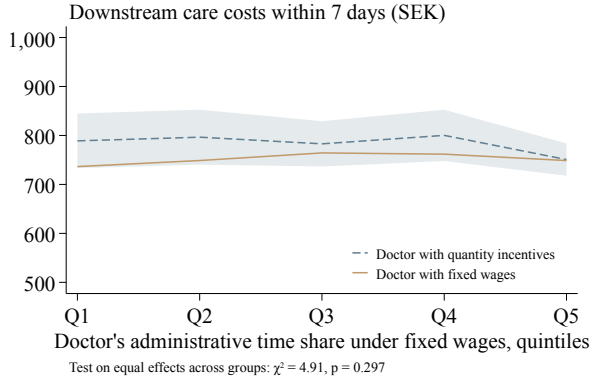
(a) Total consultation time



(b) Administrative time



(c) Patient-facing time



(d) Downstream costs

Notes: This figure reports consultation times and downstream costs under fixed wages (solid lines) and quantity incentives (dashed lines) by quintile of the physician's administrative time share. The share is the physician's total administrative time divided by total consultation time over her consultations before the physician's first month on the quantity-incentives scheme. Solid lines show observed means under fixed wages. Dashed lines add each quintile's estimated effect of quantity incentives to these means. The quintile effects are estimated using the two-stage estimator of Gardner et al. (2026) with physician and date-by-three-hour-window fixed effects and patient controls X_i in the first stage, and interactions between quantity-incentives status and quintile indicators in the second stage. The χ^2 statistics test whether these effect estimates are equal across quintiles. Panels (a), (b), and (c) report total, administrative time, and patient-facing time, respectively. Panel (d) shows downstream cost within seven days. Shaded bands show 95% confidence intervals. Standard errors are clustered at the physician level.

A Additional Institutional Details

A.1 The full payment formula

Equation (1) describes pay for video consultations at regular working hours. The full scheme also includes text-based consultations, valued at half the video consultation rate, and pay multipliers for overtime and night-time hours. These multipliers apply to both the guaranteed hourly wage and the piece rate.

Let $s \in \{1, 1.25, 1.5\}$ denote the multiplier for regular, overtime, and night-time hours, respectively. During a calendar month, a physician works L_s staffed hours at multiplier s and conducts N_s^{vid} video consultations and N_s^{txt} text-based consultations during those hours. Staffed hours include breaks. Her guaranteed monthly pay is

$$\text{BaselinePay} = \sum_s s\,\bar{W}\,L_s. \tag{A.1}$$

Her piece-rate earnings value video consultations at $p = \bar{W}/5$ and text-based consultations at $p/2$, with the applicable working-hour multiplier:

$$\text{PieceRate} = \sum_s s\,p\,(N_s^{\text{vid}} + 0.5N_s^{\text{txt}}). \tag{A.2}$$

The bonus equals the positive difference between piece-rate earnings and guaranteed pay, conditional on the physician's monthly pace exceeding five video-equivalent consultations per staffed hour:

$$\text{Bonus} = \mathbf{1}\{N/L > 5\} \cdot \max\{\text{PieceRate} - \text{BaselinePay},\ 0\}, \tag{A.3}$$

where $N = \sum_s (N_s^{\text{vid}} + 0.5N_s^{\text{txt}})$ is the number of consultations, and $L = \sum_s L_s$ is total staffed time. Total monthly pay is the sum of guaranteed pay and the bonus, BaselinePay + Bonus.

For a physician conducting only video consultations at regular working hours, guaranteed pay is $\bar{W}L$ and piece-rate earnings are pN. Since $\bar{W} = 5p$, the bonus per staffed hour is

$$\frac{\text{Bonus}}{L} = p\,\max\{n - 5, 0\} = p\,(n-5)\,\mathbf{1}\{n \geq 5\},$$

where $n = N/L$, yielding Equation (1). At exactly five consultations per hour, the bonus is zero.

B Data Appendix

Provider data. The primary data are supplied by the healthcare organization. They contain comprehensive records of all online primary care consultations within the organization for the years 2021–2024. For each consultation, we observe the consultation start date and time, the patient's self-reported symptom, the communication mode (video, text, or chat), the clinician type, and

the clinician-assigned primary diagnosis (ICD-10 code). In addition, the data record the time the physician spends in the video call or direct communication with the patient (patient-facing time), the total time she records on the consultation before signing it off (total consultation time). The difference between the total and patient-facing time is used for administrative work (administrative time). We standardize symptom names across children and adults, merge near-duplicates, and translate Swedish names into English.[17] In addition, we observe patients' age and gender. Each consultation record includes the patient's and the clinician's unique, pseudonymized personal identifiers.

For clinicians, the organization also supplies contract records, including monthly lists of all physicians working under the quantity-incentives scheme between June 2022 and May 2023. The lists are missing for employee physicians in January 2023, for whom we impute the status from the months before and after. We fill single-month gaps between months on the scheme and define treatment onset as a physician's first month on the scheme.

Register linkage. The personal identifiers enable Statistics Sweden (before pseudonymization) to link patients' consultations to individual-level information from Statistics Sweden (SCB) and other agencies. Data from the National Board of Health and Welfare (Socialstyrelsen, NBHW) capture patients' healthcare use elsewhere in the system, including urgent, hospital, and emergency care. These data also cover all prescriptions, including those from external primary care providers, but exclude consultations with these external primary care providers. For Scania and Stockholm, however, regional data cover all primary care visits, both within and outside the organization we study.

Demographic and socioeconomic data. SCB provides data from three registers: the Population Statistics Register (RTB), the Longitudinal Integrated Database for Health Insurance and Labour Market Studies (LISA), and the Education Register (UREG). For patients, we use values from 2021, the last year before the rollout. RTB provides migration background, civil status, and municipality of residence. LISA provides annual data on income, benefits, and employment status, recorded in November. UREG provides the highest completed level of education. LISA data, including region of residence, are unavailable for patients under 18, and socioeconomic variables are defined only above variable-specific age thresholds (Appendix Table A7). For physicians, RTB provides demographic information, UREG provides graduation records, and LISA provides monthly income data from each physician's three largest employers each year.

Downstream health data. Data from the NBHW cover four registers for 2013–2024: an inpatient register with hospitalizations; an outpatient register with emergency department, urgent care, and specialist visits; the cause-of-death register; and a prescription register.

[17] The symptom group *Other care processes* includes the labels Nurse Chat, Patient Visit, Phone Triage/Referral, Follow-up Survey, Video Consultation, Doctor (Language Assistance), Urgent Appointment, and Simple Care. The ICD codes recorded at these consultations indicate acute conditions such as infections.

The inpatient and outpatient registers include visit and discharge dates and ICD-10 diagnoses at the three-digit level, and the outpatient register additionally records the contact type. The cause-of-death register includes the date and cause of death. We exclude visits and admissions with any diagnosis related to pregnancy or childbirth (ICD-10 chapter O).

The prescription register includes all prescribed drugs that patients collect at pharmacies, but excludes drugs administered within healthcare facilities and over-the-counter purchases. If a prescription contains multiple drugs, each drug is recorded separately. The register reports the dispensing date, ATC codes, the cost of each prescription (borne by the national insurance system and the patient), a prescriber category (such as primary care or specialist care), and pseudonymized identifiers for the clinic which issued the prescription.

External primary care data. Primary care consultations are not collected nationwide in Sweden but held by each region. We observe external primary care visits directly in Scania (a region in southern Sweden) and Stockholm (the capital region). For adult patients registered in both regions, we observe the date of contact, pseudonymized clinic identifiers, the main ICD-10 diagnosis, the clinician type and contact type (only for Scania), and the type of care (for Stockholm). We infer primary care contacts for all other patients from their prescriptions, as the NBHW collects prescriptions nationwide.

The regional data are administratively collected and include services beyond standard primary care, such as maternity care and counseling. To ensure comparability, we restrict the Scania data to physical nurse or doctor consultations at clinics eligible for patient registration as a primary care provider, the Stockholm data to visits at primary care branches, and both regions' data to contacts outside the study organization.

C Model Appendix

This appendix sets out the setup of Section 3 formally, derives the physician's first-order conditions, and derives the responses to the quantity-incentives contract in Predictions 1, 2, and 3.

C.1 Setup

Health production and internal quality. Patient health from a consultation is $h(t_f, t_a) = h_f(t_f) + h_a(t_a)$, with h_f and h_a strictly increasing and strictly concave, such that returns to either time input are diminishing and independent of the other input. We also assume $h(0,0) < 0$, such that a consultation benefits the patient only once it receives enough time and a physician does not always want to minimize consultation length.

Internal quality is an incomplete measure of patient health:

$$q(t_f, t_a) = q_f(h_f(t_f), t_f) + q_a(t_a), \tag{C.1}$$

where patient ratings q_f are increasing in both arguments and strictly concave in patient-facing time, and clinical process quality q_a increases in administrative time only up to the verifiable minimum level $\underline{t}_a$, after which it is flat at $\underline{q}_a$. Throughout, $\partial q/\partial t_f$ denotes the total derivative of patient ratings with respect to patient-facing time. We assume that the physician chooses $t_a > \underline{t}_a$ under either contract type, so that only patient-facing time raises internal quality at the margin, whereas administrative time does not:

$$\frac{\partial q}{\partial t_f} > \frac{\partial q}{\partial t_a}, \quad \text{where } \frac{\partial q}{\partial t_a} = 0. \tag{C.2}$$

Downstream costs, payments, and minimum quality. Expected downstream costs $D(h)$ are strictly decreasing in patient health, $D' < 0$, so that any change in patient health changes downstream costs. The organization is paid a revenue of r per consultation, provided that its records and prescriptions meet the minimum standard $q_a \geq \underline{q}_a$, and pays the physician $\bar{W} + B_d(n)$ per staffed hour. For n consultations, the organization's profit is $\Pi = n\,\mathbf{1}\{q_a \geq \underline{q}_a\}r - \bar{W} - B_d(n)$. As long as the minimum standard is met and the piece rate is below the payment per consultation, $p < r$, profit increases in the number of consultations under either contract. The physician chooses time inputs such that the minimum standard is met:

$$q_a(t_a) \geq \underline{q}_a \Leftrightarrow t_a \geq \underline{t}_a. \tag{C.3}$$

By assumption, the physician chooses $t_a > \underline{t}_a$ under either contract type, so that requirement (C.3) does not bind and $q_a = \underline{q}_a$ is constant.

C.2 First-order conditions

The physician chooses time inputs t_f and t_a to maximize U_d in Equation (2), subject to the minimum quality requirement (C.3).

First, we consider the physician's cost of spending additional time on a consultation. Away from the threshold, lengthening a consultation by dT minutes changes the bonus by $\frac{dB_d(n)}{dT} = \mu\frac{dn}{dT}$, as each additional consultation per hour raises the bonus by $B_d'(n) = \mu$. The physician's pace changes by $\frac{dn}{dT} = -\frac{60}{T^2} = -\frac{n}{T}$, as $n = \frac{60}{T}$. Thus, lengthening a consultation by dT changes the physician's utility from Equation 2 by:

$$\frac{dU_d}{dT} = \mu\frac{dn}{dT} + \left(v\frac{dn}{dT} + n\frac{dv}{dT}\right) = n\left[\frac{dv}{dT} - \frac{\mu + v}{T}\right],$$

where dv/dT captures the gain in consultation value from additional time use, and $(\mu + v)/T \equiv \lambda$ captures the cost of additional time use, the shadow price of a minute of physician time. The shadow price is the same for both patient-facing and administrative time inputs, as either time input lowers the pace by the same amount.

Next, we consider the value of additional time. By Equation (C.2), the marginal value of a minute of patient-facing time is $\frac{\partial v}{\partial t_f} = \alpha\,\frac{\partial h}{\partial t_f} + \beta\,\frac{\partial q}{\partial t_f}$, and the marginal value of a minute of administrative time is $\frac{\partial v}{\partial t_a} = \alpha\,\frac{\partial h}{\partial t_a}$.

Away from the threshold, the physician optimally chooses time inputs such that the marginal value of a minute of either time input is equal to its shadow price:

$$\begin{aligned} \text{Patient-facing time: } & \alpha \frac{\partial h}{\partial t_f} + \beta \frac{\partial q}{\partial t_f} \stackrel{!}{=} \lambda, \\ \text{Administrative time: } & \alpha \frac{\partial h}{\partial t_a} \stackrel{!}{=} \lambda. \end{aligned} \tag{C.4}$$

C.3 Responses to the quantity-incentives contract

Changes in physician inputs. Let T_0 denote time per consultation under the fixed-wage contract. A physician who responds to the quantity-incentives contract chooses a pace above the threshold, so that her marginal pay changes from $\mu^0 = 0$ to $\mu^1 = p$, and Equation (C.4) holds with $\mu = p$ at her new choice.

Let κ_ℓ denote the increase in the marginal value of input $\ell \in \{f, a\}$ when it is reduced by one minute:

$$\begin{aligned} \text{Patient-facing time: } & \kappa_f \equiv -\frac{\partial^2 v}{\partial t_f^2} = -\alpha \frac{\partial^2 h}{\partial t_f^2} - \beta \frac{\partial^2 q}{\partial t_f^2} > 0, \\ \text{Administrative time: } & \kappa_a \equiv -\frac{\partial^2 v}{\partial t_a^2} = -\alpha \frac{\partial^2 h}{\partial t_a^2} > 0. \end{aligned}$$

In particular, κ_f captures that reducing patient-facing time lowers both internal quality and patient health. In contrast, κ_a contains no internal quality term, because the internal quality measures do not capture the contribution of administrative time to patient health at the margin.

Let v_ℓ^0 and v_ℓ^1 denote the marginal value of input ℓ, $\frac{\partial v}{\partial t_\ell}$, at the input choices under the fixed-wage and the quantity-incentives contract, and let λ_0 and λ_1 denote the shadow price under the two contracts. By Equation (C.4), $v_\ell^0 = \lambda_0$ and $v_\ell^1 = \lambda_1$, so that the marginal value of each input has to change by as much as the shadow price, $v_\ell^1 - v_\ell^0 = \lambda_1 - \lambda_0$. We approximate both sides to first order around the fixed-wage choice. A Taylor approximation on the marginal value gives $v_\ell^1 \approx v_\ell^0 + \frac{\partial^2 v}{\partial t_\ell^2} \Delta t_\ell = v_\ell^0 - \kappa_\ell \Delta t_\ell$. The shadow price, $\lambda = \frac{\mu + v}{T}$, depends on both time inputs and changes in its marginal pay from $\mu^0 = 0$ to $\mu^1 = p$ when moving from the fixed-wage to the quantity-incentives contract. A Taylor approximation on the shadow price gives $\lambda_1 \approx \lambda_0 + \frac{\partial \lambda}{\partial \mu}(\mu^1 - \mu^0) + \frac{\partial \lambda}{\partial t_f} \Delta t_f + \frac{\partial \lambda}{\partial t_a} \Delta t_a$, where $\frac{\partial \lambda}{\partial \mu} = \frac{1}{T}$. As $\frac{\partial T}{\partial t_\ell} = 1$, $\frac{\partial \lambda}{\partial t_\ell} = \frac{1}{T^2}[\frac{\partial v}{\partial t_\ell} T - (\mu + v)] = \frac{1}{T}[\frac{\partial v}{\partial t_\ell} - \lambda]$, which is zero at the fixed-wage choice because $v_\ell^0 = \lambda_0$. The changes in the time inputs therefore do not affect the shadow price to first order, and $\lambda_1 \approx \lambda_0 + \frac{p}{T_0}$: the quantity-incentives contract raises the shadow price of time by p/T_0. Inserting the two approximations into $v_\ell^1 - v_\ell^0 = \lambda_1 - \lambda_0$ gives $-\kappa_\ell \Delta t_\ell \approx \frac{p}{T_0}$. The physician thus reduces input ℓ by $\frac{p}{T_0}/\kappa_\ell$ minutes:

$$\Delta t_f \approx -\frac{1}{\kappa_f} \frac{p}{T_0} < 0, \quad \Delta t_a \approx -\frac{1}{\kappa_a} \frac{p}{T_0} < 0. \tag{C.5}$$

The physician reduces fewer minutes of the input with the larger κ_ℓ, as the marginal value of

that input rises more quickly when time is reduced. In particular, patient-facing time is reduced by less than administrative time if and only if $\kappa_f > \kappa_a$. Writing out κ_f and κ_a gives Condition (3). Administrative time falls at most to $\underline{t}_a$, where the minimum quality requirement (C.3) binds. This establishes Prediction 1(a).

Internal quality, patient health, and downstream costs. Clinical process quality is unchanged, $\Delta q_a = 0$, because $t_a > \underline{t}_a$ under either contract type. To first order, the change in internal quality is $\Delta q = \Delta q_f \approx \frac{\partial q}{\partial t_f} \Delta t_f < 0$, because internal quality does not respond to administrative time at the margin (Equation C.2). This establishes Prediction 1(b). To first order, the change in patient health is $\Delta h \approx \frac{\partial h}{\partial t_f} \Delta t_f + \frac{\partial h}{\partial t_a} \Delta t_a < 0$. Internal quality changes only with patient-facing time and misses the health loss from administrative time, $\frac{\partial h}{\partial t_a} \Delta t_a$. Downstream costs capture the full change in patient health, $\Delta D \approx D'(h)\, \Delta h > 0$. This establishes Prediction 2.

C.4 Physician heterogeneity

Physicians differ in the weights α and β. Let $v(n)$ denote the value of a consultation when the physician works at pace n and divides the time per consultation, $T = 60/n$, net of the given break time t_z optimally between patient-facing and administrative time:

$$v(n) = \max_{t_f + t_a = 60/n - t_z} \alpha\, h(t_f, t_a) + \beta\, q(t_f, t_a), \quad \text{subject to} \quad q_a(t_a) \geq \underline{q}_a. \tag{C.6}$$

Cost of pace. Under the fixed-wage contract, $B_0(n) = 0$, so that utility (2) is $U_0 = \bar{W} + n\, v(n)$ and the physician chooses the pace n_0 that maximizes the consultation value per staffed hour, $n\, v(n)$. To work at any faster pace $n > n_0$, she has to give up consultation value, costing

$$c(n) = n_0\, v(n_0) - n\, v(n) \geq 0. \tag{C.7}$$

Under the quantity-incentives contract, $U_1 = \bar{W} + B_1(n) + n\, v(n) = U_0(n_0) + B_1(n) - c(n)$, so that the physician chooses the pace that maximizes her gain over the fixed-wage contract,

$$G(n) = B_1(n) - c(n), \qquad B_1(n) = p\,(n - 5)\, \mathbf{1}\{n \geq 5\}. \tag{C.8}$$

This is the problem of a worker who is paid a piece rate (Lazear 2000), except that the cost $c(n)$ is the cost of the consultation value given up rather than a cost of effort. If the physician raises her pace mainly by reducing administrative time, the value given up per minute is $\alpha\, \partial h / \partial t_a$, because internal quality does not respond to administrative time at the margin (Equation (C.2)); $c(n)$ is then higher for a physician with a higher α.

Adoption. Let $n_1 = \operatorname{argmax}_{n \geq 5} G(n)$ denote the physician's optimal pace above the threshold, and $G_1 = G(n_1) = B_1(n_1) - c(n_1)$ her gain from responding. A physician with $n_0 > 5$ has $G_1 \geq G(n_0) = p\,(n_0 - 5) > 0$ and responds. A physician with $n_0 < 5$ must first raise her pace to

the threshold, incurring $c(5)$ without compensation, before she earns p per additional consultation. Her gain G_1 is larger the closer n_0 is to the threshold, which lowers $c(5)$, and the lower her cost of a faster pace $c(n)$. A physician with $G_1 \leq 0$ keeps n_0 under the quantity-incentives contract. Because $B_1(n_0) = 0$, she is equally well off under either contract, $U_1(n_0) = U_0(n_0)$, and we assume that she does not take up the quantity-incentives contract voluntarily. Under voluntary adoption, only physicians with $G_1 > 0$ change their time inputs. If s denotes their share, the average effect on time per consultation across all physicians is $s\,\mathbb{E}[T_1 - T_0 \mid G_1 > 0]$, the treatment effect on the treated times the share of voluntary adopters. This establishes Prediction 3(a).

Dispersion. Every physician with $G_1 > 0$ chooses $n_1 \geq 5$, so that her time per consultation under the quantity-incentives contract is $T_1 = 60/n_1 \leq 12$ minutes, the time per consultation at the threshold pace, so that her consultation time is $t_f + t_a \leq 12 - t_z$. Under the fixed-wage contract, $T_0 = 60/n_0$ has no upper limit. Because t_z is given, her reduction in consultation time $t_f + t_a$ equals $T_0 - T_1$ and is therefore at least $T_0 - 12$, which is larger the slower her pace under the fixed-wage contract, and the time per consultation of physicians who respond is bounded above by 12 minutes, whereas it is unbounded under the fixed-wage contract. She divides the reduction between the two inputs so that their marginal values stay equal, which to first order removes κ_f/κ_a minutes of administrative time per minute of patient-facing time (Equation C.5). Under Condition (3), the reduction therefore falls mainly on administrative time. This establishes Prediction 3(b).