Organizing Expertise with Externalities: Evidence from Primary Care

Amanda Dahlstrand[1], Shan Huang[2], Nestor Le Nestour[3], and Guy Michaels[4]

[1] *University of Zurich,* [2] *Stockholm School of Economics,* [3] *Stockholm University,* [4] *LSE*

September 9, 2026

Abstract

Many modern organizations use knowledge hierarchies, where frontline workers try to resolve tasks and pass those they cannot resolve to experts. But the implications of this organization of expertise are not well understood, especially if firm revenues depend on who resolves each task or the firm generates downstream externalities. We study a knowledge hierarchy (KH) using 500,000 patient cases from an online primary care firm, linked to outcomes and costs across Sweden's healthcare system. We use variation in assignment arising from temporary doctor congestion, controlling for date-by-three-hour windows and flexible case characteristics. Within the KH, nurses resolve 70% of cases and pass the remainder to doctors. Relative to doctor-only care, the KH modestly increases external primary care and ED visits but has no adverse effects on hospitalizations, earnings losses, or mortality. Although the mean effect of KH on systemwide (firm plus downstream) production costs is small, this masks substantial heterogeneity. Compared to using only doctors, the savings from the firm's selective use of the KH equal 1.6% of online primary care costs. However, reassigning cases between the direct-to-doctor and KH routes within 15-minute windows to minimize systemwide production costs would save 6.8% of online primary care costs.

Keywords: Division of Labor, Productivity, Healthcare

JEL codes: J24, I19, L23

An earlier title for this project was "Does Division of Labor Increase Productivity? Evidence from Primary Care". We thank Ellen Donath, Smrithi Ganesh, Emma Persson, Lisa Sennfält, and Berk Uzunonat for excellent research assistance. We also thank our discussant at NBER Organizational Economics, Carol Propper, as well as colleagues and seminar and conference participants at Aarhus University, Arne Ryde workshop in Lund, Becker Friedman Institute Health Economics conference, CEP–IFS Labour Economics Conference, Cornell-Bath Conference, CEPR Paris Symposium, CESifo Summer Institute, Columbia MAD conference, EIT Oxford, EuHEA, Frisch Centre, HEFUU, HPI-ZEW-DIW Workshop, IFAU, Imperial, Joachim Herz Foundation Add-on Fellowship Meeting in Hamburg, Johannes Kepler University, LSE, NBER Organizational Economics, RFBerlin-CEPR Annual Symposium in Labour Economics, Royal Holloway University, SSE, University of Innsbruck, University of Konstanz, University of Oslo, University of St. Gallen, University of Sussex, University of Zurich, URPP Annual Retreat in Zurich, and USI Lugano 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 the ESRC's Centre for Economic Performance, the Swiss National Science Foundation, grant 100018-228312, Handelsbankens Forskningsstiftelser, grant P25-0261, and support from the University of Zurich's Research Priority Program "Equality of Opportunity" and SSE's House of Sustainable Society. The research project was approved by the Stockholm Ethics Review Board and the Swedish Ethics Review Authority.

1 Introduction

The division of labor and occupational specialization have long been central to economic studies of productivity (Smith 1776; Becker and Murphy 1992). In modern organizations, the division of labor often centers on organizing expertise using *knowledge hierarchies*, in which frontline workers attempt to resolve tasks and pass to experts those they cannot resolve (Garicano 2000). Despite the prevalence of knowledge hierarchies, there is little empirical evidence on their productivity effects, the type of tasks assigned to them, or whether their use is efficient. This is especially true when firm revenues are contingent on who resolves a task, when task assignment creates downstream externalities (such as the shifting of tasks to experts in other organizations), or both.

Empirically evaluating knowledge hierarchies relative to a natural benchmark (the direct-to-expert route) has proven difficult for three main reasons: tasks assigned to frontline workers often do not overlap with those assigned directly to experts; even when overlaps exist, assignment is endogenous; and data rarely link within-firm task assignment to downstream quality and costs outside the firm. We study a knowledge hierarchy (KH) in a setting that allows us to overcome these empirical challenges and identify the effects of a knowledge hierarchy, including its heterogeneous effects across task types and its downstream implications.

We ask three closely related research questions. First, how do the quality of outcomes and systemwide (firm plus downstream) production costs differ between a KH and a direct-to-expert route? Second, what determines task allocation between these routes, and what are the resulting cost implications inside and outside the organization? And crucially, which tasks should be assigned to the KH, taking into account downstream externalities?

To guide our empirical analysis, we develop a parsimonious model of task assignment between a knowledge hierarchy and a direct-to-expert route when frontline workers' resolution attempts are costly, revenues depend on who resolves a task, and there are downstream externalities. The model highlights (i) how firms assign tasks; (ii) how contingent revenue structures affect whether profit maximization entails cost minimization; and (iii) how downstream externalities can lead to task misallocation.

Understanding how knowledge hierarchies are used in practice is especially important in sectors where expertise is scarce, services are often publicly financed, and demand is rising (Kelly, Stoye, and Vera-Hernández 2016). The healthcare sector, which we study, exemplifies these forces: aging populations, clinician shortages, and rapid technological change are reshaping how expertise is produced and deployed (Dahlstrand 2026; Dahlstrand, Le Nestour, and Michaels 2024; Bronsoler, Doyle, and Van Reenen 2022; Zeltzer et al. 2023). In such settings, when organizations assign tasks between a knowledge hierarchy and a direct-to-expert route, their assignment can generate downstream spillovers across other providers and payers – possibly including taxpayers (e.g., in Medicare, Medicaid, and many European health systems). We quantify these productivity and externality effects, bringing organizational design directly into the domain of public policy.

Our setting is a large Swedish firm, which provides taxpayer-financed primary healthcare services online. The tasks we analyze are patient cases with different symptoms, demographics, and medical histories. The experts are doctors, and the frontline workers are nurses. The firm's algorithm assigns cases directly to doctors or initially to nurses (the KH). This algorithmic assignment helps our identification, since it relies on a predetermined set of characteristics, rather than on subjective human judgment – leaving little room for unobservables to influence assignment. The firm maximizes its net revenue, subject to institutional constraints. We focus on "overlap" cases, for which the firm has discretion in assignment between the two routes. When the short-term availability of doctors is unconstrained, the algorithm assigns all overlap cases to doctors; however, when constrained by temporary shortages, it assigns some cases to nurses. We show that this constrained allocation also minimizes taxpayer payments to the firm (holding clinician use constant). However, it does not minimize systemwide production costs, since the firm is not incentivized by the government to internalize the downstream costs, which are dependent on the route assignment – a problem common to many contractual environments.

For our empirical analysis, we construct a novel dataset with many health outcomes within the firm and downstream. We focus on firm records for almost 500,000 overlap cases, which we link to administrative data on patients' demographic and socioeconomic characteristics and their healthcare utilization in the universal healthcare system *outside* the firm. This allows us to comprehensively characterize patients' care journeys and measure costs.

To address the sorting of cases between routes, our main identification strategy flexibly controls for fixed effects defined by three-hour windows within each calendar day and for interactions of case characteristics observed by the algorithm. We show that, conditional on these controls, case characteristics that the algorithm cannot observe are almost exactly balanced. Our main findings are also robust to using a measure of short-term doctor congestion as an instrumental variable. This congestion only causes short delays – over 94% of waiting times are less than an hour, 99.9% are less than six hours, and all are less than 24 hours. We show in placebo tests that such delays are unlikely to affect health outcomes.[1]

Our main analysis focuses on healthcare utilization and costs within seven days, when the effects are plausibly attributable to the organizational route at the initial contact. We find

[1]This is plausible in less acute primary care settings such as ours, unlike the hospital emergency department settings studied by Costantini (2025), Godøy et al. (2024), or Gruber, Hoe, and Stoye (2023).

that about 70% of the cases assigned to the KH are resolved by nurses, while the remaining 30% are sent to doctors. The effects on patient satisfaction are small. However, the KH moderately reduces diagnostic informativeness and the rate of new prescriptions relative to the direct-to-doctor route, leading us to investigate what happens downstream. The KH increases the rate of external primary care visits and emergency department visits, which increases downstream costs (discussed below).

Despite tasks shifting to experts in other organizations, the KH does not significantly change the incidence of hospitalizations or avoidable hospitalizations. Moreover, we find no adverse effects of the KH on major earnings losses in the calendar month following the initial consultation, nor on three-year mortality rates. Together, these results suggest that on average, the KH does not materially affect patient health.

We next assess the KH's effects on costs, both within the firm and in downstream healthcare services used within seven days. We estimate that the KH reduces taxpayer payments to the firm (equivalent to firm gross revenues) under current payment rules, where the firm is only paid for the clinician who resolves the case but incurs costs for all clinicians involved. This lowers the firm's revenue net of labor input (firm net revenue). Despite the increase in ED and external primary care visits, when all downstream costs borne by taxpayers are accounted for, taxpayer payments decrease by 8% under the KH relative to the direct-to-doctor route. By contrast, the proportional effect of KH on mean systemwide production costs (abstracting from transfers between taxpayers and the firm) is close to zero and, in most specifications, statistically insignificant.

However, the cost effects of the KH are heterogeneous, as some case types are easier for nurses to resolve themselves, and case types also differ in their expected downstream costs. We find that the KH almost halves systemwide production costs for almost half of patients, but on average, this cost saving is offset by increased costs for other patients higher up in the distribution. This substantial heterogeneity suggests that the assignment of tasks to the KH or directly to doctors is an important lever for healthcare productivity.

Our results also indicate that the KH's effects are not driven solely by occupational licensing differences between nurses and doctors. First, if nurses only lacked doctors' scope of practice, they could pass restricted cases to them, and the two routes should differ only in within-firm costs. Instead, both routes reduce downstream costs for different symptoms. Second, the KH lowers the probability of an informative diagnosis. Finally, the extent to which each symptom's cases receive prescriptions under direct-to-doctor assignment has limited predictive power for the probability that cases are passed to doctors when assigned to nurses. All this suggests that clinical knowledge, and not only licensing, shapes the KH.

Turning to the economic drivers of case assignment to routes, we find that when the firm is unconstrained (there is little doctor congestion) it assigns almost all cases directly to doctors. When short-term doctor congestion arises, the key characteristics that determine assignment are the patients' symptoms: symptoms with a higher relative net revenue advantage under the KH (cases that nurses are likely to resolve within the firm) are more likely to be assigned to nurses. Consistent with the model, this assignment pattern also reduces taxpayer payments to the firm, conditional on fixed clinician use. Compared to a counterfactual scenario in which only the direct-to-doctor route is used, the firm's selective use of the KH reduces systemwide production costs by 0.4% (1.6% relative to online primary care costs).

Finally, we consider allocations that assign cases to minimize systemwide production costs, while respecting short-run clinician capacity constraints. Reallocating cases on this basis within 15-minute windows without changing actual nurse or doctor use reduces systemwide production costs by 1.9% (6.8% relative to online primary care costs). The decline in the firm's net revenue in this scenario is trivial relative to the societal gain. One approach to implementing such a solution is to issue guidelines to align the firm's assignment with the planner's, although in the longer term, changes in staffing and hiring may cause new challenges. Another solution to align the firm's incentives is to use Pigouvian pricing to make it internalize the downstream externalities from its task assignment.

To demonstrate the external relevance of our findings, we combine several pieces of evidence. We show that our setting shares key institutional features with healthcare systems in other OECD countries; that there is considerable overlap between the symptoms in our setting and those seen in in-person primary care; that similar proportions of the most common symptoms are handled by nurses in in-person care; and that our main results are similar when we reweight by demographics to represent in-person care users, vary firm wedges, or conduct several robustness checks related to COVID-19. Although the magnitudes of the effects and the welfare implications may differ in other settings, the issues that we study – especially the assignment of different task types to a KH or directly to experts – are relevant for other contexts where downstream externalities are not internalized. Our setting reflects a broader problem of task allocation across skill levels when processing at different tiers generates downstream spillovers outside the decision-maker's objective. For example, in customer support, delegation may affect retention and platform demand; in infrastructure maintenance, network reliability and congestion; and in financial screening, portfolio risk borne by investors or taxpayers. In such settings, privately optimal delegation need not coincide with socially efficient assignment.

Our paper contributes to the literature by providing the first task-level causal evidence on the productivity and efficiency implications of knowledge hierarchies, which are central to many organizations. We differ from canonical studies of knowledge hierarchies, e.g., Garicano

(2000), in three ways: first, in our empirical focus; second, in the cost of knowledge hierarchies that we emphasize – where the canonical model highlights communication costs, these are limited in technology-mediated settings that are increasingly common today, and we instead focus on the costs incurred as frontline workers must attempt a task before determining whether they can resolve it; and finally, in our consideration of downstream externalities.

We also contribute to the empirical evidence on task allocation across specialized occupations in healthcare. Recent work documents the rapid expansion of mid-level occupations such as nurses (Gottlieb et al. 2025); studies occupational licensing and scope-of-practice reforms, particularly for nurse practitioners, and their effects on healthcare provision (Kleiner 2016; Alexander and Schnell 2019), and compares the productivity of clinician types performing similar tasks in emergency departments (Chan and Chen 2026). We differ from this work in three respects. First, rather than treating clinician types as substitutes or evaluating practice rights expansions, we examine a knowledge hierarchy in which nurses attempt cases first and selectively pass on those they cannot resolve – the organizational routes we study differ in how expertise is combined across occupations rather than in who supplies it. Second, we study primary care, which is less studied than emergency care, but where task allocation determines whether patients enter downstream care in the first place. Third, holding scopes of practice fixed, we ask how occupational competencies can best be deployed within organizations once costs borne beyond the focal organization (externalities) are accounted for.

A broad literature emphasizes how productivity depends on specialization and task assignment (Becker and Murphy 1992; Katz and Murphy 1992; Jones 2009; Acemoglu and Autor 2011; Ide and Talamàs 2025). Knowledge hierarchies provide a natural organizational framework for operationalizing these ideas and organizing expertise. Yet, despite their central role in theory, there is limited empirical evidence on how knowledge hierarchies affect productivity and costs in practice, or whether they are implemented efficiently in organizations. Existing empirical work, including Caliendo, Monte, and Rossi-Hansberg (2015) and Caliendo et al. (2020), documents the structure of production hierarchies and their relationship to firm productivity using firm-level data, with a macro or reduced-form perspective. Related evidence shows that managerial and organizational practices shape productivity in healthcare specifically (Bloom et al. 2015). In contrast, we study task-level assignment decisions within an organization, using quasi-experimental variation in the routing of tasks to a knowledge hierarchy or directly to experts. This micro-level perspective allows us to measure not only productivity effects, but also misallocation across tasks and the resulting efficiency losses when externality effects are not internalized, and to quantify the potential systemwide gains from reallocating tasks.

The remainder of the paper is organized as follows. Section 2 introduces the institutional background and Section 3 our data. Section 4 discusses the model that guides the empirical analysis. Section 5 describes our empirical framework. Section 6 discusses the productivity effects of the knowledge hierarchy. Section 7 discusses the firm's use of the knowledge hierarchy and how the allocation of cases can be made more productive. Section 8 concludes.

2 Institutional setting

To study the effects of division of labor in a knowledge hierarchy, we examine how a large Swedish online primary care organization, referred to as *the firm*, manages patient cases. The firm employs two clinician types, *nurses* and *doctors*, to resolve patient *cases*. Below, we describe the primary care setting in general, the particulars of the Swedish online setting, and the firm's operations. Additional institutional details are provided in Appendix A.

Primary care is the main entry point into the healthcare system for most patients: clinicians diagnose and treat common conditions and determine whether patients require downstream care. When care for a case is completed in primary care, it generates no subsequent referral, specialist, or emergency department visit. Hence, task allocation in primary care can affect costs throughout the healthcare system. Primary care is therefore viewed as a central lever for containing expenditure growth (National Academies of Sciences, Engineering, and Medicine 2021), although existing evidence remains largely associational (Basu et al. 2019).[2]

Sweden's healthcare system provides comprehensive universal coverage, which is largely tax-financed, with low patient co-pay.[3] Appendix Table A1 compares key features of Swedish healthcare to other OECD countries. In Sweden, patients register with a primary care center but may also seek online primary care from other providers. Alongside public providers, private for-profit providers operate under regulated contracts with regional authorities and account for roughly 40% of primary care centers (The Commonwealth Fund 2020).

The firm we study is a large private provider operating nationwide (and in several other countries), making it Sweden's – and Europe's – largest digital primary care provider in the study period. It employs hundreds of clinicians in Sweden, comparable in scale to large group practices in the US. Although it runs many in-person clinics for registered patients, it also provides nationwide online primary care to any citizen, which we focus on in this paper. Both types of primary care are reimbursed under the national health insurance. Online

[2]Despite the importance of primary care, economists have mostly focused on downstream settings, where emergency departments and hospitals offer high-frequency variation (Chan 2018; Silver 2021; Doyle et al. 2015), while such variation, and in particular exogenous variation, is rare in primary care settings.

[3]During our sample period, annual out-of-pocket spending for outpatient visits was capped at roughly 1,150 SEK, or about USD 125 (1 SEK = 0.1087 USD in 2020, Sveriges Riksbank 2020). Online primary care visits carried a copay of 100–200 SEK (about USD 11–22) for adults and were free for minors and the elderly; these payments counted toward the annual cap. When a copay is used, part of the administratively set fee-for-service payment to the healthcare provider is made directly by the patient.

consultations have become common in Sweden in recent years, where the use of telehealth (around a quarter of doctor consultations) is similar to many other OECD countries, including the US, and prescription treatments can be provided online (Appendix Table A1).

Like many other Swedish primary care providers, the firm is a profit-maximizer (Kry International AB 2024) whose clinical practices and reimbursement are governed by national regulation and regional agreements, and are subject to supervision. Its revenues from online care for unregistered patients (which we focus on) are determined by fee-for-service with administratively set prices, a common model internationally (Appendix Table A1). The firm is paid by regional authorities per nurse or doctor consultation according to levels intended to reflect labor costs and an assumed overhead (Sveriges Kommuner och Regioner (SKR) 2019).

The firm's task assignment is influenced by a notion of 'cost-effectiveness' in healthcare, where care is to be delivered by the lowest-cost clinician type capable of resolving the case. Accordingly, the firm is paid only for the highest-cost consultation per day.[4] As a result, the firm always assigns certain cases to nurses, others always to doctors, and the remainder ('*overlap-group cases*') to either clinician type. Cases of a given type are processed as 'first-come-first-served' (FCFS), reflecting expectations of transparency and predictability and limiting long or arbitrary waits. Although these constraints reflect the Swedish environment, similar considerations are common in many other settings.

Patients access healthcare for themselves or their children using a mobile application and an electronic identification system used across Sweden's healthcare and government institutions. This automatically informs the firm of patient data from civil records (age, gender, and region) but not of other variables available to us as researchers (e.g., socioeconomic status and previous care throughout the healthcare system).[5] The patients then report their main symptom from a predefined list and may provide some additional information. Consultations can be accessed through a drop-in queue or scheduled for a later time. When the drop-in queue is selected, the case is assigned to a clinician type for the initial consultation; after payment of any copay (which is small for all patients and exempt for many of them), the case enters a single, Sweden-wide waiting queue.

Our study focuses on *overlap cases* after spring 2019 from patients who were not registered with the firm. The firm was thus not these patients' in-person primary care provider, but an additional venue for care. Patients commonly turn to the firm and similar providers

[4]Similar revenue structures arise in other knowledge-hierarchy settings. For example, in IT and customer support, payment may be non-additive across layers, and in professional services such as law and consulting, higher-skill work is often billed at higher rates, but contracts may be similarly non-additive.

[5]Access to data containing individual-level socioeconomic information is strictly restricted and generally limited to statistical and research purposes under confidentiality protections (*Offentlighets- och sekretesslag (2009:400) [Public Access to Information and Secrecy Act]* 2009).

due to shorter wait times and the convenience of receiving care from home. The revenue structure was transparent, and these patients' assignment to nurses or doctors was governed by a nationwide algorithm. This algorithmic assignment was based primarily on the reported symptom, age, gender, and region, as well as clinician availability; the precise algorithm has evolved over time and remains proprietary. Because staffing is set in advance, short-term excess demand relative to clinician availability is common. During short periods of doctor congestion, cases are assigned to nurses to reduce waiting times, and the firm may also contact off-shift clinicians, generating variation in congestion within short time frames.[6]

We classify cases that are 'initially assigned to nurses' – who may resolve them or pass them to a doctor – as following the knowledge hierarchy route. This contrasts with cases assigned directly to doctors. Throughout the paper, *assignment* refers to the initial assignment of the case – either to the knowledge hierarchy route or directly to a doctor.

During our study period, doctors and nurses were paid flat hourly wages or monthly salaries with no bonuses for speed. Many worked part-time at the firm while also working in-person elsewhere. Nurses had completed a 3-year university degree focused on routine care and patient communication, while doctors underwent at least 7 years of formal and on-the-job medical training. Within each clinician type, assignment did not account for specialty (except for cases involving infants, which we exclude). Both nurses and doctors followed similar treatment protocols and could resolve cases with some differences noted below.

Clinicians use a digital interface that shows the consultations assigned to them, but the firm's total current congestion is harder to observe than in most in-person healthcare settings. Upon assignment, clinicians observe patients' demographics, reported symptoms, and medical histories from prior interactions with the firm. Doctor and nurse consultations may result in home-care advice, suggestions about over-the-counter medicines, and/or recommendations about further care (e.g., to visit the emergency department). Doctor consultations may also result in prescriptions, longer-term sickness absence certificates, and/or referrals to specialists outside the firm.[7] In this setting, nurses go beyond triage – they do not merely prioritize or route patients, but often complete the care episode and give health advice.

If a nurse cannot resolve a case, the nurse passes it to a doctor by scheduling a follow-up consultation for a later time. The follow-up request specifies that a doctor is required, but does not name an individual doctor. The nurse also completes a handoff note summarizing the reason for passing the case, and any relevant clinical information. The mean total time

[6] The firm's use of nurses for consultations thus has two purposes: first, to provide the lowest-cost treatment for certain cases, and second, to reduce waiting times when doctors are congested.

[7] Doctors and nurses may also order home tests (such as COVID-19 or chlamydia swabs) and, in the case of doctors, basic laboratory tests (such as simple blood tests). While we do not observe these tests, they are typically inexpensive, and the firm receives no additional payment for laboratory testing.

spent on a case by the initial clinician is 10-12 minutes (with a standard deviation of around 7 minutes), irrespective of clinician type and whether the case was resolved.[8]

3 Data

3.1 Sample construction

Data sources. Our data combine internal information on all cases handled by the healthcare firm in 2019-2020 with administrative records from national Swedish registers. Using patients' national identifiers, we link each case to (i) prescriptions and non-primary care use data from the National Board of Health and Welfare (Socialstyrelsen, 2013–2023); (ii) socioeconomic measures – including education, employment, and monthly earnings from primary employment – from Statistics Sweden (2013–2020); (iii) external primary care for patients in two regions of Sweden, Scania (2013–2020) and Stockholm (2013–2023), which together account for about 53% of cases in the analysis sample. For Scania, we additionally observe population-level data for all primary care visits in 2019, including patients' age, gender, and diagnoses – we use that for external validity checks and weighting. Details are available in Appendix B.

Analysis sample. We construct our main analysis sample as follows (see also Appendix Table A2). Starting with 1.8 million consultations in 2019 and 2020, we restrict to unscheduled online consultations with nurses or doctors between 1 April 2019 (when nurses began handling more symptoms) and 24 December 2020 (to observe revisits in the firm within 7 days). We exclude registered patients (whose primary care provider is the firm), who may have different care assignment paths and may receive in-person care from the firm; infants under the age of two, for whom different assignment rules apply; and cases preceded by consultations with the same patient within the preceding two hours, to focus on initial consultations (follow-up consultations are used for outcome variable construction). We also exclude the few cases with missing baseline characteristics. The resulting *broad drop-in sample* includes 969,483 cases assigned to nurses or doctors.

We classify cases by patient-reported symptoms, which the assignment algorithm uses. First, we classify symptoms that are almost always assigned to one clinician type: *doctor-group cases*, for which at most 1% of cases are assigned to nurses, and *nurse-group cases*, defined analogously, and both exclude rare symptoms (with fewer than 1,000 cases). Second, we define our main analysis sample as the 495,555 *overlap-group cases*, those from symptoms where between 5% and 95% of cases are initially assigned to nurses and the rest to doctors.

[8] The cases that nurses cannot resolve are plausibly more complicated than those they can resolve, explaining why the observed difference in mean consultation times between their resolved and their unresolved cases is small. The small differences that exist are discussed in the robustness analysis in Section 6.3.

These are the cases over which the firm has plausible discretion in assigning to either nurses or doctors. The remaining cases (symptoms with assignment shares of 1-5% to either clinician type and rare symptoms) form the residual group. Appendix Figure A1 reports symptoms and initial clinician counts by group. We discuss incomplete consultations in Appendix Section D.2 and address potentially selective attrition in robustness checks in Section 6.3.

3.2 Variables

We summarize key variables below, leaving precise definitions and details to Table B1 and Appendix B. Our **treatment variable** is assignment to the knowledge hierarchy, namely being *initially assigned to a nurse* rather than directly to a doctor.

Baseline controls (X_1). The firm's algorithm assigns each case to either route based on observable, predetermined characteristics. Key among those are primary symptoms selected by patients before any clinician contact, based on which we classify cases into groups, as discussed above. Within the overlap group, the commonest, *other health inquiries*, is for patients who did not select a specific symptom.[9]. To flexibly capture the algorithm's assignment rules, we interact symptoms with other variables that the algorithm observes: the patient's age, gender, and prior cases at the firm (whether the patient's most recent observed case at the firm – if any – was assigned to a doctor, and whether the patient has received a prescription written by a doctor at the firm). We also include patients' aggregated region of residence outside the interaction term. Finally, we include fixed effects for date-by–3-hour windows based on the case's login time – the moment the patient registers the case and begins waiting for the initial consultation – to absorb broader patterns in demand and availability.

Additional controls (X_2). These are (almost certainly) unobserved by the algorithm and thus cannot be used in assignment: (i) health-risk measures based on healthcare use (hospital, emergency, urgent care, and specialist care) in the three years prior to the consultation (excluding the final two days before the consultation), and comorbidities from specialist or hospital care during the same period; and (ii) socioeconomic characteristics measured in 2018 (indicators for above-median earnings, welfare benefit receipt, education level, employment, and civil and migration status) and an indicator for missing any socioeconomic data.

Quality and downstream care outcomes. To assess how the knowledge hierarchy affects quality and downstream care for patients, we study (i) within-firm outcomes, capturing whether the initial clinician passes the case to a doctor (or another doctor consultation in the case of the direct-to-doctor route), case resolution, patient satisfaction, and diagnostic informativeness; (ii) downstream healthcare use within seven days after the initial consultation

[9]The commonest (clinician-set) ICD diagnoses within *other health inquiries* are: "Other counseling/medical advice" (7.14%); "Pain, unspecified" (5.84%); and "Soft tissue disorders, unspecified" (5.18%).

date, including any new prescription and use of external primary, specialist, urgent, emergency, and hospital care; and (iii) longer-run consequences, capturing earnings declines and mortality to capture adverse events for patients outside their direct healthcare utilization.

Cost outcomes. These include taxpayer payments to the firm and in total; firm revenues; and systemwide production costs of care provision (both within-firm and downstream in the healthcare system). For cases resolved (within the firm) by nurses and doctors, the firm receives 275 SEK and 500 SEK, respectively (Sveriges Kommuner och Regioner 2024). To infer production costs, we assume a wedge between the administratively-set price and the firm's production costs of a consultation, discussed further in Section 4.2. We measure within-firm costs on the initial consultation day unless stated otherwise. We also measure downstream costs — the costs of out-of-firm care borne by taxpayers — counting up to one instance of each downstream event within seven days after the initial consultation. Downstream costs are calculated using service-specific cost estimates from Appendix Table A3. In robustness checks, we relax these assumptions and include a full aggregation of downstream events.

Congestion. We measure congestion using *log median doctor wait time* in the (up to) five preceding doctor-assigned cases 30 minutes before login, winsorized at the 99th percentile (Appendix Figures A2a and A2b show raw and transformed distributions). The probability of initial nurse assignment increases monotonically with congestion (Appendix Subfigure A2c).

Summary statistics. Table 1 reports summary statistics for the main analysis sample. Differences in initial assignment are driven primarily by symptoms, reflecting the firm's algorithm: cases assigned to nurses more often involve diffuse symptoms such as *other health inquiries*, *abdominal pain* and *fever*, while cases assigned to doctors include more women, partly reflecting more specific symptoms such as *urinary tract infection* being female-specific and sent more frequently to doctors. Differences in some other characteristics are statistically significant but typically small in magnitude. While this table shows the raw data assignment, Section 5 discusses balance, accounting for the baseline controls in our identification strategy.

The analysis sample spans a broad range of demographic characteristics, but its composition differs somewhat from both the Swedish population and patients using in-person primary care in Scania, a large region for which we observe the universe of primary-care visits. The analysis sample also has a lower share of nurse consultations than in-person primary care (Appendix Table A4). We address the demographic differences in robustness checks by reweighting the sample to match the age and gender distribution of in-person primary care. Over the study period, the firm handled on average 782 consultations per day (23% assigned to nurses); this increased sharply to 1205 (34% assigned to nurses) in the first COVID-19 wave from 1 March to 5 July 2020, before declining to 820 (18%) thereafter. We assess the sensitivity of our results to the COVID-19 period in the robustness analysis.

4 Model

The model's main role is to organize our empirical analysis and clarify mechanisms that determine the use of knowledge hierarchies. We examine when the firm's revenue structure shifts it away from cost minimization, and also analyze the case of downstream externalities.

We begin with a general model of a firm that faces a stream of heterogeneous tasks (e.g., patients, customers, orders, or applications) that require processing. We focus on short-run task assignment, abstracting from staffing and hiring decisions. Tasks can be assigned either directly to experts or to frontline workers, who resolve a subset and pass the remainder to experts (the knowledge hierarchy, KH). The model highlights (i) how firms determine task allocation; (ii) how revenue structures (e.g., differential pay for expert versus frontline worker resolution) affect whether profit maximization entails cost minimization; and (iii) how downstream externalities can lead to task misallocation. We then apply this model to our setting, where tasks are patient cases, frontline workers are nurses and experts are doctors, and the firm is paid by taxpayers. Finally, we discuss the model's implications for our empirical analysis.

Our model emphasizes the cost of frontline workers' attempts to resolve tasks even when these attempts fail. This differs from the canonical model of knowledge hierarchies (Garicano 2000), which emphasizes communication frictions as the key cost. We note that communication costs have fallen dramatically in recent years (see, e.g., Bloom et al. 2014) and are likely less important than attempt costs in many settings, including ours. We also differ in that we examine what happens when firm revenues depend on who resolves a task.

4.1 The general firm problem

This section characterizes the firm's task assignment and the conditions under which it aligns with – or diverges from – cost minimization. We consider a profit-maximizing firm that employs frontline workers (F) and experts (E). The firm faces a stream of tasks that all need to be resolved, drawn from a finite set of task types $\mathcal{T} = \{\tau_1, \ldots, \tau_S\}$. Each task i has type $\tau_i \in \mathcal{T}$ and is associated with a resolution probability $p_{\tau_i} \in (0, 1)$ – the probability that a frontline worker assigned to a task of type τ_i resolves it.[10] When a frontline worker cannot resolve a task, they pass it to an expert, and we assume that experts resolve all tasks passed or assigned to them. "Resolving" means ending the firm's processing of the task, but resolved tasks may still involve downstream (out-of-firm) costs, as we discuss below.

Let $\kappa_F > 0$ (resp. $\kappa_E > 0$) denote the firm's cost of supplying one frontline (resp. expert) worker to process a task, incurred regardless of whether the task is resolved, and assume

[10] We assume that the firm knows these parameters, leaving firm learning for future work.

$\kappa_E > \kappa_F$.[11] We assume that the task processing time per worker is constant. The firm's gross revenue per task is r_F if it is assigned to the frontline worker, who resolves it; $r_{F,E}$ if it is assigned to a frontline worker but resolved by the expert; and r_E if the task is assigned directly to an expert.

We assume that the firm's revenue and cost structure, and its staffing, are set before task assignment. Thus, when assigning tasks, the firm seeks to maximize its expected net revenue (gross revenues minus variable costs), ignoring fixed costs that do not vary with task assignment. For a task of type $\tau \in \mathcal{T}$, the firm's expected net revenue under the KH and the expert route are, respectively: $\Pi_F(\tau) \equiv p_\tau (r_F - \kappa_F) + (1 - p_\tau)(r_{F,E} - \kappa_F - \kappa_E)$ and $\Pi_E \equiv r_E - \kappa_E$. We define the *net revenue advantage of the KH route* for type τ as

$$\Delta(\tau) \equiv \Pi_F(\tau) - \Pi_E. \tag{1}$$

Since p_τ enters $\Pi_F(\tau)$ only as a mixing weight between two fixed payoffs, and all revenue and cost parameters are constants, $\Delta(\tau)$ is linear in p_τ with constant slope, given by:

$$\frac{\partial \Delta(\tau)}{\partial p_\tau} = r_F - r_{F,E} + \kappa_E. \tag{2}$$

We define the **cost minimization condition** as $\frac{\partial \Delta(\tau)}{\partial p_\tau} > 0$. Intuitively, if this condition holds and the firm has to choose how to assign two tasks between a frontline worker and an expert, it assigns the "easier" task to the frontline worker. By "easier" we mean the task with higher p_τ – the task that the frontline worker is likelier to resolve within the firm. Here and throughout the model, we assume that when indifferent, the firm breaks ties randomly.

Firm constraints. The firm faces two task-assignment constraints.

Weak first-come-first-served (FCFS) constraint. The firm must (weakly) process tasks in the order they arrive when workers are available. This allows the firm to batch-process the remaining tasks in the queue that arrived earliest (subject to worker availability), and push the remaining tasks to the next period.[12]

Staffing availability constraint. In each (short) time period t, E_t experts and F_t frontline workers are available to process tasks from a queue; the firm assigns tasks to available workers while the queue is non-empty. We assume that tasks passed from frontline workers to experts are processed later, outside the main queue. Since we focus on settings with rapid task processing, we also abstract from anticipation effects.

Solving the firm's problem. Let M_t denote the number of tasks, taken from the front of the queue, to be assigned in period t, where $M_t \leq E_t + F_t$. Any remaining tasks are moved to the front of the queue for period $t + 1$. The firm is *unconstrained* if the number of tasks

[11] In our empirics, we will show robustness to assuming that κ_F is lower for unresolved tasks.

[12] Some firms may apply FCFS within task groups. We discuss such a scenario in Appendix C.

with $\Delta(\tau) > 0$ does not exceed F_t and the number of tasks with $\Delta(\tau) < 0$ does not exceed E_t; otherwise, assignment is *constrained.*

Result 1: Firm assignment. *The* ***unconstrained*** *firm assigns all tasks with $\Delta(\tau) > 0$ to frontline workers and tasks with $\Delta(\tau) < 0$ directly to experts. The* ***constrained*** *firm allocates scarce worker types to the tasks where they have the largest net revenue advantage. The firm ranks tasks by $\Delta(\tau)$, assigns frontline workers to the highest-ranked tasks with $\Delta(\tau) > 0$ up to capacity F_t, assigns experts to the lowest-ranked tasks with $\Delta(\tau) < 0$ up to capacity E_t, and any remaining available workers to any remaining tasks.*

Due to its contingent revenue structure, the unconstrained firm does not necessarily minimize its production costs. Whether the constrained firm minimizes production costs depends on the cost minimization condition.

Result 2: Firm production cost minimization. *If $\frac{\partial \Delta(\tau)}{\partial p_\tau} > 0$, then under staffing constraints, the firm assigns frontline workers to the tasks they are most likely to resolve (the easiest tasks). In these circumstances, the constrained firm's profit maximization coincides with the minimization of firm production cost and firm gross revenues.*

Comparative advantage. When the cost minimization condition ($\frac{\partial \Delta(\tau)}{\partial p_\tau} > 0$) holds and downstream externalities are absent, the direct-to-expert cost κ_E is constant across task types, so ranking tasks by firm net revenue advantage (i.e., increasing in $\Delta(\tau)$, which itself increases in p_τ) coincides with ranking them by the cost advantage of the KH (i.e., decreasing in $(\kappa_F + (1 - p_\tau)\kappa_E)$), which in turn is equivalent to ranking them by comparative advantage (i.e., decreasing in $[\kappa_F + (1 - p_\tau)\kappa_E]/\kappa_E$). The firm's assignment – both unconstrained and constrained – then follows comparative advantage.

Perverse incentives arise when $\frac{\partial \Delta(\tau)}{\partial p_\tau} < 0$. This happens when the revenue premium from passing tasks to experts $r_{F,E} - r_F$ is large enough to offset the expert cost κ_E. In these circumstances, the constrained firm assigns frontline workers to *harder* tasks to increase the probability that these frontline workers do not resolve these tasks and pass them to experts. Under such revenue structures, net revenue-maximizing assignments need not minimize firm production costs. In other words, certain payment structures (an example of which we discuss in Appendix C) disincentivize the firm from minimizing its costs even when it is constrained. This result shows that task assignment depends not only on technological advantage (driven by p_τ) but also on the firm's revenue structure, which can distort allocation even in the absence of downstream externalities.

Non-contingent revenues. If the firm receives the same revenue in each route ($r_F = r_{F,E} = r_E$), it minimizes its production costs regardless of whether it is constrained.

The social planner's problem. Denote downstream (out-of-firm) costs for a type-τ as the random variable $C_{F,\tau}$ (where $\mathbb{E}[C_{F,\tau}] < \infty$) if the case is initially assigned to a frontline worker (the knowledge hierarchy, KH) and $C_{E,\tau}$ (where $\mathbb{E}[C_{E,\tau}] < \infty$) if directly assigned to an expert.[13] Consider a social planner who minimizes expected systemwide production costs, which are the sum of firm production costs and expected downstream costs. For a type-τ case, expected systemwide production costs are $SC_F(\tau) \equiv \kappa_F + (1 - p_\tau)\kappa_E + \mathbb{E}[C_{F,\tau}]$ under the KH, and $SC_E(\tau) \equiv \kappa_E + \mathbb{E}[C_{E,\tau}]$ under direct-to-expert assignment. The social planner's assignment rule follows the difference in expected systemwide production costs, which we call the *social planner's criterion*:

$$\Delta^{SC}(\tau) \;\equiv\; SC_F(\tau) - SC_E(\tau). \tag{3}$$

Result 3: The social planner's assignment. *When staff availability is unconstrained, the social planner assigns type-τ tasks to the KH if and only if $\Delta^{SC}(\tau)$ is negative – i.e., the planner assigns according to the KH's absolute advantage. When staff availability is constrained, the planner instead ranks task types by Equation 3, and prioritizes assignment to the KH for tasks with the lowest $\Delta^{SC}(\tau)$, subject to the constraints.*

The social planner's assignment may thus differ from the firm's, since, unlike the planner, the firm ignores downstream costs, which it neither observes nor is incentivized by the government to internalize. An approach to solving this problem is to issue guidelines to induce the firm to assign tasks as the planner would.[14] However, when these are applied, eventual changes in staffing and hiring can pose new challenges. To better align the firm with the planner, another option is to use Pigouvian fees and subsidies, which reflect the downstream externalities that the firm's staffing generates. This approach requires a pricing scheme that varies by assignment route interacted with task type, which may involve higher administrative costs. Whichever approach is taken, it is important to ensure that the firm is not induced to exit the market and continues to handle all tasks.

The social planner's assignment does not always follow comparative advantage in systemwide production costs. Because downstream costs vary by task type, the direct-to-expert cost $\mathbb{E}[C_{E,\tau}] + \kappa_E$ is not constant across tasks, so ratio and level rankings can diverge.[15] Since case quantities are fixed and the planner cannot reallocate production across task types as in standard trade models, the relevant criterion is which assignment yields the largest absolute saving in systemwide costs, not the largest proportional one. However, as we discuss in Section 7, in our setting, the two rankings (by the social planner's criterion and comparative

[13] In reality, the downstream cost distributions may vary even within type and route, e.g., by time period, but here we simplify to economize on notation, and discuss the practical implications in Section 4.3.

[14] Guidelines could be stricter (e.g., requiring that certain tasks be assigned to one route) or more lenient (e.g., guiding the firm on which tasks to assign to which route if constrained).

[15] This differs from the situation where externalities are absent, and the firm's ranking by cost advantage is equivalent to its ranking by comparative advantage.

advantage in systemwide production costs) are highly correlated.

4.2 The problem of the primary care firm

In Appendix C we characterize the problem of the primary care firm and its associated social planner's problem, leaving the question of implementation to Section 7. Here we provide a brief overview. In the firm we study, the *experts* are doctors, the *frontline workers* are nurses, and the tasks are patient cases. In each period t, the queue consists of two groups: *doctor-group* cases, for which only direct doctor assignment is suitable, and *overlap-group* cases, for which either clinician type may be assigned initially.[16] We focus on a setting in which services are paid for by taxpayers. The firm is compensated only for the clinician type that resolves the case, and we parameterize revenues using a common wedge $\theta \in (0, 1)$ between the administratively-set consultation prices and the firm's production costs for reimbursed nurse and doctor consultations. We show that the key results from Section 4.1 apply to the firm we study with only minor modifications. Specifically, the unconstrained firm assigns overlap-group cases to doctors and the constrained firm assigns the easiest cases to nurses.

4.3 Implications for the empirical analysis

Parameter values. We set $r_F = 275$ and $r_E = r_{F,E} = 500$, since the firm is only paid for the doctor when a nurse passes the case to a doctor who resolves it (see Appendix Section A). Given θ, this implies $\kappa_F = (1-\theta)r_F$ and $\kappa_E = (1-\theta)r_E$. We estimate p_τ, $\mathbb{E}[C_{E,\tau}]$, and $\mathbb{E}[C_{F,\tau}]$ directly in the data. To do so, we need to divide the cases into types. We begin our empirical analysis with a flexible definition of types and later focus on symptoms.

Since we do not know the firm's production costs, we use θ to parameterize the firm's wedge. In general, such a wedge may arise from power in labor markets (markdowns) or product markets (markups). However, since in our setting the "buyer" is the national health insurance, which sets the administrative product prices, we see little role for markups. Instead, we focus on labor market power, and assume $\theta = 0.25$, which is close to the middle of the range implied by the literature on monopsony markdowns (e.g., Azar and Marinescu 2024). This simplifying assumption of a single wedge is supported by (i) the government's revenue setting for both routes and (ii) the use of similar labor inputs (e.g., IT and administrative support) in both routes. In Section 6.3 we assess the robustness of our findings to varying θ, and allow for different wedges by clinician type.

Empirical predictions. Our model yields implications for the empirical analysis, in which

[16] Technically, there are also "nurse-group cases", i.e., those sent only to nurses, but these are outside of both the model and the empirics as they are treated as lower-urgency and processed outside the main queue. Moreover, cases passed from nurses to doctors are likewise scheduled into dedicated later time-slots and hence are not part of the model. The FCFS constraint applies separately within each case group.

we focus on the effect of the assignment of a case to a nurse (the KH) relative to its assignment to the direct-to-doctor route.

Prediction 1: Identification. In the model, types fully determine potential outcomes for firm production costs and downstream costs, so controlling for case-type fixed effects suffices for OLS identification. In practice, we face two concerns: first, we do not know exactly how types are defined in the firm's assignment algorithm, and second, even for cases of a given type, potential outcomes in downstream costs for a type may vary by seasons, days of the week, times of the day, and other transitory factors (e.g., pandemics). To address the first concern, we control flexibly for interactions of the patients' symptoms, demographics, and medical histories within the provider – the key factors that the firm's assignment algorithm observes. To address the second, we control flexibly for time windows within a calendar day. For instance, we compare cases between 12 pm and 3 pm on the 18th of June, 2019.

Prediction 1a: First-come-first-served (FCFS). The firm's FCFS constraint implies that overlap cases may be initially assigned to nurses to process the queue. The model assumes that within each case group (doctor group, nurse group, and overlap group), assignment follows FCFS. Across case groups, doctor-group cases may be delayed relative to overlap group cases, and nurse-group cases are processed later, outside the main queue.

Prediction 1b: Short-term doctor congestion. Appendix C shows that in our setting $\Delta(\tau) < 0$ for all τ, so by Result 1, the unconstrained firm assigns all overlap-group cases directly to doctors. For overlap-group cases, temporary doctor shortages (relative to case needs) shift the firm's problem from the unconstrained situation (when all cases are assigned to doctors) to the constrained situation, where some are assigned to nurses. Thus, doctor congestion increases the probability of initial nurse assignment. In a robustness check, we use waiting times for recent preceding doctor consultations to construct an instrumental variable (IV) for assignment to nurses, which we use to test the robustness of our main identification strategy.

Prediction 1c: Queue composition. If, in a congested situation, the cases immediately preceding (overlap-group) case i in the queue are doctor-group cases, then case i is more likely to be assigned to a nurse than when case i is preceded by overlap-group cases. We test this prediction but do not use queue composition as an instrument, since it only affects assignment in congested situations and is therefore weaker in predicting assignment than the congestion instrument.

Prediction 2: Mean effects of the knowledge hierarchy. Using our identification strategy (Prediction 1), we estimate the effects of the KH on care quality and costs for overlap cases. Since $\Delta(\tau) < 0$ for all τ, the model predicts that KH reduces firm net revenue and taxpayer payments to the firm. Expected effects on healthcare outcomes and downstream costs (and, thus, systemwide production costs) are theoretically ambiguous and are estimated empirically. When we consider expected total taxpayer payments (systemwide production costs plus firm net revenues, $\mathbb{E}[\Delta^{SC}(\tau) + \Delta(\tau)]$), the effects of KH are also ambiguous.

Prediction 3: Effects of the knowledge hierarchy on the cost distribution. KH increases both the probability of lower firm production costs and the probability of higher firm production costs. This follows because under direct-to-doctor assignment, the distribution of firm production costs is degenerate at κ_E. In contrast, the KH induces a two-point distribution of firm production costs for overlap cases: κ_F with probability p_τ (resolved by nurses), and $\kappa_F + \kappa_E$ with probability $1 - p_\tau$ (passed from nurses to doctors). Whether analogous patterns extend to systemwide production costs depends on downstream utilization and is assessed empirically.

Prediction 4: Firm case assignment. Since in our setting $\frac{\partial \Delta(\tau)}{\partial p_\tau} > 0$, the simpler overlap-group cases (those with higher p_τ) for which KH has a *relative net revenue advantage*, $\frac{\Pi_F(\tau)}{\Pi_E}$, are more likely to be assigned to nurses (the KH). These cases also entail lower relative *firm production costs*, defined as $\frac{\kappa_F + (1-p_\tau)\kappa_E}{\kappa_E}$, and lower *relative taxpayer payments to the firm*, defined as $\frac{p_\tau r_F + (1-p_\tau) r_{F,E}}{r_E}$. These patterns are all driven by periods in which the firm's assignment is constrained by temporary doctor shortages; when there are enough doctors, they are assigned to all overlap-group cases.

Prediction 5: Counterfactuals. Because the firm does not observe (nor internalize) downstream costs, its case assignment does not necessarily minimize the systemwide production costs. In other words, the firm may exert a downstream cost externality by allocating cases to clinicians who reduce firm production costs but increase downstream care costs. Our analysis of the social planner's objectives motivates a counterfactual case assignment that minimizes expected systemwide production costs. We focus on a setting where clinician availability is held constant within short time windows, and within each such window, the planner assigns nurses to the cases with the lowest $\Delta^{SC}(\tau)$.

5 Empirical framework

Identification. Prediction 1 motivates our identifying assumption that, conditional on the (short) time window and the case type, the initial assignment of each overlap case to a nurse (versus a doctor) is random. To translate this into an estimation strategy, we need to make assumptions about both the time windows and the types, which we do by estimating OLS specifications of the form:

$$Y_i = \delta Nurse_i + X'_{1i}\beta_1 + X'_{2i}\beta_2 + \epsilon_i, \tag{4}$$

where Y_i denotes the outcome for case i, and $Nurse_i$ denotes initial assignment to a nurse (as opposed to a doctor). The vector X_{1i} includes predetermined case characteristics likely used by the assignment algorithm (date-by–3-hour windows of patients' login time; age, gender, symptom and region; and symptom indicators interacted with indicators for patients' age and gender, whether they have any past prescription from the firm, and whether their last case at the firm, if any, was with a doctor). This specification reflects a pragmatic tradeoff: ensuring that the time windows are tight and the variables used to determine "types" are flexible, while retaining sufficient degrees of freedom. X_{2i} includes other predetermined characteristics (health risks and socioeconomic background), which, to the best of our knowledge, the algorithm cannot observe (see the description of 'Additional controls' in Section 3.2). Under conditional exogeneity, we interpret the OLS coefficient δ as a treatment-variance-weighted average over conditional treatment effects for initial nurse assignment. Since the treatment is not naturally clustered, we report robust standard errors for $\hat{\delta}$ (Abadie et al. 2023).

To support this identification strategy, we present here three pieces of evidence. First, Figure 1 examines balance in initial case assignment in X_{2i}. Panel (a) shows that, conditional on baseline case characteristics X_{1i}, assignment is balanced for most variables; where differences are statistically detectable, their magnitudes are economically negligible. Panel (b) reports the corresponding unconditional normalized differences. Even without conditioning on baseline characteristics, all imbalances fall well below 0.1, lower than thresholds (0.1 and 0.25) that commonly flag potentially problematic (Imbens and Rubin 2015; Baker et al. 2026).[17] Moreover, we also control for the variables in X_{2i}.

Second, we show that the remaining imbalance in X_2 has negligible economic impact on our main outcome of interest, systemwide production costs. We regress systemwide production costs on X_1 and X_2 and find that X_2 is strongly predictive, with a joint F-statistic of 102.9. We then use X_2 to predict systemwide production cost. Finally, we regress the predicted systemwide production costs on nurse assignment, again controlling for X_1. We find a coefficient of -1.91 SEK (s.e.: 1.04), which is economically small compared to mean systemwide production cost in direct-to-doctor cases (1,262 SEK).[18]

Third, Table 2 reports the explanatory power of case characteristics and exogenous shifters for initial nurse assignment. Baseline characteristics (X_1) alone account for a substantial share of the variation (adjusted $R^2 = 0.195$, Column 1). This contrasts with using only the

[17] Moreover, we conduct two one-sided t-tests (TOSTs) for each additional case characteristic, conditional on the baseline controls. We test for equivalence within ±5% of the standard deviation of the treatment variable (initial nurse assignment), rejecting differences outside this range at the 5% level. For comparison, a range of 20% is generally considered very conservative (Hartman and Hidalgo 2018).

[18] In Section 6.3 we also explore unobserved selection with Oster bounds and IV estimation. We can not implement Altonji and Mansfield (2018)'s group-mean control-function approach (with only one treatment and one control group, the treatment effect and sorting adjustments are not separately identified).

additional controls (X_2), which have almost no explanatory power or additional explanatory power (Columns 2 and 3). Adding doctor congestion increases the explanatory power (adjusted $R^2 = 0.200$, Column 4), but doctor-group status in prior cases does not (Column 5).

We defer to Section 6.3 for further evidence on the robustness of our empirical strategy, including shorter time windows, pre-trends, IV, and Oster bounds.

First-come-first-served (FCFS). Appendix Table A5 provides evidence consistent with a FCFS constraint (Prediction 1a). It reports results from regressing the "exit rank" (the order of the initial consultation start time) on the "login rank" (the order of requesting an appointment), controlling for date-by–3-hour login-time windows. Within the overlap and doctor-group cases (Columns 1 and 2), the coefficient of interest is very close to one, indicating that cases are processed almost exactly in the order they arrive. Doctor-group cases are delayed relative to overlap cases (Columns 3 and 4), consistent with the model's assumption regarding temporary doctor shortages, while nurse-group cases are substantially delayed (Column 4), consistent with lower urgency and processing outside the main queue.

Short-term doctor congestion. We measure doctor congestion using *log median doctor wait time* of recent patients, defined above in Section 3.2 and further discussed in Section D.1 in the Appendix. This empirical measure is motivated by the model, but does not fully capture the algorithm's case assignment, as it uses past waiting times without explicitly accounting for factors such as planned or unplanned staffing that may also shape it. Table 2 and Appendix Figure A2c test the model's Prediction 1b: that doctor congestion increases the propensity for initial nurse assignment, and show consistent evidence that there is a significant increase in nurse assignment when there is more doctor congestion. Appendix Figure A3 illustrates this pattern within symptoms: moving from the lowest to the highest congestion quartile increases nurse assignment for every symptom. Appendix Table A6 shows that higher congestion (mechanically) increases patient wait times, but is not associated with substantial changes in case processing times, ratings, attrition (measured by no-shows and technical errors), or very short consultations.[19]

A remaining concern is that congestion may affect patient outcomes directly, not only through the initial assignment. Appendix Table A7 assesses this using reduced-form regressions of case outcomes on doctor congestion for doctor-group cases. Among doctor-group cases fewer than 1% are assigned to nurses, so congestion has almost no effect through the assignment route.[20] Across within-firm outcomes, downstream healthcare use, longer-term outcomes, and

[19] A "no-show" is a label indicating that the patient did not attend the consultation, and "technical error" is a label indicating that the consultation could not be completed due to software or hardware issues.

[20] We do not perform the analogous exercise for nurse-group cases (symptoms with fewer than 1% direct-to-doctor assignments), since nearly all are chlamydia cases, which follow a different care pathway (a tracing program and home testing).

costs, nearly all estimates are small and statistically insignificant. In particular, congestion does not change the margins most plausibly affected directly, such as revisits, ratings, and diagnoses, and prescribing is tightly estimated at close to zero. These results mitigate concerns about any direct effect that congestion may have (apart from its effect on case assignment).

Queue composition. In the model, overlap cases are likelier to be assigned to nurses when they arrive with more doctor-group cases, for which doctors are needed (Prediction 1c). Appendix Figure A4 shows that overlap-group cases are indeed likelier to be assigned to nurses when preceded by doctor-group cases, with the effect fading along the queue. However, since these effects are small and add little predictive power to the assignment (Table 2, Column 5), we focus on doctor congestion as our IV in the robustness checks.

6 Results

This section uses Equation 4 to estimate the mean effects of initial assignment to a nurse (KH), relative to assignment directly to a doctor, on quality of care and costs (Prediction 2). We begin with estimates of the effects on patients' outcomes, then discuss effects on costs, and finally, the robustness of our findings.

6.1 Effects of the knowledge hierarchy on patient outcomes

Table 3 shows OLS estimates for the mean effects of KH on within-firm outcomes (Panel A), downstream healthcare use (Panel B), and longer-term consequences (Panel C). Detailed descriptions of our outcome variables are provided in Appendix B.

As Table 3 shows, within the firm, KH increases the probability of a clinician-booked doctor appointment within seven days by 31 percentage points (p.p.), from 1.4% in the direct-to-doctor route (the baseline). This implies that nurses resolve almost 70% of the cases themselves. The effect of KH on achieving a top satisfaction score in the route's final (within-firm) consultation is small and negative when we include all patients, but small and positive (coef.: 0.0230; s.e.: 0.0023) when we restrict to the 233,593 consultations in which the patient gave a satisfaction score (since KH slightly reduces the probability of any satisfaction score). Our measure of unresolved cases that patients come back to the firm with, unbooked (re)visits with the same symptom, increases only slightly, by 0.8 p.p. from a baseline of 5.8%. The KH barely increases the rate of no-shows at the initial consultation (coef.: 0.00043; s.e.: 0.00014) and slightly reduces the rate of technical errors (coef.: -0.0024; s.e.: 0.00025). In sum, these internal measures suggest there are no large differences in patient satisfaction with the KH compared to the direct-to-doctor route. Yet, the KH reduces the probability of receiving an informative diagnosis (a non-missing ICD code that is not a symptom or health status update) by 9 p.p. (from a baseline of 79%) in the final consultation (as a comparison, 60.21% of diagnoses are informative in the universe of Region Scania's in-person primary care). This suggests an effect on patients from a different organization of expertise, even if they do not notice it at first. Hence, we turn downstream.

We measure downstream healthcare use within 7 days of the initial consultation day and test alternative aggregations in robustness analyses. KH decreases the probability that a patient receives a new prescription by 10 p.p. (from a baseline of 44%). New prescriptions include both those issued by the firm and by external care providers, for drugs not purchased in the six months prior (to exclude recurrent treatments). Nevertheless, a relatively large share of cases sent to KH still end up with a new prescription within seven days, and much of that seems to be due to these patients being passed on to doctors.[21]

We find evidence on some shifting to experts in other organizations. Initial nurse assignment increases external primary care consultations by 2.5 percentage points relative to a baseline of 16%. External primary care is observed directly for patients registered in Scania and Stockholm and inferred from prescription data for patients in other regions; results are similar when restricting to Scania and Stockholm only (coef.: 0.035, s.e.: 0.0022). The rate of visiting urgent care centers – out-of-hours primary care facilities to alleviate pressure on emergency care – increases by 0.2 percentage points from a baseline of .9% and emergency department visits rise by 0.6 p.p. from a baseline of 3.7%. Effects on specialist visits are fairly small (0.3 p.p. from a baseline of 3.5%).

While patients do use some additional external care, we find no increase in high-stakes adverse health events: hospitalizations are unaffected, and so is an alternative measure of care quality, avoidable hospitalizations, (not in table; coef.: 0.00019; s.e.: 0.00015). Moreover, the KH does not worsen longer-term employment outcomes. If anything, KH reduces the probability of earnings declines from primary employment in the calendar month following the initial consultation, relative to average earnings in the preceding three months (this can only be measured for patients with stable prior employment). In particular, KH reduces the incidence of earnings declines greater than 20% by 2.1 percentage points from a baseline of 22%, and reduces the probability of zero earnings by 0.3 percentage points from 5%, which may be due to the fact that doctors, rather than nurses, need to sign off longer sick leaves.[22]

[21]Descriptively, the new prescription rate for cases passed to doctors in the KH, 42%, is comparable to that among cases assigned directly to doctors. 18% of cases that nurses resolve (within the firm) result in new prescriptions, likely reflecting prescriptions in downstream care. These prescription rates are descriptive and are not estimates from our causal identification strategy.

[22]Observed earnings reductions may reflect sick leave, as employer-paid earnings fall to 80% for the first 14 days and to zero thereafter, at which point the Swedish Social Insurance Agency pays sickness benefits. Employees can self-certify sickness for the first seven days, but longer sick leaves need to be signed off by doctors; thus, the slightly lower rate of earnings losses after KH may partly reflect lower access to doctors to

As in most economics papers on health care, we acknowledge that there may be effects on patients that do not show up in health care utilization. We attempt to capture such effects through the employment measures above, but we also add an even more long-term, high-stakes measure – 3-year mortality. Importantly, while our point estimate is a precise zero, the rarity of the outcome means we cannot rule out modest effects on three-year mortality. While the KH does not seem to have major adverse effects on patients, we have shown evidence of some increase in care-seeking outside the organization (spillovers) that we account for below.

6.2 Effects of the knowledge hierarchy on costs

Next, we examine the cost effects of the KH on patients' healthcare journeys. Panel D in Table 3 reports OLS estimates of the mean effects of initial assignment to a nurse (the KH) on within-firm payments and on costs including downstream care. Our cost measures are constructed using taxpayer payments for the categories listed in Appendix Table A3 and the model assumptions described in Section 4.

The KH substantially reduces taxpayer payments to the firm (which equal firm gross revenue). Over a seven-day window (when including up to one additional within-firm visit), payments in the KH route are 149 SEK lower, a reduction of ≈27% relative to the baseline mean. The results are almost identical when we restrict to payments on the day of the initial consultation (as in the model); henceforth, we focus on within-day payments to estimate firm production cost effects. As predicted by the model, these taxpayer savings are offset by losses to the firm, since initial nurse assignment reduces firm net revenue by 108 SEK, relative to a baseline of 125 SEK under doctor assignment. From the taxpayer's perspective, the KH therefore delivers sizable cost savings, but these come at the expense of lower firm revenues.

Within-firm, the KH reduces production costs by 40 SEK, more than 10% relative to a baseline mean of 375 SEK. To calculate total taxpayer payments and systemwide production costs, we consider payments and costs on the day of the initial consultation and count at most one instance of each out-of-firm downstream event within seven days, assuming that downstream costs are all borne by taxpayers. Total taxpayer payments reflect public healthcare spending, while systemwide production costs net out transfers between taxpayers and the firm. The KH reduces mean total taxpayer payments by 112 SEK, or about 8% relative to the baseline mean of 1,387 SEK. In contrast, mean systemwide production costs – which abstract from transfers between taxpayers and the firm – are 1,262 SEK under direct doctor assignment, and the mean effect of initial nurse assignment on them is very close to zero. Taken together, these results show that on average, KH redistributes from the firm to the taxpayer, but it does not affect systemwide production costs of care.

certify sick leaves (Försäkringskassan 2025).

The distribution of costs. Here we examine the model's Prediction 3, that KH increases the probabilities of both low and high firm costs. Figure 2 plots cumulative distribution functions (CDFs) constructed from regressions using Equation 4, where each outcome is an indicator for costs being below a given threshold, evaluated over a grid of thresholds. The CDF for cases assigned to doctors reflects the baseline share of cases below each threshold, while the CDF for cases assigned to nurses captures the estimated effect of the KH at each threshold. Ensuring that the estimated CDFs are monotonic requires fully saturating the control set X_1 and X_2, which is computationally infeasible. To address this, we follow Chernozhukov, Fernandez-Val, and Galichon (2009) in rearranging the estimated CDFs.

Subfigure 2a presents the CDFs of firm production costs on the initial consultation day for cases assigned to doctors and to nurses. Consistent with Prediction 3, the KH route is likelier to result in lower firm production costs (corresponding to cases that nurses resolve) but also likelier to result in higher firm production costs (when nurses do not resolve cases).

Subfigure 2b shows that this pattern is inherited by systemwide production costs, which additionally incorporate downstream healthcare use. The KH almost halves costs for almost half of patients, but on average, this saving is offset by higher costs for other patients.[23]

6.3 Robustness

In this section, we assess potential threats to the validity of our main result on systemwide production costs, as well as on patient health outcomes. We assess identification, alternative sample definitions, selective attrition before the initial consultation, robustness to cost construction and pricing assumptions, and external validity across settings and time periods.

Alternative controls and unobserved selection. Our baseline OLS specification conditions on fine-grained time windows and a rich set of predetermined case characteristics likely used by the assignment algorithm. Because the algorithm's exact functional form is unknown, residual endogeneity could remain if assignment depends on nonlinear combinations of observables imperfectly captured by our controls, or on omitted variables.

We probe this concern by varying controls and the assignment environment. Appendix Table A8 shows that estimated effects on systemwide production costs are mostly unchanged when we (i) control only for baseline case characteristics (X_1), (ii) use even finer-grained date-by-1-hour windows, (iii) control for symptoms $\times$ 3-hour windows $\times$ weekday $\times$ quarter, (iv) explicitly control for doctor congestion using congestion-quartile indicators, or (v) restrict

[23]However, the mean effect of KH on systemwide production costs is similar for patients aged 25–65 with high socioeconomic status (measured as those with above-median earnings) and those with low socioeconomic status (measured as those with below-median earnings). We estimate a pooled regression interacting initial nurse assignment with an indicator for above-median earnings among patients aged 25–65 (N = 282,067), using full controls (X_1 and X_2), and cannot reject the null of equal effects (p-value = 0.566).

to repeat patients and include patient fixed effects to net out time-invariant unobserved patient traits, such as patient assertiveness.

Appendix Figure A5 shows outcome dynamics as a complementary diagnostic. Patients assigned initially to nurses have similar pre-consultation healthcare use as those assigned directly to doctors. In the week starting with the consultation, the KH increases the likelihood of any unbooked revisit, reduces prescribing, and raises external primary, specialist, and emergency care use relative to direct-to-doctor assignment, with these effects disappearing in the week after. The absence of pre-trends and occurrence of effects only in the consultation week support our identification strategy and our use of a seven-day care window.

We further assess unobserved selection using two approaches. First, we calculate Oster (2019) bounds: the interval between the baseline (fully controlled) coefficient and Oster's bias-adjusted estimate, assuming equal selection on observed and unobserved controls ($\delta = 1$) and a maximum R^2 of 1.3 times the fully controlled R^2. The implied interval for the mean effect on systemwide production costs, -31.3 to -4.8 SEK, remains small.

Alternative overlap sample definitions. Our main overlap sample comprises symptoms with 5–95% of cases initially assigned to nurses. Appendix Table A8 varies this cutoff. The mean effect on systemwide production costs remains small when we restrict to the most common overlap symptom, *other health inquiries* (Appendix Figure A1), or to symptoms with 10–95% of cases initially assigned to nurses. At a 20–95% cutoff, the effect turns significantly negative: this narrower sample consists mainly of case types for which the KH lowers costs (as we will show in Section 7), whereas our main estimate averages over a broader set of cases.

Differential attrition. Selective attrition before the initial consultation could threaten our identification. We address it in two complementary approaches: Lee bounds, and restricting to subsamples in which incentives or opportunities for selective attrition are even more limited. Appendix Section D.2 sets out institutional details and the available data on attrition.

First, Appendix Table A8 computes Lee (2009) bounds assuming monotone selection. Within each symptom group, we trim the cost distribution in the lower-attrition group to match the higher-attrition group. As some types of dropouts are unobserved in our sample, we use symptom-specific dropout rates from a later period and calculate Lee bounds conditional on symptoms. Differential attrition in that data is small, with a 2.49 percentage-point higher average dropout rate for cases assigned to nurses. Attrition directly observable in our sample, no-shows and technical errors, runs in the opposite direction, with a 0.32 percentage-point higher attrition in the direct-to-doctor route. Since healthcare costs exhibit fat right tails, and we do not use log costs as outcome variables for transparency, Lee bounds will be very conservative. Nonetheless, they remain consistent with a small mean effect of the KH on systemwide production costs.

Second, Appendix Table A8 reports two additional checks. We restrict to cases with patients who have had nurse consultations in the past – patients who have revealed that they do not select out of initial nurse assignment. We also exploit a change in the assignment process in mid-2019: before, clinician type was assigned extremely shortly before the consultation, leaving even less room for differential attrition. We estimate differences across that policy change by interacting the nurse indicator with a second-period indicator (April-June versus July-September 2019).[24] Neither suggests that differential attrition drives our findings.

Cost construction and pricing. We examine the robustness of our mean and distributional results under alternative assumptions on how downstream events map into costs. First, Appendix Table A8 reports mean systemwide production cost estimates when restricting the sample to cases in Stockholm and Scania, for which downstream care including all external primary care is fully observed (i.e., using no imputation based on prescriptions); using aggregate costs, defined as the summed costs for all downstream events within 7 or 30 days instead of indicators for the first event; and assuming that the firm's nurse costs are 20% lower when the nurse passes a case to a doctor (keeping the nurse cost as in the baseline when the nurse resolves). These alternative assumptions do not affect our main conclusions.

Second, Appendix Figure A6 shows that systemwide production costs are robust to allowing for differential markdowns by clinician types. Our main specifications assume a common wedge of $\theta = 0.25$ between the price and the firm's production costs of doctor or nurse consultations. Given the administrative price-setting procedure, the wedge is unlikely to vary much by clinician type.[25] Existing monopsony markdown estimates suggest $\theta \in [0.13, 0.33]$ (Azar and Marinescu 2024) (see Section 4.3). Varying wedges slightly beyond these estimates, from 0.1 to 0.35, or letting the wedge differ by ± 0.1 between nurses and doctors within the range of $0.1 - 0.35$, changes the proportional cost effects by only $\approx 5\%$.

External validity. We assess sensitivity to two features of our setting: (i) the online-care context and (ii) the timing of our sample, which partly overlaps with the pandemic. First, Appendix Table A12 shows that most of the ten most common diagnoses in our sample were also common in the universe of in-person primary care in Scania in 2019 (the only region for which we have such data), and that in-person care has quite similar shares of nurse consultations as online care does for many of the diagnoses. A reason for differences in the condition mix between online and in-person is that cases in online care are likelier to involve younger and female patients than those in physical primary care (Appendix Table A4). To address these demographic differences, we reweight our analysis sample to match the

[24] While KH cost estimates are higher in both these periods than in later periods, this difference is unlikely to be driven by the pandemic, which we discuss separately below.

[25] The prices are administratively set based on calculations of each clinician type's labor costs and approximations of firms' other operating costs (Sveriges Kommuner och Regioner (SKR) 2019).

age-by-gender distribution of in-person primary care visits in Scania in 2019. We focus on age and gender, as these are the only predetermined characteristics available in both the Scania population and our analysis sample; patient-determined symptoms are not observed in in-person care, and although diagnoses are observed, Table 3 shows that they are outcomes affected by the KH. We construct sampling weights from a probit model on the pooled sample of the analysis sample and all nurse and doctor consultations in physical primary care, using age interacted with gender as predictors. The estimate under demographic reweighting in Appendix Table A8 supports our main conclusion.

Second, our sample spans multiple phases of the pandemic – the onset and first wave (spring 2020), an interlude period with low incidence (summer 2020), and the second wave (fall 2020 onward) – as well as around a year before the pandemic, from spring 2019. Sweden did not impose lockdowns and physical care remained accessible during the pandemic, but public guidance to stay home may have affected care-seeking. However, Appendix Table A8 shows limited COVID-19 related heterogeneity in the KH's effect on systemwide production cost, suggesting that sharp pandemic-related changes do not affect our main conclusions.

Effects on the cost distribution. Although the mean effect of the KH on systemwide production costs is small and insignificant, the KH meaningfully reshapes the cost distribution, and this pattern is stable across our sensitivity exercises: Oster and Lee bounds, IV strategy, alternative cost definitions and windows, and demographic reweighting (Appendix Figure A7). The KH increases the probability of low-cost realizations driven by cases resolved at the initial consultation without downstream care, and it increases some medium-cost realizations, while leaving costs in the upper tail unchanged.

Robustness of additional patient outcomes. Appendix Table A11 shows that the KH's estimated effects on care quality and downstream use change little when we probe the identifying assumption (baseline controls X_1 only; Oster bounds), assess differential attrition (cases with past nurse consultations; before versus after the assignment process change in mid-2019), or examine COVID-19 related heterogeneity.

7 Task assignment

Our evidence thus far suggests that while KH barely changes mean systemwide production costs, it significantly widens the distribution. This suggests that the selective assignment of tasks to KH may yield savings. This section, therefore, explores task assignment to KH – first by the firm, and second, by a social planner.

The analysis in this section uses symptoms as case types. We do this because, as Appendix Figure A8a shows, symptoms are the most important determinant of assignment across routes, and we want to ensure that we have a sufficiently large number of observations per route within case types. We index the symptoms by k.

7.1 Task assignment by the firm

Prediction 4 is that within narrow time windows, overlap-group cases, for which KH has a *relative net revenue advantage* (the easier cases, with higher p_τ) are (at least weakly) more likely to be assigned to nurses. This assignment minimizes *taxpayer payments to the firm*, conditional on staff use. This, however, does not take into account any downstream externalities, which we discuss in Section 7.2.

Estimation. To study what drives task assignment, we use regressions to estimate the mean KH effect on outcomes (e.g., costs, revenues, or payment), by symptom.[26] Using these estimates, we calculate the ratios of mean outcomes (for the KH route over the direct-to-doctor route): (i) Firm net revenues, and (ii) Taxpayer payments to the firm (i.e., firm gross revenues). While the relationship between the firm's task assignment and downstream costs is theoretically ambiguous, we also use as outcome ratios that include downstream costs: (iii) Total taxpayer payments, and (iv) Systemwide production costs. We begin by modifying Equation 4 to allow the estimated effect of initial nurse assignment to vary by symptom:

$$Y_i = \sum_{k \in \text{Symptoms}} \delta_k \, \text{Nurse}_i \mathbf{1}\{\text{Symptom}_i = k\} + \tilde{X}'_{1i}\beta_1 + X'_{2i}\beta_2 + \varepsilon_i, \tag{5}$$

where Symptom_i is the symptom of case i, Nurse_i is an indicator for initial nurse assignment, and δ_k captures the KH's heterogeneous effects by symptom. The vector $\tilde{X}_{1i}$ contains the baseline controls in X_{1i}, including symptom-group indicators, but omits interactions between symptom indicators and other predetermined case characteristics. Using the estimated coefficients, we predict symptom-specific outcomes under initial nurse assignment, $\hat{Y}_{i,k}(n) \equiv (\hat{\delta}_k + \tilde{X}'_{1i}\hat{\beta}_1 + X'_{2i}\hat{\beta}_2)\mathbf{1}\{\text{Symptom}_i = k\}$, and under direct doctor assignment, $\hat{Y}_{i,k}(d) \equiv (\tilde{X}'_{1i}\hat{\beta}_1 + X'_{2i}\hat{\beta}_2)\mathbf{1}\{\text{Symptom}_i = k\}$. We use these to construct ratios of sums, $\sum_i \hat{Y}_{i,k}(n)/\sum_i \hat{Y}_{i,k}(d)$.

Finally, we relate these ratios to the firm's propensity to assign symptoms to the KH using linear regressions weighted by symptom frequency. Standard errors are obtained by bootstrapping the full procedure with 100 replications. In each bootstrap draw, we re-estimate the symptom-specific mean KH effects δ_k, the predicted outcomes under each route, the ratios of the sum of predicted outcomes, and the weighted regressions; the bootstrap standard errors are then constructed as the standard deviations of the weighted regression coefficients across all draws.

Results. Before turning to the firm's assignment decisions, we first verify a key mechanism

[26] For firm net revenues, this corresponds to estimating the net revenue advantage of the KH route for symptom k, $\Delta(k)$.

implied by the model: that higher ratios (KH over direct-to-doctor) of firm net revenues reflect nurses' better ability to resolve cases without passing them to doctors (higher p_τ). To do so, we regress the KH's relative firm net revenue advantage on the observed symptom-level rate of nurses resolving cases themselves ($\hat{p}_k$), and obtain a regression coefficient of 1.212 (bootstrap s.e: 0.007, R^2: 0.993, observations: 495,555). In the model, the corresponding derivative is exactly 1.2; the slight difference arises because the empirical measure of firm net revenue includes the costs of additional same-day drop-in consultations, which the model measure excludes.

Figure 3 relates the firm's case assignment to revenues, taxpayer payments, and costs. Consistent with Prediction 4, when comparing overlap cases of different symptoms, the firm is likelier to assign a case initially to a nurse when the KH has a relative firm net revenue advantage (Subfigure 3a), relatively lower firm costs (Subfigure 3b), and relatively *lower* taxpayer payments to the firm (Subfigure 3c). Subfigure 3d shows that the firm's assignment also reduces relative systemwide production costs, although this relationship is less tight than the other three, since the firm neither observes nor is incentivized to target downstream costs.

Next, Figure 4 examines how these relationships vary with doctor congestion. In line with Prediction 4, regressions of initial nurse assignment propensity on relative firm net revenue advantage reveal significantly steeper slopes at higher quartiles of doctor congestion (Subfigure 4a).[27] Following the model, at higher congestion quartiles, the firm is likelier to assign to nurses cases for which the KH has relatively lower firm costs (Subfigure 4b) and taxpayer payments to the firm (Subfigure 4c). In contrast, for systemwide production costs, the relationship varies less by congestion quartile (Subfigure 4d) – there was no such prediction by the model for this variable, since the firm neither observes nor is incentivized to target downstream costs.

Taken together, the evidence in this section shows that, under capacity constraints, firm and taxpayer incentives are aligned, if one abstracts from downstream costs. However, absent information on (and incentives to internalize) downstream costs, the firm's task assignment may not minimize systemwide production costs.

[27] As noted above, the ratio of predicted firm net revenues is a near-linear function of the nurse resolution rate $\hat{p}_k$ (estimated slope: 1.212; R^2: 0.993), so plotting initial nurse assignment against $\hat{p}_k$ produces a figure almost identical to Subfigure 4a: assignment rises with nurse resolution probability, and does so more steeply with doctor congestion. Also consistent with the model, around 90% of cases in the lowest congestion quartile are sent to doctors, since absolute firm net revenue is higher under direct doctor assignment for all cases. The fact that about 10% of cases are initially assigned to nurses in the lowest congestion quartile may be because our congestion measure is limited – for example, the algorithm may use information that we do not observe (e.g., staffing changes).

7.2 Counterfactual assignment of tasks

This section compares the firm's observed task assignment to two sets of counterfactuals: (i) non-optimized assignments (counterfactual "random" assignments to the KH or the direct-to-doctor route, given clinician availability, and counterfactual assignments to only one route); and (ii) assignments that maximize the social planner's objective with or without staffing constraints, following the social planner's criterion (Prediction 5). Note that we use the estimated *causal* effects of the KH to construct the counterfactuals, so that potential case selection into the KH versus the direct-to-doctor route is accounted for.

Characterizing the social planner's objective. For each task type, the social planner assigns according to Expression 3 (the social planner's criterion), which we estimate using Equation 5, with systemwide production costs as the dependent variable. KH has an *absolute advantage* in producing care for symptom k if $\hat{\delta}_k < 0$. When unconstrained, the social planner assigns each case by its absolute advantage. When constrained, the social planner assigns the cases with the lowest $\hat{\delta}_k$ to the KH.

Figure 5 summarizes our main findings. Subfigure 5a ranks symptoms by the social planner's criterion. The leftmost symptom, *abdominal pain*, is the one for which the KH reduces systemwide production costs the most. The KH has an absolute advantage for the symptoms in which its effect is negative, and the direct-to-doctor route has an absolute advantage in the rest. The figure shows that for most symptoms, the effect of KH assignment on systemwide production costs is driven mostly by its effects on downstream costs rather than on firm production costs.

Subfigure 5b ranks symptoms by the ratio of estimated systemwide production costs for the KH route over the direct-to-doctor route, $\sum_i \hat{Y}_{i,k}(n)/\sum_i \hat{Y}_{i,k}(d)$, reflecting the KH's *comparative advantage*, in the spirit of Dornbusch, Fischer, and Samuelson (1977). KH has the strongest comparative advantage in the leftmost symptom, *COVID-19*. In practice, the social planner's criterion (KH effect $\hat{\delta}_k$) and the comparative advantage criterion (the ratio of systemwide production costs) give an almost identical ordering of the 16 symptoms, with a rank correlation of 0.96.

To provide intuition, we discuss key drivers of differential downstream costs under the KH vs. the direct-to-doctor route. For symptoms where direct-to-doctor assignment has the largest comparative advantage (Figure 5b), we find patterns consistent with direct-to-doctor assignment reducing the tail risk of high-cost downstream care through faster or more certain access to antibiotics, which are commonly prescribed after online visits with these symptoms. For example, KH assignment decreases the probability of a new prescription for cases with the symptom *urinary tract infection* (coef.: -0.183; SE: 0.0085), while for cases with the symptom *sore throat*, the KH increases the probability of later emergency department visits within seven days (coef.: 0.019; SE: 0.0026), along with hospitalizations within seven days (coef.: 0.005; SE: 0.0014), with common diagnoses reflecting complications from bacterial infections (e.g., "Peritonsillar abscess" and "Acute tonsillitis").[28]

In contrast, for symptoms where downstream costs are higher under direct-to-doctor assignment (Figure 5a), KH assignment may reduce costs by limiting ineffective prescribing. This is particularly relevant when prescription drugs may provide limited benefit or have harmful side effects, such as when symptoms are diffuse (e.g., *diarrhea or vomiting*) or treatment knowledge is evolving (e.g., *COVID-19*). In such cases, care advice from a nurse, or combined guidance from a nurse and doctor, may be more effective.[29] Such nurse advice or combined advice may be lowering urgent care use (coef.: -0.005; SE: 0.0019) after KH for *abdominal pain* symptoms, and lower emergency department visits (coef.: -0.018; SE: 0.0049) and hospitalizations (coef.: -0.006; SE: 0.0027) after KH for *diarrhea or vomiting* symptoms.[30]

KH is more than occupational licensing. The differences between nurses and doctors reflect not only differences in training and skills, but also occupational licensing, which restricts the tasks that workers in each occupation are allowed to perform. Here, however, we counter the notion that if nurses were given the same "rubber stamps" (e.g., they could prescribe as doctors do), they could fully substitute doctors. First, if nurses only lacked doctors' scope of practice, they could simply pass the relevant tasks to doctors, and the two routes would differ only in within-firm costs. But Figure 5a shows that both routes reduce downstream externalities for some outcomes. Second, KH lowers the probability of an informative diagnosis by 9 percentage points (Table 3). Finally, we construct a symptom-level "prescription need" – the fraction of cases that resulted in prescriptions among those assigned initially to doctors.[31] We then consider all cases initially assigned to nurses and examine the

[28]These estimates were obtained from OLS regressions of any new prescription, seven-day hospitalization and emergency department visit rates on initial clinician assignment, fully interacted with symptom categories, using controls for $\tilde{X}_1$ and X_2.

[29]Among cases assigned to the KH, we compare outcomes for cases resolved by nurses with those passed to doctors, and find that ED and external primary care visits are rarer among cases referred to doctors. However, this comparison is descriptive and should not be interpreted as causal.

[30]The most common diagnosis among hospitalizations following an online consultation for the symptom *diarrhea or vomiting* is "Other gastroenteritis and colitis of infectious and unspecified origin", closely followed by "Acute appendicitis". The former diagnosis suggests cases where hospitalization may be prevented through appropriate advice (e.g., on fluid intake), which could be delivered by a nurse or a doctor, or both, online. The latter diagnosis is for a hospitalization that most likely cannot be prevented by online care. The role of the primary care nurse and/or doctor for these cases is instead to decide which of the patients with the symptom *diarrhea or vomiting* should go to the ED urgently with suspected "Acute appendicitis", as a physical examination and surgery may be needed.

[31]We focus on prescriptions, which are by far the most commonly used action where nurses are far more restricted than doctors. Referrals and sick notes are rarer and measured less precisely in our data.

extent to which "prescription need" predicts whether cases are passed to doctors, controlling for X_1 and X_2. This regression has $R^2 = 0.100$, and "prescription need" alone has a within-R^2 of only 0.020. Taken together, these three pieces of evidence show that a knowledge hierarchy of nurses and doctors reflects not only occupational licensing, but also occupational differences in knowledge and how that knowledge is combined.

Defining counterfactual case assignments. Using our estimates from Figure 5a, we define several counterfactual assignment rules. The counterfactual assignments we consider are all deterministic (given our estimates above). For example, if at time t, there are D_t doctors and N_t nurses to assign to $D_t + N_t$ cases with the same symptom, we assign each the baseline mean cost (under direct-to-doctor assignment) plus $N_t/(D_t + N_t)$ times the estimated KH effect.

We begin by considering several non-optimized assignments as benchmarks to the actual assignment. First, we examine assignments that reflect actual staffing, but reassign nurses and doctors within short time windows. We measure clinician availability in two ways: based on the number of cases fully processed by each clinician type in the 15 minutes or 3 hours preceding each case's login; and based on the number of cases assigned within each fixed 15-minute or 3-hour interval to each clinician type. Second, we consider counterfactuals in which all cases are assigned to a single route (either initially to nurses or directly to doctors).

We then consider counterfactual assignments that minimize systemwide production costs, using symptoms to define case types. If relevant heterogeneity remains within symptoms, the savings achieved under this symptom-level partition are a lower bound on those achievable under a finer partition.[32] In the counterfactual that maximizes the social planner's objective without constraints, we ignore staff availability constraints, and instead assume that clinicians can be assigned without affecting wages or prices. This allows assignment by *absolute advantage*: all cases of a symptom are assigned to the KH if expected systemwide production costs are lower (left of the dashed line in Subfigure 5a), and otherwise to doctors.

In counterfactuals with a constrained social planner, we respect short-run clinician availability within fixed time windows (15 minutes or 3 hours), following the ranking of symptoms in Subfigure 5a. Within each time window, we reassign cases by symptom between clinician types to minimize systemwide production costs, according to that ranking. These counterfactuals assign cases to the available nurses, starting with those symptoms that yield the largest cost savings, and moving down the ranking until all nurses are assigned.

Results. Table 4 summarizes the implications of the different counterfactuals for various

[32] This follows because any assignment rule based on symptom averages is feasible under a finer, symptom-nested partition, so optimizing over the finer partition can only weakly lower systemwide costs. The bound is strict only if some sub-type's individual comparative advantage differs enough from its symptom average to change that sub-type's optimal assignment.

outcomes. Panel A considers the non-optimized assignments that ignore the KH's heterogeneous cost effects across symptoms. Allocations 1)-4) assign cases proportionally to clinician availability. These assignments increase systemwide production costs by between 0.52% and 0.55%, equivalent to between 1.83% and 1.93% of online primary care production costs. Assignment 5), which uses only the KH, increases systemwide production costs by 1.5% (equivalent to 5.3% in firm production costs). Finally, using only doctors increases systemwide production costs by 0.4% (equivalent to 1.6% increase in firm production costs) – in other words, compared to using only doctors, the firm's actual (selective) use of KH generates meaningful savings.

Panel B implements the social planner's objective. Counterfactual 1) shows that in the unconstrained optimum, the use of the KH route increases to 61% relative to its actual use of around 23% in the analysis sample. This reduces systemwide production costs by 3.5%, but also involves redistribution, as it reduces the firm's net revenue by almost 40%. Because this counterfactual assumes no change in factor prices, it can be interpreted as an upper bound on the savings achievable through changes in staffing — for instance, by expanding nurse training. Instead, counterfactuals 2) and 3) respect actual clinician use within 3-hour or 15-minute windows, which still yield significant savings, with minimal implications for the firm's revenues.[33] The reassignment within 15-minute slots reduces systemwide production costs by 1.9%, which is equivalent to a 6.8% savings in firm production costs.

Appendix Table A13 reports robustness tests on the final counterfactual result from Table 4. First, bootstrapped standard error estimates indicate that our main estimate is precise. Next, we show that our estimate is robust to cost aggregations over 7 days and 30 days; using only cases from Stockholm and Scania; estimating effects using odd calendar days and constructing the counterfactual using even days; setting the KH effect to zero for symptoms for which it is not significant; using comparative advantage as in Subfigure 5b; excluding the COVID-19 symptom; and allowing for additional costs of re-assigning cases to the KH.[34]

This potential to make substantial cost savings with existing staff arises from a lack of information and inadequate incentives for the firm. A simple corrective is to issue guidelines that steer the allocation of cases toward lower systemwide production costs. It is common for healthcare payers to issue guidelines to prioritize cost-effective care, such as those mentioned in Section 2, or the UK NICE guidelines (Claxton et al. 2015; National Institute for Health

[33]Firm net revenues fall slightly because fewer cases are resolved by nurses without involving a doctor.

[34]The final row examines the possibility that the firm assigns to nurses cases that are easier to resolve within each symptom. Specifically, we assume that nurses are 5 percentage points less likely to resolve (within the firm) cases that the planner reassigns to them, and subtract these additional costs from our baseline results. This simple calculation, which does not account for social planner re-optimization, still yields large savings relative to the firm's assignment.

and Care Excellence 2024). In the firm we study, such guidelines could be easily implemented by adjusting the algorithm that assigns cases, without further administrative burden, and with only trivial losses of net revenue for the firm, which can easily be compensated for to achieve a Pareto improvement. Another approach for better aligning the firm's assignment with those of the social planner is to consider Pigouvian fees and subsidies when setting taxpayer payments to the firm. Eventually, such changes may induce the firm to change staffing or hiring; in our setting, they will likely lead to the hiring of more nurses.

8 Conclusion

In knowledge-intensive sectors where expertise is important and scarce, knowledge hierarchies are often used. However, despite their ubiquity, evidence on the productive effects of knowledge hierarchies has been limited. We contribute by studying knowledge hierarchies in a setting where cost and productivity pressures are particularly acute: primary healthcare.

Specifically, we study the effects of a nurse-doctor knowledge hierarchy (KH) in a Swedish primary healthcare provider, where we observe downstream healthcare outcomes and costs, allowing us to speak to externalities and efficiency. Using quasi-experimental algorithmic variation in the assignment of patient cases, we find that the KH reduces diagnostic informativeness and new prescriptions and slightly increases external care such as emergency department visits. However, effects on higher-stakes outcomes — including hospitalizations, labor market earnings losses, and three-year mortality — are close to zero. The KH reduces overall taxpayer payments for healthcare, but also the firm's net revenue, and the mean net effect on systemwide production costs is close to zero. However, these mean effects mask substantial heterogeneity, as the KH increases the likelihood of both low costs (when nurses resolve cases) and costs higher up in the distribution (when nurses pass cases to doctors). Moreover, the effects of the KH vary significantly by symptom. We show that the firm's selective use of the KH saves costs. However, we also show that productivity can be improved by changing the assignment of tasks between the KH and the direct-to-doctor route. Our proposed improvement takes into account the differential externality of the KH by symptom on downstream healthcare outcomes.

Profit-maximizing assignment of tasks need not be socially efficient when processing by different tiers generates downstream spillovers. We show that from a systemwide perspective, the costs involved can be nontrivial. Fixing this problem requires neither new technology nor additional staff. What is required is a better organization of information and incentives to ensure that task assignment between routes accounts for costs borne beyond the organization.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2023). "When Should You Adjust Standard Errors for Clustering?". *Quarterly Journal of Economics* 138 (1): 1–35.

Acemoglu, D. and Autor, D. (2011). "Skills, Tasks and Technologies: Implications for Employment and Earnings". In: *Handbook of Labor Economics*. Vol. 4. Elsevier, pp. 1043–1171.

Alexander, D. and Schnell, M. (2019). "Just What the Nurse Practitioner Ordered: Independent Prescriptive Authority and Population Mental Health". *Journal of Health Economics* 66, 145–162.

Azar, J. and Marinescu, I. (2024). "Chapter 10 — Monopsony Power in the Labor Market". In: *Handbook of Labor Economics*. Ed. by C. Dustmann and T. Lemieux. Vol. 5. Elsevier, pp. 761–827.

Baker, A., Callaway, B., Cunningham, S., Goodman-Bacon, A., and Sant'Anna, P. H. C. (2026). "Difference-in-Differences Designs: A Practitioner's Guide". *Journal of Economic Literature* 64 (2): 498–557.

Basu, S., Berkowitz, S. A., Phillips, R. L., Bitton, A., Landon, B. E., and Phillips, R. S. (2019). "Association of Primary Care Physician Supply with Population Mortality in the United States, 2005–2015". *JAMA Internal Medicine* 179 (4): 506–514.

Becker, G. S. and Murphy, K. M. (1992). "The Division of Labor, Coordination Costs, and Knowledge". *Quarterly Journal of Economics* 107 (4): 1137–1160.

Bloom, N., Garicano, L., Sadun, R., and Van Reenen, J. (2014). "The Distinct Effects of Information Technology and Communication Technology on Firm Organization". *Management Science* 60 (12): 2859–2885.

Bloom, N., Propper, C., Seiler, S., and Van Reenen, J. (2015). "The Impact of Competition on Management Quality: Evidence from Public Hospitals". *Review of Economic Studies* 82 (2): 457–489.

Bronsoler, A., Doyle, J., and Van Reenen, J. (2022). "The Impact of Health Information and Communication Technology on Clinical Quality, Productivity, and Workers". *Annual Review of Economics* 14, 23–46.

Caliendo, L., Mion, G., Opromolla, L. D., and Rossi-Hansberg, E. (2020). "Productivity and Organization in Portuguese Firms". *Journal of Political Economy* 128 (11): 4211–4257.

Caliendo, L., Monte, F., and Rossi-Hansberg, E. (2015). "The Anatomy of French Production Hierarchies". *Journal of Political Economy* 123 (4): 809–852.

Chan, D. and Chen, Y. (2026). "The Productivity of Professions: Evidence from the Emergency Department". *American Economic Review*.

Chan, D. C. (2018). "The Efficiency of Slacking Off: Evidence from the Emergency Department". *Econometrica* 86 (3): 997–1030.

Chernozhukov, V., Fernandez-Val, I., and Galichon, A. (2009). "Improving Point and Interval Estimators of Monotone Functions by Rearrangement". *Biometrika* 96 (3): 559–575.

Claxton, K., Martin, S., Soares, M., Rice, N., Spackman, E., Hinde, S., Devlin, N., Smith, P. C., and Sculpher, M. (2015). "Methods for the Estimation of the National Institute for Health and Care Excellence Cost-Effectiveness Threshold". *Health Technology Assessment* 19 (14): 1–503.

Costantini, S. (2025). "How Do Mental Health Treatment Delays Impact Long-Term Mortality?". *American Economic Review* 115 (5): 1672–1707.

Dahlstrand, A. (2026). "Defying Distance? The Provision of Medical Services in the Digital Age". *American Economic Review* 116 (9): 3464–3509.

Dahlstrand, A., Le Nestour, N., and Michaels, G. (2024). "Online versus In-Person Services: Effects on Patients and Providers" (Discussion Paper No. 17177, Institute of Labor Economics).

Dornbusch, R., Fischer, S., and Samuelson, P. A. (1977). "Comparative Advantage, Trade, and Payments in a Ricardian Model with a Continuum of Goods". *American Economic Review* 67 (5): 823–839.

Doyle Jr., J. J., Graves, J. A., Gruber, J., and Kleiner, S. A. (2015). "Measuring Returns to Hospital Care: Evidence from Ambulance Referral Patterns". *Journal of Political Economy* 123 (1): 170–214.

Försäkringskassan (2025). *Sjuklön: När behöver medarbetaren läkarintyg?* Försäkringskassan.

Garicano, L. (2000). "Hierarchies and the Organization of Knowledge in Production". *Journal of Political Economy* 108 (5): 874–904.

Godøy, A., Haaland, V. F., Huitfeldt, I., and Votruba, M. (2024). "Hospital Queues, Patient Health, and Labor Supply". *American Economic Journal: Economic Policy* 16 (2): 150–181.

Gottlieb, J. D., Mahoney, N., Rinz, K., and Udalova, V. (2025). "The Rise of Healthcare Jobs" (NBER Working Paper 33583, National Bureau of Economic Research).

Gruber, J., Hoe, T. P., and Stoye, G. (2023). "Saving Lives by Tying Hands: The Unexpected Effects of Constraining Health Care Providers". *Review of Economics and Statistics* 105 (1): 1–19.

Hartman, E. and Hidalgo, F. D. (2018). "An Equivalence Approach to Balance and Placebo Tests". *American Journal of Political Science* 62 (4): 1000–1013.

Ide, E. and Talamàs, E. (2025). "Artificial Intelligence in the Knowledge Economy". *Journal of Political Economy* 133 (12): 3762–3800.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

Jones, B. F. (2009). "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?". *Review of Economic Studies* 76 (1): 283–317.

Katz, L. F. and Murphy, K. M. (1992). "Changes in Relative Wages, 1963–1987: Supply and Demand Factors". *Quarterly Journal of Economics* 107 (1): 35–78.

Kelly, E., Stoye, G., and Vera-Hernández, M. (2016). "Public Hospital Spending in England: Evidence from National Health Service Administrative Records". *Fiscal Studies* 37 (3-4): 433–459.

Kleiner, M. M. (2016). "Battling over Jobs: Occupational Licensing in Health Care". *American Economic Review* 106 (5): 165–170.

Kry International AB (2024). *Annual Report and Consolidated Accounts*.

Lee, D. S. (2009). "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects". *Review of Economic Studies* 76 (3): 1071–1102.

National Academies of Sciences, Engineering, and Medicine (2021). *Implementing High-Quality Primary Care: Rebuilding the Foundation of Health Care*. Washington, DC: The National Academies Press.

National Institute for Health and Care Excellence (2024). *NICE Guidance*.

Offentlighets- och sekretesslag (2009:400) [Public Access to Information and Secrecy Act] (2009).

Oster, E. (2019). "Unobservable Selection and Coefficient Stability: Theory and Evidence". *Journal of Business & Economic Statistics* 37 (2): 187–204.

Saez, E., Schoefer, B., and Seim, D. (2019). "Payroll Taxes, Firm Behavior, and Rent Sharing: Evidence from a Young Workers' Tax Cut in Sweden". *American Economic Review* 109 (5): 1717–1763.

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

Smith, A. (1776). *An Inquiry Into the Nature and Causes of the Wealth of Nations*.

Sveriges Kommuner och Regioner (2024). *Ekonomi- och verksamhetsstatistik inom hälso- och sjukvården 2015–2025*.

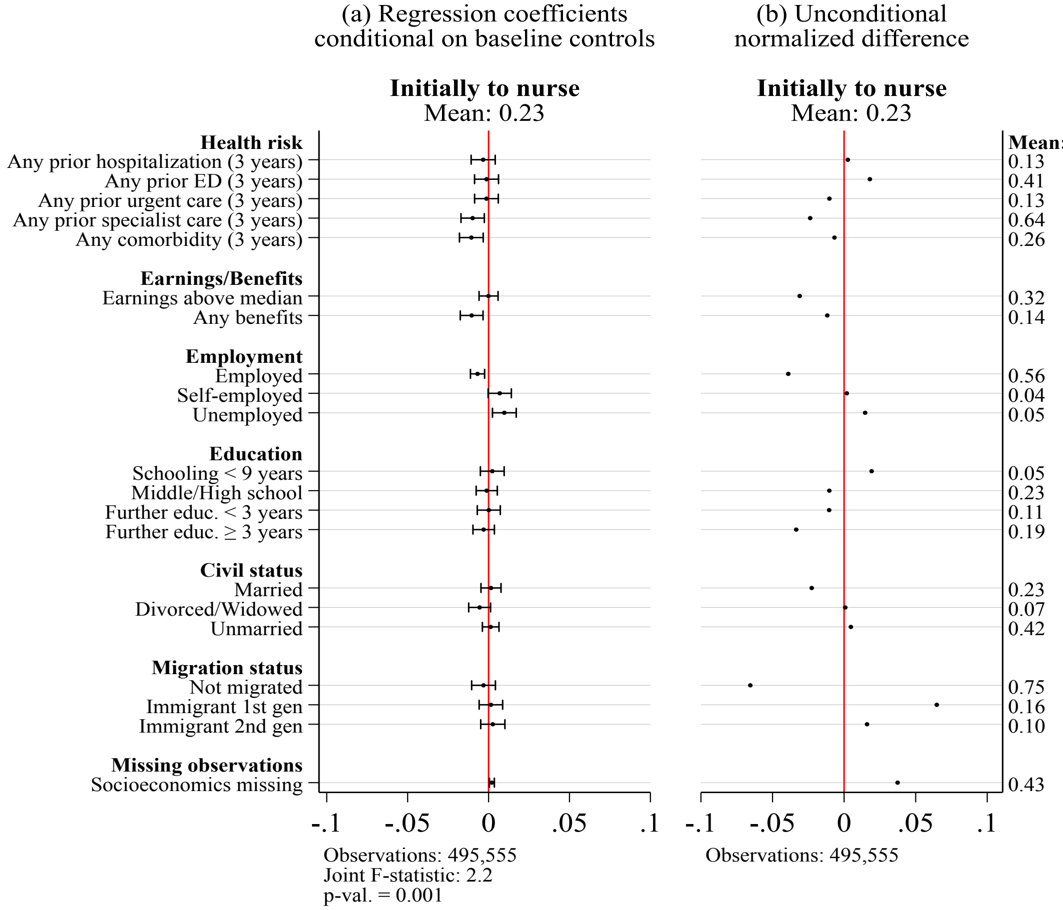
Sveriges Kommuner och Regioner (SKR) (2019). "Meddelande från styrelsen — Utvecklingen av digitala vårdtjänster" (Sveriges Kommuner och Regioner (SKR)).

Sveriges Riksbank (2020). *Exchange Rates — SEK to USD (Daily Rates 2020)*. Sveriges Riksbank statistics.

The Commonwealth Fund (2020). *Sweden | International Health Care System Profiles*.

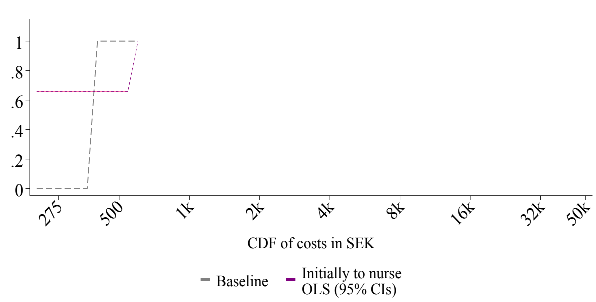
Zeltzer, D., Einav, L., Rashba, J., and Balicer, R. D. (2023). "The Impact of Increased Access to Telemedicine". *Journal of the European Economic Association*.

Figure 1. Balance of the treatment in case characteristics

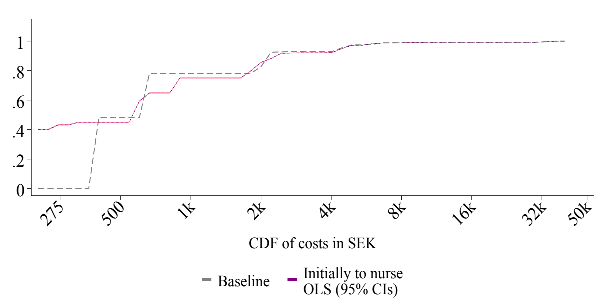


Notes: This figure presents balance tests for the treatment variable (*Nurse*). Each row shows a separate regression of the treatment on the characteristic listed on the vertical axis (each from the additional characteristics X_2). Panel (a) reports coefficients conditional on all baseline controls (X_1). Panel (b) reports 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). The F-statistic is a joint F-test of the treatment on all additional case controls (X_2), conditional on all baseline controls (X_1). The numbers on the right show sample means for each characteristic.

Figure 2. The effect of initial nurse assignment on the distribution of costs



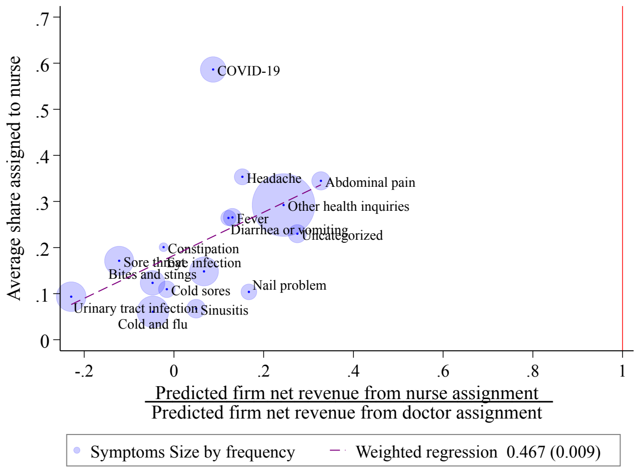
(a) Distribution of firm production costs



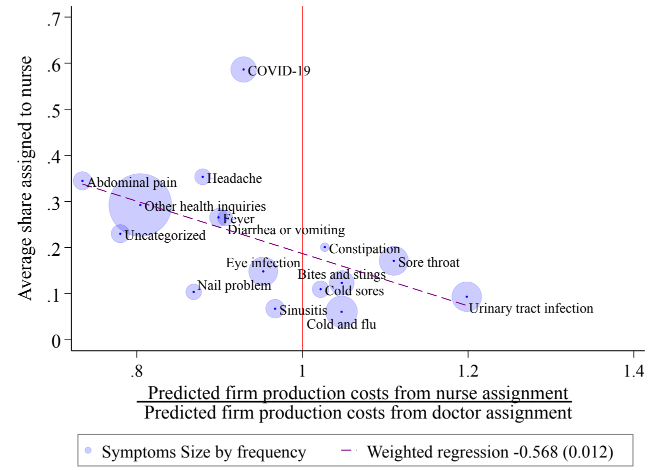
(b) Distribution of systemwide production costs

Notes: This figure plots cumulative distribution functions (CDFs) constructed from regressions using Equation 4, where each outcome is an indicator for firm production costs (Panel a) or systemwide production costs (Panel b) being weakly below the threshold shown on the x-axis (in SEK, logarithmic scale). The y-axis shows estimated probabilities of costs falling below each threshold at 0.1 intervals, with the full set of controls (X_1 and X_2). *Baseline* plots the CDF for patients initially assigned to doctors, while *Initially to nurse* plots the CDF for KH-assigned patients, constructed as the baseline mean plus the OLS coefficient. The pink shaded areas around the *Initially to nurse* line show 95% confidence intervals. CDF monotonicity is ensured following the method of Chernozhukov, Fernandez-Val, and Galichon (2009).

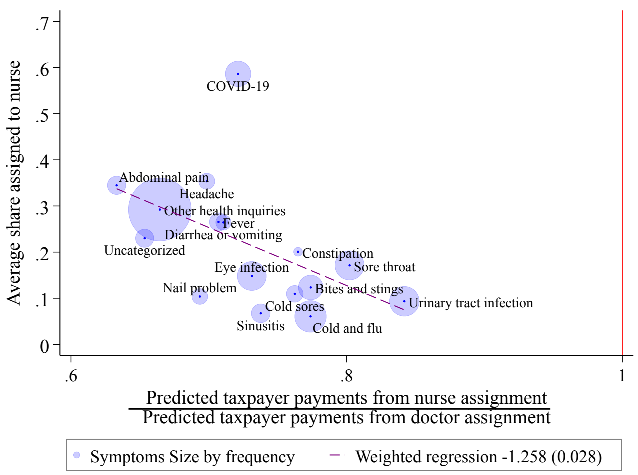
Figure 3. Case assignment and ratios of firm net revenues, taxpayer payments and costs



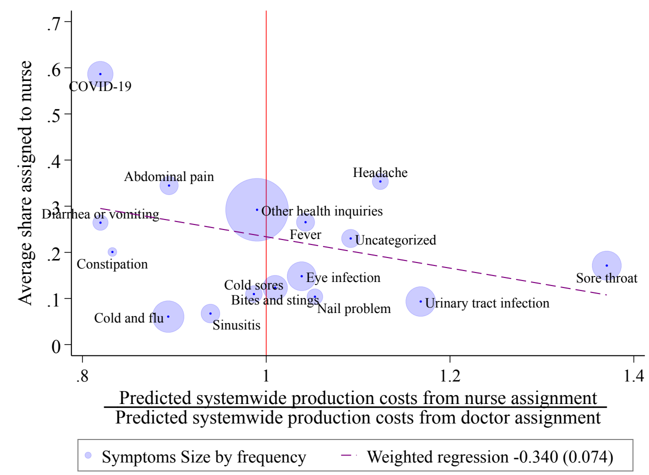
(a) Firm net revenues



(b) Firm production costs



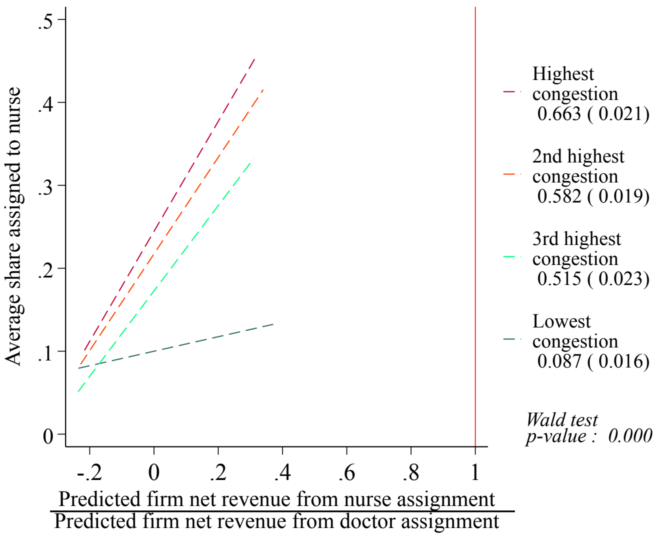
(c) Taxpayer payments to the firm



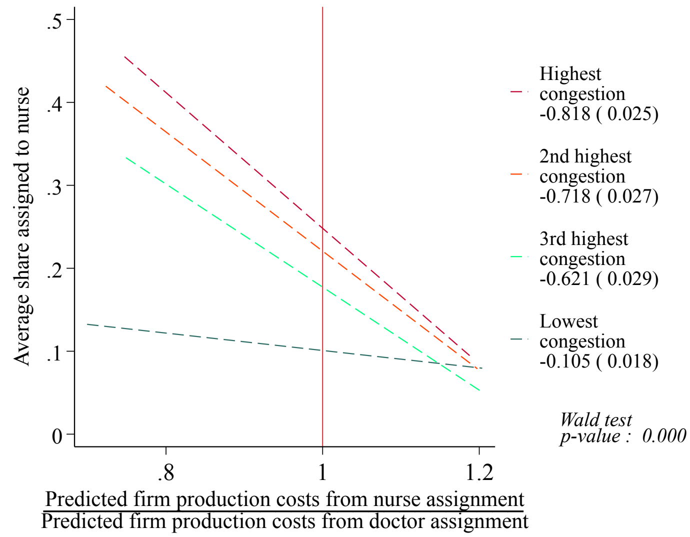
(d) Systemwide production costs

Notes: Each panel shows the firm's case assignment by the ratios of firm net revenues, firm production costs, taxpayer payments, and systemwide production costs under initial nurse over initial doctor assignment across symptom groups. The y-axes plot the sample analogues of $\mathbf{Pr}(\text{Nurse}_i = 1 \mid \text{Symptom}_i = k)$, the share of cases in symptom group k initially assigned to nurses. The x-axes plot, for each symptom group k, the estimated ratio of total outcomes under universal *knowledge hierarchy* assignment to total outcomes under universal direct-to-doctor assignment, $\sum_i \hat{Y}_{i,k}(n)/\sum_i \hat{Y}_{i,k}(d)$. Circle sizes reflect symptom frequencies. Fitted lines are weighted OLS regressions (weighted by symptom frequency), and their coefficients are reported in the legends, with standard errors that are calculated from 100 bootstrap replications (resampled at the consultation level). The vertical red line indicates equal costs or revenues. Negative (positive) slope in cost (revenue) panels suggests that nurses are more likely to be assigned cases where they have a predicted advantage over doctors. All regressions use the analysis sample with full controls (X_1 and X_2).

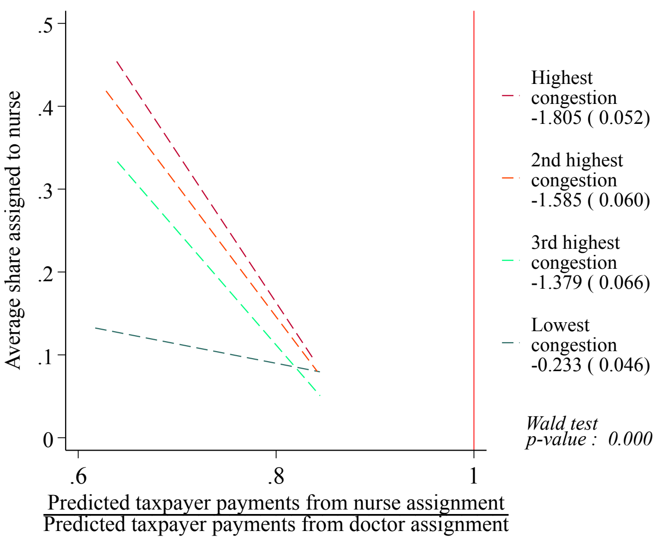
Figure 4. Case assignment and ratios of revenues, taxpayer payments and costs, by congestion



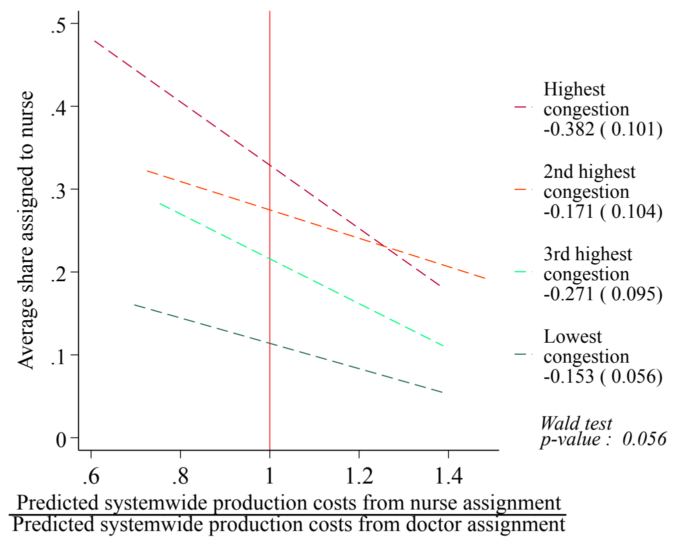
(a) Firm net revenues



(b) Firm production costs



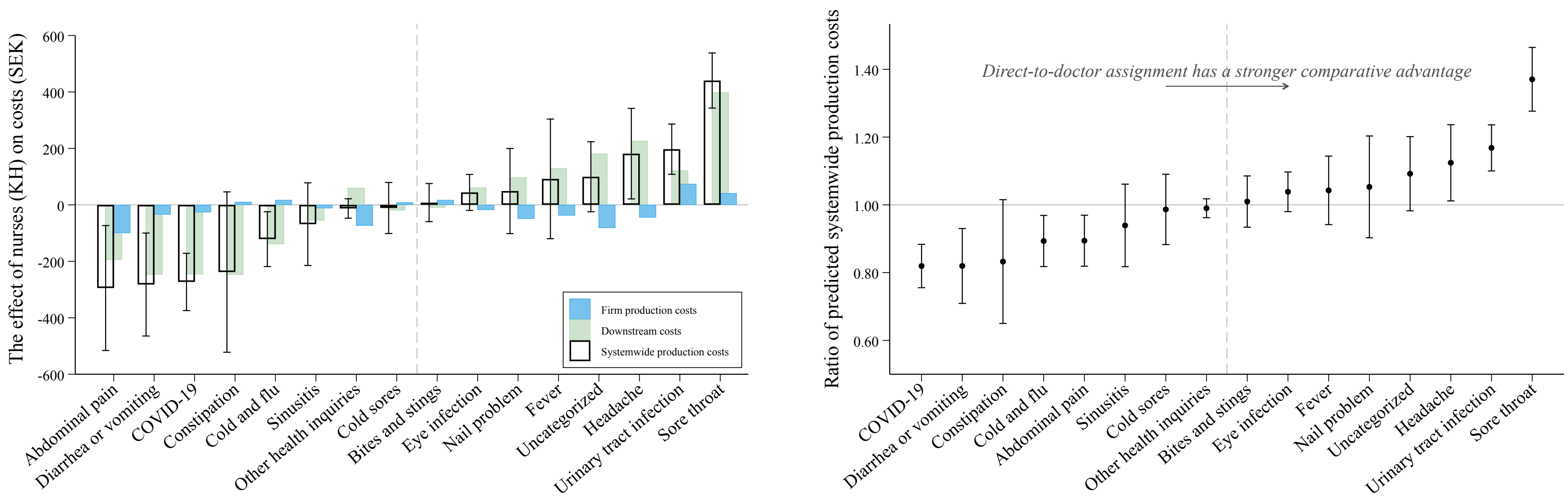
(c) Taxpayer payments to the firm



(d) Systemwide production costs

Notes: These figures show the ratios of firm net revenue, firm production costs, taxpayer payments to the firm, and systemwide production costs of initial nurse over initial doctor assignment across symptom groups, by congestion quartile. Congestion quartiles are based on the log median waiting time of the five most recent doctor consultations within 30 minutes of patient login. The y-axes plot the sample analogues of $\mathbf{Pr}(\text{Nurse}_i = 1 \mid \text{Symptom}_i = k, Q_i = q)$, the share of cases in symptom group k and congestion quartile q initially assigned to nurses, where Q_i denotes the congestion quartile of case i. The x-axes plot, for each symptom group k, the estimated ratio of total outcomes under universal *knowledge hierarchy* assignment to total outcomes under universal direct-to-doctor assignment, $\sum_i \hat{Y}_{i,k}(n)/\sum_i \hat{Y}_{i,k}(d)$. Dashed lines are weighted OLS regression fits (weighted by symptom frequency), and their coefficients are reported in the legend below each quartile, with standard errors calculated from 100 bootstrap replications (resampled at the consultation level). The vertical red line marks equal predicted outcomes under nurse and doctor assignment. The Wald test reported in each legend is for the difference of the slopes between the lowest and highest congestion quartiles.

Figure 5. KH effect and average predicted cost by symptoms



(a) The social planner's criterion and its constituent parts

(b) Comparative advantage

Notes: These figures show heterogeneous effects of the *knowledge hierarchy* across symptom groups estimated using Equation 5. The left panel decomposes systemwide production costs into downstream costs and firm costs. The right panel shows, for each symptom group k, the estimated ratio of total systemwide production costs under universal KH assignment to systemwide production costs under universal direct-to-doctor assignment, $\sum_i \hat{Y}_{i,k}(n)/\sum_i \hat{Y}_{i,k}(d)$, with symptoms sorted by this ratio. For the first eight symptom groups, predicted costs are lower under initial nurse assignment (nurse absolute advantage); the vertical line marks where this advantage reverses. Error bars in the left panel denote 95% confidence intervals based on analytical robust standard errors, whereas those in the right panel are based on bootstrapped standard errors with 100 replications.

Table 1. Characteristics of the analysis sample

	Full sample		Initially to nurse		Directly to doctor		Diff. p -value
Symptoms							
Abdominal pain	0.030	[0.17]	0.046	[0.21]	0.025	[0.16]	0.000
Cold and flu	0.089	[0.28]	0.024	[0.15]	0.11	[0.31]	0.000
Cold sores	0.025	[0.16]	0.012	[0.11]	0.029	[0.17]	0.000
Constipation	0.0070	[0.083]	0.0062	[0.078]	0.0072	[0.085]	0.000
COVID-19	0.058	[0.23]	0.15	[0.36]	0.031	[0.17]	0.000
Diarrhea or vomiting	0.021	[0.14]	0.024	[0.15]	0.020	[0.14]	0.000
Eye infection	0.075	[0.26]	0.049	[0.22]	0.083	[0.28]	0.000
Fever	0.029	[0.17]	0.034	[0.18]	0.028	[0.16]	0.000
Headache	0.023	[0.15]	0.036	[0.19]	0.019	[0.14]	0.000
Nail problem	0.022	[0.15]	0.010	[0.10]	0.026	[0.16]	0.000
Other health inquiries	0.35	[0.48]	0.45	[0.50]	0.32	[0.47]	0.000
Bites and stings	0.054	[0.23]	0.030	[0.17]	0.062	[0.24]	0.000
Sinusitis	0.031	[0.17]	0.0093	[0.096]	0.038	[0.19]	0.000
Sore throat	0.076	[0.26]	0.057	[0.23]	0.081	[0.27]	0.000
Uncategorized	0.030	[0.17]	0.030	[0.17]	0.029	[0.17]	0.419
Urinary tract infection	0.079	[0.27]	0.032	[0.18]	0.092	[0.29]	0.000
Demographics							
Female	0.63	[0.48]	0.59	[0.49]	0.64	[0.48]	0.000
Patient age	29.2	[16.5]	29.3	[16.6]	29.2	[16.5]	0.569
Health risk factors							
Any prior hospitalization (3 years)	0.13	[0.34]	0.14	[0.34]	0.13	[0.34]	0.443
Any prior ED (3 years)	0.41	[0.49]	0.41	[0.49]	0.40	[0.49]	0.000
Any prior urgent care (3 years)	0.13	[0.34]	0.13	[0.33]	0.13	[0.34]	0.002
Any prior specialist care (3 years)	0.64	[0.48]	0.63	[0.48]	0.64	[0.48]	0.000
Any comorbidity (3 years)	0.26	[0.44]	0.26	[0.44]	0.26	[0.44]	0.048
Observations	495,555		112,641		382,914		

Note: This table reports summary statistics for the analysis sample. The displayed variables are components of baseline controls X_1 and additional controls X_2, described in section 3.2. Standard deviations are in brackets. The final column reports p-values from t-tests comparing the means between cases initially assigned to nurses and those directly assigned to doctors.

Table 2. Predictors of initial assignment to nurse

	Initial assignment to nurse				
	(1)	(2)	(3)	(4)	(5)
X_1	✓		✓	✓	✓
X_2		✓	✓	✓	✓
Log median doctor wait time				✓	✓
Doctor-group status in prior cases					✓
R-squared	0.216	0.002	0.216	0.221	0.221
Adjusted R-squared	0.195	0.002	0.195	0.200	0.200
Observations	495,555	495,555	495,555	495,555	495,555

Notes: This table reports the predictive quality from regressions of initial nurse assignment on case characteristics and exogenous assignment shifters. Baseline controls X_1 and additional controls X_2 are defined in Section 3.2. The exogenous assignment shifters include congestion (*log median doctor wait time*) and indicators for the doctor-group status for the 50 previous cases, ordered by login time among cases with either doctor- or overlap-group symptoms. Kleibergen–Paap F-statistics are 3,726 in column (4) and 3,730 in column (5).

Table 3. Effects of the knowledge hierarchy on patient outcomes and costs

Panel A: Within-firm outcomes

	Any clinician-booked doctor visit (7 days)	Any unbooked revisit (7 days)	Top score rating (final)	Informative diagnosis (final)
Initially to nurse	0.31	0.0078	-0.015	-0.093
	(0.0015)	(0.00093)	(0.0018)	(0.0017)
X_1 and X_2	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555
Baseline mean	0.014	0.058	0.40	0.79

Panel B: Downstream healthcare use

	Any new prescription (7 days)	Any external PCP consultation (7 days)	Any urgent care (7 days)	Any specialist care (7 days)
Initially to nurse	-0.10	0.025	0.0022	0.0032
	(0.0017)	(0.0014)	(0.00038)	(0.00073)
X_1 and X_2	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555
Baseline mean	0.44	0.16	0.0086	0.035

	Any ED (7 days)	Any hospitalization (7 days)
Initially to nurse	0.0056	-0.00048
	(0.00081)	(0.00038)
X_1 and X_2	✓	✓
Observations	495,555	495,555
Baseline mean	0.037	0.0082

Panel C: Long-term consequences

	Earnings decrease > 20 pct. (cal. month after)	Zero earnings (cal. month after)	Death (3 years)
Initially to nurse	-0.021	-0.0031	-0.000013
	(0.0024)	(0.0013)	(0.00018)
X_1 and X_2	✓	✓	✓
Observations	214,674	214,674	495,555
Baseline mean	0.22	0.050	0.0021

Panel D: Cost outcomes

	Taxpayer payments to the firm (within-day)	Taxpayer payments to the firm (7 days)	Firm net revenue (within-day)	Firm production costs
Initially to nurse	-147.9	-149.1	-107.6	-40.3
	(0.33)	(0.63)	(0.22)	(0.55)
X_1 and X_2	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555
Baseline mean	500	544.6	125	375

	Total taxpayer payments	Systemwide production costs
Initially to nurse	-112.5	-4.84
	(13.2)	(13.2)
X_1 and X_2	✓	✓
Observations	495,555	495,555
Baseline mean	1386.8	1261.8

Note: This table presents regression results for our main specifications on the effects of initial case assignment to a nurse (the *knowledge hierarchy*) rather than directly to a doctor. Each subpanel corresponds to a different outcome. Outcomes are defined in Table **??**. All earnings measures are defined for a subsample with stable earnings of employed patients earning more than 3,533 SEK (2010 values) per month in each of the three prior months, based on an annual threshold of 42,400 SEK (Saez, Schoefer, and Seim 2019). The columns in each subpanel report the OLS estimate for *Nurse* based on Equation 4. All regressions control for baseline characteristics X_1 and additional controls X_2 as described in the main text. The baseline mean is the average of the dependent variable for cases directly assigned to a doctor. Robust standard errors are in parentheses.

Table 4. Percentage change in outcomes and inputs from counterfactuals relative to the actual assignment

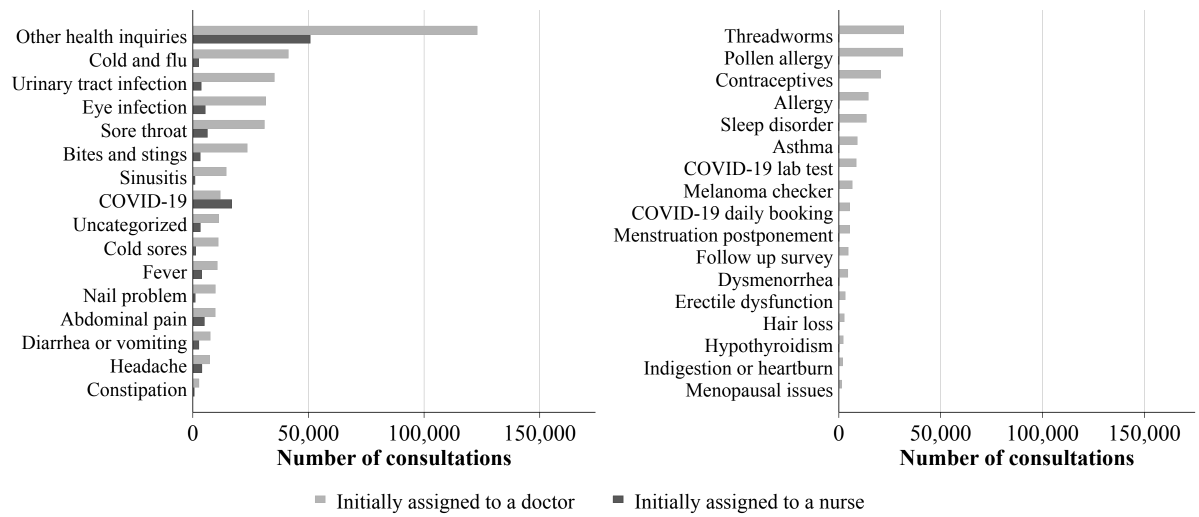
	$\frac{\Delta\text{ Systemwide production costs}}{\text{Systemwide production costs}}$	$\frac{\Delta\text{ Systemwide production costs}}{\text{Firm production costs}}$	$\frac{\Delta\text{ Total taxpayer payments}}{\text{Total taxpayer payments}}$	$\frac{\Delta\text{ Firm net revenue}}{\text{Firm net revenue}}$	Share of cases assigned initially to nurses
	(1)	(2)	(3)	(4)	(5)
A: Switching from actual assignment to non-optimized assignments					
1) Cases assigned uniformly to nurse/doctor proportional to clinician shares in *prior* 15 minutes	0.55%	1.92%	0.42%	−1.22%	23%
2) Cases assigned uniformly to nurse/doctor proportional to clinician shares in *fixed* 15-minute windows	0.52%	1.83%	0.39%	−1.31%	23%
3) Cases assigned uniformly to nurse/doctor proportional to clinician shares in *prior* 3 hours	0.53%	1.87%	0.50%	0.057%	21%
4) Cases assigned uniformly to nurse/doctor proportional to clinician shares in *fixed* 3-hour windows	0.55%	1.93%	0.41%	−1.40%	23%
5) All cases are initially assigned to nurses	1.50%	5.25%	−5.05%	−88.3%	100%
6) All cases are initially assigned to doctors	0.45%	1.56%	2.16%	24.0%	0%
B: Switching from actual assignment to assignments that minimize systemwide production costs					
1) Unconstrained minimization of systemwide production costs	−3.46%	−12.1%	−6.08%	−39.5%	61%
2) Constrained minimization of systemwide production costs (Nurses constant within 3-hour windows)	−2.43%	−8.50%	−2.30%	−0.71%	23%
3) Constrained minimization of systemwide production costs (Nurses constant within 15-minute windows)	−1.94%	−6.81%	−1.82%	−0.27%	23%

Note: This table shows the percentage change in outcomes from counterfactual assignments relative to the actual assignment. To avoid randomization, the counterfactuals assign each case the cost associated with a "fractional clinician": the baseline mean cost (under direct-to-doctor assignment) plus $N_t/(D_t + N_t)$ times the estimated nurse–doctor cost difference, where D_t and N_t are the number of doctors and nurses available at time t. Only the assignments in Panel A rows 2 and 4 and Panel B rows 2 and 3 hold total clinician assignments fixed as in reality; the remaining counterfactuals entail using more or fewer nurses as indicated in the final column. The number of actual nurse assignments is 112,641.

Online Appendix

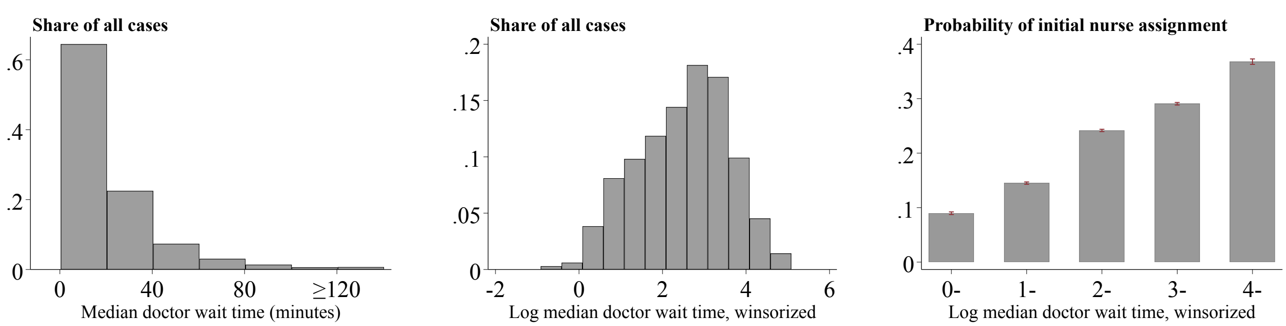
Organizing Expertise with Externalities: Evidence from Primary Care

Figure A1. Case groups of symptoms



Notes: This figure shows the number of consultations by self-reported symptom category, split by whether the case was initially assigned to a nurse or directly to a doctor. Panel (a) shows the overlap group (the analysis sample) and panel (b) the doctor group; see Appendix Table A2 for the definition of each group and the full set of sample restrictions. The nurse group is not shown separately: it contains 52,477 consultations, of which 50,502 report *chlamydia* and 1,975 report *follow up* as the symptom.

Figure A2. Congestion measure



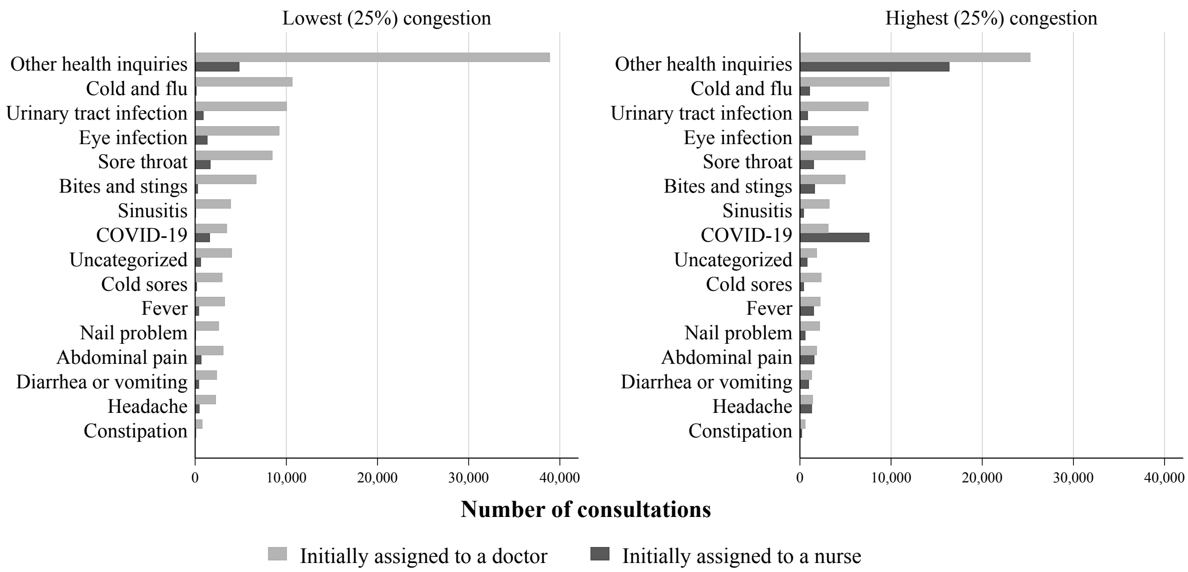
(a) Wait time before transformation

(b) Wait time after transformation

(c) Congestion and initial nurse assignment

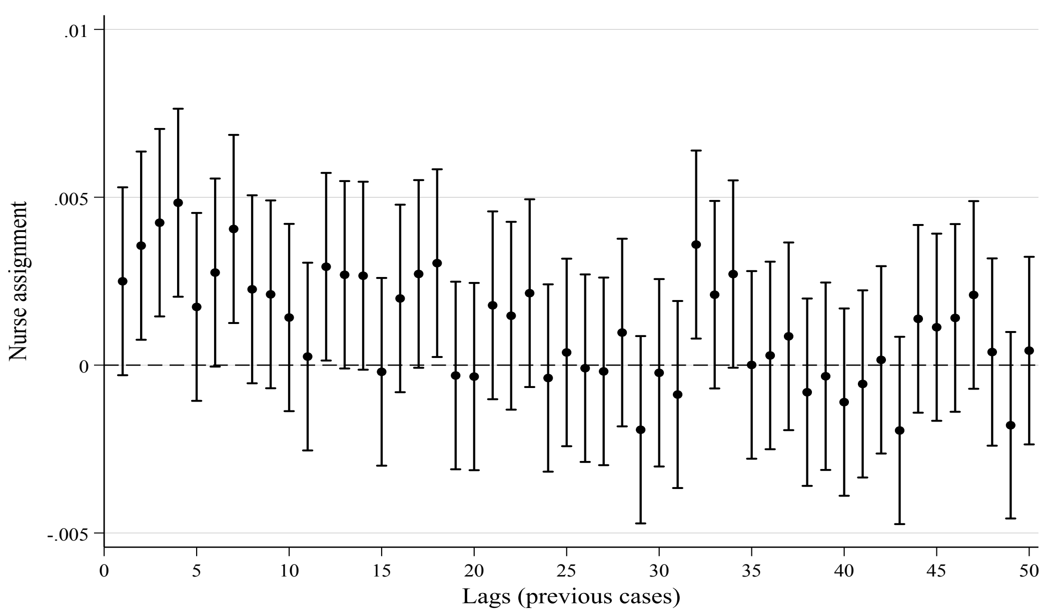
Notes: Panel (a) shows the distribution of the untransformed congestion measure in minutes, defined as the median doctor waiting time among the up to 5 prior patients whose consultations began within 30 minutes of each patient's login time. Panel (b) shows the distribution of *Log median doctor wait time* and Panel (c) shows the probability of initial nurse assignment across congestion levels.

Figure A3. Assignment of symptoms in the lowest and highest quartiles of congestion



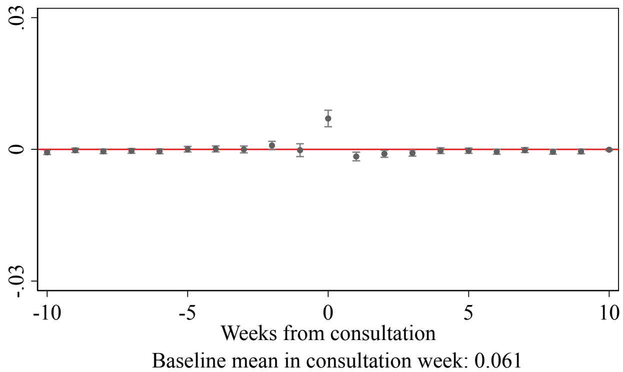
Notes: This figure shows the symptom categories of patients initially assigned to nurses or doctors in the highest and lowest quartiles of doctor congestion in the analysis sample. Congestion is measured as the log median waiting time of the five most recent doctor consultations within 30 minutes of patient login. The lowest and highest quartiles correspond to the bottom and top 25% of the congestion distribution in the analysis sample.

Figure A4. Effects of lagged doctor group case on case assignment

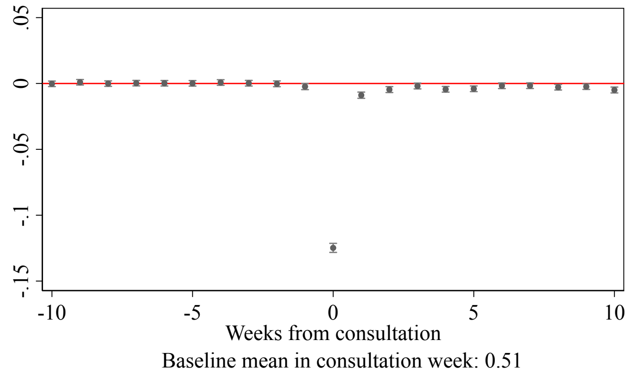


Notes: This figure reports estimates from regressions of initial nurse assignment on indicators for lagged doctor-group cases, ordered by patient login time. The doctor group consists of symptoms that are assigned to a doctor in at least 99% of cases. Each point corresponds to the coefficient from a separate OLS regression using the case at lag n. All regressions include full controls (X_1 and X_2) and use the analysis sample of overlap-group cases. Vertical lines denote 95% confidence intervals.

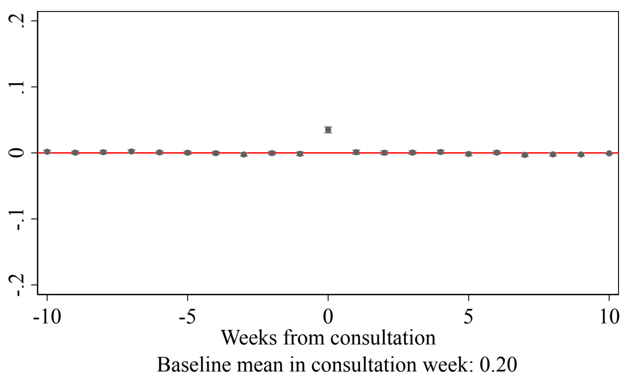
Figure A5. Effect of an initial nurse assignment on lags and leads of patient outcomes



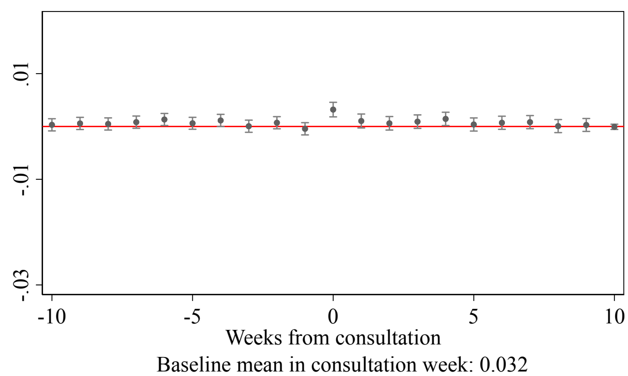
(a) Any unbooked revisit



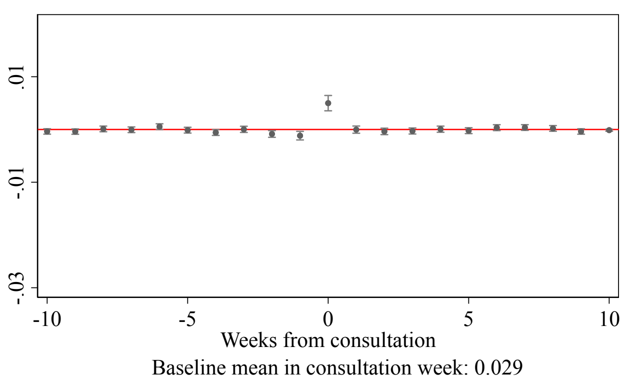
(b) Any prescription



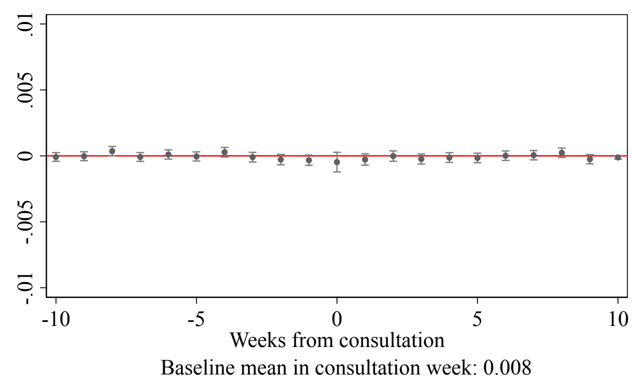
(c) Any external PCP consultation (Stockholm & Scania)



(d) Any specialist care



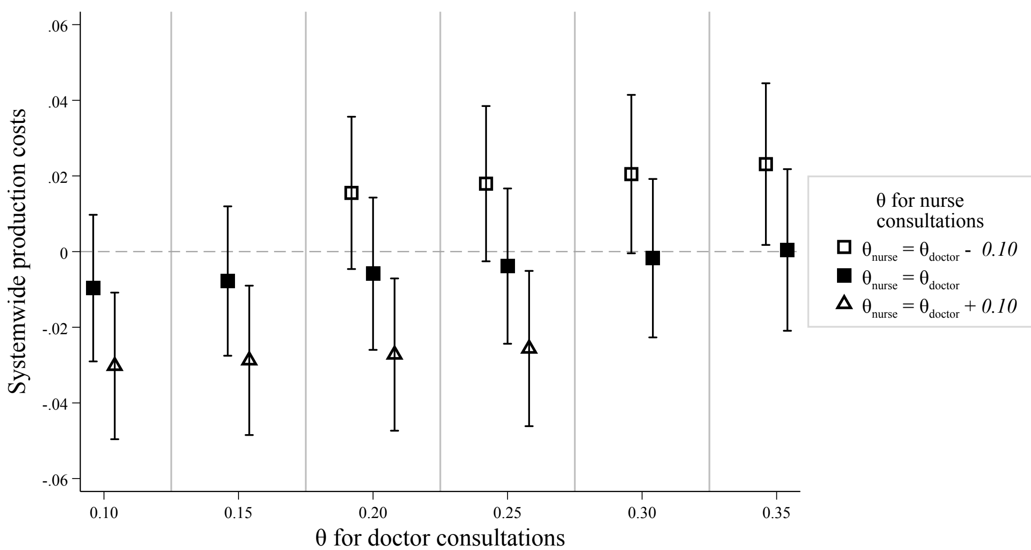
(e) Any ED



(f) Any hospitalization

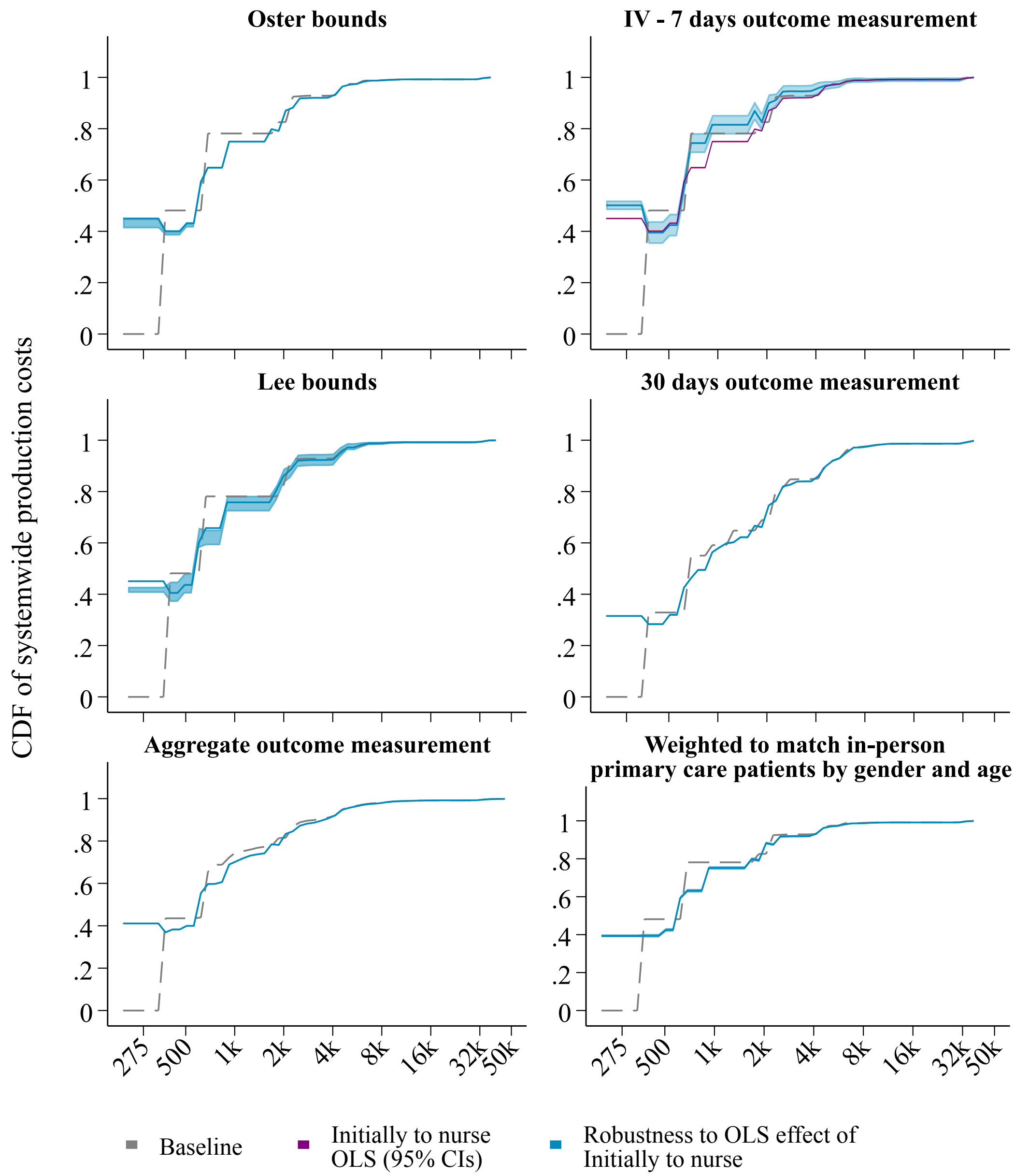
Notes: These figures show the estimated effect of initial assignment to a nurse (the *knowledge hierarchy*) on outcomes in the 10 weeks before and after the initial consultation. "Week 0" marks the week starting with the initial consultation date. Each estimate is from a separate OLS regression of a lagged or lead outcome on *Nurse*, based on Equation 4 with the full set of controls and date-by-3-hour fixed effects. The baseline mean is the average of the week-0 outcome for cases directly assigned to a doctor. Vertical lines denote 95% confidence intervals. Subfigure A5c uses only cases from Stockholm and Scania (for whom external primary care is fully observed), comprising 262,144 consultations. All other figures use the main analysis sample, consisting of 495,555 observations.

Figure A6. Systemwide production costs when markdowns θ vary between nurses and doctors



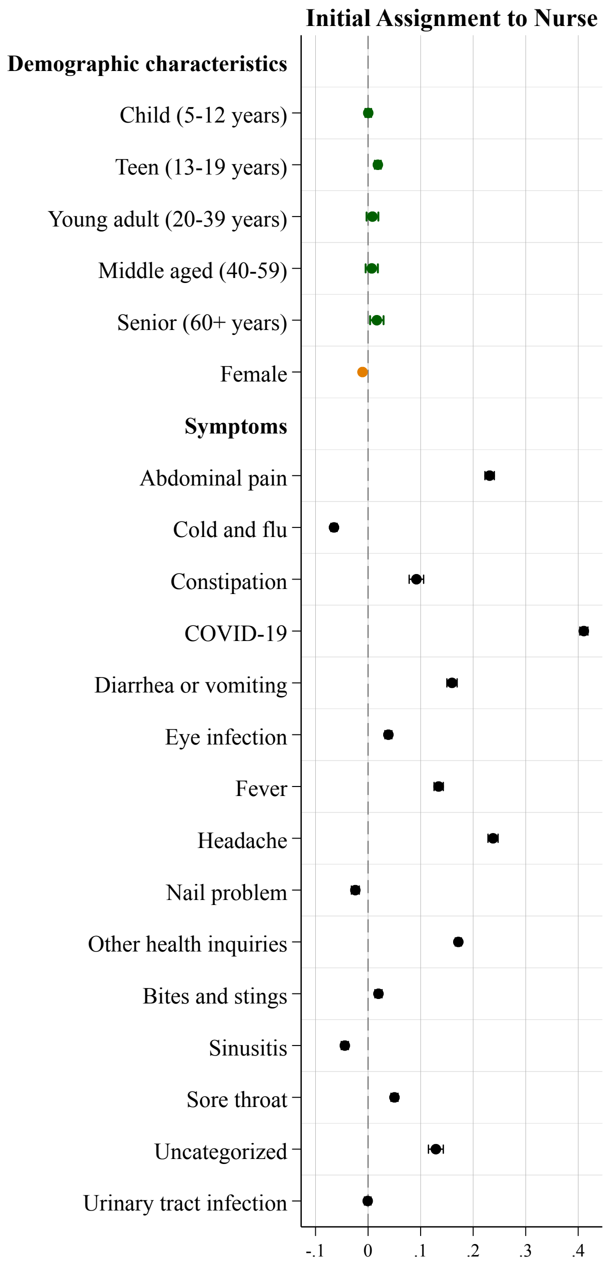
Notes: This figure shows the estimated proportional effect of initial nurse assignment on systemwide production costs across combinations of clinician-type-specific markdowns. The x-axis shows the markdowns for consultations resolved by doctors, θ_{doctor}, over the range $[0.1, 0.35]$, which spans existing monopsony estimates of $\theta \in [0.13, 0.33]$ (Azar and Marinescu 2024). The markers vary the markdown for consultations resolved by nurses relative to the markdown for doctor consultations: equal ($\theta_{\text{nurse}} = \theta_{\text{doctor}}$), lower ($\theta_{\text{nurse}} = \theta_{\text{doctor}} - 0.10$), or higher ($\theta_{\text{nurse}} = \theta_{\text{doctor}} + 0.10$). Our rule of including plus/minus 0.10 compared to equal markdowns lead to different number of coefficients being shown in the window 0.10 to 0.35. Vertical bars show standard errors. The full set of controls X_1 and X_2 is included.

Figure A7. Robustness on the distribution of systemwide production costs

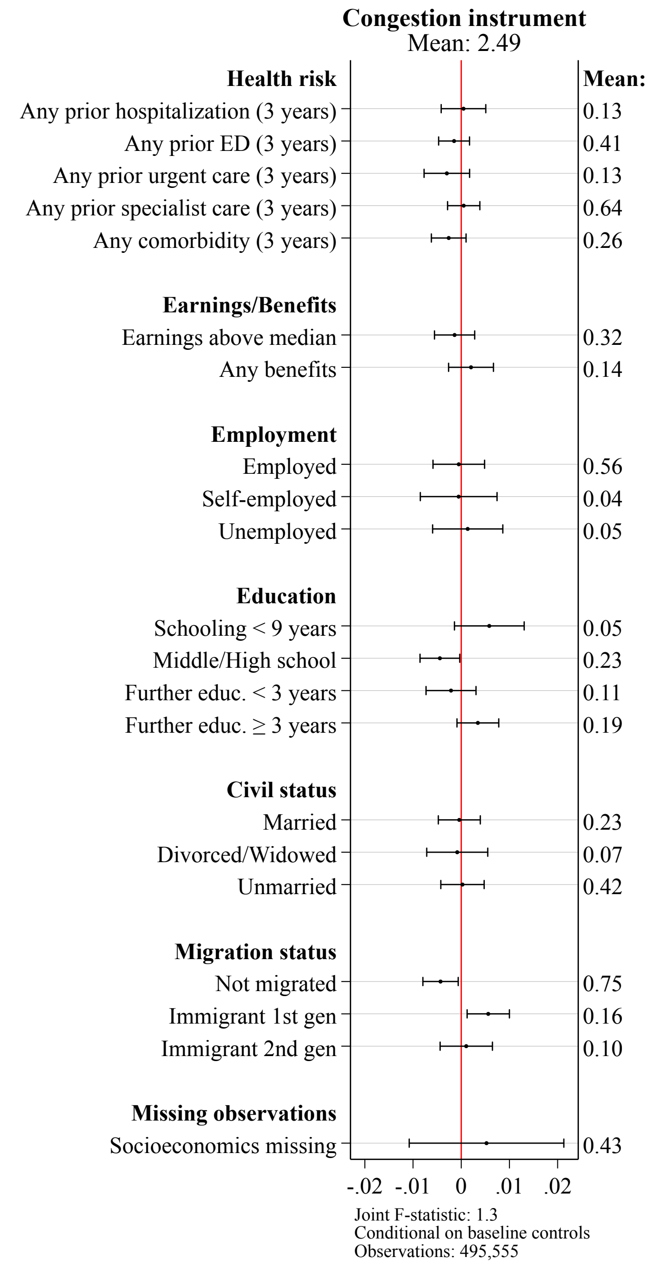


Notes: These panels provide robustness checks for the Cumulative Distribution Function (CDF) estimates in Figure 2b. The top-left panel shows Oster bounds following Oster (2019). The grey dashed line reports the baseline specification, while the blue line shows the bias-adjusted estimates assuming the maximum R^2 is 1.3 times the controlled R^2. The top-right panel presents IV estimates using the congestion instrument with 7-day cost outcomes. The middle-left panel shows Lee bounds, constructed using symptom-specific conversion rates for treatment and control groups as described in Section 6.3. The middle-right panel reports results using an alternative outcome window, measuring costs over 30 days. The bottom-left panel uses an alternative outcome definition based on aggregated costs over 7 days. The bottom-right panel reports estimates reweighted to match the distribution of in-person primary care patients by age and gender. Across panels, the grey dashed line indicates the baseline, the purple line shows the OLS estimate for *initially to nurse* with 95% confidence intervals, and the blue line reports the corresponding robustness specification. All regressions include the full set of controls X_1 and X_2.

Figure A8. Nurse assignment heterogeneity and instrument balance



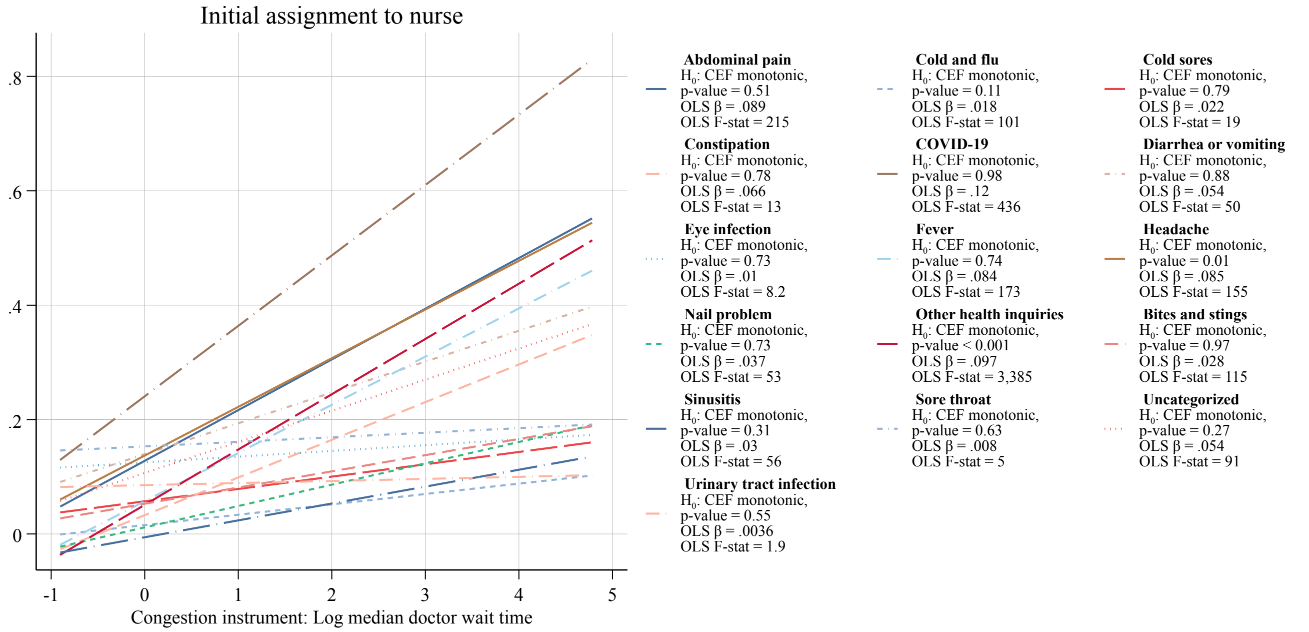
(a) Nurse assignment heterogeneity



(b) Balance of the congestion instrument in case characteristics

Notes: Panel (a) shows heterogeneity in nurse assignment. Coefficients are estimated by regressing nurse assignment on the additional controls (X_2) and a modified version of the baseline controls (X_1). In this modified version, age-group, patient gender, and symptom-group indicators enter separately, while the interaction term includes only whether the patient's last case at the firm was with a doctor and whether the patient has a past prescription from the firm. Date-by-3-hour windows and the patient's aggregated region enter as in the baseline specification. Panel (b) presents balance tests for the congestion instrument (*log median doctor wait time*). Each row reports a separate regression of the instrument on the corresponding characteristic from X_2, conditional on X_1 as defined in the baseline specification. The reported F-statistic tests the joint association between the instrument and all additional case characteristics in X_2, conditional on X_1. Numbers on the right report the sample mean of each characteristic. Horizontal lines show 95% confidence intervals based on robust standard errors in both panels.

Figure A9. Instrument monotonicity test



Notes: This figure shows non-parametric estimates of the conditional expectation function (CEF) for initial nurse assignment as a function of the congestion instrument, *log median doctor wait time*, estimated separately within symptom subsamples. In each subsample, the CEF is shown as a smoothed line over a binscatter. All estimates condition on the baseline controls X_1. Bins are chosen using the MSE-optimal procedure of Cattaneo et al. (2024). A p-value close to 1 indicates no evidence against monotonicity, while a low p-value indicates evidence of a monotonicity violation. The reported OLS β coefficients and F-statistics are obtained from linear regressions estimated separately within each subsample, conditional on the same X_1 controls.

Table A1. Healthcare comparison

	Indicator	Countries
(1)	Telehealth share in 2019 (14 countries with data)	DK 29.9%; EE 20.6%; PT 18%; IL 15.6%; FI 8.3%; ES 4.1%; HU 3.7%; CZ 3.7%*; NO 2.2%; SI 1.5%; LT 1%; FR 0%*; AU 0%; GR 0%
(2)	Telehealth share in 2020 (19 countries with data)	PT 37.5%; US 36-48%; EE 35.9%; ES 30.6%; DK 29.9%; IL 29.5%; **SE** 27.6%; LT 26.5%; NO 17.8%; BE 13.9%; AU 9.5%; FI 9.1%; CZ 8.6%*; FR 5.7%*; CL 4.5%; HU 4.2%; SI 3.7%; DE 1.1%*; GR 0%
(3)	Telehealth share in 2023 (22 countries with data)	EE 35.6%; IL 29.2%; PT 26.4%; DK 25%; **SE** 25%; US 25-27%; ES 22.4%; AU 18.9%; NO 18.4%; SI 14.8%; LT 12.4%; HU 12.4%; BE 10.2%; PL 8.3%; CZ 8%*; FI 6.8%; FR 3.6%*; LU 1.9%; DE 1%*; CL 0%; GR 0%; KR 0%
(4)	Ratio of nurses to doctors in 2019 (38 countries with data)	JP 4.7; FI 4.4; US 4.3; CH 4.1; LU 3.9; IE 3.9; IS 3.9; NO 3.6; CA 3.6; BE 3.5; FR 3.3; DE 3.2; SI 3.2; KR 3.2; AU 3.2; NZ 3; NL 2.9; GB 2.8; **SE** 2.5; DK 2.4; CZ 2.1; PL 2.1; AT 2; HU 1.9; EE 1.8; LT 1.7; GR 1.6; SK 1.6; IT 1.5; IL 1.5; ES 1.3; PT 1.3; LV 1.3; TR 1.2; MX 1.2; CR 1.1; CL 1.1; CO 0.6
(5)	Ratio of nurses to doctors in 2023 (26 countries with data)	FI 4.4; US 4.4; CA 3.7; KR 3.6; IE 3.6; IS 3.4; NZ 3.2; NO 3.1; AU 3.1; SI 3.0; NL 2.8; DE 2.6; **SE** 2.5; DK 2.3; CZ 2.1; GB 2.0; AT 1.9; EE 1.9; LT 1.6; IL 1.6; PL 1.5; HU 1.5; ES 1.3; IT 1.3; LV 1.2; MX 1.1
(6)	Prescriptions can be issued in teleconsultations (30 countries with data)	*YES*: AU; AT; BE; CA; CR; CZ; DE; EE; FR; GB-ENG; HU; IS; IE; IL; JP; KR; LV; LT; LU; MX; NL; NZ; NO; PL; PT; SI; **SE**; CH; TR; US
(7)	Primary care model may include fee-for-service (25 countries with data)	*YES*: AU; AT; BE; CA; CZ; DE; DK; EE; FI; FR; GB; GR; IT; LU; MX; NL; NZ; NO; PL; PT; SI; **SE**; US *NO*: LV; ES
(8)	Teleconsultations may be provided without prior contact (32 countries with data)	*YES*: AT; BE; CA; CH; CR; DE; DK; FI; FR; GB-ENG; HU; IS; IE; IL; IT; JP; LV; LT; NZ; NO; PL; PT; SI; **SE**; TR; US *NO*: AU; CZ; NL; MX; EE; LU

Notes: This table compiles health system indicators from multiple sources, focusing on OECD countries where data are available. For each indicator, we list country abbreviations with their associated shares, ratios, or yes/no status, as relevant for each indicator. The country abbreviations are ISO 3166-1 alpha-2 codes. Rows (1), (2), and (3) use data from OECD (2025a), National Center for Health Statistics (2023), and Centers for Medicare and Medicaid Services (CMS) (2025). Row (4) uses data from OECD (2021a). Row (5) uses data from OECD (2025c) and OECD (2025b). Row (6) and (8) use data from OECD (2021b). Row (7) uses data from the latest WHO Health System Summary (European Observatory on Health Systems and Policies 2026a) or Health System Review (European Observatory on Health Systems and Policies 2026b) for each country. Figures pertain to 2023 unless otherwise specified. *Teleconsultations include only those reimbursed through the national healthcare system.

Table A2. Restrictions on the main analysis sample

	Number of observations	Share assigned to nurses (%)
All observed consultations	1,814,706	
1. Keep only consultations that involve either doctors or nurses	1,668,303	14.54%
2. Exclude scheduled or in-person consultations	1,239,527	19.45%
3. Exclude observations outside the analysis time window	1,126,662	21.11%
4. Exclude patients registered with the provider	1,054,985	20.57%
5. Exclude infants (ages 0-1)	995,177	21.20%
6. Exclude follow-ups within two hours by the same patient	970,277	21.18%
7. Exclude consultations with missing characteristics → *Broad drop-in sample*	969,483	21.16%
a) *Main analysis sample*		
· Restrict broad drop-in sample to overlap group	496,002	22.73%
· Exclude singleton observations in the full control scheme	495,555	22.73%
b) *Doctor group*		
· Restrict broad drop-in sample to doctor group	168,098	0.06%
· Exclude singleton observations in the full control scheme	167,050	0.06%
c) *Nurse group*		
· Restrict broad drop-in sample to nurse group	52,477	99.99%
d) *Residual group*		
· Restrict broad drop-in sample to residual group	252,906	15.76%

Note: This table presents the number of observations remaining after each sample restriction - columns show the total number of consultations, and the share of consultations initially assigned to a nurse. Restrictions 1–3 limit the sample to unscheduled online consultations with doctors or nurses from 1 April 2019 to 24 December 2020, when nurse consultations started to take place for overlap cases and follow-up consultations within 7 days are observable. Restrictions 4 and 5 exclude patients registered with the firm and infants (age $<$ two years), as different internal protocols apply to them. Restriction 6 excludes repeat consultations within 2 hours; Restriction 7 excludes consultations where baseline characteristics (login time, patient age, gender, region) are missing yielding the *broad drop-in sample*. Patient-reported symptoms define four subgroups: The *main analysis sample* is the overlap group where 5% – 95% of consultations are assigned to nurses. A symptom belongs to the *doctor group* (*nurse group*) if assigned to doctors (nurses) in at least 99% of cases. The *residual group* includes symptoms that are neither in the main analysis sample nor in the doctor or nurse group, along with rare symptoms (fewer than 1,000 observations in broad drop-in sample).

Table A3. Costs of healthcare services covered by Sweden's public healthcare system

Service	Cost in SEK	Year	Source	Link
Online primary care consultation				
Doctor	500	2019	Recommendation on remuneration for digital healthcare services to healthcare providers.	Sveriges Kommuner och Regioner (SKR)[1] Last access: 19 May 2026
Nurse	275	2019	Recommendation on remuneration for digital healthcare services to healthcare providers.	Sveriges Kommuner och Regioner (SKR)[1] Last access: 19 May 2026
External primary care consultation				
Physical consultation	1452.35	2019/2020	Weighted average cost: 0.396×614 SEK + 0.604×2002 SEK = 1452.35 SEK. Cost is calculated based on: A doctor visit cost of 2002 SEK (national average for 2019–2020: 0.5×1838 SEK + 0.5×2166 SEK = 2002 SEK), a nurse visit cost of 614 SEK (Region Östergötland estimate), and weights of 39.6% nurse and 60.4% doctor consultations (own calculations from Region Scania data*).	Myndigheten för vård- och omsorgsanalys[2] Last access: 19 May 2026 Own calculation. Data from Region Scania.
Other healthcare				
Prescription	260	2019/2020	Average cost to the region over all prescriptions in 2019 or 2020 among patients in our analysis sample.	Own calculation Data from Socialstyrelsen.
Specialist visit	3930.68	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	3991.5	2019/2020	Average costs in Southern Sweden, (3963 SEK + 4020 SEK) $\times$ 0.5 = 3991.5 SEK.	Södra Sjukvårdsregionen[4] Last access: 5 Nov 2024
Urgent care center visit	2002	Costs from 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 2002 SEK.	Region Stockholm (1177)[5] Last access: 5 Nov 2024
Hospitalization	28487.5	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/
*Region Scania data refers to a sample of 327,462 observations restricted to primary care patient's initial and follow-up visits.

Table A4. Comparison of sample characteristics to the Swedish population

	Analysis sample	Swedish Population	Physical primary care in Scania
Demographic characteristics			
Patient Age	29.2	41.8	52.2
Toddler (2-4 years)	0.056	0.036	0.049
Child (5-12 years)	0.11	0.098	0.054
Teen (13-19 years)	0.12	0.081	0.041
Young adult (20-39 years)	0.46	0.27	0.17
Middle aged (40-59 years)	0.21	0.26	0.22
Senior (60+ years)	0.047	0.26	0.47
Female	0.63	0.50	0.55
Regions			
Middle Sweden	0.18	0.25	
Norrland	0.034	0.087	
Småland + the Islands	0.042	0.084	
South Sweden	0.095	0.15	
Stockholm	0.44	0.23	
West Sweden	0.20	0.20	
Socioeconomic characteristics			
Further Education	0.45	0.39	
Employed	0.56	0.67	
Any migration background	0.25	0.26	
Annual earnings (median, 1000s SEK)	294.7	370.8	
Married	0.29	0.42	
Clinician type			
Nurse consultation	0.23		0.48
Observations	495,555	10,327,589	2,953,269

Notes: This table compares average case characteristics in the analysis sample (N=495,555) to the Swedish population. Column (1) reports means of our sample. The means in Column (2) are the means for Sweden in 2019, and those in Column (3) are the means across all physical nurse and doctor consultations at primary care clinics in Scania in 2019, excluding the firm we study. "Further education" is reported for people aged 23 and above and indicates any post-secondary education, "Married" is reported for people aged 18 and above, and "Any migration" is an indicator for individuals who were either born outside Sweden (first generation migrants) or whose parents were both born outside Sweden (2nd generation migrant). "Annual earnings" includes annual earnings from wages and self-employment in SEK and is reported for individuals strictly over age 20.

Table A5. Exit rank on login rank regressions

	Exit rank in overlap group (1)	Exit rank in doctor group (2)	Exit rank in overlap and doctor group (3)	Exit rank in broader drop-in sample, singletons excluded (4)
Login rank	0.98	0.99	0.98	0.99
	(0.00067)	(0.0020)	(0.00088)	(0.0011)
Doctor group			11.0	8.90
			(0.14)	(0.16)
Nurse group				217.7
				(1.31)
Date-by-3 hour FEs	✓	✓	✓	✓
Observations	495,555	167,050	662,605	969,482
R-squared	1.00	1.00	1.00	1.00

Notes: This table reports estimates from regressions of exit rank on login rank among patients in the drop-in queue. Exit rank is defined as the rank when sorted by the meeting start time and login rank is the rank when sorted by patient's login time. In column (4), singletons are excluded by the regression, which explains why it contains one less observation than the broad drop-in sample in Table A2. All regressions use date-by-3-hour fixed effects. Robust standard errors are in parentheses.

Table A6. Differences in consultation characteristics of the highest and lowest congestion quartiles

	High congestion	Low congestion	T-test		Observations	
	Mean	Mean	Difference	p-val.	High congestion	Low congestion
Patient waiting time in min.	44.3	5.62	38.6	0.00	123,871	124,073
	(36.1)	(13.9)				
Nurse total time	10.8	11.3	-0.50	0.00	39,506	13,483
	(6.68)	(8.32)				
Doctor total time	11.7	11.8	-0.069	0.02	84,365	110,590
	(6.45)	(6.62)				
Rating: top score	0.38	0.40	-0.022	0.00	123,871	124,073
	(0.49)	(0.49)				
No show	0.00090	0.0013	-0.00039	0.00	123,871	124,073
	(0.030)	(0.036)				
Technical error	0.0046	0.0075	-0.0029	0.00	123,871	124,073
	(0.068)	(0.086)				
Consultation time below one minute	0.047	0.051	-0.0038	0.00	123,871	124,073
	(0.21)	(0.22)				
Consultation time below 30 seconds	0.015	0.016	-0.00051	0.30	123,871	124,073
	(0.12)	(0.13)				

Notes: This table presents summary statistics for various consultation characteristics during low and high congestion. For low (high) congestion, we consider the lower (upper) quartile of the congestion instrument, *Log median doctor wait time*. First and second columns show the means and standard deviations in parenthesis. The third and fourth columns perform t-tests on the difference in means between consultation characteristics at low and high congestion. Last two columns show the number of observations in each test group.

Table A7. Reduced-form regressions in the doctor group

Panel A: Within-firm outcomes

	Any clinician-booked doctor visit (7 days)	Any unbooked revisit (7 days)	Top score rating (final)	Informative diagnosis (final)
Log median doctor wait time	-0.00016	-0.00011	-0.0035	-0.00051
	(0.00051)	(0.00099)	(0.0023)	(0.0016)
X_1 and X_2	✓	✓	✓	✓
Observations	167,050	167,050	167,050	167,050
Baseline mean	0.012	0.045	0.46	0.79

Panel B: Downstream healthcare use

	Any new prescription (7 days)	Any external PCP consultation (7 days)	Any urgent care (7 days)	Any specialist care (7 days)
Log median doctor wait time	-0.0028	-0.00058	0.00028	0.00015
	(0.0021)	(0.0012)	(0.00025)	(0.00074)
X_1 and X_2	✓	✓	✓	✓
Observations	167,050	167,050	167,050	167,050
Baseline mean	0.54	0.065	0.0026	0.027

	Any ED (7 days)	Any hospitalization (7 days)
Log median doctor wait time	-0.00052	-0.00032
	(0.00056)	(0.00024)
X_1 and X_2	✓	✓
Observations	167,050	167,050
Baseline mean	0.014	0.0028

Panel C: Long-term consequences

	Earnings decrease $>$ 20 pct. (cal. month after)	Zero earnings (cal. month after)	Death (3 years)
Log median doctor wait time	-0.000067	0.0040	0.00014
	(0.0031)	(0.0017)	(0.00019)
X_1 and X_2	✓	✓	✓
Observations	67,226	67,226	167,050
Baseline mean	0.19	0.049	0.0016

Panel D: Cost outcomes

	Taxpayer payments to the firm (within-day)	Taxpayer payments to the firm (7 days)	Firm net revenue (within-day)	Firm production costs
Log median doctor wait time	-0.0060	-0.48	-0.0090	0.0030
	(0.021)	(0.61)	(0.013)	(0.020)
X_1 and X_2	✓	✓	✓	✓
Observations	167,050	167,050	167,050	167,050
Baseline mean	500	538.4	125	375

	Total taxpayer payments	Systemwide production costs
Log median doctor wait time	-11.7	-11.7
	(8.88)	(8.88)
X_1 and X_2	✓	✓
Observations	167,050	167,050
Baseline mean	979.2	854.2

Note: This table presents regression results of case outcomes on doctor congestion for the doctor group, which consists of consultations in symptom categories for which at most 1% of cases are initially assigned to a nurse (see Appendix Table A2). Each subpanel corresponds to a different outcome. All outcomes are defined in Table B1. All earnings measures are defined for a subsample with stable earnings of employed patients earning more than 3,533 SEK (2010 values) per month in each of the three prior months, based on an annual threshold of 42,400 SEK (Saez, Schoefer, and Seim 2019). The coefficient reports the association between the outcome and the congestion measure, *Log median doctor wait time*, defined as the winsorized natural logarithm of the median wait time among the up to 5 prior doctor consultations within 30 minutes. All regressions control for baseline characteristics (X_1) and additional controls (X_2) as described in the main text. The baseline mean is the average of the dependent variable in the doctor group. Robust standard errors are in parentheses.

Table A8. Robustness checks: Systemwide production costs

Panel A: Alternative controls and unobserved selection

	A1. Controlling for X_1 only	A2. Full controls with date-by-1 hour windows	A3. Full controls with symptom controls $\times$ 3-hour windows $\times$ weekday $\times$ quarter	A4. Full controls including congestion quartiles	A5. Full controls including patient fixed effects	A6. Oster bounds
Initially to nurse	-6.76	-0.76	1.04	-4.53	9.38	-4.84
	(13.3)	(13.4)	(13.3)	(13.2)	(23.4)	(13.2)
X_1	✓					
Date-by-1 hour FEs		✓				
Congestion quartiles				✓		
Patient FEs					✓	
X_1 and X_2		✓	✓	✓	✓	✓
Observations	495,555	494,819	495,485	495,555	228,694	495,555
Bias-adjusted β						-31.3
Baseline mean	1261.8	1261.0	1261.8	1261.8	1283.4	1261.8

Panel B: Alternative overlap sample definitions (B1-B3) and differential attrition (B4-B6)

	B1. Other health inquires	B2. Share initially assigned to nurse: 10–95%	B3. Share initially assigned to nurse: 20–95%	B4. Lee bounds	B5. Cases with past nurse consultations	B6. Difference in assignment timing (April-June vs. July-September 2019)
Initially to nurse $\times$ second period						62.4
						(65.4)
Initially to nurse	5.43	-4.63	-57.9		-48.5	88.5
	(19.1)	(14.3)	(17.8)		(39.6)	(45.1)
Lower bound				-262.5		
Upper bound				258.1		
Symptom controls				✓		
X_1 and X_2	✓	✓	✓		✓	✓
Observations	173,565	396,845	271,852	495,555	52,107	104,138
Baseline mean	1296.1	1296.7	1448.0	1261.8	1374.5	1166.0

Panel C: Cost calculations

	C1. Stockholm & Scania sample	C2. Systemwide production costs (30 days)	C3. Aggregate systemwide production costs (7 days)	C4. Aggregate systemwide production costs (30 days)	C5. Differential frontline worker cost
Initially to nurse	-0.84	-5.89	-3.74	-17.2	-22.3
	(18.7)	(17.0)	(16.6)	(26.1)	(13.2)
X_1 and X_2	✓	✓	✓	✓	✓
Observations	262,144	495,555	495,555	495,555	495,555
Baseline mean	1349.5	2024.8	1495.2	2668.4	1261.8

Panel D: External validity

	D1. Weighted to match in-person primary care	D2. Excluding COVID-19 symptom	D3. Pre COVID-19 onset + interlude period	D4. First and second COVID-19 wave	D5. Difference between 2 months before and 2 months after COVID-19 onset
Initially to nurse $\times$ after onset					-5.77
					(51.2)
Initially to nurse	35.8	28.8	25.7	-23.3	85.1
	(56.0)	(13.5)	(18.0)	(20.1)	(42.6)
X_1 and X_2	✓	✓	✓	✓	✓
Observations	495,555	466,724	296,647	197,806	125,790
Baseline mean	1261.8	1253.2	1238.4	1298.0	1281.8

Notes: This table presents robustness checks for the effect of initial case assignment to a nurse on systemwide production costs. Each cell shows the OLS estimate for *Initially to nurse*, under alternative model specifications. In columns B6 and D5, *Initially to nurse* is interacted with an indicator for the later period, so that the row *Initially to nurse* reports the effect in the earlier period and the interaction row the difference between the two. Baseline controls (X_1) include symptom categories fully interacted with demographics and with prior provider use, and region. Additional controls (X_2) include health risk and socioeconomic variables. Patient fixed effects are included where indicated. The baseline mean is the average of the dependent variable for cases directly assigned to a doctor. Robust standard errors are in parentheses.

Table A9. Instrumental variables robustness: Effects of the knowledge hierarchy on healthcare and cost outcomes

Panel A: Within-firm outcomes

	Any clinician-booked doctor visit (7 days)		Any unbooked revisit (7 days)		Top score rating (final)		Informative diagnosis (final)	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Initially to nurse	0.31	0.25	0.0078	-0.018	0.059	0.054	-0.093	-0.032
	(0.0015)	(0.010)	(0.00093)	(0.011)	(0.0020)	(0.025)	(0.0017)	(0.019)
X_1 and X_2	✓	✓	✓	✓	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555	472,275	472,275	495,555	495,555
First-stage K-P F-statistic		3,726		3,726		3,397		3,726
Baseline mean	0.014	0.014	0.058	0.058	0.40	0.40	0.79	0.79

Panel B: Downstream healthcare use

	Any new prescription (7 days)		Any external PCP consultation (7 days)		Any urgent care (7 days)		Any specialist care (7 days)	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Initially to nurse	-0.10	-0.026	0.025	-0.015	0.0031	0.000017	0.0032	-0.010
	(0.0017)	(0.022)	(0.0014)	(0.017)	(0.00051)	(0.0060)	(0.00073)	(0.0088)
X_1 and X_2	✓	✓	✓	✓	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555	495,555	495,555	495,555	495,555
First-stage K-P F-statistic		3,726		3,726		3,726		3,726
Baseline mean	0.44	0.44	0.16	0.16	0.016	0.016	0.035	0.035

	Any ED (7 days)		Any hospitalization (7 days)	
	OLS	IV	OLS	IV
Initially to nurse	0.0049	-0.0068	-0.00049	0.00095
	(0.00074)	(0.0085)	(0.00038)	(0.0044)
X_1 and X_2	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555
First-stage K-P F-statistic		3,726		3,726
Baseline mean	0.030	0.030	0.0082	0.0082

Panel C: Long-term consequences

	Earnings decrease > 20pct. (cal. month after)		Zero earnings (cal. month after)		Death (3 years)	
	OLS	IV	OLS	IV	OLS	IV
Initially to nurse	-0.021	0.0010	-0.0031	0.0059	-0.000013	-0.0045
	(0.0024)	(0.030)	(0.0013)	(0.016)	(0.00018)	(0.0022)
X_1 and X_2	✓	✓	✓	✓	✓	✓
Observations	214,674	214,674	214,674	214,674	495,555	495,555
First-stage K-P F-statistic		1,492		1,492		3,726
Baseline mean	0.22	0.22	0.050	0.050	0.0021	0.0021

Panel D: Cost outcomes

	Taxpayer payments to the firm (within-day)		Taxpayer payments to the firm (7 days)		Firm net revenue (withing-day)		Firm production costs	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Initially to nurse	-147.9	-161.5	-149.1	-172.0	-107.6	-98.6	-40.3	-62.9
	(0.33)	(2.02)	(0.63)	(6.81)	(0.22)	(1.34)	(0.55)	(3.36)
X_1 and X_2	✓	✓	✓	✓	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555	495,555	495,555	495,555	495,555
First-stage K-P F-statistic		3,726		3,726		3,726		3,726
Baseline mean	500	500	544.6	544.6	125	125	375	375

	Total taxpayer payments		Systemwide production costs	
	OLS	IV	OLS	IV
Initially to nurse	-113.4	-229.8	-5.74	-131.2
	(13.1)	(153.6)	(13.1)	(153.7)
X_1 and X_2	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555
First-stage K-P F-statistic		3,726		3,726
Baseline mean	1374.0	1374.0	1249.0	1249.0

Note: This table presents OLS and instrumental variable estimates for the effects of initial case assignment to a nurse (the *knowledge hierarchy*) rather than directly to a doctor. Each subpanel corresponds to a different dependent variable. Outcomes are defined as in Table B1. "Top score rating (final)" indicates whether the patient gave a top satisfaction rating (5 out of 5) at the final within-firm consultation — the clinician-booked follow-up where there was one, and the initial consultation otherwise; "Informative diagnosis (final)" indicates whether the patient receives a non-missing diagnosis excluding ICD-10 chapters R (symptoms, e.g., cough, rash) and Z (factors influencing health status and health service contacts) at that same consultation. The left column in each subpanel presents the OLS estimate for *Nurse* based on Equation 4. The right column provides the IV estimate using *Log median doctor wait time* as an instrument for *Nurse*, estimated via Two-Stage Least Squares (2SLS). *Log median doctor wait time* is defined as the winsorized natural logarithm of the median wait time among the up to 5 prior doctor consultations within 30 minutes. All regressions control for X_1 and X_2 as described in the main text. The baseline mean is the average of the dependent variable for cases directly assigned to a doctor. The K-P F-statistic is the Kleibergen-Paap F-statistic. Robust standard errors are in parentheses.

Table A10. Mean characteristics of compliers

	Complier mean	Sample mean	Complier mean minus sample mean	Equality of means p-value
Symptoms				
Abdominal pain	0.049	0.030	0.019	0.00
Cold and flu	0.028	0.089	-0.061	0.00
Cold sores	0.0074	0.025	-0.018	0.00
Constipation	0.0053	0.0070	-0.0017	0.27
COVID-19	0.10	0.058	0.044	0.00
Diarrhea or vomiting	0.025	0.021	0.0041	0.17
Eye infection	0.011	0.075	-0.064	0.00
Fever	0.041	0.029	0.012	0.00
Headache	0.032	0.023	0.0086	0.01
Nail problem	0.013	0.022	-0.0094	0.00
Other health inquiries	0.60	0.35	0.25	0.00
Bites and stings	0.032	0.054	-0.022	0.00
Sinusitis	0.014	0.031	-0.017	0.00
Sore throat	0.0081	0.076	-0.068	0.00
Uncategorized	0.030	0.030	0.00077	0.82
Urinary tract infection	0.0047	0.079	-0.074	0.00
Demographics				
Female	0.59	0.63	-0.041	0.00
Patient age	28.8	29.2	-0.35	0.24
West Sweden	0.21	0.20	0.014	0.07
Stockholm	0.42	0.44	-0.021	0.02
Middle Sweden	0.19	0.18	0.0013	0.86
South Sweden	0.099	0.095	0.0042	0.46
Norrland	0.029	0.034	-0.0054	0.11
Småland + the islands	0.047	0.042	0.0051	0.18
Health risk				
Any prior hospitalization (3 years)	0.13	0.13	-0.0072	0.26
Any prior ED (3 years)	0.41	0.41	0.0027	0.77
Any prior urgent care (3 years)	0.13	0.13	-0.0026	0.67
Any prior specialist care (3 years)	0.62	0.64	-0.019	0.04
Any comorbidity (3 years)	0.24	0.26	-0.016	0.05
Past provider use				
Last provider consultation with doctor	0.39	0.38	0.0067	0.45
Prior provider prescription	0.32	0.35	-0.030	0.00
Socio-economic variables				
Earnings above median	0.31	0.32	-0.018	0.04
Any benefits	0.14	0.14	-0.0037	0.57
Schooling < 9 years	0.054	0.049	0.0052	0.19
Middle/High school	0.21	0.23	-0.022	0.00
Further educ. < 3 years	0.11	0.11	0.0035	0.53
Further educ. ≥ 3 years	0.18	0.19	-0.0075	0.30
Employed	0.53	0.56	-0.034	0.00
Self-employed	0.040	0.041	-0.0013	0.72
Unemployed	0.054	0.050	0.0040	0.34
Married	0.21	0.23	-0.016	0.03
Divorced/Widowed	0.076	0.074	0.0011	0.81
Unmarried	0.42	0.42	-0.0027	0.77
Not migrated	0.71	0.75	-0.035	0.00
Immigrant 1st gen	0.18	0.16	0.021	0.00
Immigrant 2nd gen	0.11	0.096	0.013	0.03
Socioeconomics missing	0.46	0.43	0.024	0.01
Covid periods				
During Covid periods (First or Second Wave)	0.51	0.40	0.11	0.00
Clock hours				
0-3	0.00025	0.012	-0.012	0.00
3-6	0.0078	0.011	-0.0033	0.11
6-9	0.13	0.13	-0.00052	0.93
9-12	0.25	0.25	0.00022	0.98
12-15	0.17	0.20	-0.027	0.00
15-18	0.12	0.17	-0.054	0.00
18-21	0.20	0.16	0.038	0.00
21-24	0.12	0.066	0.059	0.00

Note: This table reports average characteristics for the overall analysis sample and the population of compliers. The first column shows the estimated average of a case characteristic (x_i) in the complier population. The second column reprints the overall sample mean. The third column reports the difference between the complier mean and sample mean. The fourth column reports the p-value from a test of the null hypothesis that the complier mean equals the sample mean. We follow the procedure described in Frandsen, Lefgren, and Leslie (2023) based on Abadie (2003) and report the mean characteristic among compliers as the estimate of $\frac{E[\omega_i X_i]}{E[\omega_i]}$, where ω_i is the weight given to case i by the IV. We estimate complier characteristics using Two-Stage Least Squares (2SLS) regressions of the interaction between a predetermined case characteristic and treatment *Initial assignment to nurse* ($x_i \times Nurse_i$) on the treatment ($Nurse_i$), instrumented by our congestion instrument *log median doctor wait time* and the baseline time controls (login date-by-3-hour fixed effects).

Table A11. Robustness checks on additional outcomes

Panel A: Informative diagnosis

	Main specification	Controlling for X_1 only	Oster bounds	Cases with past nurse consultations	Difference in assignment timing (April-June vs. July-September 2019)
Initially to nurse × second period					0.026
					(0.0073)
Initially to nurse	-0.093	-0.094	-0.093	-0.11	-0.059
	(0.0017)	(0.0017)	(0.0017)	(0.0053)	(0.0052)
Bias-adjusted β			-0.068		
X_1 and X_2	✓		✓	✓	✓
Observations	495,555	495,555	495,555	52,107	104,138
Baseline mean	0.79	0.79	0.79	0.74	0.81

	Excluding COVID-19 symptom	Pre COVID-19 onset + interlude period	First and second COVID-19 wave	Difference between 2 months before and 2 months after COVID-19 onset
Initially to nurse × after onset				-0.0057
				(0.0067)
Initially to nurse	-0.08	-0.086	-0.1	-0.09
	(0.0017)	(0.0023)	(0.0025)	(0.0057)
X_1 and X_2	✓	✓	✓	✓
Observations	466,724	296,647	197,806	125,790
Baseline mean	0.79	0.79	0.78	0.77

Panel B: Any new prescription

	Main specification	Controlling for X_1 only	Oster bounds	Cases with past nurse consultations	Difference in assignment timing (April-June vs. July-September 2019)
Initially to nurse × second period					0.012
					(0.0084)
Initially to nurse	-0.102	-0.102	-0.102	-0.0861	-0.0515
	(0.0017)	(0.0017)	(0.0017)	(0.0052)	(0.0062)
Bias-adjusted β			-0.074		
X_1 and X_2	✓		✓	✓	✓
Observations	495,555	495,555	495,555	52,107	104,138
Baseline mean	0.44	0.44	0.44	0.37	0.44

	Excluding COVID-19 symptom	Pre COVID-19 onset + interlude period	First and second COVID-19 wave	Difference between 2 months before and 2 months after COVID-19 onset
Initially to nurse × after onset				0.0337
				(0.0067)
Initially to nurse	-0.105	-0.097	-0.109	-0.126
	(0.0018)	(0.0024)	(0.0025)	(0.0057)
X_1 and X_2	✓	✓	✓	✓
Observations	466,724	296,647	197,806	125,790
Baseline mean	0.45	0.45	0.42	0.43

Panel C: Any ED

	Main specification	Controlling for X_1 only	Oster bounds	Cases with past nurse consultations	Difference in assignment timing (April-June vs. July-September 2019)
Initially to nurse × second period					-0.0077
					(0.004)
Initially to nurse	0.0056	0.0056	0.0056	0.0009	0.009
	(0.00081)	(0.00081)	(0.00081)	(0.0026)	(0.0028)
Bias-adjusted β			0.0029		
X_1 and X_2	✓		✓	✓	✓
Observations	495,555	495,555	495,555	52,107	104,138
Baseline mean	0.037	0.037	0.037	0.048	0.035

	Excluding COVID-19 symptom	Pre COVID-19 onset + interlude period	First and second COVID-19 wave	Difference between 2 months before and 2 months after COVID-19 onset
Initially to nurse × after onset				-0.0069
				(0.0032)
Initially to nurse	0.0076	0.0069	0.0049	0.016
	(0.00084)	(0.0011)	(0.0012)	(0.0027)
X_1 and X_2	✓	✓	✓	✓
Observations	466,724	296,647	197,806	125,790
Baseline mean	0.036	0.036	0.038	0.035

Panel D: Any hospitalization

	Main specification	Controlling for X_1 only	Oster bounds	Cases with past nurse consultations	Difference in assignment timing (April-June vs. July-September 2019)
Initially to nurse × second period					0.0023
					(0.0019)
Initially to nurse	-0.00048	-0.0005	-0.00048	-0.00095	0.00075
	(0.00038)	(0.00038)	(0.00038)	(0.0011)	(0.0013)
Bias-adjusted β			-0.0014		
X_1 and X_2	✓		✓	✓	✓
Observations	495,555	495,555	495,555	52,107	104,138
Baseline mean	0.0082	0.0082	0.0082	0.0092	0.0057

	Excluding COVID-19 symptom	Pre COVID-19 onset + interlude period	First and second COVID-19 wave	Difference between 2 months before and 2 months after COVID-19 onset
Initially to nurse × after onset				-0.00078
				(0.0015)
Initially to nurse	0.00016	0.00042	-0.00091	0.0013
	(0.00038)	(0.00051)	(0.00058)	(0.0012)
X_1 and X_2	✓	✓	✓	✓
Observations	466,724	296,647	197,806	125,790
Baseline mean	0.0079	0.0072	0.0099	0.0089

Notes: This table presents robustness checks for the effect of initial case assignment to a nurse. Each panel corresponds to a different dependent variable. Columns report alternative specifications and samples. In the columns reporting differences across periods, *Initially to nurse* is interacted with an indicator for the later period, so that the row *Initially to nurse* reports the effect in the earlier period and the interaction row the difference between the two. The column headed *Controlling for X_1 only* includes only X_1 controls; all other columns include X_1 and X_2 controls unless noted. The baseline mean is the average of the dependent variable for cases directly assigned to a doctor. Robust standard errors are in parentheses.

Table A12. Ten most common diagnoses in the analysis sample, compared to in-person care consultations of the Scania subsample

ICD	Description	Share of consultations (%) Analysis Sample	Share of consultations (%) In-person care in Scania	Share handled by nurses (%) Analysis sample	Share handled by nurses (%) In-person care in Scania
J06	Acute upper respiratory infections of multiple and unspecified sites	10.4	4.05	15.0	10.7
N30	Cystitis	5.25	1.93	5.34	9.22
J03	Acute tonsillitis	4.51	1.67	16.2	8.69
T14	Injury of unspecified body region	4.33	0.075	24.2	27.3
Z71	Other counseling/medical advice, not elsewhere classified	4.25	1.96	46.6	39.8
H10	Conjunctivitis	3.23	0.76	14.1	20.3
Z53	Persons encountering health services for specific procedures, not carried out	2.69	0.033	16.2	1.80
R52	Pain, not elsewhere classified	2.69	3.08	43.7	25.4
N39	Other disorders of urinary system	2.59	0.52	19.0	13.6
J01	Acute sinusitis	2.35	0.52	4.23	4.71

Note: This table presents the ten most common ICD-10 codes in our main analysis sample. Column 1 reports the ICD-10 codes and Column 2 reports their descriptions. Columns 3 and 4 show how frequently these codes appear in our sample and in in-person primary care visits in the Scania region, respectively. The final two columns show the share of cases handled by a nurse in each sample.

Table A13. Robustness: Percentage change in systemwide production costs

	$\frac{\Delta \text{ Systemwide production costs}}{\text{Systemwide production costs}}$	$\frac{\Delta \text{ Systemwide production costs}}{\text{Firm production costs}}$
1) Main specification	-1.94%	-6.81%
	(0.19)	(0.67)
2) Aggregate systemwide production costs (7 days)	-2.05%	-8.27%
3) Aggregate systemwide production costs (30 days)	-1.50%	-8.57%
4) Sample restricted to Stockholm and Scania	-1.85%	-6.93%
5) Split sample using even and odd days	-1.94%	-6.79%
6) Significant symptoms	-1.76%	-6.19%
7) Relative systemwide cost advantage	-1.87%	-6.54%
8) Excluding the COVID-19 symptom	-1.85%	-6.47%
9) Additional reassignment cost (5 percentage points)	-1.76%	-6.17%

Note: This table shows the percentage change in outcomes from counterfactual assignments relative to the actual assignment for different systemwide production cost specifications. Standard errors are bootstrapped 100 times and reported only for the main specification due to computational constraints. The two columns differ in their denominator: Column 1 normalises the change in systemwide production costs by total systemwide production costs, while Column 2 normalises by firm production costs. Row 1 (main specification) uses 7-day systemwide costs for Column 1 and initial-day firm production costs for Column 2. Rows 2 and 3 use aggregated costs for both columns over 7-day and 30-day horizons, respectively. Row 4 restricts the sample to patients in the Stockholm and Scania regions. Row 5 estimates the comparative advantage of symptoms using odd-numbered calendar days, defined within each month so that the count resets from 1 to 31, and then applies the resulting counterfactual assignment rule to even-numbered calendar days, thereby providing an out-of-sample assignment rule. Row 6 restricts the counterfactual reassignment to symptoms for which nurses and doctors differ in costs at the 5 percent significance level. Row 7 uses the relative systemwide cost advantage of nurses compared to doctors instead of the net systemwide cost advantage. Row 8 restricts the sample to patients who do not report COVID-19 as the symptom. Row 9 increases the cost of cases reassigned to nurses under the Row 1 specification by accounting for the expected cost of a five-percentage-point increase in the probability that nurses pass these cases to doctors.

Table B1. Description of key case characteristics and outcomes

Variable	Description
Baseline controls (X_1)	
Age	Set of indicator variables for patient age, defined for each integer year from 2 to 69, with one combined indicator for ages 70 and above. (Source: Provider)
Age group	Set of indicator variables for the patient's age group, as one of the following categories: Toddler (2-4 years); Child (5-12 years); Teen (13-19 years); Young adult (20-39 years); Middle aged (40-59 years); Senior (60+ years). (Source: Provider)
Female	Indicates whether the patient is female. (Source: Provider)
Patient's region	A set of indicator variables for the patient's regional area of residence, as reported by national agencies in November of the year prior to the initial consultation. We consider six regional areas based on Sweden's national areas (riksområden): West Sweden, Stockholm, Middle Sweden (combining East Middle Sweden and West Middle Sweden), South Sweden, Norrland (combining Middle Norrland and Upper Norrland), and Småland and the islands. (Source: SCB)
Time fixed effects	Fixed effects for the login time of a case when a consultation is requested. Login time is constructed as calendar date-by-3 hour window: 0 am - 3 am; 3 am - 6 am; 6 am - 9 am; 9 am - 12 pm; 12 pm - 15 pm; 15 pm - 18 pm; 18 pm - 21 pm; 21 pm - 24 pm of a given date, such as April 1, 2020, 12 am - 3 am. (Source: Provider)
Symptom category	Set of indicator variables for a case's main symptom, as indicated by the patient before the consultation. The full list of symptoms is shown in the Table 1. (Source: Provider)
Last visit within the firm was with a doctor	Indicates whether the patient's most recent consultation with the firm, prior to the initial consultation, was initially assigned to a doctor; coded 0 if no prior consultation exists. (Source: Provider)
Prior prescription from the firm	Indicates whether a patient has filled any prescription from the firm prior within the 3 years prior to the initial consultation, excluding visits in the 2 days immediately preceding it. (Source: NBHW)
Treatment and Congestion	
Doctor congestion	Log median waiting time of the previous at most 5 doctor consultations within 30 minutes of login time of the patient, winsorized at the 99th percentile, missing values set to the smallest waiting time. (Source: Provider)
Initially to nurse	Indicates whether the initial consultation of a case is assigned to a nurse, with value one; and zero if the case is directly assigned to a doctor. (Source: Provider)
Outcomes: Quality and downstream care measures	
Any clinician-booked doctor visit (7 days)	Indicates whether the patient is directed by a nurse or a doctor to a doctor working with the provider within 7 days after their initial consultation. (Source: Provider)
Any ED (7 days)	Indicates any emergency department visit within 7 days of the initial consultation, excluding visits with ICD codes in category O related to childbirth. (Source: NBHW)
Any external PCP consultation (7 days)	Indicates whether the patient visited a primary care provider outside the firm within 7 days. For patients registered in Stockholm or Scania, visits are observed directly; for others, external visits are inferred from any external primary care prescriptions within 7 days. (Source: NBHW)
Any hospitalization (7 days)	Indicates any hospitalization within 7 days of the initial consultation, excluding ICD codes in category O related to childbirth. (Source: NBHW)
Any urgent care (7 days)	Indicates any urgent care center visit within 7 days of the initial consultation, excluding visits with ICD codes in category O related to childbirth. (Source: NBHW)
Any new prescription (7 days)	Indicates whether the patient fills a new prescription within 7 days of the initial consultation. A prescription is defined as "new" if the same ATC code (at the 4 digit level) has not been filled in the previous 6 months. (Source: NBHW)
Any specialist care (7 days)	Indicates any specialist outpatient care (excluding ED) within 7 days of the initial consultation, excluding visits with ICD codes in category O related to childbirth. (Source: NBHW)
Any unbooked revisit (7 days)	Indicates whether the patient has an unscheduled return visit with the same symptom category within 7 days. (Source: Provider)
Clinician total time in min. (7 days)	Total time spent on the patient by the clinicians, including the initial consultation, subsequent doctor and nurse consultations and administrative time, recorded in minutes. (Source: Provider)
Death (3 years)	Indicates whether the patient passed away within 3 years of the initial consultation(Source: NBHW)
Doctor total time in min. (7 days)	Total time doctors spend on the patient, including the initial doctor assignment (if any), follow-up consultations, and administrative tasks. (Source: Provider)

Table B1. Description of key case characteristics (Continued)

Variables	Descriptions
Informative diagnosis (final)	Indicates whether the patient receives a non-missing diagnosis belonging to ICD-10 chapters other than R (symptoms, e.g., cough, rash) and Z (factors influencing health status and health service contacts) at the final within-firm consultation, taken from the clinician-booked follow-up where there was one and from the initial (nurse) consultation otherwise. (Source: Provider)
Earnings decrease > 20 pct. (cal. month after)	Indicator for whether the patient's earnings in the given calendar month are more than 20% below their average monthly earnings over the three calendar months immediately preceding the initial consultation. Defined only for patients with stable pre-meeting income (earnings above the minimum income threshold 3,533 SEK (2010 values) in each of the three months prior, based on the annual threshold of 42,400 SEK). (Source: SCB)
Nurse total time in min. (7 days)	Total time spent on the patient by the nurses, including the initial consultation, any subsequent nurse consultation and administrative time. (Source: Provider)
Rating: top score (final)	Indicates whether the patient gave a top satisfaction rating (5 out of 5) at the final within-firm consultation, taken from the clinician-booked follow-up where there was one and from the initial (nurse) consultation otherwise. (Source: Provider)
Zero earning (cal. month after)	Indicates whether the patient had no earnings from their main income source in the calendar month following the initial consultation. Defined for a subsample with stable earnings consisting of employed patients who earn more than 3,533 SEK (2010 values) in each of the three months prior, based on an annual threshold of 42,400 SEK (Saez, Schoefer, and Seim 2019). (Source: SCB)
Outcomes: Cost measures	
Systemwide production costs	The total costs of healthcare provision within 7 days after the initial assignment, calculated as total taxpayer payment minus the firm's net revenues created using the indicator approach. 30 days in parenthesis refer to the version that counts within 30 days healthcare provisions(Source: NBHW)
Aggregate systemwide production costs	The systemwide production costs incurred within 7 days after the initial assignment, calculated as total costs minus the firm's net revenues created using all the healthcare events within 7 days, rather than the indicator approach. (Source: NBHW)
Total taxpayer payments	The total taxpayer cost of healthcare services incurred within 7 days of the initial consultation. This includes costs at the firm on the calendar day of the initial consultation and the cost of up to one occurrence of each downstream event within 7 days, including an emergency department visit, specialist visit, urgent care visit, hospitalization, or prescription. (Source: NBHW)
Taxpayer payments to the firm	The taxpayer's incurred costs for consultations at the firm, including up to one consultation within-day (per reimbursement rules) and up to one additional follow-up consultation within 7 days of the initial assignment (per the indicator approach discussed in 6.2). Within the week, two consultations may occur if both a nurse and a doctor are involved; otherwise, at most one consultation is included. (Source: Provider)
Taxpayer payments to the firm (within-day)	The taxpayer's incurred costs for within-day consultations at the firm, covering up to one consultation (per reimbursement rules). (Source: Provider)
Downstream costs	The taxpayer's incurred costs of total healthcare provision within 7 days of the initial assignment, covering only downstream, out of firm spending. These include costs of up to one occurrence of each downstream event (emergency department visit, specialist or urgent care visit, hospitalization, or prescription) within 7 days. (Source: NBHW)
Firm net revenue	The firm's net revenue from this case on the calendar day of the initial assignment, defined as gross revenues minus costs of provision. Gross revenues are earned for up to one consultation per day (per reimbursement rules). However, costs would be incurred for all consultations. Costs are measured as within-day taxpayer payments of the initial and potential follow-up consultations, multiplied by multiplied by $(1-\theta)$, where θ denotes the firm's markdown. (Source: Provider)
Firm production costs	Consultations that are not reimbursed are treated as costs to the firm. Costs are measured as within-day taxpayer costs of the initial consultations, multiplied by (1 - θ), where θ denotes the firm's markdown. (Source: Provider)
Additional controls (X_2)	
Education level	Set of indicator variables for the highest educational degree reported in Nov 2018, as one of the following categories: schooling below 9 years; middle/high school; secondary education < 3 years; secondary education $\geq$ 3 years. Educational degrees are reported for patients $\geq$ age 25. (Source: SCB)

Table B1. Description of key case characteristics (Continued)

Variables	Descriptions
Employment Status	Set of indicator variables for the main employment type reported in Nov 2018, as one of the following categories: employed; self-employed; unemployed. Employment status is reported for patients aged 20 to 67. (Source: SCB)
Civil status	Set of indicator variables for the patient's civil status reported in Nov 2018, as one of the following categories: married; divorced or widowed; unmarried. Civil status is reported for patients > age 18. (Source: SCB)
Any comorbidity	Indicates an Elixhauser comorbidity index of 1 or higher pre-2019, as diagnosed within specialty or hospital care within the 3 years prior to the initial consultation, excluding visits in the 2 days immediately preceding it. The Elixhauser list of comorbidities is adapted from 4-digit to 3-digit ICD codes to fit the number of digits in the reporting from Socialstyrelsen. (Source: NBHW)
Any prior ED	Indicates any ED visits within the 3 years prior to the initial consultation, excluding visits in the 2 days immediately preceding it. Excludes visits with ICD codes in category O related to childbirth. (Source: NBHW)
Any prior hospitalization	Indicates any inpatient hospitalization within the 3 years prior to the initial consultation, excluding visits in the 2 days immediately preceding it. Excludes visits with ICD codes in category O related to childbirth. (Source: NBHW)
Any prior specialist care	Indicator for any specialist visits within the 3 years prior to the initial consultation, excluding visits in the 2 days immediately preceding it. Excludes visits with ICD codes in category O related to childbirth. (Source: NBHW)
Any prior urgent care	Indicates any urgent care center visit within the 3 years prior to the initial consultation, excluding visits in the 2 days immediately preceding it. Excludes visits with ICD codes in category O related to childbirth. (Source: NBHW)
Any benefits	Indicates receipt of at least one social insurance or housing related benefit in 2018, including youth disability insurance, disability insurance, sickness benefits, rehabilitation benefits, or housing allowance. Benefits are reported for individuals aged 20 or older in November 2018. (Source: SCB)
Earnings above median	Indicates whether the patient's annual earning in November 2018 is greater than the in-sample median of non-missing earnings (293,300 SEK (approximately 31,900 USD using the 2020 average exchange rate: 1 SEK = 0.1087 USD in 2020 (Sveriges Riksbank 2020)). Earnings is reported for patients aged 20 or older. (Source: SCB)
Migration status	Set of indicator variables for the migration status of the patient, as one of the following categories: Not migrated; 1st generation immigrant (born outside of Sweden); 2nd generation immigrant (born in Sweden to two foreign-born parents).
Socioeconomics missing	Indicates that any socio-economic variables (education level, civil status, migration status, earnings, or benefits) contain missing values. (Source: SCB)
Additional outcomes	
Any new ext. PCP prescription (7 days)	Indicates whether the patient fills a new prescription written by a primary care provider (PCP) external ("ext.") to the online firm. A prescription is defined as "new" if the same ATC code (at the 4 digit level) has not been filled in the previous 6 months. Primary care providers are defined by Socialstyrelsen. (Source: NBHW)
No show	Indicates whether the initial consultation was recorded as a "no show". (Source: Provider)
Technical error	Indicates whether the initial consultation had a technical error. (Source: Provider)
Any avoidable hospitalizations (7 days)	Indicates whether the patient had an avoidable hospitalization in the 7 days following the initial consultation with the provider. We classify a hospitalization as avoidable if its ICD-10 diagnosis code is listed in Table A1 of Page et al. (2007)
Case characteristics used in robustness	
COVID-19 periods	Set of indicator variables classifying COVID-19 spread into four periods: pre COVID-19 (before 1 March 2020), first wave (1 March–5 July 2020), summer lull (6 July–23 October 2020), and second wave (24 October 2020 onward), based on Folkhälsomyndigheten (2020) (Source: The Public Health Agency of Sweden)

A Additional institutional details

Cost-effectiveness. The firm's choice of a clinician type (such as nurse or doctor) for a case is guided by cost-effectiveness. The regional authorities determine payments to the firm according to levels proposed by the Swedish Association of Local Authorities and Regions, which are intended to reflect the firm's production costs for its services (Sveriges Kommuner och Regioner (SKR) 2019). Cost-effectiveness is formalized as a requirement to deliver care at the 'lowest effective care level' (referred to as the *LEON* principle, based on the Swedish *Lägsta Effektiva Omhändertagande-Nivå*), where nurses are lower cost than doctors, and thus represent the lower level of care when clinically appropriate (Socialdepartementet 2016). Similar principles exist elsewhere, e.g., in Norway, the Netherlands, or England (NHS England 2026; Ministry of Health, Welfare and Sport 2018; Royal Norwegian Ministry of Health and Care Services 2006). The firm is only paid for one consultation with the highest-level clinician type per day, so it is not paid for additional same-day consultations (we refer to this as the 'resolving-clinician' payment rule).[35]

Publicly funded providers must also assess patients before contact with a clinician to ensure that care is delivered at the appropriate level (Sveriges Kommuner och Regioner (SKR) 2019). The firm uses an algorithmic assessment before a full clinical evaluation by a human for this purpose, which leads to the case groups described in the main paper.

First-come-first-served (FCFS). The firm implements FCFS as a simple, transparent rule that aligns with expectations from both patients and oversight authorities and reduces the risk of complaints. As a publicly funded provider operating across several European countries, it is reviewed by national inspection bodies (IVO in Sweden), which emphasize short waiting times and fair, accessible services (Kry International AB 2024). The scheduling literature similarly highlights transparency and procedural fairness as key motivations for FCFS (Ala and Chen 2022). In Sweden, these considerations are further reinforced by the Healthcare Law, which requires equal treatment unless documented medical urgency justifies prioritization (Socialdepartementet 2017). Because primary care typically handles non-urgent conditions and urgency is hard to assess before clinical evaluation, the firm applies FCFS once the lowest effective level of care is determined.

Nurses and doctors. The firm employs nurses in addition to doctors in order to prioritize doctor consultations to where they make the most difference. The firm also employs other clinician types but these are not included in our sample (Kry International AB 2024).

Training and scope of practice. The clinician types differ in training in a similar way to other countries' doctors and nurses. Doctors typically complete an 11-semester medical

[35] See further regional payment guidelines, e.g., Region Kronoberg 2025 or Region Skåne 2023.

degree, followed by an 18-month internship (Allmäntjänstgöring, AT) to obtain their license, and five or more years of residency training to become specialists (Specialiseringstjänstgöring, ST), amounting to seven or more years of formal and on-the-job training. Nurses complete a three-year degree for their license and may, after at least one year of experience, pursue a 1.5-year specialization (such as district nursing, midwifery, or anesthesia), although most nurses do not.[36] Nurses and doctors also differ in their scope of practice. Nurses typically handle patient communication, routine check-ups, and triage, whereas only doctors can prescribe most medications, issue specialist referrals, and issue longer sick notes.

Wages and labor markets. In Sweden in 2020, average monthly salaries were about 38,700 SEK for nurses and 79,200 SEK for doctors (Statistics Sweden (SCB) 2020). While others have noted that doctor incomes in Sweden are lower than in other high-income countries (Buehler et al. 2026), this is not specific to doctors but reflects an overall compression of the wage distribution in Nordic countries (Mogstad, Salvanes, and Torsvik 2025). Buehler et al. (2026) also note that relatively low doctor wages are consistent with monopsony markdowns, as we also model. Despite this, Sweden faces shortages of healthcare workers, particularly among specialized doctors and specialized nurses (Socialstyrelsen 2022). The nurse-to-doctor ratio is about 2.5 nurses to one doctor, which is close to the OECD average ratio (Table A1).

B Data Appendix

Provider data. The primary data is supplied by a large healthcare firm in Sweden. Their data contain comprehensive records of all online primary care consultations within the firm for the years 2019 and 2020. For each consultation, we observe the date, duration, and symptom of the consultation; the clinician type, the contact method, and primary diagnosis set by the clinician in the form of the ICD-10 codes, the standardized system of disease classification.[37] We standardize symptom names across children and adults, merge near-duplicates, and translate Swedish names into English. In addition, we observe patients' demographic characteristics, such as age and gender. For clinicians, demographic information is limited. Each consultation record includes patients' and clinicians' unique personal identifiers, which have been pseudonymized for us.

The personal identifiers enable Statistics Sweden (before pseudonymization) to link patients' consultations to additional individual-level information from Statistics Sweden (SCB) and other agencies. Through data from the National Board of Health and Welfare (Socialstyrelsen, NBHW), we observe patients' healthcare utilization in the rest of the

[36] In our sample, about 10% of nurses hold a specialty qualification, compared with more than 75% of doctors who are in or have completed specialty training.

[37] Diagnoses are rarely missing in our sample (0.33%), as they are required for payment.

healthcare system, except consultations from external primary care; prescriptions from external primary care are included. In addition, Regions Scania and Stockholm provide data on all primary care visits — not only within the firm we study but also with other providers — in their respective regions. Below, we describe our data sources in more detail.

Demographic and socioeconomic data. Statistics Sweden (SCB) provides us with information from three registries: the Population Statistics Register (RTB), the Longitudinal Integrated Database for Health Insurance and Labour Market Studies (LISA), and the Education Register (UREG). RTB covers patients' immigration background as well as, from 2013 to 2018, their civil status and municipality of residence. LISA provides annual socioeconomic and demographic data for patients, including income, rehabilitation benefits, employment status, paid sickness, and income-related benefits, from 2013 to 2020, with additional income and employment data available from 2019 to 2022. Finally, UREG offers data on patients' educational backgrounds from 2013 to 2020.

We observe monthly gross earnings from the main employer. To ensure we do not capture employees changing jobs or seasonal workers, we restrict the analysis to cases where patients are employed at the end of the year (excluding self-employment), whose income in the three months preceding the consultation exceeds 3,533.33 SEK each month in 2010 values, based on the annual threshold of 42,400 SEK used by Saez, Schoefer, and Seim (2019). We focus on employed patients because monthly income from self-employment is calculated as the annual income divided by twelve in our data.

Information at SCB are collected at different points during the year: Income and employment status are measured in October each year, educational attainment is measured at the end of each spring semester, immigration background does not change over time, and the remaining variables are measured typically in November each year. We use the 2018 values for all control variables to ensure they are recorded before the earliest consultation starts. When education is missing in 2018, we impute it by backward filling: we first use 2017 values, and if those are unavailable, we use 2016 values.

Downstream health data. Data from the National Board of Health and Welfare (NBHW) contains prescriptions, mortality, outpatient, and inpatient healthcare data. The prescription data, covering 2013 to 2023, includes medications collected by patients at pharmacies but excludes medical drugs administered within healthcare facilities and over-the-counter drug purchases. We include both prescriptions from the firm and external prescriptions. To exclude prescriptions for chronic conditions, we consider only drugs on the chemical subgroup level (Anatomical Therapeutic Chemical level 4) that patients have not yet purchased at pharmacies within the 6 months prior to the initial consultation. Mortality data, available for 2019 to 2023, include the date and cause of death. Outpatient and inpatient data, spanning

2013 to 2023, include visit and admission dates, ICD-10 diagnostic codes, and, for outpatient data, additional information on contact types.

Healthcare data except primary care consultations are provided by the National Board of Health and Welfare and cover the years 2013-2023 in four datasets: an inpatient care dataset contains information on hospitalizations; an outpatient care dataset includes data on ED visits, urgent care center visits and specialist visits; a dataset includes information about deceased patients; and a prescription dataset includes all drug prescriptions that patients have collected over the year. Inpatient and outpatient datasets include information about visit date, discharge date, and the ICD-10 diagnostic codes with three digits; the dataset on deceased patients includes the cause of death and the date. The prescription dataset includes all drug prescriptions that patients have collected over the years 2013-2023. If a prescription contains multiple drugs, each drug is recorded in a separate observation. The dataset also contains unique anonymized identifiers for each clinic that issued the prescription. Moreover, the National Board of Health and Welfare provides the costs of each prescription (incurred by the national insurance system and any additional patient costs), ATC codes, and a prescriber category (such as primary care, specialist care, school nursery etc.).

Primary care data. Primary care consultation data is not collected nationwide in Sweden, but held by each region. We collect data on external primary care visits directly in Scania and Stockholm, and infer primary care consultations for patients registered in other regions from their prescriptions, as primary care prescription data is collected nationwide by the NBHW. Scania is a region in southern Sweden with about 13% of the country's population; Stockholm is the large capital region with about 20% of the country's population. Data from Region Scania cover 2013–2020, and data from Region Stockholm cover 2013–2023, linked to our other data via unique patient identifiers. For both regions we observe the date of contact, a clinic identifier (pseudonymized in Stockholm), and the main diagnosis as an ICD-10 code; for Scania we additionally observe clinician type and contact type, and for Stockholm the care branch.

Data from Regions Scania and Stockholm are administratively collected and include services beyond standard primary care (e.g., maternity care and counseling). To ensure comparability, for Scania we restrict the data to physical nurse or doctor consultations at clinics eligible for patient registration as a primary care provider, for Stockholm to visits at primary care branches, and for both regions to contacts excluding the study provider.

Details on primary care costs and downstream care costs. Cost estimates are obtained from the sources presented in Appendix Table A3. Online primary care consultations are based on the public payers' recommended reimbursement provided by SKR (Myndigheten för vård- och omsorgsanalys 2022, p. 127; Södra sjukvårdsregionen 2020, p. 91). External primary care consultations are assumed to be 1452 SEK. This number is based on the weighted average of in-person nurse and doctor visits in the Scania sample. Weights are based on the frequency of nurse and doctor visits in the sample.

Prescription costs are the average cost of each drug collected by patients within our time frame, and specialist care costs are the median cost of selected specialists' visit costs provided by the Northern healthcare region, calculated across the following specialties: internal medicine (excluding cardiology), rheumatology, pulmonology, infectious diseases, dermatology, surgery, oncology, urology, orthopedics, ophthalmology, otolaryngology, pediatrics, gynecology, obstetrics, rehabilitation, psychiatry, and child psychiatry (Norra sjukvårdsregionförbundet 2018). Hospitalization costs are computed in two steps. For each clinical specialty, we sum four components: the one-time admission fee, admission costs, daily medical care, and daily nursing care, where the two daily components are multiplied by the sample average length of stay of three days. We then take the median of these specialty-level totals. Detailed hospitalization costs are provided by the Southern Healthcare Region (Södra sjukvårdsregionen 2020), and admission and medical care costs include doctor consultations.

Matching firm initial consultations to their later booked visits. To match initial consultations to booked revisits, we link each clinician-booked visit to its originating consultation within the preceding seven days. We use two strategies to identify clinician-booked doctor visits. First, we match each doctor consultation that is labeled by the firm as a "booked revisit" to a preceding consultation with the same patient and symptom. Second, we match each doctor consultation labeled as "revisit" or "phone triage" to the closest preceding consultation. When an initial consultation can be linked to multiple clinician-booked doctor visits, we prioritize matches from the first strategy and the clinician-booked doctor visit that is closest in time. Appendix Table B1 presents definitions of the case characteristics used in our main analysis.

C Model Appendix

Here we apply the general framework of Section 4.1 to our setting. Section 4.2 summarizes the main takeaways from this model appendix. In the setting we study here, the *experts* are doctors and the *frontline workers* are nurses. We retain the general notation but note that κ_E and κ_F now denote the per-case costs of a doctor and a nurse consultation, respectively, with $\kappa_E > \kappa_F > 0$ reflecting the higher training and wage costs of doctors. The *tasks* are patient cases.

Queue structure. In each period t, the queue consists of two groups: *doctor-group* cases, for which only direct doctor assignment is suitable, and *overlap-group* cases, for which either clinician type may be assigned initially. Let $Q_t^D \geq 0$ and $Q_t^{\mathrm{OV}} \geq 0$ denote the number of doctor-group and overlap-group cases, respectively, to be assigned in period t.[38]

Doctors are first assigned to doctor-group cases. Define $\tilde{E}_t \equiv E_t - \min(E_t, Q_t^D) \geq 0$ as the number of doctors remaining after doctor-group cases have been assigned in period t; any unserved doctor-group cases carry over to period $t+1$. The $\tilde{E}_t$ remaining doctors and F_t nurses are then available for overlap-group cases. Let $M_t^{\mathrm{OV}} \equiv \min(\tilde{E}_t + F_t, Q_t^{\mathrm{OV}})$ denote the number of overlap-group cases assigned in period t; unassigned overlap-group cases carry over to period $t+1$. Note that M_t^{OV} corresponds to M_t from Section 4.1, restricted to overlap-group cases. We assume that cases passed from nurses to doctors are processed in later slots outside the main queue, so $\tilde{E}_t$ is calculated before such cases are added back to a doctor's workload.

Revenue structure. The firm's per-case revenues are administratively set and follow a *resolving-clinician* payment rule: the firm is compensated only for the clinician type that resolves the case. Because it is the doctor who resolves the case in both the direct-to-doctor route and in the knowledge hierarchy route when a case is passed from a nurse to a doctor, this rule implies $r_{F,E} = r_E$. We parameterize revenues using a common wedge $\theta \in (0,1)$ between administratively set prices and the firm's production costs: $r_E = \frac{\kappa_E}{1-\theta}$ and $r_F = \frac{\kappa_F}{1-\theta}$, so that each payment to the firm covers its production cost with a proportional margin $\theta/(1-\theta)$. Together, these expressions give the full revenue vector $(r_E, r_F, r_{F,E}) = \left(\frac{\kappa_E}{1-\theta}, \frac{\kappa_F}{1-\theta}, \frac{\kappa_E}{1-\theta}\right)$, with $r_{F,E} = r_E > r_F$ since $\kappa_E > \kappa_F$.[39]

Unconstrained assignment. Substituting the firm's revenue structure into Equation (1), the net revenue advantage of the knowledge hierarchy simplifies to $\Delta(\tau) = p_\tau \cdot \frac{\kappa_F - \theta\kappa_E}{1-\theta} - \kappa_F$. Since costs are such that $\kappa_E > \kappa_F$ and $\Delta(\tau)$ is linear in p_τ with bounds

$$\Delta(\tau)\big|_{p_\tau=0} = -\kappa_F < 0 \quad \text{and} \quad \Delta(\tau)\big|_{p_\tau=1} = -\frac{\theta(\kappa_E - \kappa_F)}{1-\theta} < 0,$$

we see that $\Delta(\tau) < 0$ for all τ. And as in Result 1, the unconstrained firm therefore assigns all overlap-group cases directly to doctors.

Constrained assignment. When doctor availability binds ($M_t^{\mathrm{OV}} > \tilde{E}_t$), the firm must assign some overlap cases to nurses. Evaluating Condition 2 under the firm's revenue structure gives

[38] Nurse-group cases are treated as lower-urgency and processed outside the main queue; cases passed from nurses to doctors are likewise scheduled into dedicated later time-slots. The FCFS constraint applies separately within each case group.

[39] As in the general model, we assume that task processing time per worker is constant and we could equivalently model the number of slots for each worker type in each period. We discussed in Section 2 that task processing durations vary little across worker types in our setting.

$$\frac{\partial \Delta(\tau)}{\partial p_\tau} \;=\; r_F - r_{F,E} + \kappa_E \;=\; \frac{\kappa_F - \theta\kappa_E}{1-\theta}, \tag{6}$$

which is positive if and only if $\kappa_F > \theta\kappa_E$. This condition holds for any $\theta < \kappa_F/\kappa_E$, and is satisfied by the parameters in our setting (see Section 4.3). By Result 2, the constrained firm therefore assigns to the available nurses the $M_t^{\text{OV}} - \tilde{E}_t$ overlap cases with the *highest* p_τ—that is, the cases nurses are most likely to resolve themselves. Given the firm's revenue structure, the cases with the highest p_τ are also those in which nurses have the highest relative net revenue advantage.

An alternative (hypothetical) revenue structure that generates perverse incentives. Here we consider a hypothetical departure from the firm's revenue structure that induces perverse incentives.

Flat payment per consultation. Suppose the firm is paid ρ per *consultation*, so that if both a nurse and a doctor are consulted, the firm is paid for both. Because passing on an unresolved case to a doctor generates revenues from two consultations (one with the nurse, one with the doctor), while a directly resolved case generates only one paid consultation, the revenue vector is $(r_E,\, r_F,\, r_{F,E}) \;=\; (\rho,\, \rho,\, 2\rho)$. Then $\partial\Delta(\tau)/\partial p_\tau = \rho - 2\rho + \kappa_E = \kappa_E - \rho$. If $\kappa_E < \rho$, condition (2) is violated: $\Delta(\tau)$ *decreases* in p_τ, and the constrained firm optimally assigns lower-p_τ (harder) cases to nurses to increase the likelihood of escalation and thus earn 2ρ. This revenue structure, therefore, incentivizes the firm to *maximize* rather than minimize production costs when constrained. The government's restriction of paying for no more than one consultation per patient per day avoids this problem.

Social planner's problem. The social planner's objective in our setting is a natural counterpart of Equation (3). Because virtually all downstream care in Sweden is publicly provided, we assume that all downstream costs are borne by taxpayers and that taxpayer payments directly reflect production costs (i.e. transfers from taxpayers to the firm net out). The social planner, therefore, minimizes the total production costs of healthcare:

$$\Delta^{SC}(\tau) \equiv \text{SC}_F(\tau) - \text{SC}_E(\tau) \;=\; [\kappa_F + (1-p_\tau)\kappa_E + \mathbb{E}[C_{F,\tau}]] - [\kappa_E + \mathbb{E}[C_{E,\tau}]], \tag{7}$$

where $C_{F,\tau}$ and $C_{E,\tau}$ are the downstream cost random variables defined in Section 4.1.[40] The planner assigns type-τ cases to the knowledge hierarchy whenever (7) is negative, and to doctors otherwise. When clinician availability is constrained, the planner ranks case types by (7) and fills nurse capacity from the top of this ranking, exactly as in Section 4.1. Because

[40] We expect patient ill health and suffering to be positively correlated with downstream healthcare costs: more severe health outcomes generally require more intensive and expensive care. Minimizing systemwide production costs should therefore largely align with minimizing patient suffering, though the two objectives need not coincide exactly. For instance, a hospitalization both reflects a serious health problem and is costly; conversely, some high-cost downstream care may reflect over-treatment rather than poor health.

the firm neither observes $C_{F,\tau}$ and $C_{E,\tau}$ for individual patients nor internalizes them under the current contract, its constrained assignment need not align with this social optimum—a source of misallocation we quantify in Section 7.

D Additional robustness

D.1 Alternative identification: Instrumental variable strategy

The OLS coefficients are informed by our understanding of the inputs into the assignment algorithm. However, since the exact functional form of the algorithm is unknown, we may not fully account for potential endogeneity in the initial assignment even with our rich set of control variables. To address this concern, we complement our OLS approach with an instrumental variable (IV) strategy that leverages exogenous variation in doctor congestion (Prediction 1b). While the IV strategy imposes weaker structural assumptions on the algorithm, it may yield imprecise results.

Estimating equation. For the IV strategy, we specify the first-stage regression as follows:

$$Nurse_i = \gamma \text{Doctor Congestion}_i + X'_{1i}\lambda_1 + X'_{2i}\lambda_2 + \nu_i, \tag{8}$$

where $\text{Doctor Congestion}_i$ denotes our congestion instrument, defined as *log median doctor wait time* (see Section 3). The vectors of baseline case characteristics and additional controls, X_{i1} and X_{i2}, are defined as above, and ν_i is an error term. We estimate a two-stage least squares (2SLS) model represented by Equation 4 and Equation 8 with robust standard errors.

In the IV setup, δ in Equation 4 represents the local average treatment effect (LATE) of being initially assigned to a nurse as opposed to direct assignment to a doctor, for cases where the initial assignment is affected by congestion. To interpret δ as the LATE, we require that our congestion instrument satisfies the four standard IV assumptions: relevance, conditional independence, monotonicity, and exclusion.

Validity of the congestion instrument. Validity of the congestion instrument. First, the congestion instrument must be relevant by shifting the initial assignment. Figure A2c shows that the probability of initial nurse assignment increases with doctor congestion. The first-stage coefficient from Equation 8 is 0.056, with an F-statistic of 3,726.

Second, we require that potential outcomes are independent of the congestion instrument conditional on controls. In the absence of the exact assignment algorithm, we condition on baseline case characteristics X_1 and test whether additional characteristics related to health risk and socioeconomic status, which could influence potential outcomes, are balanced with respect to congestion. Appendix Figure A8b shows that, conditional on baseline controls, the congestion instrument is largely balanced against these additional characteristics, both individually and jointly.

Third, we require monotonicity: higher congestion should not reduce the likelihood of initial nurse assignment, and cases assigned to nurses under low congestion should also be assigned under high congestion. Appendix Figure A9 shows that, for almost every symptom, the share of cases initially assigned to nurses weakly increases with congestion. Across the subsamples, coefficients are generally positive, statistically significant, and similar in magnitude, providing support for the monotonicity assumption.

Finally, for a causal interpretation of the IV estimates, congestion must affect outcomes only through initial assignment. As discussed in Section 5 (see also Appendix Table A6), congestion is associated only with mechanical increases in waiting times and small changes in consultation duration, with no evidence of meaningful effects on care provision or patient outcomes. While patient wait times mechanically increase with congestion, even in the highest congestion quartile, the mean wait time is below 45 minutes. This is unlikely to meaningfully affect care quality or costs, especially when compared to physical primary care, where over 25% of patients wait more than 24 hours for an initial doctor appointment.

D.2 Attrition

We observe consultations that did not occur due to patient no-shows or technical failures. However, no patient information is recorded for cases which do not complete their registration, as is common on online platforms. After submitting the symptom form, patients are shown their assigned clinician type and may cancel while waiting, which we define as one of several types of dropout. Differential attrition—such as higher cancellation when patients expect a doctor but are assigned to a nurse—could bias our estimates.

However, information on registration incompletion rates during a later period allows us to bound potential differential attrition also on this margin, and we also perform additional robustness checks regarding dropouts, discussed in Section 6.3.

Appendix References

Abadie, A. (2003). "Semiparametric Instrumental Variable Estimation of Treatment Response Models". *Journal of Econometrics* 113 (2): 231–263.

Ala, A. and Chen, F. (2022). "Appointment Scheduling Problem in Complexity Systems of the Healthcare Services: A Comprehensive Review". *Journal of Healthcare Engineering* 2022, 5819813.

Buehler, A., Gottlieb, J. D., Hicks, J., Laun, L., Palme, M., Polyakova, M., Udalova, V., and Ventura, M. (2026). "International Comparison of Physician Incomes". *AEA Papers and Proceedings* 116, 144–149.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Feng, Y. (2024). "On Binscatter". *American Economic Review* 114 (5): 1488–1514.

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

European Observatory on Health Systems and Policies (2026a). *Health System Summaries*.

European Observatory on Health Systems and Policies (2026b). *Health Systems Reviews*.

Folkhälsomyndigheten (2020). *Veckorapport om covid-19, vecka 27*.

Frandsen, B., Lefgren, L., and Leslie, E. (2023). "Judging Judge Fixed Effects". *American Economic Review* 113 (1): 253–277.

Ministry of Health, Welfare and Sport (2018). "The Right Care in the Right Place" (tech. rep., Ministry of Health, Welfare and Sport (Netherlands)).

Mogstad, M., Salvanes, K. G., and Torsvik, G. (2025). "Income Equality in the Nordic Countries: Myths, Facts, and Lessons". *Journal of Economic Literature* 63 (3): 791–839.

Myndigheten för vård- och omsorgsanalys (2022). *Besök via nätet*. Stockholm.

National Center for Health Statistics (2023). *2023 National Health Interview Survey (NHIS)*.

NHS England (2026). *Principle 3 — Right Care, Right Time, Right Place for Carers*.

Norra sjukvårdsregionförbundet (2018). *Price Lists of Northern Healthcare Region*.

OECD (2021a). "Health at a Glance" (OECD).

OECD (2021b). "The COVID-19 Pandemic and the Future of Telemedicine" (OECD).

OECD (2025a). *Health Statistics Data Explorer: Doctor Consultations*.

OECD (2025b). *Health Statistics Data Explorer: Physicians*.

OECD (2025c). *OECD Health Statistics Data Explorer: Practising Nurses*.

Page, A., Ambrose, S., Glover, J., and Hetzel, D. (2007). "Atlas of Avoidable Hospitalisations in Australia" (Public Health Information Development Unit, University of Adelaide).

Region Kronoberg (2025). *Riktlinje för registrering av vårdkontakter*.

Region Skåne (2023). *Regionala riktlinjer för registrering och klassificering i Region Skåne*.

Royal Norwegian Ministry of Health and Care Services (2006). "National Health Plan for Norway (2007–2010)" (tech. rep., Royal Norwegian Ministry of Health and Care Services).

Socialdepartementet (2016). "Effektiv vård: Slutbetänkande av En nationell samordnare för effektivare resursutnyttjande inom hälso- och sjukvården" (SOU 2016:2, Statens offentliga utredningar).

Socialdepartementet (2017). *Hälso- och sjukvårdslag (2017:30)*.

Socialstyrelsen (2022). *Bedömning av tillgång och efterfrågan på legitimerad personal i hälso- och sjukvård samt tandvård*.

Södra sjukvårdsregionen (2020). *Reports About Healthcare Costs*.

Statistics Sweden (SCB) (2020). *Wages and Salaries by Occupation*. Statistics Sweden Statistical Database.