Does Division of Labor Increase Productivity? 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; University of Copenhagen,* [3] *Stockholm University,* [4] *London School of Economics*

February 20, 2026
PRELIMINARY - PLEASE DO NOT POST OR CIRCULATE

Abstract

Understanding the determinants and consequences of the division of labor is central to economics. Modern organizations often divide labor using knowledge hierarchies in which frontline workers handle simpler tasks and pass tasks they cannot handle to experts. We study the effects of a knowledge hierarchy (KH) in almost 500,000 patient cases from an online primary care firm, linked to outcomes and costs across Sweden's healthcare system. The firm's algorithm quasi-randomly assigns similar patients either to a nurse–doctor hierarchy or directly to doctors, depending on short-term variation in the availability of doctors. Within the KH, nurses resolve 70% of cases and pass the remainder to doctors. Relative to doctor-only care, the KH modestly increases some follow-up care outside the firm, but has no adverse effects on patient satisfaction and higher-stakes outcomes, including hospitalizations, earnings losses, and mortality. The KH does not change mean total healthcare costs, though it raises costs for some tasks and reduces them for others. Compared to using only doctors, the firm's *selective* use of the KH for different symptoms yields total healthcare savings equivalent to 2.33% of online primary care costs. However, the firm does not observe the downstream outcomes of individual patients, and the incentives the firm faces do not make it internalize the externalities it generates on the healthcare system. We show that reassigning cases according to the comparative advantage in total healthcare costs (including downstream costs) would yield additional savings equivalent to 6.7% of online primary care costs.

Keywords: Division of Labor, Productivity, Healthcare
JEL codes: J24, I19, L23

We thank Ellen Donath, Emma Persson, and Berk Uzunonat for excellent research assistance. We also thank many colleagues and seminar and conference participants at Aarhus University, Arne Ryde workshop in Lund, Becker Friedman Institute Health Economics conference, CEPR Paris Symposium, CESifo Summer Institute, EIT Oxford, Frisch Centre, HEFUU workshop in Uppsala, Imperial College London, Joachim Herz Foundation Add-on Fellowship Meeting in Hamburg, Johannes Kepler University, London School of Economics, RFBerlin-CEPR Annual Symposium in Labour Economics, Royal Holloway University, Stockholm School of Economics, University of Innsbruck, University of Konstanz, University of Oslo, University of St. Gallen, 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 granting access to their data. We thankfully acknowledge financial support from the ESRC's Centre for Economic Performance, the Swiss National Science Foundation, grant number 100018-228312, Handelsbankens Forskningsstiftelser (Program grant P25-0261),and support from the University of Zurich's Research Priority Program "Equality of Opportunity". The research project was approved by the Stockholm Ethics Review Board and the Swedish Ethics Review Authority.

The greatest improvement in the productive powers of labour... seem to have been the effects of the division of labor. — (Smith 1776)

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, labor is often divided using *knowledge hierarchies*, in which frontline workers handle simpler tasks and pass harder tasks to experts (Garicano 2000). Despite the prevalence of knowledge hierarchies, there is little empirical evidence on their productivity effects, the tasks to which they are assigned, or whether their use is efficient, especially when they generate downstream externalities, for instance through affecting the content of the service, which leads to shifting some tasks to experts in other organizations.

Empirically evaluating knowledge hierarchies relative to the direct-to-expert route has proven difficult for three main reasons: tasks assigned initially to frontline workers often do not overlap with those assigned directly to experts; even when overlaps exist, assignment is endogenous; and available data rarely link within-firm task-level assignment to downstream quality and costs measured outside the firm. We study a knowledge hierarchy (KH) in healthcare, in which nurses resolve some cases and pass other cases to doctors, and compare it to the direct-to-doctor route. Our setting allows us to overcome the empirical challenges and identify the effects of a knowledge hierarchy.

Understanding the productivity effects of knowledge hierarchies is especially important in sectors where expertise is scarce, demand is rising, and services are publicly financed. Healthcare exemplifies these forces: aging populations, clinician shortages (Maier and Aiken 2016), and rapid technological change are reshaping how expertise is produced and deployed (Dahlstrand 2024; Dahlstrand et al. 2024; Bronsoler et al. 2022; Rajpurkar et al. 2022). In such settings, delegation is not merely an internal managerial choice–it determines costs, quality, and who ultimately bears them. When organizations decide whether to rely on knowledge hierarchies or direct-to-expert routes, their internal task allocation can generate downstream spillovers across other providers and payers (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.

To study these issues, we leverage algorithmic assignment that generates plausibly exogenous variation in whether patient cases are routed through a knowledge hierarchy or directly to a doctor. Using this variation, we ask three questions. First, how do total healthcare costs and health outcomes 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? And finally, which cases should be assigned to the KH, taking into account downstream externalities on the healthcare system?

Our setting is a large Swedish firm that provides taxpayer-financed primary healthcare services online, which is a common mode of care delivery in many countries (see Table 1). The tasks we analyze are patient cases with different symptoms, demographics, and medical histories, which are handled by the firm. To guide our empirical analysis, we model the firm's algorithm, which assigns

cases either directly to doctors or initially to nurses (the KH). The firm maximizes its revenues, net of costs, subject to institutional constraints. When the short-term availability of doctors is unconstrained, the algorithm assigns all cases to doctors; however, when constrained by temporary shortages, it assigns some cases to nurses. Although this constrained allocation also minimizes taxpayer payments to the firm (holding constant clinician use), it does not minimize total healthcare costs because downstream healthcare costs are neither observed nor internalized by the firm given the contract structure, which is not an uncommon situation. In our model, we show that efficient assignment of clinicians follows the comparative advantage of the KH in total healthcare costs by case type. We use the model's predictions to guide our empirical strategy and construct counterfactual assignments of patient cases to quantify the efficiency gains from better-targeted use of the KH.

For our empirical analysis, we construct a novel dataset that allows us to observe many health outcomes, both within the firm and downstream. We focus on firm records for almost 500,000 "overlap" cases, which can be initially assigned (by the algorithm) to either doctors or nurses. These cases are linked to detailed 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 calculate costs.

To address algorithmic sorting of cases between routes, our main identification strategy controls for fixed effects for three-hour windows within each calendar day and flexibly for interacted case characteristics that the algorithm observes. We show that, conditional on these controls, case characteristics that the algorithm cannot observe are almost exactly balanced. Our findings are also generally robust to the use of a measure of short-term doctor congestion as an instrumental variable. The delays involved are short – more than 90% of cases are handled within an hour and more than 99.9% within a day, which speaks against concerns about delay having a direct health effect as in Costantini (2025) and Godøy et al. (2024).

Our main analysis focuses on downstream healthcare utilization and outcomes within seven days, when the effects are directly observable and attributable to the organizational route at the initial contact. We find that 70% of cases in the KH are resolved by nurses themselves, while the remaining 30% are sent to doctors. This KH moderately reduces diagnostic specificity and the rate of new prescriptions. Yet, we find that patients are no less satisfied with the KH – if anything, patients give more top ratings. And while the KH modestly increases the rate of external primary care visits and emergency department visits, the incidence of hospitalizations is unchanged. We also find no adverse effects of the KH on major income 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 under current payment rules, where the firm is only paid for the clinician who resolves a case but incurs costs for all clinicians involved, which lowers the firm's net revenue. When downstream costs borne by taxpayers are accounted for, taxpayer payments decrease by 9% under the KH. By contrast,

the KH does not affect the mean total healthcare costs when abstracting from transfers between taxpayers and the firm. Overall, the KH maintains care quality but does not affect total production costs on average.

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 costs for almost half of patients, while leaving costs in the upper tail (driven by downstream care) largely unchanged.

When the firm is unconstrained in the sense of having no doctor congestion, it sends almost all the cases to doctors. But the key characteristics that determine case assignment when there are short-term doctor shortages are the patients' symptoms. Symptoms with a higher relative net revenue advantage under the KH (cases that nurses are likely to resolve themselves) are more likely to be assigned to nurses. This assignment pattern also reduces taxpayer payments to the firm, conditional on clinician availability. Compared to a counterfactual scenario in which only the direct-to-doctor route is used, the firm's selective use of the KH reduces total healthcare costs by 0.7% (2.33% relative to online primary care costs).

Finally, we consider counterfactual case allocations which implement the social planner's objective by assigning cases according to the KH's symptom-specific comparative advantage in total healthcare costs, while respecting short-run clinician capacity constraints. Reallocating cases on this basis within 15-minute windows without changing actual nurse or doctor utilization reduces total healthcare costs by 2.1% (6.7% relative to online primary care costs). The decline in the firm's net revenues in this scenario is trivial relative to the societal gain. These results suggest that simple corrective measures, such as clear guidelines on when to use the KH, can meaningfully lower total healthcare costs. Moreover, it suggests that the KH itself is a useful organizational tool to reduce total healthcare costs, when applied to the right tasks. Implementing it for the right tasks – which nurses are likely to resolve by themselves (as is already done), and for which nurse resolution does not lead to higher downstream healthcare costs (which our analysis identifies) – essentially creates a new division of labor between nurses and doctors.

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 in in-person primary care; and that our main results are similar when we reweight by demographics or focus on the period before Covid-19. Although these do not guarantee that our point estimates generalize exactly, the mechanisms that we study – especially the systematic assignment of task types to a KH or directly to experts, and the lack of internalization of downstream externalities – are relevant for many other settings.

Our paper contributes to the literature by providing novel, task-level causal evidence on the productivity and efficiency implications of knowledge hierarchies, which are central to many organizations. We differ from the canonical model of KH (Garicano (2000)) in three ways: in our empirical focus, in our focus on a different cost of KH (we emphasize the need for frontline workers to try a

task before determining whether they can resolve a task, rather than communication costs), and in our consideration of downstream externalities.

Our study also differs from research that compares occupations head-to-head. This includes the important contribution of Chan and Chen (2022), from which we differ in our focus on a KH with an organizational design lens, rather than a head-to-head contrast between doctors and less specialized clinicians; and in our setting (primary care with nurses rather than emergency room with nurse practitioners), and perhaps most importantly, in our study of downstream externalities.

A broad literature emphasizes how productivity depends on specialization, task assignment, and comparative advantage (Becker and Murphy 1992; Katz and Murphy 1992; Jones 2009; Acemoglu and Autor 2011). 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, most notably Caliendo, Monte & Rossi-Hansberg (2015) and Caliendo, Mion, Opromolla & Rossi-Hansberg (2020), documents the structure of production hierarchies and their relationship to firm productivity using rich firm-level data, with a macro or reduced-form perspective on organization. In contrast, we study task-level assignment decisions within an organization, exploiting quasi-experimental variation in whether identical tasks are routed through 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 gains from reallocating tasks according to comparative advantage in a cost measure that accounts for these externalities.

We also contribute to growing 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). A related literature studies occupational licensing and scope-of-practice reforms, particularly for nurse practitioners, and their effects on healthcare provision (Kleiner 2016; Traczynski and Udalova 2018; Alexander and Schnell 2019; McMichael 2023; Guo et al. 2024; Dillender et al. 2024). Rather than estimating the effects of expanding practice rights, we study task allocation within organizations holding scopes of practice fixed. We advance the existing healthcare literature by analyzing how long-standing occupational competencies can best be deployed in organizations, while accounting for downstream effects beyond the focal organization.

The remainder of the paper is organized as follows. Section 2 introduces the institutional background and Section 3 our data. Section 4 discusses our conceptual framework for guiding 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 productivity 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 outline the firm's operations; additional institutional details are provided in Appendix C.

2.1 Background

Sweden's publicly funded healthcare system provides comprehensive universal coverage. The system is largely tax financed, and patient cost sharing is low.[1] Table 1 compares key features of Swedish healthcare to other OECD countries. Primary care serves as the main entry point into the system: clinicians diagnose and treat common conditions and determine whether patients require downstream care. Patients register with a primary care center but may also seek online primary care from other providers. Care is delivered by both public and private centers, with private for-profit providers operating under regulated contracts with regional authorities and accounting for roughly 40% of primary care (The Commonwealth Fund 2020).

The firm we study is a large private provider operating nationwide (and in several other countries). It employs hundreds of clinicians, comparable in scale to large group practices in the US. Although it runs in-person clinics for registered patients, its core activity is nationwide online primary care, making it Sweden's – and Europe's – largest digital healthcare provider. As shown in Table 1, online consultations have become common in Sweden in recent years, where the use of telehealth (around a quarter of all healthcare consultations) is similar to many other OECD countries, including the US, and prescription treatments can be provided online.

The firm explicitly aims to maximize profits. Its revenues are determined by fee-for-service with administratively set prices, a model also common internationally (Table 1). The firm is paid by regional authorities according to payment levels intended to reflect the firm's production costs for its services (Sveriges Kommuner och Landsting SKL). Its core business focuses on doctor consultations; however, it also employs other clinicians, including nurses.

The firm faces constraints on both the choice of clinician type and on the case ordering. The clinician type choice is guided by cost-effectiveness in Swedish 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, while additional same-day consultations are not paid ("resolving clinician" rule). As a result, the firm assigns some cases always to nurses, some always to doctors, and the remaining set of *overlap-group cases* to either clinician type. Case processing follows a first-come-first-served (FCFS) approach, reflecting expectations of transparency and predictability and limiting long or arbitrary waits. Although these constraints reflect the Swedish environment,

[1] During our sample period, annual out-of-pocket spending for outpatient visits was capped at roughly 1,150 SEK (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.

similar scheduling and cost-effectiveness considerations are common across many healthcare systems.

2.2 Cases at the firm

Patients access healthcare for themselves or their children in a mobile application through an electronic identification system used across Sweden's healthcare and government institutions. The firm automatically receives patients' age, gender, and region from civil records. Patients then report their main symptom from a predefined list and may provide additional information through questionnaires. 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, it enters a waiting queue.

Our analysis focuses on *overlap cases* after spring 2019 from patients unregistered with the firm, for which either a nurse or a doctor may be assigned initially. Prior to 2019, all consultations were handled by doctors, with nurses limited to triage and a chlamydia screening program. From spring 2019, assignment to a nurse or a doctor is determined by an algorithm based primarily on the reported symptom, age, gender, and region, while also taking into account clinician availability. The precise algorithm has evolved over time and is proprietary. Because staffing schedules are 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.

Doctors and nurses are paid flat hourly or monthly wages, and many work part-time at the firm while also working in-person elsewhere. Nurses have completed a three-year university degree focused on patient communication and routine care, whereas physicians undergo at least seven years of formal and on-the-job training. Within each clinician type, the firm makes no further distinctions by specialty (except for cases involving infants, which we exclude). Both nurses and doctors follow similar treatment protocols and can generally resolve cases; however, only doctors may prescribe most medications, refer patients to specialists, or issue longer-term sickness absence certificates.

Clinicians use a digital interface that shows their personal assigned consultations. While the overall number of waiting patients and staffed clinicians is findable in principle, it is not directly observable during consultations, limiting clinicians' awareness of congestion relative to many in-person healthcare settings. Upon assignment, clinicians see the patient's demographics, reported symptoms, and medical history from prior interactions with the firm. Consultations may result in home-care advice (which both doctors and nurses can make), prescriptions, or recommendations (which also nurses can make, e.g. to visit the ED) and referrals (direct letters to specialists, which only doctors can make) to downstream care outside the firm. Doctors or nurses may also order home tests (such as Covid 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.

If a nurse cannot resolve a case, the nurse passes it to a doctor by scheduling a follow-up consultation in a later time slot. The follow-up consultation specifies that a doctor is required, but

not which individual doctor. The nurse also completes a mandatory handoff note summarizing symptoms, the reason for passing the case, and any relevant clinical information.

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

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 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. Further details are available in the Appendix D.

Analysis sample. We construct our main analysis sample as follows (see also Appendix Table A1). From 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 to handle a broader set of cases, and 24 December 2020, to observe revisits at the firm during 7 days (Restrictions 1–3). We exclude patients registered with the provider as their primary care clinic, who may receive in-person care from the firm and may have different care assignment paths (Restriction 4), infants under age two, for whom different assignment rules apply (Restriction 5), and cases preceded by another consultation with the same patient within two hours to focus on initial consultations (Restriction 6). We also exclude the few cases with missing baseline characteristics (Restriction 7). The resulting *broad drop-in sample* includes about 969,000 cases assigned to nurses or doctors.

We then classify cases by patient-reported symptoms, which are used by the assignment algorithm. First, we identify symptoms with effectively deterministic initial assignment: *doctor-group cases*, for which at most 5% of cases are assigned to nurses, and *nurse-group cases*, defined analogously, both excluding rare symptoms with fewer than 1,000 cases. Second, we define our main analysis sample as the remaining approximately 495,000 *overlap group cases*, which the firm may assign to either nurses or doctors. Figure A1 lists symptoms and initial clinician counts by group.

We discuss the institutional details regarding attrition and the data we have on that in Appendix Section F.2. We show robustness to attrition in the robustness part of the results, Section 6.3.

3.2 Variables

We summarize the key variables used in our analysis. Additional details are provided in Appendix D.

Treatment variable. Our treatment variable is assignment to the knowledge hierarchy, defined as a case being *initially assigned to a nurse* rather than directly to a doctor.

Baseline case characteristics. The firm's algorithm assigns each case to a nurse or a doctor based on observable, predetermined characteristics. Most importantly, patients select a primary symptom prior to any clinician contact, which places cases into doctor, nurse, or overlap groups. We display the symptoms by case group in Figure A1. The most common overlap symptom, *other health inquiries*, is a catch-all category for cases where patients have not picked another symptom, while the remaining symptoms reflect common conditions such as infections. Appendix Table A2 shows that, apart from mental health conditions, diagnoses in our sample largely overlap with those in the universe of in-person primary care in Scania in 2019.

To flexibly capture the assignment rules, we interact symptoms with the patient's age, gender, and prior cases at the firm, including whether the most recent observed case was assigned to a doctor and whether the patient has received a prescription from the firm. We also include patients' aggregated region of residence in the year prior to the consultation. Finally, we include fixed effects for date-by–3-hour windows based on the case's login time – the moment the case is registered and begins waiting for the initial consultation – to absorb broader patterns in demand and availability. After accounting for these characteristics, the remaining variation in assignment reflects exogenous short-term congestion.

Additional controls. We observe a set of additional predetermined characteristics. These variables are unobserved by the firm and unlikely to be used by the algorithm. They include (i) health risk measures based on healthcare use in the three years before the consultation excluding the 30 days prior (hospital, emergency, urgent care, and specialist care) and pre-2019 comorbidities from specialist or hospital care; and (ii) socioeconomic background in 2018, including indicators for above-median income, benefits receipt, education level, and employment, civil, and migration status.

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 (different) doctor, case resolution, patient satisfaction, and diagnostic informativeness; (ii) downstream healthcare use within seven days after the initial consultation date, including any new prescription and use of external primary, urgent (often out-of-hour centers), specialist, emergency, and hospital care; and (iii) longer-run consequences, capturing income declines and mortality to capture adverse events that for patients outside their direct healthcare utilization.

Cost outcomes. We also examine how the knowledge hierarchy affects taxpayer payments, firm revenues, and total healthcare costs of care provision, both within-firm and downstream in the healthcare system. Within-firm, cases resolved by nurses and doctors generate fixed taxpayer payments of 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 in detail in Section 4. We measure within-firm outcomes on the initial consultation day unless stated otherwise. For downstream costs, 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 A11. In robustness checks, we vary both the assumptions on production costs and the aggregation of downstream events.

Congestion. We measure congestion by *log median doctor wait time*: the log of the median wait time among the five most recent doctor-assigned cases in the 30 minutes before login, winsorized at the 99th percentile, and with missing values set to the lowest observed value. Figure A2 summarizes this measure: raw wait times are right-skewed, with over 60% of cases below 20 minutes and fewer than 10% above 60 minutes (Subfigure A2a), and the probability of initial nurse assignment increases monotonically with congestion (Subfigure A2b).

Summary statistics. Table 2 reports summary statistics for the main analysis sample of overlap-group cases. About 63% of cases involve female patients, and the average age is approximately 30. 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 a slightly higher share of women, partly reflecting more specific symptoms such as urinary tract infection being female-specific and sent more frequently to doctors. Differences in other characteristics are statistically significant in some cases but small in magnitude.

The analysis sample is younger and more female than both the Swedish population and patients in physical primary care in Scania, and it has a lower share of nurse consultations than in-person primary care (Appendix Table A3). We address the demographic differences in robustness checks by reweighting the sample to match the age and gender distribution of physical care. Appendix Figure A3 shows the evolution of nurse and doctor consultations over time, highlighting the firm's increasing reliance on nurses over 2019 and a sharp rise in consultations during the Covid-19 pandemic in spring 2020. We assess the sensitivity of our results to the Covid-19 period in the robustness analysis.

The total time spent on a case is between 10 and 12 minutes and the standard deviation is around 7 for doctors and nurses, whether resolving or not.

4 Model

This section introduces a model whose objective is to organize our empirical analysis and clarify the forces governing task assignment in knowledge hierarchies. The model highlights how firms trade off worker cost differentials, resolution probabilities, and possibly downstream externalities when allocating tasks across skill levels, and how alternative revenue structures shape these incentives.

We consider a firm that faces a stream of heterogeneous tasks (e.g., patients, customers, orders, applications, material inputs). Tasks can be assigned either directly to experts or to frontline workers, who resolve a subset and escalate the remainder to experts (the knowledge hierarchy, KH). The firm

chooses a profit-maximizing task assignment subject to its constraints.

We then map this general model to our empirical setting, in which tasks correspond to patient cases, frontline workers to nurses, and experts to doctors. The framework delivers a characterization of the firm's optimal assignment rule as well as the social planner's efficient assignment, and highlights the reasons for misallocation. Finally, we discuss the implications for our empirical analysis.

Conceptually, we depart from canonical KH models by emphasizing a different friction: Rather than communication costs (Garicano 2000), we focus on the costs incurred when frontline workers handle tasks they ultimately cannot resolve. In settings such as ours, communication has become much cheaper in recent years (e.g., Bloom et al. (2014)) and is unlikely to remain the dominant cost. We also differ in considering when the firm's revenue structure induces an assignment of tasks (holding staffing constant), which minimizes the firm's production costs. Finally, we differ in examining the role of downstream externalities.

4.1 The general firm problem

We consider a profit-maximizing firm that employs frontline workers (F) and experts (E). The firm faces a stream of tasks 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.[2] When a frontline worker cannot resolve a task, they escalate it to an expert; we assume experts resolve all tasks assigned to or escalated to them. "Resolving" means that the firm's processing of the task ends; however, resolved tasks may still involve downstream (out-of-firm) costs, as we discuss towards the end of this section.

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 $\kappa_E > \kappa_F$. We assume that the task processing time per worker is constant. The firm's gross revenue per task is:

- r_F if the task is assigned to and resolved by the frontline worker;
- $r_{F,E}$ if assigned to a frontline worker but resolved by the expert;
- r_E if assigned directly to an expert.

We assume the firm's revenue and cost structure, and its staffing, are set before task assignment. We focus on expected net firm revenues (gross revenues minus variable costs), ignoring fixed costs that cannot be changed during the assignment process. For a task of type $\tau \in \mathcal{T}$, the firm's expected net revenues under the KH and the expert route are, respectively:

$$\Pi_F(\tau) \equiv p_\tau (r_F - \kappa_F) + (1 - p_\tau)(r_{F,E} - \kappa_F - \kappa_E), \tag{1}$$

$$\Pi_E(\tau) \equiv r_E - \kappa_E. \tag{2}$$

We define the *net revenue advantage of the KH route* for type τ as

$$\Delta(\tau) \equiv \Pi_F(\tau) - \Pi_E(\tau). \tag{3}$$

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

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

$$\frac{\partial \Delta(\tau)}{\partial p_\tau} \;=\; r_F - r_{F,E} + \kappa_E. \tag{4}$$

The sign of this expression determines whether the firm prefers to assign easier or harder tasks to frontline workers, as characterized in Result 3 below.

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 within each task group. This allows the firm to process multiple tasks within a group simultaneously if they arrive sequentially.

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 also assume that tasks escalated from frontline workers to experts are processed later, outside the main queue.[3] We focus on short-run task assignment, abstracting from staffing and hiring decisions. 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 $M_t \leq E_t + F_t$.

Unconstrained vs. constrained assignment. The firm is *unconstrained* if the number of tasks with $\Delta(\tau) > 0$ does not exceed F_t and the number of tasks with $\Delta(\tau) \leq 0$ does not exceed E_t. Otherwise, assignment is *constrained*.

Result 1 (Unconstrained assignment)**.** *The unconstrained firm assigns all tasks with $\Delta(\tau) > 0$ to frontline workers and all remaining tasks directly to experts.*

Result 2 (Constrained assignment)**.** *The constrained firm proceeds in three steps.* (i) *If tasks with $\Delta(\tau) > 0$ exist, it assigns frontline workers to those tasks in decreasing order of $\Delta(\tau)$, until either all such tasks or all F_t frontline workers are assigned.* (ii) *If tasks with $\Delta(\tau) \leq 0$ exist, it assigns experts to those tasks in increasing order of $\Delta(\tau)$ (i.e., most negative first), until either all such tasks or all E_t experts are assigned.* (iii) *Any remaining workers are then assigned to the remaining tasks.*

We next characterize when this constrained assignment minimizes firm costs.

Result 3 (Firm cost minimization)**.** *If*

$$\frac{\partial \Delta(\tau)}{\partial p_\tau} \;=\; r_F - r_{F,E} + \kappa_E \;>\; 0, \tag{5}$$

[3]We assume task processing durations are fixed and show below that they vary little in our setting. Therefore, we could equivalently model the number of slots for each worker type in each period.

the constrained firm assigns to the available frontline workers the $M_t - E_t$ tasks with the highest p_τ (the easiest tasks).

This condition is violated when the revenue premium from passing to an expert $r_{F,E} - r_F$ is large enough to offset the expert cost κ_E. In that case, the firm prefers to assign frontline workers to harder tasks, increasing the likelihood that these tasks are passed to experts. In other words, certain pay structures (an example of which we discuss in Appendix E) disincentivize the firm from minimizing its costs even when it is constrained.

The social planner's problem We consider a social planner who focuses on minimizing the total production costs, which is the sum of firm costs and downstream costs.

Let downstream costs for a type-τ task be the random variable $C_{F,\tau}$ if the case is initially assigned to a frontline worker (the knowledge hierarchy, KH) and $C_{E,\tau}$ if directly assigned to an expert. Note that transfers from taxpayers to the firm do not enter the planner's objective unless they affect resource use.

For a type-τ case, expected total healthcare 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 these costs:

$$\Delta^{SC}(\tau) \;\equiv\; SC_F(\tau) - SC_E(\tau). \tag{6}$$

With unconstrained clinician availability, the social planner assigns type-τ cases to the KH if and only if $\Delta^{SC}(\tau)$ is negative – i.e., it assigns according to the KH's absolute advantage. If clinician availability is constrained, the planner instead ranks case types by Equation 6, and assigns each case subject to clinician constraints, assigns nurses to the cases with the lowest $\Delta^{SC}(\tau)$.

4.2 The problem of the primary care firm

In Appendix E we characterize the problem of the primary care firm we study. The *experts* are doctors and the *frontline workers* are nurses. 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.[4] 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 prices and the firm's production costs. We also consider two alternative pay structures. We show that the key results from Section 4.1 apply to the firm we study with only minor modifications.

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

4.3 Implications for the empirical analysis

Parameter values. We set $r_F = 275$ and $r_E = r_{F,E} = 500$, since the firm only gets paid for the expert when a frontline worker passes the case to an expert who resolves it (see Section 3.2). 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 cases into types. In practice, we find that symptoms are key to task allocation, so we equate types with symptoms and provide evidence to support our approach below.

Since we do not have data on the firm's costs, the wedge θ plays an important role in our model. In general, this wedge may arise from labor market power (markdowns) or product market (markups). However, since in our setting the "buyer" is the government who 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). In Section 6.3 we assess the robustness of our findings to varying θ, and allow for different wedges by clinician type.

Our model yields implications for the empirical analysis, in which 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: Empirical framework. In the model, the initial assignment of overlap-group cases to nurses (as opposed to doctors) depends on their type and predetermined factors observable to the algorithm. Conditional on these, there is no systematic selection into the treatment on other case characteristics. This provides a testable implication, which justifies our main (OLS) identification strategy that relies on conditional random assignment.

Prediction 1a: First-come-first-served (FCFS). The model assumes that within each case group (doctor group, nurse group, and overlap group of symptoms), assignment follows FCFS, that doctor-group cases may be delayed relative to overlap group cases, and that nurse-group cases are processed later.

Prediction 1b: Short-term congestion. For overlap-group cases, temporary doctor shortages (relative to cases) shift the algorithm from the unconstrained situation (when all cases are assigned to doctors) to the constrained situation, where some are assigned to nurses. Thus, lower doctor availability 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 for assignment to nurses, which we use to test the robustness of our main identification strategy.

Prediction 1c: Queue composition. For overlap-group cases, queue composition also exogenously shifts case assignment. Specifically, if the cases immediately preceding case i (an overlap group case) in the queue are doctor-group cases, case i is more likely to be assigned to a nurse. Since queue composition affects assignment only during doctor congestion, we do not use it as an instrument, but we provide evidence consistent with this prediction.

Prediction 2: Mean effects of the knowledge hierarchy. Using our identification strategy

outlined above, we estimate the effects of the knowledge hierarchy on care quality and costs for overlap cases. The model predicts that KH reduces firm net revenues and taxpayer-to-firm payments. Effects on healthcare outcomes and downstream costs (and thus on total taxpayer costs and total healthcare costs) are theoretically ambiguous and are estimated empirically.

Prediction 3: Distributional effects of the knowledge hierarchy. KH increases both the probability of lower firm costs and the probability of higher firm costs. This follows because under direct to expert assignment, firm costs are degenerate at κ_E. In contrast, KH induces a two-point distribution of firm real resource costs for overlap cases: κ_F with probability p_τ (resolved by frontline) and $\kappa_F + \kappa_E$ with probability $1 - p_\tau$ (escalated). Whether analogous distributional patterns extend to total healthcare costs depends on downstream utilization and is assessed empirically.

Prediction 4: Firm case assignment. Overlap-group cases, for which KH has a *relative net revenue advantage* (higher p_τ) are more likely to be assigned to nurses (the KH). These cases also entail lower *relative taxpayer payments to the firm.*

Prediction 5: Counterfactuals. Because the firm does not observe (nor internalize) downstream costs, its case assignment does not necessarily minimize the total healthcare costs of producing care. In other words, the firm may exert a downstream cost externality by allocating cases to clinicians who reduce firm costs but increase downstream care costs. Our analysis of the social planner's objectives motivates a counterfactual case assignment that minimizes expected total healthcare costs by allocating cases according to comparative advantage. 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. To identify the quality and cost effects of the knowledge hierarchy compared to the direct-to-doctor route, we must overcome selective assignment of cases across routes. We do so by estimating OLS specifications of the form:

$$Y_i = \delta_i Nurse_i + X_{i1}\beta_{1i} + X_{i2}\beta_{2i} + \epsilon_i, \tag{7}$$

where Y_i denotes the outcome of interest for case i, and $Nurse_i$ denotes the initial assignment to a nurse (as opposed to a doctor). The vector X_{i1} includes predetermined case characteristics likely used by the assignment algorithm (symptoms; symptoms interacted with age, gender, region, past prescriptions and whether prior case at the firm was with a doctor; and date-by–3-hour windows of patients' log in), while X_{i2} includes other predetermined characteristics (health risk and socioeconomic background), which we think the algorithm cannot observe. All the variables are described in Section 3.2. δ_i captures the (potentially heterogeneous) effect of initial nurse assignment – the knowledge hierarchy. ϵ_i is an error term. Since cases are not naturally clustered, we report

robust standard errors (Abadie and Cattaneo 2018) [5]

Prediction 1 motivates our identifying assumption that, conditional on factors likely affecting the algorithm's assignment (X_{i1}), the actual assignment is effectively random, implying $\mathbb{E}[\epsilon_i|X_{i1}] = 0$. To support this identification strategy and empirical specification (Prediction 1), we present two pieces of evidence. First, Figure 1 examines balance in initial case assignment across additional health risk and socioeconomic characteristics X_{i2}, conditional on baseline case characteristics X_{i1}. Assignment is balanced for most variables, and even when it is not, the differences are economically negligible.[6] Second, Table 3 reports the explanatory power of case characteristics and exogenous shifters for initial nurse assignment. Baseline characteristics X_{i1} alone account for a substantial share of the variation (adjusted $R^2 = 0.1945$, Column 1). Adding doctor congestion increases the explanatory power (adjusted $R^2 = 0.1998$, Column 2), but additional controls or exogenous assignment shifters do not (Columns 3 and 4).

First-come-first-served (FCFS). Appendix Table A4 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 modestly 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, as discussed in Section F.1 in the Appendix. Figure A2 and Table 3 test the model's (Prediction 1b) that doctor congestion increases the propensity for initial nurse assignment. Appendix Figure A4 illustrates this pattern within symptoms: moving from the lowest to the highest congestion quartile increases nurse assignment for every symptom. Appendix Table A5 shows that higher congestion (mechanically) increases patient wait times, but is not associated with substantial changes in case processing times, ratings, no-shows, technical errors, or very short consultations. These results support that residual variation in nurse assignment is driven by short-term congestion and motivate its use as an instrumental variable in our robustness analysis.

Queue composition. In the model, when there is a shortage of clinicians, overlap cases may be assigned to nurses if they arrive at a time t with more doctor-group cases, for which doctors are needed (Prediction 1c). Appendix Figure A5 shows that overlap-group cases are indeed more likely to be assigned to a nurse 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

[5] The results are similar when we cluster the standard errors by login-time windows.

[6] Moreover, we conduct two one-sided t-tests (TOSTs) for each additional case characteristic, conditional on the baseline controls. We can reject effects larger and smaller than 5% of the standard deviation of the treatment (initial nurse assignment) on the 5% significance level. A range of 20% is generally considered very conservative (Hartman and Hidalgo 2018).

(Table 3, Column 4), we focus on doctor congestion as our instrumental variable.

6 Results

This section uses Equation 7 to estimate the mean effects of case 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 4 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 D.

Within the firm, KH increases the probability of a clinician-booked doctor appointment within seven days by 31 percentage points from a baseline rate of 1.4%. This implies that nurses resolve almost 70% of the cases themselves. As a measure of unresolved cases, unbooked (re)visits of patients with the same symptom increase only slightly, by 0.8 percentage points from a baseline of 5.8%. KH does not reduce patient-reported satisfaction; if anything, the probability of receiving a top-score rating after the initial consultation increases by 1.2 percentage points, from a baseline of 40%. Appendix Table A6 supports these findings, showing that KH also slightly increases the probability of getting a top satisfaction rating within 7 days; that the rate of no-shows at the initial consultation barely increases; and the rate of technical errors decreases. KH reduces the probability of receiving an informative diagnosis, defined as a non-missing ICD code that is not a symptom or health status update, by 14 percentage points (from a baseline of 78%) in the initial consultation, or by 6.4% (from a baseline of 80%) within 7 days (Appendix Table A6).

We measure downstream healthcare use within seven 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 percentage points (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 chronic treatments. Nevertheless, 34% 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.[7]

Initial nurse assignment also increases external primary care consultations by 2.2 percentage points relative to a baseline of 11%. 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 the sample to Scania and Stockholm only (Appendix Table A6). The rate of

[7]The unconditional new prescription rate for cases passed to doctors in the KH, 42%, is comparable to that in cases assigned to doctors. In contrast, only 18% of cases that nurses resolve (within the firm) result in new prescriptions; these may be due to downstream care rather than to the nurses themselves.

visiting urgent care centers – out-of-hours primary care facilities to alleviate pressure on emergency care – increases by 0.3 percentage points from a baseline of 3.5%. Effects on specialist visits are small (0.3 percentage points from a baseline of 1.6%). We find limited increase in high-stakes downstream care: while emergency department visits rise by 0.5 percentage points from a baseline of 3.0%, hospitalizations are unaffected.

We find that KH does not worsen longer-term outcomes, measured by declines in labor income and mortality. If anything, KH reduces the probability of income declines from primary employment in the calendar month following the initial consultation, relative to average earnings in the preceding three months, for patients with stable prior employment. KH reduces the incidence of income declines of 20% or more by 2.1 percentage points from a baseline of 22%, and reduces the probability of zero income by 0.3 percentage points from a baseline of 5%. We estimate a precise zero effect on three-year mortality.[8]

6.2 Effects of the knowledge hierarchy on costs

Next, we examine the cost effects of the KH on patients' healthcare journeys. Table 5 reports OLS estimates of the mean effects of initial assignment to a nurse (the KH) on within-firm payments (Panel A) and on costs including downstream care (Panel B). Our cost measures are constructed using taxpayer payments for the categories listed in Appendix Table A11 and the model assumptions described in Section 4.

Within-firm costs. The KH substantially reduces taxpayer payments to the firm. 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 $\approx 28\%$ 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), and henceforth, we focus on those for the within-firm 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.

Costs including downstream care. For costs including downstream care, we consider the within-firm costs on the day of the initial consultation and count at most one instance of out-of-firm downstream event within seven days, assuming that downstream costs are all borne by taxpayers. Taxpayer payments reflect public healthcare spending, while total healthcare costs net out transfers between taxpayers and the firm. The KH reduces mean total taxpayer payments by 117 SEK, or about 9% relative to the baseline mean of 1,304.8 SEK. In contrast, mean total healthcare costs – which abstract from transfers between taxpayers and the firm – are 1179.8 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 within-firm costs or total healthcare production costs of care.

[8] Observed income reductions reflect sick leave, as employer-paid income falls to 80% for the first 14 days and to zero thereafter, when 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 earning losses after KH may partly reflect lower access to doctors (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 7, 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_{1i} and X_{2i}, which is computationally infeasible. To address this, we follow Chernozhukov et al. (2009) in reordering the estimated CDFs.

Subfigure 2a presents the CDFs of within-firm 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 costs (corresponding to cases that nurses resolve) but also likelier to result in higher firm costs (when nurses do not resolve cases).

Subfigure 2b shows that this pattern is inherited by total healthcare costs, which additionally incorporate downstream healthcare use. The KH almost halves costs for almost half of patients, while leaving costs in the upper tail largely unchanged.

6.3 Robustness

In this section we assess potential threats to the validity of our results and summarize corresponding robustness checks. We provide evidence consistent with conditional independence of initial nurse assignment given baseline case characteristics, examine the role of selective attrition before the initial consultation, assess robustness to cost construction and pricing assumptions, and evaluate robustness across settings and time periods. We also summarize the robustness of our distributional results for total healthcare costs.

Assessing identification. Our baseline OLS specification conditions on a rich set of predetermined case characteristics likely used by the assignment algorithm, and fine-grained time controls. 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 first probe this concern by varying controls and the assignment environment. Panel A of Appendix Table A8 shows that estimated effects on total healthcare costs are mostly unchanged when we (i) control only for baseline case characteristics, (ii) use finer date-by–1-hour fixed effects, and (iii) explicitly control for doctor congestion via congestion-quartile indicators. To address selection on unobserved patient traits, such as patient assertiveness, we restrict the sample to repeat patients and include patient fixed effects; this nets out time-invariant patient traits that could correlate with assignment. We see little change in estimates.

Appendix Figure A6 provides a complementary diagnostic based on outcome dynamics. Patients assigned directly to nurses have similar pre-consultation healthcare use as those assigned directly to doctors. In the week of the consultation, the KH increases the likelihood of a subsequent consultation,

reduces prescribing, and raises emergency care use relative to direct-to-doctor assignment, with these effects disappearing in the week after. The absence of pre-trends and the occurrence of effects only in the consultation week support our identification strategy and our preferred specification of a seven-day care window.

We assess remaining concerns about unobserved selection using two approaches. First, in Panel A of Appendix Table A8 we calculate Oster bounds. These bounds assess robustness to residual selection on unobservables by extrapolating coefficient and R^2 movements from the uncontrolled to the fully controlled specification to obtain a bias-adjusted estimate; the implied interval for the mean effect on total healthcare costs, –38.5 to 81 SEK, includes zero and remains small. Second, Appendix Table A7 complements these checks with an IV strategy based on Prediction 1b that leverages short-run fluctuations in doctor congestion. This strategy does not require our controls to capture the correct functional form of the assignment mechanism and therefore provides a useful check for misspecification of our OLS strategy. The IV estimates are less precise but deliver the same qualitative message as OLS: the KH changes within-firm case processing, while effects on downstream care and total healthcare costs remain small and statistically insignificant. Taken together, these results suggest that any unobserved selection between initial nurse assignment and potential outcomes is unlikely to be large enough to overturn our main findings.

Selective attrition before consultation. A potential threat to identification is selective attrition before the initial consultation. Appendix Section F.2 discusses this and the data we have on attrition in more detail. Panel B of Appendix Table A8 addresses this concern using two complementary approaches: bounding the maximum bias under a transparent selection structure, and probing whether the dropout channel is empirically important in settings where selection incentives or opportunities are limited.

First, we compute Lee bounds (Lee 2009) under the assumption of monotone selection. Within each case type, we trim the outcome distribution in the lower-attrition group to match the higher-attrition group. Because dropouts are unobserved in our data, we use administrative case-type-specific dropout rates from a later period and calculate the Lee bounds conditional on case types rather than our usual controls, X_1. Differential attrition is small in the separate sample where we measure all dropouts during registration, with a 2.68 percentage-point higher dropout rate on average for cases assigned to nurses. In our data, the attrition we measure is no shows and technical errors, and there the pattern is the opposite: a 0.3 percentage point higher attrition in the direct to doctor route. Still, since healthcare costs data have fat right tails, and we use levels instead of logs (which goes against us), unlike many other papers on healthcare, the Lee bounds will be very harsh. The calculated Lee bounds for the KH's mean total healthcare cost effects remain consistent with our baseline conclusion that average total healthcare cost effects are small.

Second, Panel B reports two restrictions designed to limit the scope for differential attrition. Restricting the sample to patients with prior nurse consultations—who have revealed that they do not select out of initial nurse assignments—yields similar mean effects on total healthcare costs. In

addition, exploiting a change in the assignment process around mid-2019, when clinician type was assigned only shortly before the consultation, we find comparable estimates before (April–June 2019: 103.7 SEK) and after the change (July–September 2019: 97.4 SEK), with overlapping confidence intervals.[9] Taken together, these results suggest that differential attrition is unlikely to drive our findings.

Cost construction and pricing. Our cost measures require choices about (i) which downstream utilization events to include and over what horizon, and (ii) how to map these events into SEK. Appendix Table A8 examines the robustness of our mean and distributional results under alternative assumptions and specifications.

Alternative cost aggregation. In our main specification, we aggregate downstream costs by including at most one additional within-firm visit and up to one of each out-of-firm care event within seven days of the initial consultation. Panel C of Appendix Table A8 reports mean total healthcare cost estimates under alternative cost aggregation schemes: (i) restricting the sample to cases in Stockholm and Scania, for whom downstream care including all external primary care is fully observed; (ii) extending the aggregation window to 30 days; and (iii) using cumulative costs, defined as the summed costs for all downstream events within 7 or 30 days. These alternative aggregations do not affect our main conclusions.

Alternative markdown assumptions. 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, consistent with existing monopsony markdown estimates implying $\theta \in [0.13, 0.33]$ (Section 3.2). Under this assumption, the KH has no significant mean effect on total healthcare costs. We assess the robustness of this result to alternative markdown assumptions by (i) varying the markdown and (ii) allowing markdowns to differ between nurses and doctors.

Panel D of Appendix Table A8 reports estimates for alternative values of θ: 0.1, 0.2, and 0.3. The results are almost unchanged, reflecting that within-firm costs constitute only a small share of total healthcare costs and most costs accrue in downstream care. Appendix Figure A7 shows robustness to differential markdowns by clinician types. Since the prices the firm receives are administratively set, it is implausible that the wedges differ greatly between clinician types. In practice, even varying wedges from $0.1 - 0.35$ and allowing them to differ by ± 0.1 only changes the proportional cost effects by $\approx 5\%$.

External validity. We assess sensitivity to two features of our setting: (i) the online-care context and (ii) the timing of our sample during the Covid-19 pandemic.

First, cases in online care are younger and more likely to be with female patients than those in physical primary care (Appendix Table A3). To address these demographic differences, we reweight our analysis sample to match the age-by-gender distribution of in-person primary care visits in Scania in 2019. Sampling weights are constructed from a probit model estimated on the combined

[9] While KH cost estimates are higher in this period than in the main analysis, this difference is unlikely to be driven by the Covid-19 pandemic, which we discuss separately below.

sample of the analysis sample and all nurse or doctor consultations in physical primary care centers, using age interacted with gender as predictors. Panel E of Appendix Table A8 and Panel F of Appendix Figure A8 report the weighted estimates, which are similar to our main results.

Second, our sample spans multiple phases of the Covid-19 pandemic: the onset and first wave (spring 2020), an interlude period with low incidence (summer 2020), and the second wave (fall 2020 onward). Although Sweden did not impose lockdowns and physical care remained accessible, public guidance to stay home may have affected care-seeking. Panel E of Appendix Table A8 reports estimates by pandemic phase. While the estimates vary across periods and indicate larger cost increases under the KH before the pandemic onset, these differences are unlikely to be driven by the pandemic itself: focusing on the two months before and after the onset yields nearly identical estimates, suggesting that temporal variation is not driven by sharp pandemic-related changes.

Distributional effects. Our baseline estimates imply a small and statistically insignificant mean effect of the KH on total healthcare costs, but the KH meaningfully reshapes the distribution of costs. Appendix Figure A8 summarizes the distributional patterns under our robustness checks. Across the IV estimates and sensitivity exercises (Oster and Lee bounds, alternative cost definitions and windows, and demographic reweighting), the distributional effects of the KH remain stable. The KH increases the probability of low-cost realizations—driven by cases resolved at the initial consultation without downstream care—while leaving costs in the upper tail largely unchanged.

7 Assignment of tasks

We have discussed the economic factors shaping the firm's use of the knowledge hierarchy (KH) and its quality and cost implications relative to the direct-to-doctor route. We now turn to case allocation: (i) how the firm's allocation of overlap-group cases between the two routes aligns with firm and taxpayer incentives, and (ii) how alternative allocations could improve productivity and reduce total healthcare costs of care provision.

7.1 The firm's assignment of tasks

First, we examine the firm's case assignment across case types in the overlap group, which in our setting corresponds to symptoms. Prediction 4 implies that the firm's use of the KH across symptoms is driven by relative net revenue advantage: initial assignment to nurse (the KH) should be more likely with a higher net revenue advantage of the KH over the direct-to-doctor route and lower relative taxpayer payments to the firm. The model further predicts that this relationship strengthens with doctor congestion, as capacity constraints make such allocation more profitable for the firm and more closely aligned with taxpayer incentives.

For each case type (symptom), we measure the KH's relative advantage over the direct-to-doctor route for four outcome measures: (i) *total healthcare costs*; (ii) *firm net revenues*; (iii) *taxpayer payments to the firm*; (iv) *total taxpayer payments*. We relate these measures to the firm's propensity

to assign a given symptom to the KH by plotting predicted symptom-level outcomes against predicted assignment probabilities.

Estimation. To estimate the KH's symptom-specific relative advantage, we allow the effect of initial nurse assignment to vary by symptom in Equation 7:

$$Y_{i,k(i)} = \delta_k \, Nurse_i \times Symptom_{k(i)} + X_{1i}\beta_1 + X_{2i}\beta_2 + \varepsilon_i, \tag{8}$$

where $Symptom_{k(i)}$ denotes indicators for symptom k and δ_k captures the KH's heterogeneous effects by symptom. $Y_{i,k(i)}$ denotes the cost or revenue outcome for a case i with symptom k's.

Using the estimated coefficients, we predict symptom-specific outcomes under initial nurse assignment, $\hat{Y}_{i,k}(n)$, and under direct doctor assignment, $\hat{Y}_{i,k}(d)$. For a given outcome, we construct as measure of the KH's relative advantage, the ratio of the sum of values under initial nurse assignment versus the sum under doctor assignment, $\sum \hat{Y}_{i,k}(n) / \sum \hat{Y}_{i,k}(d)$.

To measure the firm's propensity to assign a symptom to the KH, we regress nurse assignment on doctor congestion and the full set of controls (the first-stage specification for the congestion instrument, see Equation 19 in the Appendix), and compute the average predicted probability of initial nurse assignment for each symptom. Finally, we relate symptom-level assignment probabilities to relative advantages using linear regressions weighted by symptom frequency. Standard errors are obtained by bootstrapping the full procedure with 100 replications. In each draw, we re-estimate relative advantages, predicted assignment probabilities, and the weighted regression of assignment on relative advantage; 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 implied by the model: that the KH's relative advantage in firm net revenues is driven by nurses' ability to resolve cases without passing them to doctors (p_τ). For cases initially assigned to a nurse, we regress the KH's relative firm net revenue advantage on the observed symptom-level rate of nurses resolving cases themselves ($\hat{p}_\tau$), and obtain a regression coefficient of 1.212 (bootstrap SE: 0.006, R^2: 0.993).[10]

Next, Figure 3 relates the firm's case assignment to costs and revenues. Subfigure 3a shows that the firm is more likely to assign a symptom initially to a nurse when the KH entails lower relative total healthcare costs. Consistent with Prediction 4, the firm's assignment of symptoms is driven by relative net revenue advantage under the KH (Subfigure 3b), which is also associated with lower taxpayer payments to the firm (Subfigure 3c). Subfigure 3d further reveals that, empirically, the firm's assignment reduces total taxpayer payments, thus not increasing downstream care costs. Importantly, the firm's assignment follows *relative* revenue advantages: in absolute terms, firm revenues are higher under direct doctor assignment for all cases.

Figure 4 examines how the alignment between the firm's incentives and taxpayer objectives varies with doctor congestion. Regressions of the initial nurse assignment propensity on relative

[10] In the model, the corresponding derivative is exactly 1.2; the slight difference arises because the empirical measure of firm net revenues includes the costs of same-day drop in consultations.

net revenue advantage by congestion quartile show significantly steeper slopes at higher congestion levels, indicating that the firm relies more heavily on relative net revenue advantages when doctor capacity binds. For total taxpayer payments, we observe a similar but weaker pattern, consistent with the firm not internalizing downstream care costs in its case assignment.

Overall, limited incentives and information about downstream costs lead the firm to prioritize absolute net revenues except during periods of doctor congestion. When the firm is constrained by doctor congestion, its assignment follows relative net revenue advantages, reducing firm net revenues and aligning more closely with the minimization of taxpayer payments.

7.2 Counterfactual assignment of tasks

We next study how alternative allocation rules for overlap-group cases affect total healthcare costs, focusing on absolute and comparative advantage across case types (symptoms) in allocating cases to the two routes, initial nurse assignment (the knowledge hierarchy, KH) or direct-to-doctor assignment. The firm's observed assignment reflects short-run capacity constraints and within-firm incentives and does not internalize downstream care costs. We compare this observed allocation to two sets of counterfactuals: (i) non-optimized assignments that ignore comparative advantage across case types; and (ii) cost-minimizing assignments that follow a social planner's objective by internalizing downstream costs and allocating cases by absolute or comparative advantage (Prediction 5).

Setting up the social planner's objective. Condition 6 characterizes the social planner's assignment rule: a symptom should be assigned to the KH only if and only if its expected total healthcare costs are lower than under direct-to-doctor assignment. To operationalize this objective, we first estimate expected total healthcare costs under each route for every symptom. Specifically, we estimate Equation 8, allowing heterogeneous effects on costs of care production by symptom, using *firm costs*, *out-of-firm downstream costs*, and *total healthcare costs* as outcomes. We then average predicted costs across all cases of a given symptom under each route.

We interpret the KH route as having an *absolute advantage* in producing care for a symptom, if it yields lower production costs (that is, lower total healthcare costs) than the expert route. To capture comparative advantage, we compute symptom-level ratios of total healthcare costs across routes and rank symptoms by their relative cost savings under the KH.

Figure 5 shows substantial heterogeneity across symptoms in the cost difference between the KH and the direct-to-doctor route. Subfigure 5a shows the ranking of symptoms by the ratio of predicted total healthcare costs for assignment to nurses (the KH) relative to doctors. This figure can be thought of as reflecting the comparative advantage of KH, in the spirit of Dornbusch et al. (1977). Subfigure 5b reports the absolute effects of assignment to nurses (the KH), and its breakdown into the effects on firm costs and downstream costs. In most cases, the downstream cost effects dominate, suggesting an important role for changing the initial reassignment of tasks between nurses and doctors to utilize comparative advantage, to mitigate the downstream externalities.

To provide intuition, we examine common downstream cost drivers. For symptoms such as sore throat and urinary tract infection – where doctors have the largest comparative advantage

and downstream costs are higher under the KH – assignment to the KH significantly increases the probability of ED visits and hospitalizations within seven days relative to direct-to-doctor assignment. The most frequent ED and hospitalization diagnoses after such online symptoms include peritonsillar abscess, acute tonsillitis, and acute tubulo-interstitial nephritis, conditions associated with bacterial infection. Consistent with this mechanism, antibiotic prescriptions (e.g., beta-lactam antibacterials) are common following online visits with sore throat and UTI symptoms. These patterns suggest that faster or more certain access to a doctor – and thus to antibiotic treatment – reduces the risk of escalation to emergency or inpatient care for these symptoms.

On the other side of the spectrum, on the left hand side of Figure 5, there is Covid-19 and diarrhea/vomiting, which have higher downstream costs when seen online directly by a doctor. Covid-19 and diarrhea/vomiting have significantly lower ED and hospitalizations after assigned to the KH compared to when assigned directly to doctors. One reason could be that there is also a significantly lower likelihood of being given a prescription drug when going through the KH. Prescription drugs could have side effects, and in particular for diseases with evolving knowledge about how to treat them, like Covid-19, might not be effective. Care advice by a nurse, or even more care advice by both nurse and doctor combined, could be more effective.

Defining counterfactual case assignments. Using our estimates of expected total healthcare costs, we define several counterfactual assignment rules.

As benchmarks, we consider non-optimized case allocations that ignore cost or revenue differences across symptoms. First, we examine *allocations proportional to actual nurse and doctor availability*, in which cases are assigned uniformly across symptoms to the KH or the expert route in proportion to observed nurse and doctor availability within short time windows. We measure availability in two ways: (i) prior to a new case, based on the number of cases fully processed (with consultation and administrative work completed) by each clinician type in the 15 minutes or 3 hours preceding a case's login; and (ii) fixed windows, based on the number of cases assigned within fixed 15-minute or 3-hour intervals to each clinician type. In each case, the observed nurse share defines the probability that a new case is assigned to a nurse. These allocations capture scenarios in which the allocation of cases is effectively random rather than determined by the firm's relative net revenue advantage across case types.

Second, we consider extreme *allocations in which all cases are assigned to a single route*, either initially to nurses (the KH) or directly to doctors (the expert route). Assigning all cases initially to nurses corresponds to a scenario in which the firm is forced to process all cases in the KH route. Assigning all cases to doctors instead eliminates the KH entirely, corresponding to the baseline of assigning all cases to the expert route from Section 6.

We then consider case re-allocations that minimize total healthcare costs. In an unconstrained scenario, we assume that nurse capacity can be expanded without affecting wages or prices. This allows assignment by *absolute advantage*: all cases of a symptom are assigned to the KH if its expected total healthcare cost under KH is lower than under the expert route, and to doctors otherwise – that is, the cost ratio in Subfigure 5a is below one. This directly implements the social

planner's objective.

In a constrained scenario, we respect short-run clinician capacity within fixed time windows (15 minutes; 3 hours). This yields an assignment according to *comparative advantage*. Within each fixed time window, we fix the total number of cases initially assigned to nurses to match observed nurse availability and reallocate cases across symptoms to minimize total healthcare costs, according to the ranking of symptoms by their relative total healthcare cost advantage of KH over the doctor route in Subfigure 5a. This allocation assigns cases to the available nurse capacity, starting with those symptoms that yield the largest cost savings, and moving down the ranking until nurse capacity is exhausted. When nurse capacity binds, one symptom within each window is assigned probabilistically.

Results. To compare each counterfactual assignment to the firm's observed assignment, we summarize productivity differences as percentage changes in aggregate costs and revenues. Specifically, we consider aggregate: (i) total healthcare costs; (ii) total taxpayer costs; and (iii) firm net revenues. For a given outcome, we compute the percentage change as $\sum \Delta\tilde{Y}_i / \sum \hat{Y}_i$, where $\Delta\tilde{Y}_i = \tilde{Y}_i - \hat{Y}_i$ denotes the difference in outcome predictions under case i's counterfactual ($\tilde{Y}_i$) and observed ($\hat{Y}_i$) assignment. The numerator aggregates these differences across all overlap-group cases, while the denominator normalizes by aggregate predicted costs or revenues under the observed assignment. We additionally report changes in aggregate total healthcare costs normalized by aggregate firm costs, which we interpret as cost changes relative to the total production costs of online primary care.

Table 6 reports the calculated percentage changes in costs and revenues. Panel A considers non-optimized assignments, in which the KH's heterogeneous cost effects across symptoms are ignored. Allocations 1)-4) assign cases proportionally to clinician availability. These allocations increase total healthcare costs between 0.55-0.58%.[11]

Allocations 5)-6) assign all cases to one of the routes. Assigning all cases to the KH route increases the number of initial nurse assignments by roughly 340%. This raises total healthcare costs by 0.79%. Assigning all cases to the expert route also increases total healthcare costs: Eliminating the KH route would increase total healthcare costs by 0.71% (2.33 percent of online primary care costs).

Panel B implements the social planner's objective. Allocation (1) shows the unconstrained optimum: if the firm could expand nurse capacity holding all else fixed, implementing absolute advantage increases use of the KH route by about 170%. This allocation would lower total healthcare costs by 4.07%, reducing taxpayer payments but also firm net revenues. Allocations (2) and (3) impose short-run capacity constraints by fixing nurse assignments within 3-hour or 15-minute windows. Even under these constraints, reallocating cases across symptoms by comparative advantage yields sizable gains. A reallocation within 15-minute windows reduces total healthcare costs by about 2.1% (6.7% relative to online primary care costs), cutting taxpayer payments by 1.91% and modestly reducing firm net revenues by 0.28%.

[11]When availability is defined using fixed windows, the number of nurse-assigned cases is held fixed; defining availability based on prior processing allows this number to vary. Given otherwise similar results across the two definitions of clinician availability, we use fixed windows in the optimized constrained assignments.

We construct total healthcare costs by combining firm costs on the initial consultation date with downstream care costs within seven days, counting at most one event per care category. Table 5 shows that KH effects on within-firm care are similar when outcomes are aggregated over seven days rather than measured on the initial consultation date. Appendix Table A9 shows that the estimated percentage changes are essentially unchanged when all downstream care events are counted individually; if anything, the productivity gains from any KH use and from the optimized assignments become larger.

Appendix Figure A9 illustrates how the optimal allocation is implemented in practice by relating the KH's comparative advantage to both the observed and the optimal constrained case allocations. Symptoms with high predicted relative costs under the KH, such as sore throat or urinary tract infection, are reallocated such that they have a lower probability of initial nurse assignment, while symptoms with low predicted relative costs, such as constipation and abdominal pain, are reallocated to a higher probability of initial nurse assignment.

Discussion. Taken together, the counterfactual analysis shows that the firm's observed use of the knowledge hierarchy already yields meaningful productivity effects. Not assigning cases selectively, given the existing clinician capacity, would increase total healthcare costs, as well as making both the firm and taxpayer worse off, reflecting that the firm's actual assignment is partly aligned with taxpayer incentives (Figure 3). Eliminating the knowledge hierarchy entirely would further reduce productivity: Relative to only using the expert route, the firm's assignment saves 0.7% of total healthcare costs (2.33% of online primary care costs). Nonetheless, additional savings are possible: Reallocating cases by comparative advantage across symptoms – while compensating the firm for lost revenues – could reduce total healthcare costs by a further 2.1% (6.7% relative to online primary care costs).

The potential for productivity gains suggested by our counterfactual analysis arises partly because the firm lacks information about patients' downstream outcomes. Even if such information were available, the firm would not have incentives to internalize downstream cost effects under the current administratively set price structure.

A simple corrective is to issue guidelines that steer the allocation of cases toward lower total healthcare costs. It is not unusual for healthcare payers to issue guidelines to prioritize cost-effective care, such as those mentioned in Section 2, or the UK NICE guidelines. In the firm we study, such guidelines could be implemented easily by adjusting the algorithm that initially assigns cases to nurses or doctors. Hence, these guidelines would not burden clinicians with more checklists or administrative tasks. In the short run, with staffing held fixed, the constrained optimization prioritizes case types by comparative advantage in total healthcare costs within each time window. Figure 5 indicates sizeable differences in comparative advantage over symptoms, which could be codified into general symptom-specific guidelines. Such guidelines may, in principle, affect the firm's hiring or staffing incentives; however, assessing such changes is beyond the scope of this paper.

In the longer run, changing the relative availability of clinicians, such as by expanding the training of nurses in public universities, could move the allocation closer to the unconstrained solution, in

which a larger share of cases is assigned first to nurses.

8 Conclusion

In knowledge-intensive sectors where expertise is important and scarce, knowledge hierarchies are often used. However, despite the ubiquity of KH, evidence on the productive effects has been limited. We contribute by studying KH in a setting where cost 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. Using quasi-experimental algorithmic variation in the assignment of patient cases, we find that KH reduces the rates of diagnoses and new prescriptions and slightly increases ED visits, but its mean effects on patient satisfaction and high-stakes outcomes — including hospitalizations, labor market earnings losses, and three-year mortality — are close to zero. KH reduces overall taxpayer payment, but also the firm's net revenues, so the mean net effect on total healthcare costs is zero. However, these mean effects mask substantial heterogeneity, as the KH increases the likelihood of both low costs (when nurses resolve cases) and higher costs (when nurses pass cases to doctors). Moreover, the effects of KH vary significantly by symptom. We show that the firm's selective use of KH saves some 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 KH by symptom on downstream healthcare outcomes.

References

Abadie, Alberto and Matias D. Cattaneo, "Econometric Methods for Program Evaluation," *Annual Review of Economics*, August 2018, *10* (1), 465–503.

Acemoglu, Daron and David Autor, "Skills, Tasks and Technologies: Implications for Employment and Earnings," in "Handbook of Labor Economics," Vol. 4, Elsevier, 2011, pp. 1043–1171.

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

Alexander, Diane and Molly Schnell, "Just what the nurse practitioner ordered: Independent prescriptive authority and population mental health," *Journal of Health Economics*, July 2019, *66*, 145–162.

Azar, José and Ioana Marinescu, "Chapter 10 - Monopsony power in the labor market," in Christian Dustmann and Thomas Lemieux, eds., *Handbook of Labor Economics*, Vol. 5, Elsevier, January 2024, pp. 761–827.

Becker, Gary S. and Kevin M. Murphy, "The Division of Labor, Coordination Costs, and Knowledge," *The Quarterly Journal of Economics*, 1992, *107* (4), 1137–1160. Publisher: Oxford University Press.

Bhuller, Manudeep, Gordon B Dahl, Katrine V Løken, and Magne Mogstad, "Incarceration, Recidivism, and Employment," *Journal of Political Economy*, 2016.

Bloom, Nicholas, Luis Garicano, Raffaella Sadun, and John Van Reenen, "The Distinct Effects of Information Technology and Communication Technology on Firm Organization," *Management Science*, 2014, *60* (12), 2859–2885.

Bronsoler, Ari, Joseph Doyle, and John Van Reenen, "The Impact of Health Information and Communication Technology on Clinical Quality, Productivity, and Workers," *Annual Review of Economics*, August 2022, *14* (Volume 14, 2022), 23–46. Publisher: Annual Reviews.

Caliendo, Mion, Opromolla & Rossi-Hansberg, "Productivity and Organisation in Portuguese firms," *Journal of Political Economy*, 2020, *128* (11), 4211–4257.

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

Cattaneo, Matias D., Richard K. Crump, Max H. Farrell, and Yingjie Feng, "On Binscatter," *American Economic Review*, May 2024, *114* (5), 1488–1514.

Centers for Medicare & Medicaid Services (CMS), "Medicare Telehealth Trends — CMS Data," 2025. Accessed on October 9, 2025.

Chan, David and Yiqun Chen, "The Productivity of Professions: Evidence from the Emergency Department," *NBER Working Paper*, October 2022.

Chernozhukov, V., I. Fernandez-Val, and A. Galichon, "Improving point and interval estimators of monotone functions by rearrangement," *Biometrika*, September 2009, *96* (3), 559–575.

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

Dahlstrand, Amanda, "Defying Distance? The provision of services in the digital age," *Working Paper*, 2024.

_ , Nestor Le Nestour, and Guy Michaels, "Online versus In-Person Services: Effects on Patients and Providers," 2024. Mimeo.

Dillender, Marcus, Anthony T. Lo Sasso, Brian J. Phelan, and Michael Richards, "Occupational Licensing and the Healthcare Labor Market," *Journal of Human Resources*, February 2024, pp. 0722–12450R2.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang, "The Effects of Pre-Trial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, February 2018, *108* (2), 201–240.

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

European Observatory on Health Systems and Policies, "Health System Summaries," n.d. Accessed on January 12, 2026. https://eurohealthobservatory.who.int/publications/health-systems-reviews/health-system-summaries.

_ , "Health Systems Reviews," n.d. Accessed on January 12, 2026. https://eurohealthobservatory.who.int/publications/health-systems-reviews.

Folkhälsomyndigheten, "Veckorapport om covid-19, vecka 27," https://www.folkhalsomyndigheten.se/globalassets/statistik-uppfoljning/smittsamma-sjukdomar/veckorapporter-covid-19/2020/covid-19-veckorapport-vecka-27-final.pdf July 2020. Public Health Agency of Sweden weekly report on COVID-19 situation in Sweden.

Försäkringskassan, "Sjuklön: När behöver medarbetaren läkarintyg?," 2025. Accessed: 2026-01-26.

Garicano, Luis, "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, 2000, *108* (5), 874–904. Publisher: The University of Chicago Press.

Godøy, Anna, Venke F. Haaland, Ingrid Huitfeldt, and Mark Votruba, "Hospital Queues, Patient Health, and Labor Supply," *American Economic Journal: Economic Policy*, May 2024, *16* (2), 150–81.

Gottlieb, Joshua D, Neale Mahoney, and Victoria Udalova, "The Rise of Healthcare Jobs," 2025.

Guo, Jiapei, Angela E. Kilby, and Mindy S. Marks, "The impact of scope-of-practice restrictions on access to medical care," *Journal of Health Economics*, March 2024, *94*, 102844.

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

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

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

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

Kry International AB, "Annual Report and Consolidated Accounts," 2024. Accessed on January 13, 2026.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 2009, *76* (3), 1071–1102. Publisher: [Oxford University Press, The Review of Economic Studies, Ltd.].

Maier, Claudia B. and Linda H. Aiken, "Task shifting from physicians to nurses in primary care in 39 countries: a cross-country comparative study," *European Journal of Public Health*, December 2016, *26* (6), 927–934.

McMichael, Benjamin J., "Supply-side health policy: The impact of scope-of-practice laws on mortality," *Journal of Public Economics*, June 2023, *222*, 104901.

Ministry of Health, Welfare and Sport, "The Right Care in the Right Place," Technical Report, Ministry of Health, Welfare and Sport (Netherlands), The Hague 2018.

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

Myndigheten för vård- och omsorgsanalys, "Besök via nätet," 2022. Rapport 2022:1.

National Center for Health Statistics, "2023 National Health Interview Survey (NHIS)," 2023. Accessed on October 28, 2024.

NHS England, "Principle 3 – Right care, right time, right place for carers," n.d. Accessed: 2026-01-20.

Norra sjukvårdsregionförbundet, "Price Lists of Northern Healthcare Region," https://www.norrasjukvardsregionforbundet.se/halso-och-sjukvard/avtal-och-priser/arkiv/ 2018. Last accessed December 16, 2024.

OECD, "The COVID-19 Pandemic and the Future of Telemedicine," 2021. Accessed on October 9, 2025.

OECD, "Health at a Glance," 2021. Accessed on October 9, 2025.

_ , "Health Statistics Data Explorer: Doctor Consultations," 2025. Accessed on October 9, 2025.

_ , "Health Statistics Data Explorer: Physicians," 2025. Accessed on October 9, 2025.

_ , "OECD Health Statistics Data Explorer: Practising nurses," 2025. Accessed on October 9, 2025.

Rajpurkar, Pranav, Emma Chen, Oishi Banerjee, and Eric J. Topol, "AI in health and medicine," *Nature Medicine*, January 2022, *28* (1), 31–38. Publisher: Nature Publishing Group.

Region Kronoberg, "Riktlinje för registrering av vårdkontakter," April 2025.

Region Skane, "Regionala riktlinjer för registrering och klassificering i Region Skåne," January 2023.

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

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

Smith, Adam, *An Inquiry Into the Nature and Causes of the Wealth of Nations* 1776.

Socialdepartementet, "Effektiv vård: Slutbetänkande av En nationell samordnare för effektivare resursutnyttjande inom hälso- och sjukvården," 2016. Published 13 January 2016.

_ , "Hälso- och sjukvårdslag (2017:30)," February 2017.

Socialstyrelsen, "Bedömning av tillgång och efterfrågan på legitimerad personal i hälso- och sjukvård samt tandvård," February 2022.

Statistics Sweden (SCB), "Wages and salaries by occupation," Statistics Sweden Statistical Database 2020. Accessed 13 January 2026.

Sveriges Kommuner och Landsting (SKL), "Meddelande från styrelsen - Utvecklingen av digitala vårdtjänster," June 2019.

Sveriges Kommuner och Regioner, "Ekonomi- och verksamhetsstatistik inom hälso- och sjukvården," 2024. Accessed on October 28, 2024.

Sveriges Riksbank, "Exchange rates – SEK to USD (daily rates 2020)," Sveriges Riksbank statistics 2020. Accessed 13 January 2026.

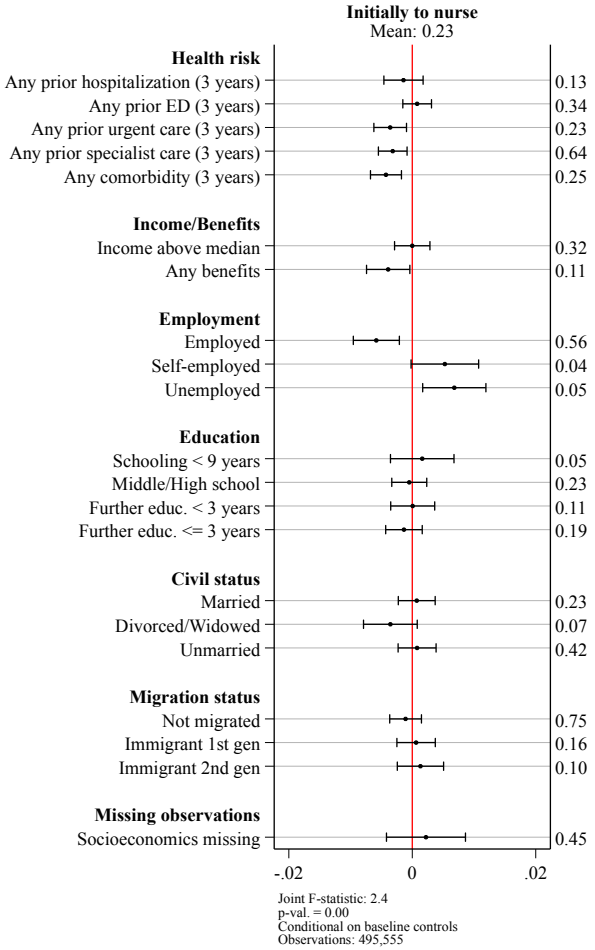
Södra sjukvårdsregionen, "Reports About Healthcare Costs," https://sodrasjukvardsregionen.se/verksamhet/avtal-priser/regionala-priser-och-ersattningar-foregaende-ar/ 2020. Last accessed February 13, 2024.

The Commonwealth Fund, "Sweden | International Health Care System Profiles," 2020. Accessed on January 13, 2026.

Traczynski, Jeffrey and Victoria Udalova, "Nurse practitioner independence, health care utilization, and health outcomes," *Journal of Health Economics*, March 2018, *58*, 90–109.

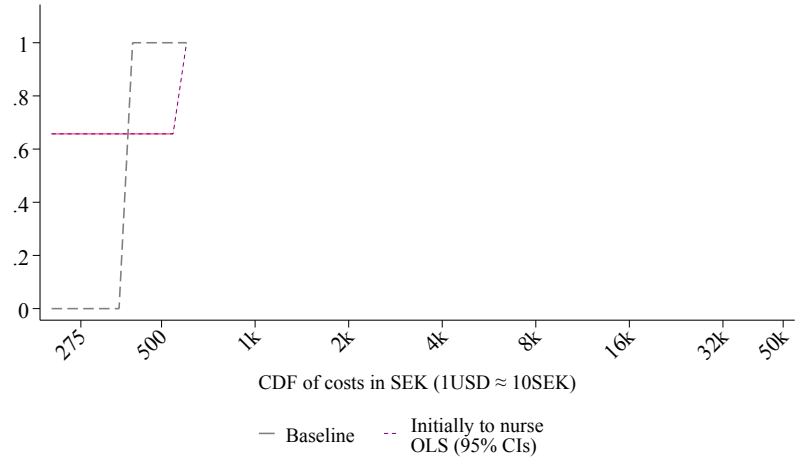
Main Figures

Figure 1. Balance of the treatment in case characteristics

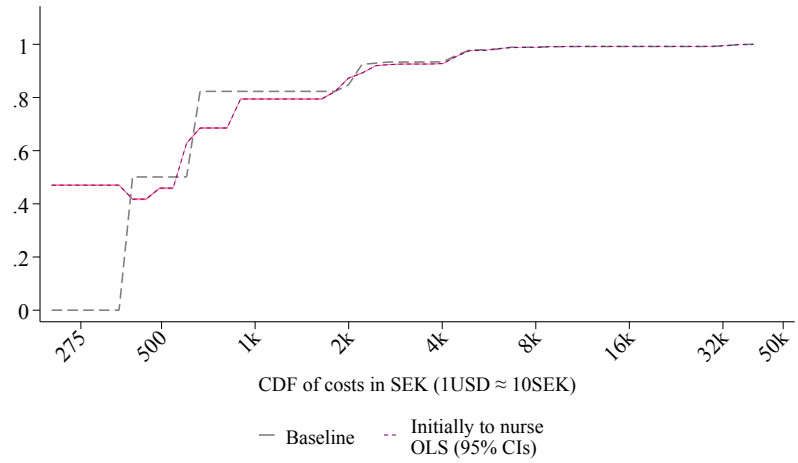


Notes: This figure presents balance tests for the treatment variable (*Nurse*). Each row shows the results of a bivariate regression where the treatment is regressed on a specific case characteristic. All regressions account for baseline case characteristics: login time, main symptom, and patient demographics and the demographics × past firm use interaction term. Login time indicates date-by-3 hours fixed effects. Symptom categories refer to indicators for the main symptom category reported by the patient when requesting a consultation. Demographics include indicators for patient gender, age categories, and aggregated regions. Interaction term includes age, gender, symptom, indicators for previous doctor consultation and prescription pickup from provider The horizontal lines represent 95% confidence intervals based on robust standard errors. The F-statistics reflect the joint F-tests for the treatment when regressed on all control variables, conditioned on login time. The numbers on the far right display the sample means for each case characteristic.

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



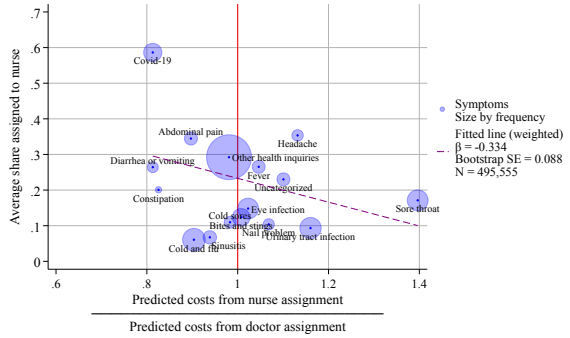
(a) Distribution of firm costs



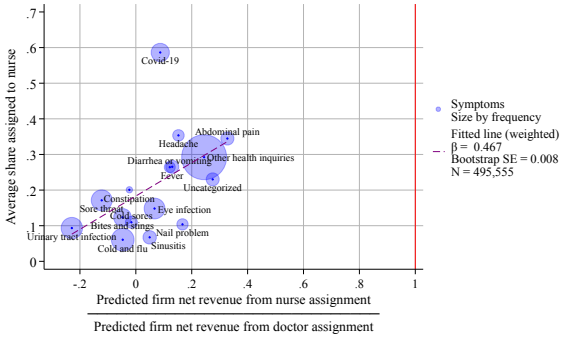
(b) Distribution of total healthcare cost

Notes: These figures presents the probabilities that the firm and total healthcare cost fall below the threshold shown on the x-axis. The y-axis shows estimated probabilities from an OLS regression based on Equation 7, with full controls where the dependent variable is an indicator exceeding various ln(total healthcare cost/firm costs) thresholds at 0.1 intervals. The 95% confidence intervals are also visualized around the estimates.

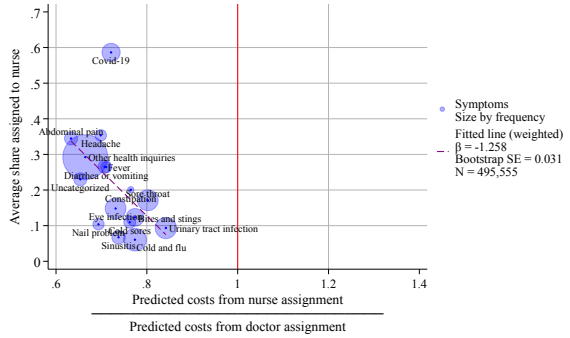
Figure 3. Relative revenue and cost advantage across case types



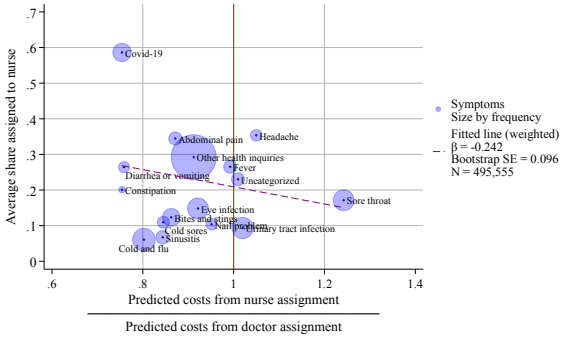
(a) Total healthcare costs



(b) Firm net revenue



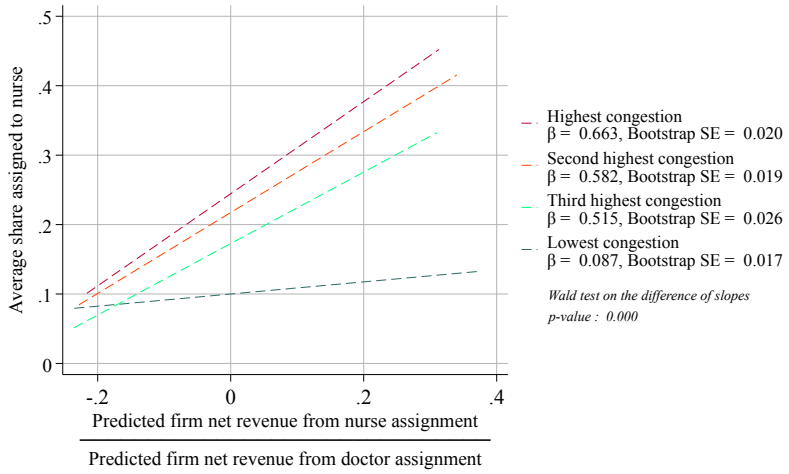
(c) Taxpayer payments to the firm



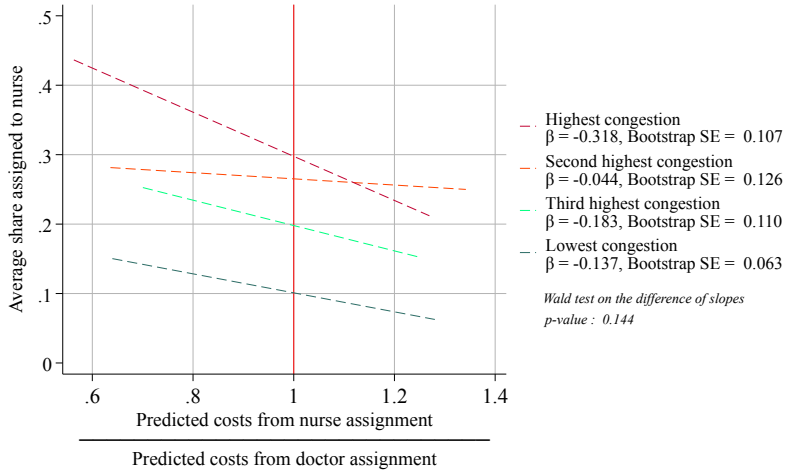
(d) Total taxpayer payments

Notes: Each panel shows the firm's case assignment by relative revenue or cost advantage of an initial nurse versus initial doctor assignment across symptom groups for the following outcomes: total healthcare costs (Panel 3a), the firm's net revenue (Panel 3b), taxpayer payments to the firm (Panel 3c), and total taxpayer payments (Panel 3d). The y-axis plots the average probability of initial nurse assignment, which is equivalent to a linear prediction from the first-stage regression (Equation 19). The x-axis plots the relative revenue or cost advantage under initial nurse relative to initial doctor assignment based on predicted outcomes, derived from a modified version of Equation 7 interacting symptom groups with the nurse indicator $Nurse_i$. The relative cost or revenue advantage is constructed by summing predicted outcomes within each symptom group and taking the ratio. Circles represent symptom frequencies. Fitted lines are weighted OLS regressions. Standard errors are computed from 100 bootstrap replications with resampling at the individual consultation level. The vertical red line indicates equal costs or revenues across assignment routes. All regressions use the full set of controls and the analysis sample.

Figure 4. Relative revenue and cost advantage across case types, by congestion



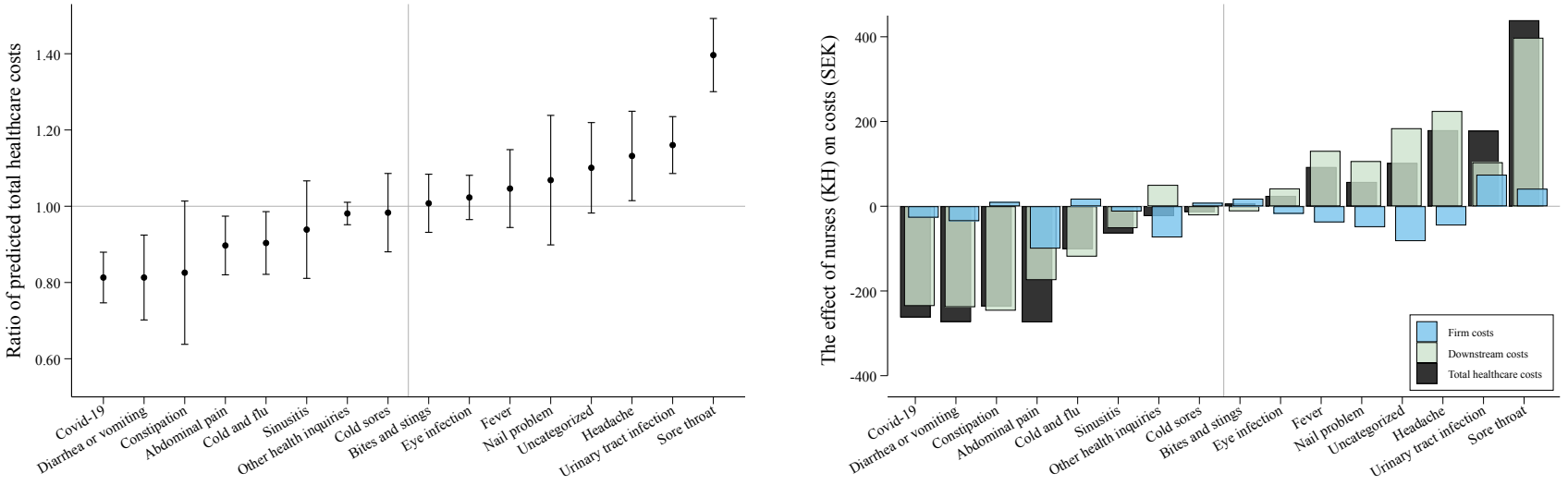
(a) Relative firm net revenue advantage



(b) Relative total taxpayer payment advantage

Notes: These figures show the relative revenue and cost advantage of an initial nurse versus initial doctor assignment across symptom groups by congestion quartile. The y-axis plots the predicted probability of initial nurse assignment from the first-stage regression (Equation 19). The x-axis plots the relative revenue or cost advantage under initial nurse relative to initial doctor assignment based on predicted outcomes, derived from a modified version of Equation 7 that interacts symptom groups with the nurse indicator $Nurse_i$. Relative advantages are constructed by summing predicted outcomes within each symptom group and taking ratios. Dashed lines show weighted regression fits, using symptom frequencies as weights. Standard errors are computed from 100 bootstrap replications with resampling at the individual consultation level.

Figure 5. KH effect and average predicted cost by symptoms



(a) Comparative advantage

(b) Breakdown of total healthcare costs

Notes: These figures show heterogeneous treatment effects of KH across symptom groups. The first two panels show the effects of nurse assignment (KH) on firm costs and downstream costs, respectively, estimated by interacting symptom indicators with nurse assignment. The third figure shows the predicted total healthcare costs for initial nurse and initial doctor assignments for each symptom group, derived from a modified version of Equation 7 that interacts symptom groups with the nurse indicator $Nurse_i$. The predicted costs are then averaged within each symptom group. For the first eight symptom groups, the predicted costs are higher with initial doctor assignment than initial nurse assignment, indicating that nurses have an absolute advantage in these cases. The vertical line marks the point at which this absolute advantage in costs reverses.

Main Tables

Table 1. Healthcare comparison

	Indicator	Countries
(1)	Telehealth share in 2019 (15 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 (23 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; SL 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 (31 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

Note: This table compiles health-system indicators from multiple sources, focusing on OECD countries where data is available. Rows (1), (2) and (3) use data from the OECD (2025a), National Center for Health Statistics (2023), and the Centers for Medicare & Medicaid Services (CMS). Row (4) uses data from OECD (2021b). Row (5) uses data from the OECD (2025c) and OECD (2025b). Row (6) and (8) use information from OECD (2021a). Row (7) uses information from the latest WHO Health System Summary (European Observatory on Health Systems and Policies n.d.a) or Health System Review (European Observatory on Health Systems and Policies n.d.b) for each country. Countries are identified using ISO 3166-1 alpha-2 codes. Figures pertain to 2023 unless otherwise specified.*Teleconsultations include only those reimbursed through the national healthcare system.

Table 2. Characteristics of the analysis sample

	Full sample		Initially to nurse		Directly to doctor		Diff. p-value
Symptom categories							
Abdominal pain	0.030	[0.17]	0.046	[0.21]	0.025	[0.16]	0.00
Cold and flu	0.089	[0.28]	0.024	[0.15]	0.11	[0.31]	0.00
Cold sores	0.025	[0.16]	0.012	[0.11]	0.029	[0.17]	0.00
Constipation	0.0070	[0.083]	0.0062	[0.078]	0.0072	[0.085]	0.00
Covid-19	0.058	[0.23]	0.15	[0.36]	0.031	[0.17]	0.00
Diarrhea or vomiting	0.021	[0.14]	0.024	[0.15]	0.020	[0.14]	0.00
Eye infection	0.075	[0.26]	0.049	[0.22]	0.083	[0.28]	0.00
Fever	0.029	[0.17]	0.034	[0.18]	0.028	[0.16]	0.00
Headache	0.023	[0.15]	0.036	[0.19]	0.019	[0.14]	0.00
Nail problem	0.022	[0.15]	0.010	[0.10]	0.026	[0.16]	0.00
Other health inquiries	0.35	[0.48]	0.45	[0.50]	0.32	[0.47]	0.00
Bites and stings	0.054	[0.23]	0.030	[0.17]	0.062	[0.24]	0.00
Sinusitis	0.031	[0.17]	0.0093	[0.096]	0.038	[0.19]	0.00
Sore throat	0.076	[0.26]	0.057	[0.23]	0.081	[0.27]	0.00
Uncategorized	0.030	[0.17]	0.030	[0.17]	0.029	[0.17]	0.42
Urinary tract infection	0.079	[0.27]	0.032	[0.18]	0.092	[0.29]	0.00
Demographics							
Female	0.63	[0.48]	0.59	[0.49]	0.64	[0.48]	0.00
Patient age	29.2	[16.5]	29.3	[16.6]	29.2	[16.5]	0.57
West Sweden	0.20	[0.40]	0.20	[0.40]	0.20	[0.40]	0.07
Stockholm	0.44	[0.50]	0.45	[0.50]	0.44	[0.50]	0.00
Middle Sweden	0.18	[0.39]	0.18	[0.39]	0.19	[0.39]	0.00
South Sweden	0.095	[0.29]	0.093	[0.29]	0.095	[0.29]	0.01
Norrland	0.034	[0.18]	0.031	[0.17]	0.035	[0.18]	0.00
Småland + the islands	0.042	[0.20]	0.042	[0.20]	0.042	[0.20]	0.25
Health risk factors							
Any prior hospitalization (3 years)	0.13	[0.34]	0.14	[0.34]	0.13	[0.34]	0.44
Any prior ED (3 years)	0.34	[0.47]	0.35	[0.48]	0.34	[0.47]	0.00
Any prior urgent care (3 years)	0.23	[0.42]	0.22	[0.41]	0.23	[0.42]	0.00
Any prior specialist care (3 years)	0.64	[0.48]	0.63	[0.48]	0.64	[0.48]	0.00
Any comorbidity (3 years)	0.25	[0.43]	0.25	[0.43]	0.25	[0.43]	0.01
Socioeconomic characteristics							
Income above median	0.32	[0.47]	0.31	[0.46]	0.33	[0.47]	0.00
Any benefits	0.11	[0.31]	0.11	[0.31]	0.11	[0.32]	0.00
Schooling $<$ 9 years	0.049	[0.21]	0.052	[0.22]	0.048	[0.21]	0.00
Middle/High school	0.23	[0.42]	0.23	[0.42]	0.24	[0.42]	0.00
Further educ. $<$ 3 years	0.11	[0.31]	0.10	[0.31]	0.11	[0.31]	0.00
Further educ. $<=$ 3 years	0.19	[0.39]	0.18	[0.39]	0.20	[0.40]	0.00
Employed	0.56	[0.50]	0.55	[0.50]	0.57	[0.50]	0.00
Self-employed	0.041	[0.20]	0.042	[0.20]	0.041	[0.20]	0.58
Unemployed	0.050	[0.22]	0.052	[0.22]	0.049	[0.22]	0.00
Married	0.23	[0.42]	0.22	[0.41]	0.23	[0.42]	0.00
Divorced/Widowed	0.074	[0.26]	0.075	[0.26]	0.074	[0.26]	0.80
Unmarried	0.42	[0.49]	0.42	[0.49]	0.42	[0.49]	0.16
Not migrated	0.75	[0.43]	0.73	[0.45]	0.75	[0.43]	0.00
Immigrant 1st gen	0.16	[0.36]	0.17	[0.38]	0.15	[0.36]	0.00
Immigrant 2nd gen	0.096	[0.29]	0.100	[0.30]	0.095	[0.29]	0.00
Socioeconomics missing	0.45	[0.50]	0.47	[0.50]	0.45	[0.50]	0.00
Observations	495,555		112,641		382,914		495,555

Note: This table provides summary statistics for case characteristics used as control variables in the main analysis sample. "Symptom categories" include the main symptom reported by patients when requesting a consultation. "Demographics" include patients' gender and age at consultation, and the region of residence as of November 2018. "Health risk factors" include healthcare use during the prior 3 years, excluding the 30 days immediately before consultation, and any comorbidity diagnosed in specialist care. "Socioeconomic characteristics" include indicators for above-median income (where the median is 294,700 SEK among patients over the age of 20), receipt of welfare benefits, employment, education, civil status, and migrant background, all as of November 2018. Standard deviations are shown in square brackets. The final column reports p-values from t-tests comparing the means between patient cases initially assigned to a nurse and those directly assigned to a doctor.

Table 3. Predictors of initial assignment to nurse

	Initial assignment to nurse			
	(1)	(2)	(3)	(4)
Log median doctor wait time		0.0555***	0.0555***	0.0556***
		(0.00091)	(0.00091)	(0.00091)
Baseline characteristics	✓	✓	✓	✓
Additional controls			✓	✓
Doctor-group status among previous cases in queue				✓
R-squared	0.2156	0.2208	0.2208	0.2208
Adjusted R-squared	0.1945	0.1998	0.1998	0.1998
Observations	495,555	495,555	495,555	495,505

Notes: This table reports the predictive quality from regressions of the treatment (*Initial assignment to nurse*) on the case characteristics and exogenous assignment shifters. Baseline characteristics ($X1_i$) and additional controls ($X2_i$) are described in the main text. The exogenous assignment shifters include congestion (*Log median doctor wait time*) and indicators for the doctor-group status for the previous 50 cases with either doctor-group or overlap-group symptoms. K-P F-statistics are 3726.36 for column (2), 3725.22 for column (3) and 3670.24 for column (4). Robust standard errors are in parentheses.

Table 4. Quality and downstream care of KH

Panel A: Within-firm outcomes

	Any clinician-booked doctor visit (7 days)	Any unbooked revisit (7 days)	Top score rating (initial)	Informative diagnosis (initial)
Initially to nurse	0.31***	0.0078***	0.012***	-0.14***
	(0.0015)	(0.00094)	(0.0019)	(0.0017)
Baseline characteristics	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555
Baseline mean	0.014	0.058	0.40	0.78

Panel B: Downstream healthcare use

	Any external PCP consultation (7days)	Any new prescription (7 days)	Any urgent care (7 days)	Any specialist care (7 days)
Initially to nurse	0.022***	-0.10***	0.0033***	0.0031***
	(0.0013)	(0.0017)	(0.00074)	(0.00051)
Baseline characteristics	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555
Baseline mean	0.11	0.44	0.035	0.016

	Any ED (7 days)	Any hospitalization (7 days)
Initially to nurse	0.0049***	-0.00049
	(0.00075)	(0.00038)
Baseline characteristics	✓	✓
Observations	495,555	495,555
Baseline mean	0.030	0.0082

Panel C: Long-term consequences

	Income decrease > 20 pct. (cal. month after)	Zero income (cal. month after)	Death (3 years)
Initially to nurse	-0.021***	-0.0031**	-0.000012
	(0.0025)	(0.0013)	(0.00018)
Baseline characteristics	✓	✓	✓
Observations	214,674	214,674	495,555
Baseline mean	0.22	0.050	0.0021

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 as the dependent variable. Outcomes are defined as follows: "Any clinician-booked doctor visit (7 days)" indicates whether the patient is directed by a nurse or a doctor to a doctor at the firm within 7 days of the initial consultation; "Any unbooked revisit" indicates whether the patient has an unscheduled return visit with the same symptom category within 7 days; "Top score rating" indicates whether the patient gave a top satisfaction rating (5 out of 5) for any consultation within the initial consultation day; "Informative diagnosis" 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) within the initial consultation day; "Any new prescription (7 days)" indicates whether the patient fills a new prescription (at the ATC 4th level) within 7 days; "Any external PCP consultation" indicates whether the patient visited a primary care provider outside the firm within 7 days; "Any urgent care" indicates any urgent care center (närakuter) visit within 7 days; "Any specialist care" indicates any specialist care within 7 days; "Any ED" indicates any emergency department visit within 7 days; "Any hospitalization" indicates any hospitalization within 7 days; "Income decrease" indicates whether the patient's income fell by at least 20% relative to their average income over the previous three months; "Zero income" indicates whether the patient had no earnings from their main source in the calendar month following the initial consultation; "Death (3y)" indicates whether the patient passed away within 3 years of the initial consultation. All income measures are 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 et al. 2019). The columns in each subpanel reports the Ordinary Least Squares (OLS) estimate for the treatment variable (*Nurse*) based on Equation 7. All regressions control for symptom categories, demographics, region, health risk factors, and socioeconomic factors, along with login date-by-3-hour fixed effects. The baseline mean represents the average of the dependent variable for cases directly assigned to a doctor. Robust standard errors are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 5. Costs of KH

	Taxpayer payments to the firm (within-day)	Taxpayer payments to the firm (7 days)	Firm net revenue (within-day)
Initially to nurse	-147.9***	-149.1***	-107.6***
	(0.33)	(0.63)	(0.22)
Baseline characteristics	✓	✓	✓
Additional controls	✓	✓	✓
Observations	495,555	495,555	495,555
Baseline mean	500	544.6	125
	Firm costs	Total taxpayer payments	Total healthcare costs
Initially to nurse	-40.3***	-117.4***	-9.72
	(0.55)	(13.2)	(13.2)
Baseline characteristics	✓	✓	✓
Additional controls	✓	✓	✓
Observations	495,555	495,555	495,555
Baseline mean	375	1304.8	1179.8

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 as the dependent variable. Outcomes are defined as follows: "Taxpayer payments to the firm (7 days)" from administratively set prices: 500 SEK for doctor consultation and 275 for nurse consultation from day 0 to day 7 for each patient. Firm is reimbursed for only one consultation per day (highest applicable) on day 0. "Taxpayer payments to the firm (within-day)" are the adminsitratively set prices for the first day a patient has a consultation (day 0). "Firm net revenue" is the net revenue from doctor (θ x 500 SEK) and nurse (θ x 275 SEK) consultations, calculated using a margin of $\theta = 0.25$. "Total taxpayer payments" reflect public spending on healthcare including the downstream expenses of any prescriptions, specialist visits, emergency department visits, urgent care visits, and hospitalizations. "Total healthcare costs" are calculated as public spending on healthcare - firm net revenue. The sample includes patients in Scania or Stockholm, for whom external primary care is observed. Estimates for each cost category are provided in Table A11. The column in each subpanel reports the Ordinary Least Squares (OLS) estimate for the treatment variable (*Nurse*) based on Equation 7. All regressions control for symptoms, demographics, region, health risk factors, and socioeconomic variables, along with login date-by-3-hour fixed effects. The baseline mean represents the average of the dependent variable for cases directly assigned to a doctor. Robust standard errors are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 6. Percentage change in outcomes and inputs from counterfactuals relative to the actual assignment

	$\frac{\Delta \text{ Healthcare costs}}{\text{Healthcare costs}}$	$\frac{\Delta \text{ Healthcare costs}}{\text{Firm costs}}$	$\frac{\Delta \text{ Taxpayer payments}}{\text{Taxpayer payments}}$	$\frac{\Delta \text{ Firm net revenue}}{\text{Firm net revenue}}$	$\frac{\Delta \text{ Cases to nurses}}{\text{Cases to nurses}}$
	(1)	(2)	(3)	(4)	(5)
A: Percentage changes in outcome from switching from actual assignment to assignments to non-optimized assignments					
1) Cases assigned uniformly to nurse/doctor proportional to clinician shares in *prior* 15 minutes	0.58%	1.90%	0.44%	−1.21%	−0.73%
2) Cases assigned uniformly to nurse/doctor proportional to clinician shares in *fixed* 15-minute windows	0.55%	1.81%	0.41%	−1.30%	0.00%
3) Cases assigned uniformly to nurse/doctor proportional to clinician shares in *prior* 3 hours	0.58%	1.89%	0.54%	0.06%	−5.66%
4) Cases assigned uniformly to nurse/doctor proportional to clinician shares in *fixed* 3-hour windows	0.58%	1.91%	0.43%	−1.39%	0.00%
5) All cases are initially assigned to nurses	0.79%	2.58%	−6.12%	−88.29%	339.94%
6) All cases are initially assigned to doctors	0.71%	2.33%	2.52%	24.01%	−100.00%
B: Percentage changes in outcome from switching from actual assignment to assignments that minimize total healthcare costs					
1) Unconstrained minimization of total healthcare costs	−4.07%	−13.32%	−6.81%	−39.43%	169.26%
2) Constrained minimization of total healthcare costs (Nurses constant within 3-hour windows)	−2.55%	−8.36%	−2.41%	−0.71%	0.00%
3) Constrained minimization of total healthcare costs (Nurses constant within 15-minute windows)	−2.05%	−6.71%	−1.91%	−0.28%	0.00%

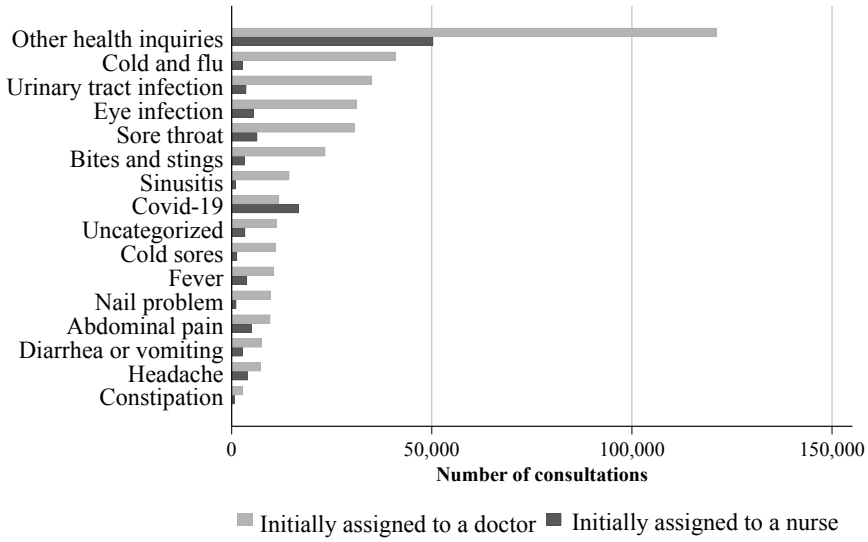
Note: This table shows the percentage change in outcomes from counterfactual assignments relative to the actual assignment. To avoid randomization, the counterfactuals in this table assign the costs associated with "fractional clinicians": q × the predicted cost of a nurse meeting plus (1-q) × the predicted cost of a doctor meeting, where q is in the interval [0,1]. Only the assignments in Panel A lines 2 and 4 and Panel B lines 2 and 3 hold total clincian assignments fixed as in reality; the rest entail using more (or fewer) nurses as indicated in the final column.

Appendix

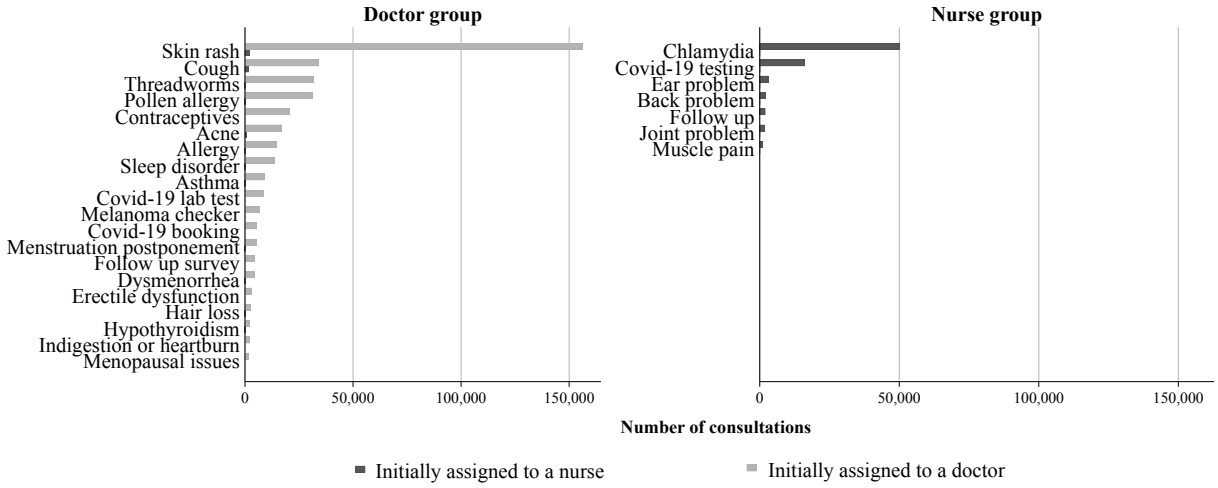
Does Division of Labor Increase Productivity? Evidence from Primary Care

A Appendix Figures

Figure A1. Case groups



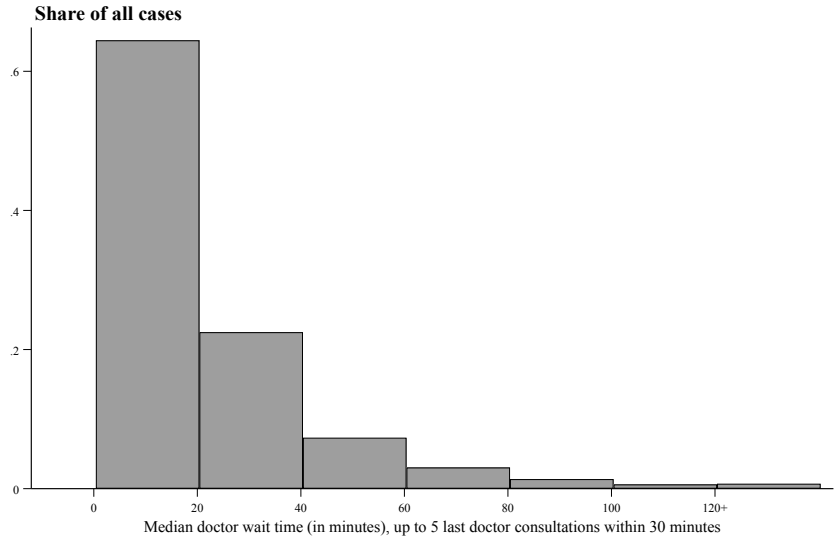
(a) Overlap group



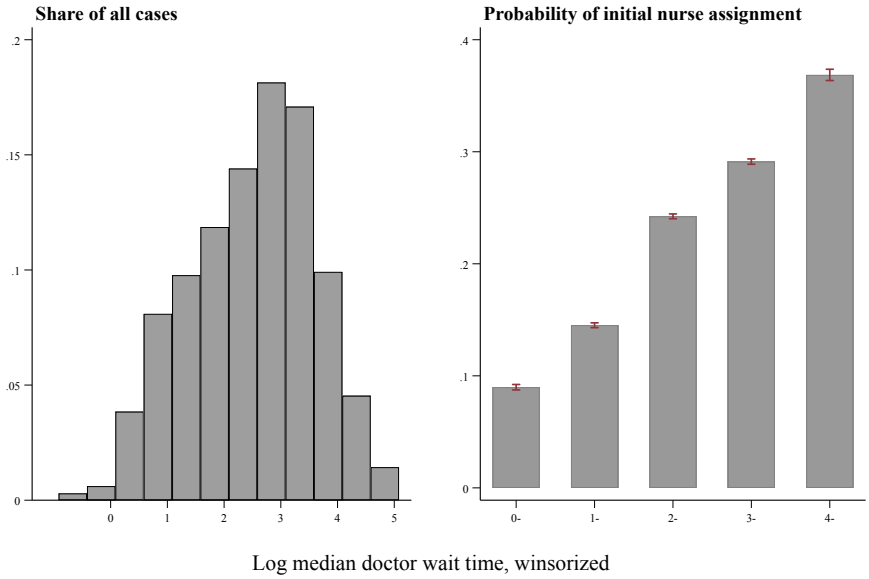
(b) Doctor group and nurse group

Notes: Panel (a) shows the distribution of marginal symptom categories in the analysis sample (N = 495,555). Marginal symptoms are defined as those assigned to a clinician type in no more than 95% of cases. Panel (b) presents deterministic symptom assignments, defined as symptoms assigned to a single clinician type (nurse or doctor) for at least 95% of visits in the broader provider sample.

Figure A2. Congestion measure



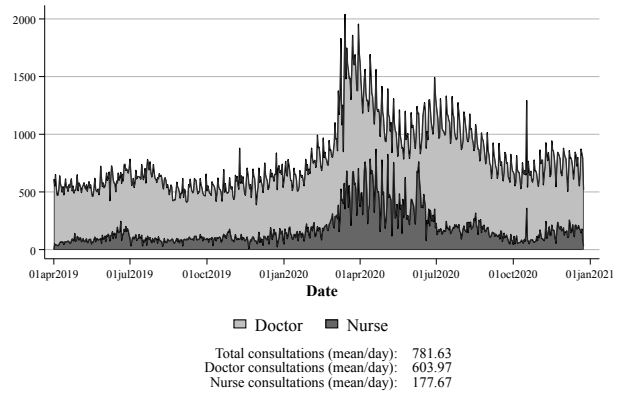
(a) Doctor wait time before transformation



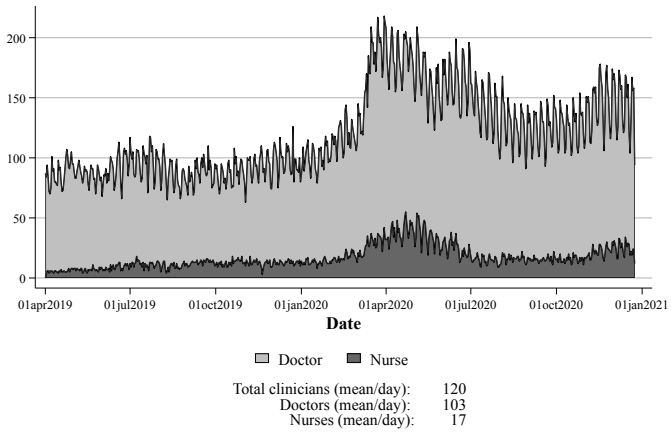
(b) Congestion measure and initial nurse assignment

Notes: Panel (a) shows the untransformed congestion measure in minutes, defined as the median doctor waiting time among the last up to five patients whose consultations began within 30 minutes before each patient's login time in the analysis sample. Panel (b) presents descriptives of the congestion measure, *Log median doctor wait time*. The first plot shows the distribution of congestion; the second shows the probability of initial nurse assignment across congestion levels.

Figure A3. Number of consultations and clinicians per calendar day



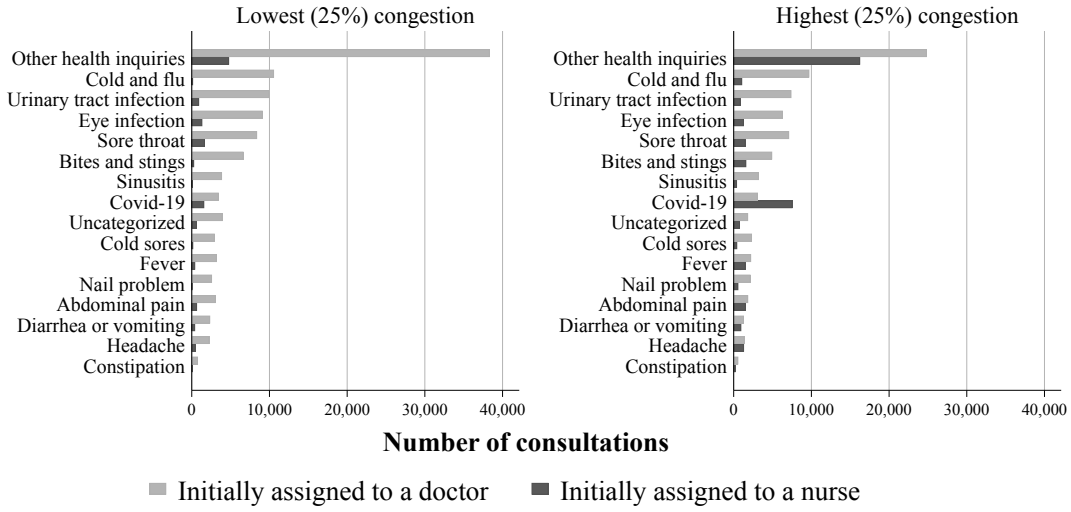
(a) Number of consultations



(b) Number of clinicians

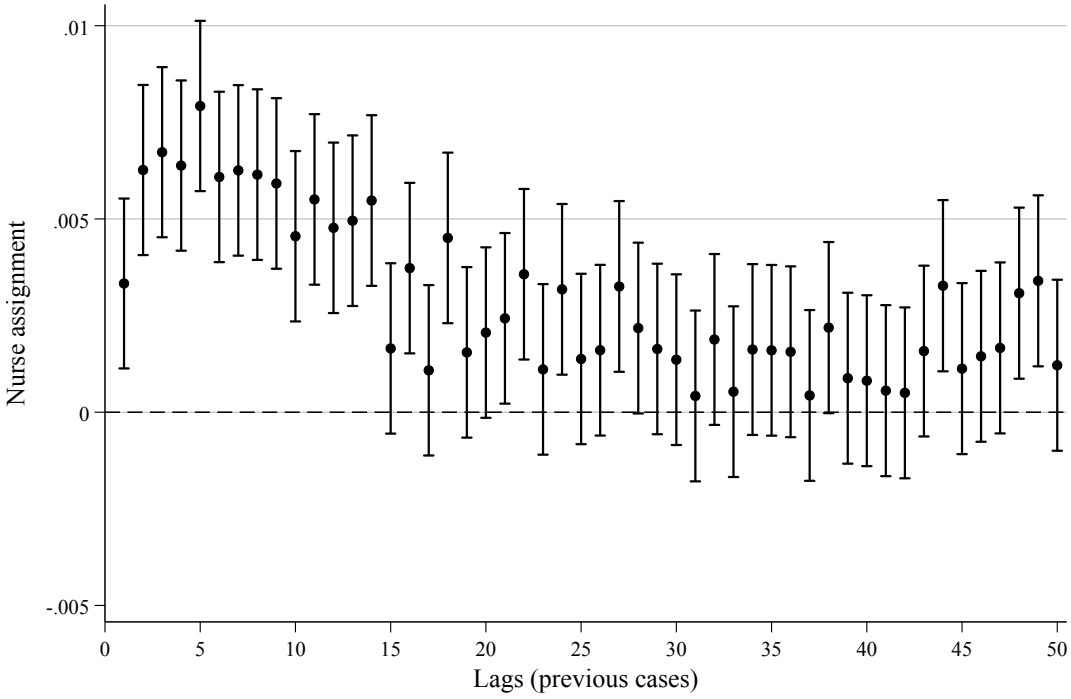
Notes: Panel (a) shows the number of consultations per calendar day for nurses and doctors in the analysis sample. Panel (b) shows the number of active clinicians per calendar day for nurses and doctors in the analysis sample.

Figure A4. Assignment of symptoms in the lowest and highest quartiles of congestion



Notes: This figure shows the symptom categories of patients who are initially assigned to nurses or doctors by the highest and the lowest quartiles of doctor congestion in the analysis sample.

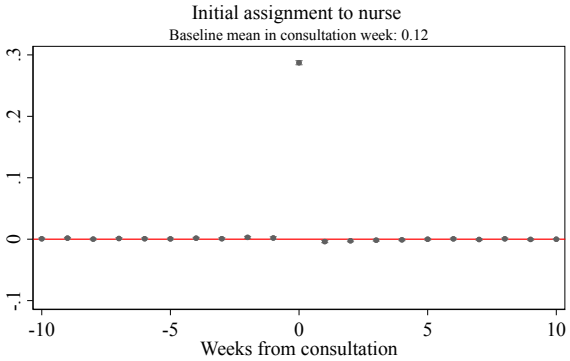
Figure A5. Effects of lagged doctor group case on case assignment



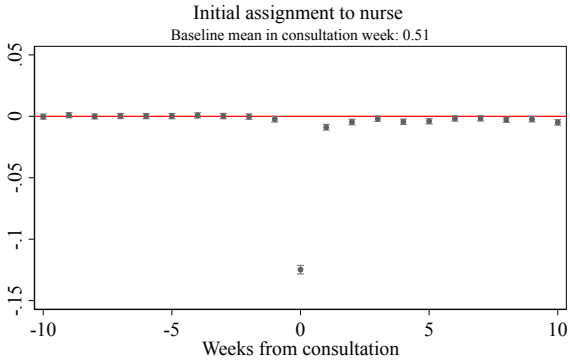
Notes: The figure reports estimates from regressions of initial nurse assignment on indicators for lagged doctor-group cases. Each point corresponds to the coefficient from a separate OLS regression using the case at lag $n-1$, where lags are defined by ordering all cases in the broad drop-in sample (excluding rare symptoms) by login time. All regressions include date-by-3-hour fixed effects and use the analysis sample of overlap-group cases. Vertical lines denote 95% confidence intervals. Lags refer to preceding cases in calendar time, based on patients' login times.

Figure A6. Effect of an initial nurse assignment on lags and leads of patient outcomes

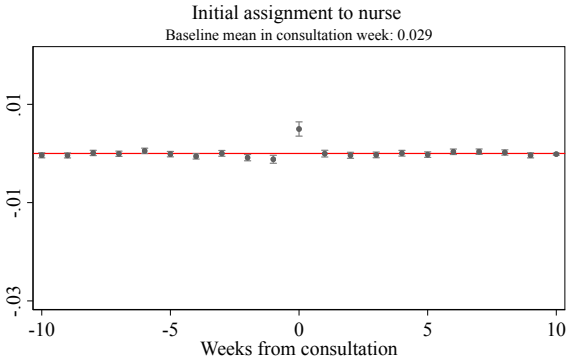
(a) Any consultation at the provider



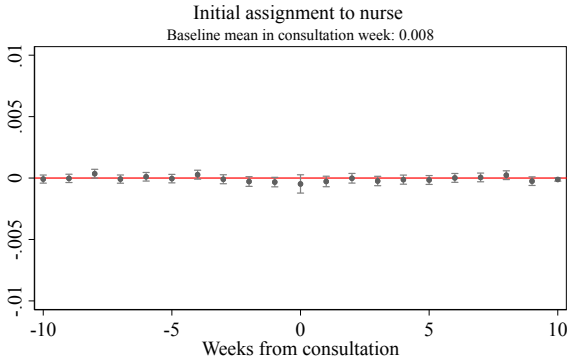
(b) Any prescription



(c) Any emergency department visit

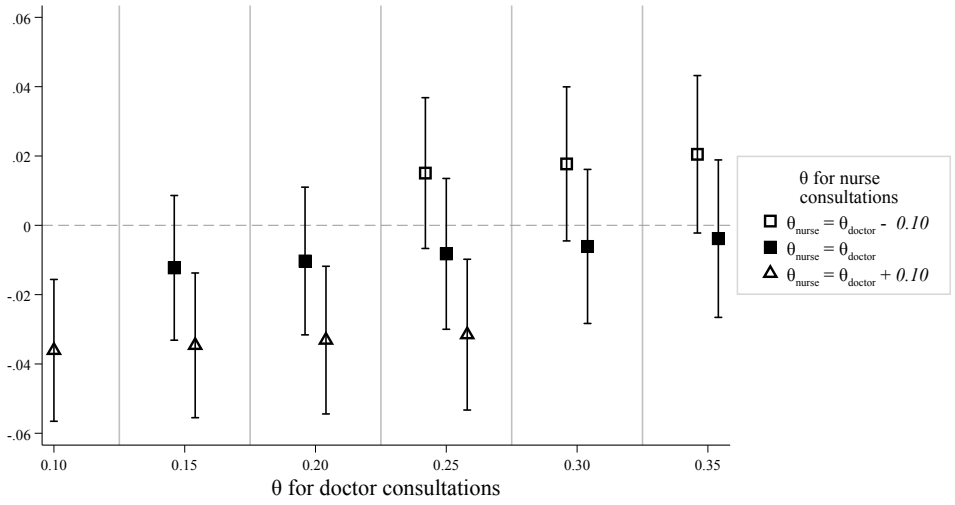


(d) Any hospitalization



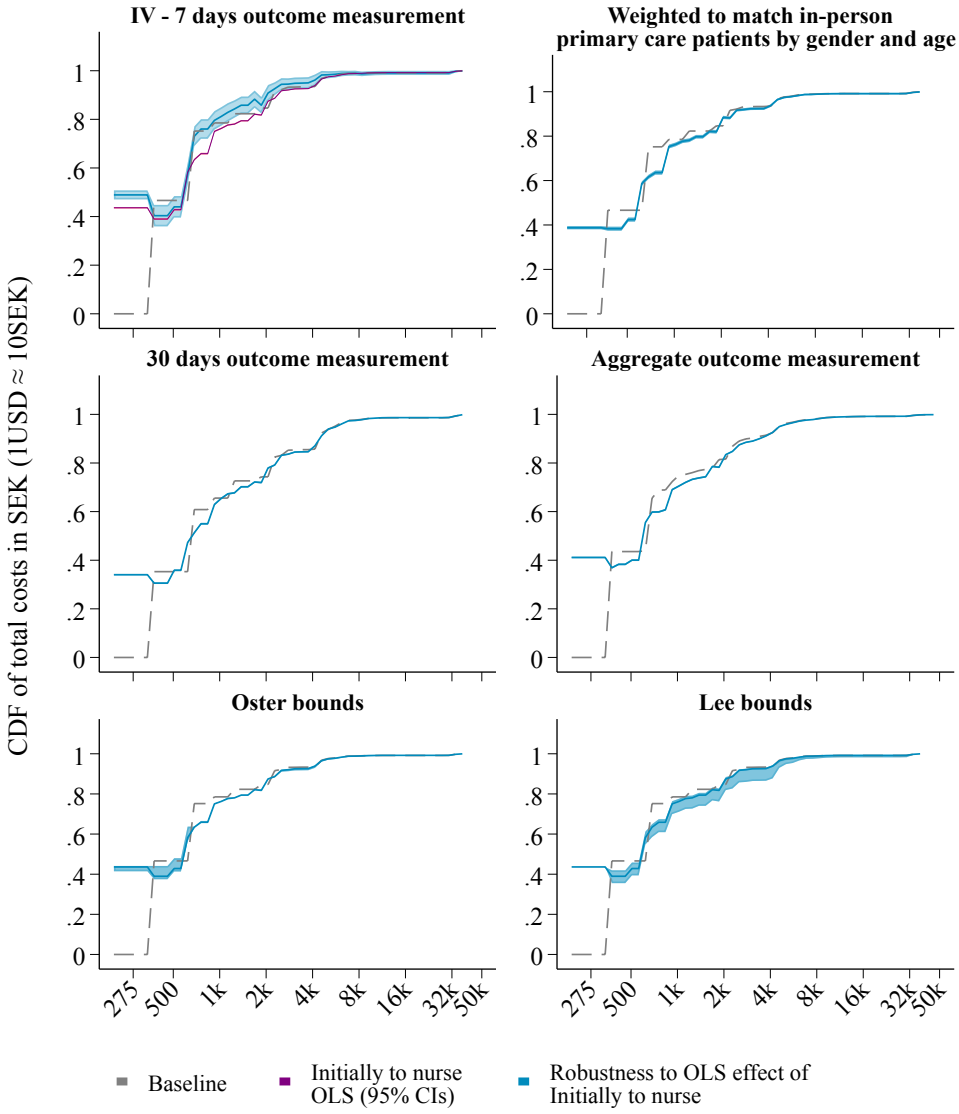
Notes: These figures show the estimated effect of an initial assignment to a nurse (the *knowledge hierarchy*) for outcomes in the 10 weeks prior to or after the initial consultation. "Week 0" marks the week starting with the initial consultation date. Each estimate is based on a separate regression of a lagged or lead outcome on the treatment variable (*Nurse*). Estimates are obtained by Ordinary Least Squares (OLS) based on Equation 7 using *Nurse staff share, past 60 min.* as an instrument for *Nurse* based on Equation 7. All regressions control for the full set of case characteristics, along with login date-by-4-hour fixed effects. The baseline mean represents the average of an outcome in week 0 for cases directly assigned to a doctor.

Figure A7. Total healthcare costs when markdowns θ vary between nurses and doctors



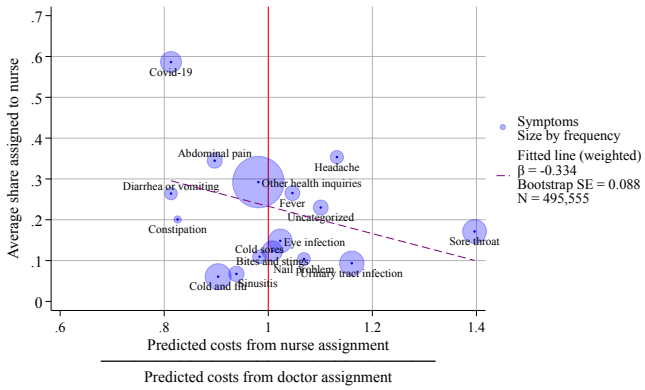
Notes: This figure shows coefficients and 95% confidence intervals from regressions of total healthcare costs on initial nurse assignment using the full set of controls, divided by baseline healthcare costs. Results are shown for varying doctor consultation wedges on x-axis. For each doctor wedge, nurse wedges are displayed at -0.10, equal to, and +0.10 relative to the doctor wedge.

Figure A8. Robustness on the distribution of total healthcare costs

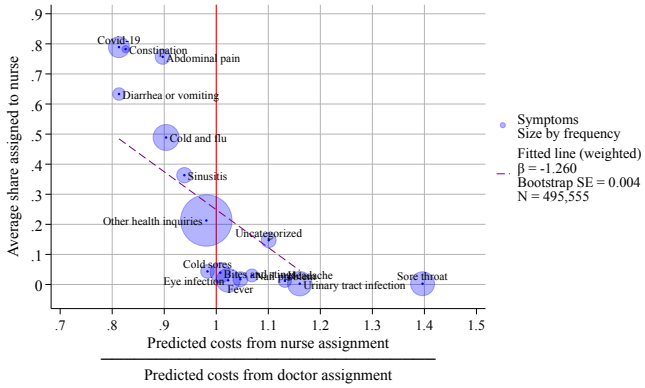


Notes: These panels provide robustness checks for the CDF estimates in Figure 2b. The first row applies alternative identification using the congestion instrument in Equation 19 and alternative weighting by using the main identification Equation 7, but reweighting by primary care patient age and gender. The second row reports results under alternative outcome measurements, measuring cost indicators over 30 days and measuring aggregated costs over 7 days after the intial consultation. The third row reports Lee and Oster bounds. The Lee bounds are constructed using the conversion rates of the treatment (72.2%) and control (77.5%) groups, implying a trimming rate of $(0.775 - 0.722)/0.775 = 0.0684$. Since a larger share of doctor consultations remains in the sample, trimming is applied to doctor consultations at both extremes of the cost distribution: most expensive (lower bound) and least expensive (upper bound). The 95% confidence intervals are shown around the OLS estimates. The Oster bounds are following Oster (2016). The green line shows the estimate from an OLS regression with no controls, providing a bound for a biased estimate. The blue line represents the bias-adjusted estimates under the assumption that the maximum R^2 is 1.3 times the R^2 from the controlled regression, providing an opposite-sign bound to the uncontrolled estimate. All panels use the full set of controls. 95 percent confidence intervals are shown for the first two rows.

Figure A9. Relative cost advantage in the actual and counterfactual case allocation



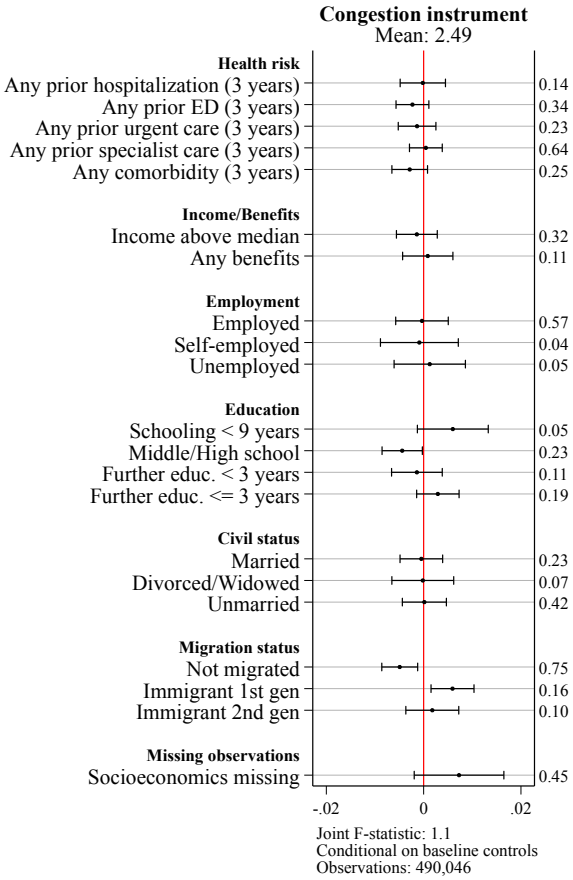
(a) Actual allocation



(b) Counterfactual allocation (constrained optimum)

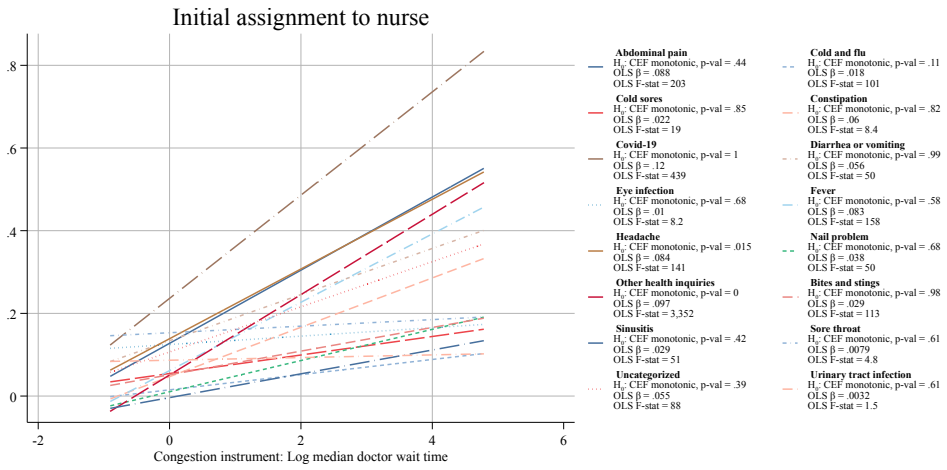
Notes: The figures show firm's case assignment by relative cost advantage of an initial nurse versus initial doctor assignment across symptom groups, using the actual clinician allocation and the counterfactual clinician allocation. The counterfactual assignment reduces total healthcare costs while keeping the number of nurses constant within 15-minute windows. The y-axis plots the predicted probability of initial nurse assignment from the first-stage regression (Equation 19). The x-axis plots the relative revenue or cost advantage under initial nurse relative to initial doctor assignment based on predicted outcomes, derived from a modified version of Equation 7 interacting symptom groups with the nurse indicator $Nurse_i$. The relative cost or revenue advantage is constructed by summing predicted outcomes within each symptom group and taking the ratio. Circles represent symptom frequencies. Fitted lines are weighted OLS regressions. Standard errors are computed from 100 bootstrap replications with resampling at the individual consultation level. The vertical red line indicates equal costs or revenues across assignment routes. All regressions use the full set of controls and the analysis sample.

Figure A10. Balance of the congestion instrument in case characteristics



Notes: This figure presents balance tests for the congestion instrument (*Log median doctor wait time*). Each row shows the results of a bivariate regression where the instrumental variable is regressed on a specific case characteristic. All regressions account for baseline case characteristics: login time, main symptom, and patient demographics and the demographics × past firm use interaction term. Login time indicates date-by-3 hours fixed effects. Symptom categories refer to indicators for the main symptom category reported by the patient when requesting a consultation. Demographics include indicators for patient gender, age categories, and aggregated regions. Interaction term includes age, gender, symptom, indicators for previous doctor consultation and prescription pickup from provider The horizontal lines represent 95% confidence intervals based on robust standard errors. The F-statistics reflect the joint F-tests for the treatment or instrument variable when regressed on all control variables, conditioned on login time. The numbers on the far right display the sample means for each case characteristic

Figure A11. Instrument monotonicity test



Notes: This figure shows non-parametric estimates within subsamples of the conditional expectation function (CEF) for the treatment, initial nurse assignment, given our congestion instrument, *Log doctor wait time of at most 5 patients in past 30 minutes*. The CEF is estimated within each subsample as a smoothed line over a binscatter. The estimates are conditional on our baseline set of controls (login time, symptom, demographic controls and interactions), and bins are selected to minimize the integrated mean square error (Cattaneo et al. 2024). The OLS β and F-statistics are estimated with linear regressions in each subsample conditional on our baseline set of controls.

B Appendix Tables

Table A1. Restrictions on the main analysis sample

	Nurse	**Total**
All observed consultations		1,814,706
1. Keep only consultations that involve either doctors or nurses	14.5%	1,668,303
2. Exclude scheduled or in-person consultations	19.5%	1,239,527
3. Exclude observations outside the analysis time window	21.1%	1,126,662
4. Exclude patients registered with the provider	20.6%	1,054,985
5. Exclude infants (ages 0-1)	21.2%	995,177
6. Exclude follow-ups within two hours by the same patient	21.2%	970,277
7. Exclude consultations with missing characteristics → *Broad drop-in sample*	21.2%	969,483
a) *Doctor group*		
· Restrict broad drop-in sample to doctor group	1.2%	381,791
· Exclude singleton observations in the full control scheme	1.2%	380,901
b) *Nurse group*		
· Restrict broad drop-in sample to nurse group	98.9%	85,729
· Exclude singleton observations in the full control scheme	99.0%	84,610
c) *Rare symptom group*		
· Restrict broad drop-in sample to rare symptoms	53.1%	5,961
· Exclude singleton observations in the full control scheme	58.4%	3,592
d) *Main analysis sample*		
· Restrict broad drop-in sample to overlap group	22.7%	496,002
· Exclude singleton observations in the full control scheme	22.7%	495,555

Note: This table presents the number of observations after applying our sample restrictions. The columns show the share of consultations with a nurse (instead of with a doctor), as well as the total number of consultations. Each row introduces an additional sample restriction: Restrictions 1, 2 and 3 limit the sample to unscheduled online consultations with doctors or nurses from 1 April 2019 to 24 December 2020, during which nurse consultations started to take place for overlap cases and follow-up consultations within 7 days can be observed. Restrictions 4 and 5 exclude patients registered with the firm and infant patients (age strictly below two years), as different internal protocols apply to them. Restriction 6 excludes consultations that involve the same patient within two hours. Restriction 7 excludes consultations where baseline characteristics (login time, patient age, gender, region) are missing and defines the *broad drop-in sample*. We then create three subsamples from step 7 based on symptom assignment. A symptom belongs to the doctor group (nurse group) if assigned to doctors (nurses) in at least 95% of cases. Rare symptoms refer to symptom categories with fewer than 1,000 observations in the *broad drop-in sample*. The *analysis sample* consists of overlap group where between 5% and 95% of consultations are assigned to nurses.

Table A2. Ten most common diagnoses in the analysis sample, compared to physical primary care consultations of the Scania subsample

		Share of consultations (%)		Share of consultations handled by nurses (%)	
ICD codes	Description	Analysis Sample	Physical primary care in Scania	Analysis sample	Physical primary 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
R52	Persons encountering health services for specific procedures, not carried out	2.69	3.08	43.7	25.4
N39	Pain, not elsewhere classified	2.59	0.52	19.0	13.6
J01	Other disorders of urinary system	2.35	0.52	4.23	4.71
Z53	Acute sinusitis	2.69	0.033	16.2	1.80

Note: This table presents the most common ICD codes in our main analysis sample. Column 1 reports the ICD codes and 2 reports the description. Column 3 and 4 shows how frequent these ICD codes are in our sample and in-person primary care visits in Scania region. Last two columns show the percent of the cases handled by each nurse in both samples.

Table A3. Comparison of sample characteristics to the Swedish population

	Analysis sample	Swedish Population	Physical primary care visits 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 income (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. "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 income" includes annual earnings from wages and self-employment in SEK and is reported for individuals strictly over age 20.

Table A4. First-in First-out rank regressions

	Exit rank in overlap group	Exit rank in doctor group	Exit rank in overlap and doctor group	Exit rank in broader drop-in sample
	(1)	(2)	(3)	(4)
Login rank	0.98***	0.99***	0.99***	0.99***
	(0.00064)	(0.0010)	(0.00071)	(0.0011)
Doctor group			8.79***	3.61***
			(0.090)	(1.32)
Nurse group				89.2***
				(6.41)
Date-by-3 hour FEs	✓	✓	✓	✓
Observations	495,555	380,901	876,456	966,622
R-squared	1.00	1.00	1.00	1.00

Notes: All regressions use the full set of controls. 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. Panel A replaces the usual time fixed effects by one hour fixed effects and panel B uses the baseline fix effects setting. The third and fourth column include dummy variables indicating deterministic tasks for nurses and doctors. Robust standard errors are in parenthesis

Table A5. 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	124,071
	(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	110,588
	(6.45)	(6.62)				
Rating: top score	0.38	0.40	-0.022	0.00	123,871	124,071
	(0.49)	(0.49)				
No show	0.00090	0.0013	-0.00039	0.00	123,871	124,071
	(0.030)	(0.036)				
Technical error	0.0046	0.0075	-0.0029	0.00	123,871	124,071
	(0.068)	(0.086)				
Consultation time below one minute	0.047	0.051	-0.0038	0.00	123,871	124,071
	(0.21)	(0.22)				
Consultation time below 30 seconds	0.015	0.016	-0.00052	0.30	123,871	124,071
	(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 A6. Additional Outcomes: OLS

	Top score rating (7 days)	Informative diagnosis (7 days)	Any external PCP consultation (7 days), Stockholm & Scania Sample	No show (initial)	Technical error (initial)
Initially to nurse	0.047***	-0.064***	0.030***	0.00043**	-0.0024***
	(0.0019)	(0.0016)	(0.0018)	(0.00014)	(0.00026)
Baseline characteristics	✓	✓	✓	✓	✓
Additional controls	✓	✓	✓	✓	✓
Observations	495,555	495,555	262,144	495,555	495,555
Baseline mean	0.42	0.80	0.11	0.0010	0.0066

Note: This table presents regression results for additional outcomes examining 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 as the dependent variable. Outcomes are defined as follows: Top Score rating (initial), Informative diagnosis (initial), and Any external PCP consultation (7 days) in the Stockholm & Scania sample. The columns reports the Ordinary Least Squares (OLS) estimate for the treatment variable (*Nurse*) based on Equation 7. All regressions control for symptom categories, demographics, region, health risk factors, and socioeconomic variables, along with login date-by-3-hour fixed effects. The baseline mean represents the average of the dependent variable for cases directly assigned to a doctor. Robust standard errors are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A7. Effect of an initial nurse assignment on healthcare costs

Panel A: Within-firm outcomes

	Any clinician-booked doctor visit (7 days)		Any unbooked revisit (7 days)		Top score rating (initial)		Informative diagnosis (initial)	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Initially to nurse	0.31***	0.25***	0.0078***	-0.018	0.012***	-0.0012	-0.14***	-0.064***
	(0.0015)	(0.010)	(0.00094)	(0.011)	(0.0019)	(0.023)	(0.0017)	(0.019)
Baseline characteristics	✓	✓	✓	✓	✓	✓	✓	✓
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555	495,555	495,555	495,555	495,555
First-stage K-P F-statistic		3,670		3,670		3,670		3,670
Baseline mean	0.014	0.014	0.058	0.058	0.40	0.40	0.78	0.78

Panel B: Downstream healthcare use

	Any external PCP consultation (7 days)		Any new prescription (7 days)		Any urgent care (7 days)		Any specialist care (7 days)	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Initially to nurse	0.022***	-0.021	-0.10***	-0.026	0.0033***	-0.010	0.0031***	-0.000066
	(0.0013)	(0.015)	(0.0017)	(0.022)	(0.00074)	(0.0088)	(0.00051)	(0.0061)
Baseline characteristics	✓	✓	✓	✓	✓	✓	✓	✓
Additional controls	✓	✓	✓	✓	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555	495,555	495,555	495,555	495,555
First-stage K-P F-statistic		3,670		3,670		3,670		3,670
Baseline mean	0.11	0.11	0.44	0.44	0.035	0.035	0.016	0.016

	Any ED (7 days)		Any hospitalization (7 days)	
	OLS	IV	OLS	IV
Initially to nurse	0.0049***	-0.0070	-0.00049	0.00087
	(0.00075)	(0.0085)	(0.00038)	(0.0045)
Baseline characteristics	✓	✓	✓	✓
Additional controls	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555
First-stage K-P F-statistic		3,670		3,670
Baseline mean	0.030	0.030	0.0082	0.0082

Panel C: Long-term consequences

	Income decrease > 20pct. (cal. month after)		Zero income (cal. month after)		Death (3 years)	
	OLS	IV	OLS	IV	OLS	IV
Initially to nurse	-0.021***	0.00034	-0.0031**	0.0056	-0.000012	-0.0045**
	(0.0025)	(0.031)	(0.0013)	(0.017)	(0.00018)	(0.0022)
Baseline characteristics	✓	✓	✓	✓	✓	✓
Additional controls	✓	✓	✓	✓	✓	✓
Observations	214,674	214,674	214,674	214,674	495,555	495,555
First-stage K-P F-statistic		1,457		1,457		3,670
Baseline mean	0.22	0.22	0.050	0.050	0.0021	0.0021

Panel D: Firm costs

	Taxpayer payments to the firm (within-day)		Taxpayer payments to the firm (7 days)		Firm net revenue (within-day)	
	OLS	IV	OLS	IV	OLS	IV
Initially to nurse	-147.9***	-161.5***	-149.1***	-172.0***	-107.6***	-98.6***
	(0.33)	(2.03)	(0.63)	(6.87)	(0.22)	(1.35)
Baseline characteristics	✓	✓	✓	✓	✓	✓
Additional controls	✓	✓	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555	495,555	495,555
First-stage K-P F-statistic		3,670		3,670		3,670
Baseline mean	500	500	544.6	544.6	125	125

Panel E: Costs including downstream care

	Total taxpayer payments		Total healthcare costs	
	OLS	IV	OLS	IV
Initially to nurse	-117.4***	-242.8	-9.72	-144.2
	(13.2)	(154.2)	(13.2)	(154.2)
Baseline characteristics	✓	✓	✓	✓
Additional controls	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555
First-stage K-P F-statistic		3,670		3,670
Baseline mean	1304.8	1304.8	1179.8	1179.8

Note: This table presents regression results for instrumental variable and reduced form estimates for our main outcomes 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 as the dependent variable. Outcomes are defined as follows: Outcomes are defined as follows: "Costs of provider visits" includes only the costs associated with online consultations at the provider. Costs, excl. ext. PCP" represents the estimated care costs, including costs at the provider and the downstream expenses of any prescriptions, specialist visits, emergency department visits, urgent care visits, and hospitalizations within seven days of the initial consultation, but exclude the costs of external primary care consultations. "Costs, incl. ext. PCP" includes the costs of external primary care consultations. The sample includes patients in Scania or Stockholm, for whom external primary care is observed. Estimates for each cost category are provided in Table A11. The left column presents the reduced-form regression estimates for the congestion instrument, the natural logarithm of the median wait time in the up to 5 prior doctor consultations within 30 minutes, winsorized at the 1st and 99th percentile (*Log median doctor wait time*). The right column provides the Instrumental Variables specification, using *Log median doctor wait time* as an instrument for *Nurse*, estimated via Two-Stage Least Squares (2SLS), following Equations 7 and 19. All regressions control for symptoms, demographics, region, health risk factors, and socioeconomic variables, along with login date-by-3-hour fixed effects. The baseline mean represents the average of the dependent variable for cases directly assigned to a doctor. The first-stage K-P F-statistic refers to the Kleibergen-Paap F-statistic. Robust standard errors are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A8. Robustness checks: Total healthcare costs

Panel A: Alternative controls and unobserved selection

	Only baseline controls	Full controls with date-by-1 hour windows	Full controls including congestion quartiles	Full controls including patient fixed effects	Oster bounds
Initially to nurse	-11.3	-5.68	-9.43	4.90	-9.72
	(13.1)	(13.3)	(13.1)	(23.2)	(13.1)
Full controls		✓	✓	✓	✓
Date-by-1 hour FEs		✓			
Baseline controls	✓				
Congestion quartiles			✓		
Patient FEs				✓	
Observations	495,555	494,819	495,555	228,694	495,555
Bias-adjusted β					-35.6
Uncontrolled β					76.5
Baseline mean	1179.8	1179.0	1179.8	1189.2	1179.8

Panel B: Differential attrition

	Lee bounds	Consultations in April, May & June 2019	Consultations in July, August & September 2019	Cases with prior nurse consultations
Initially to nurse	-22.9*	104.5**	95.6*	-43.5
	(12.6)	(47.4)	(52.1)	(39.4)
Full controls		✓	✓	✓
Symptom controls	✓			
Observations	495,555	50,852	51,786	52,107
Lower bound	-253.7			
Upper bound	263.7			
Baseline mean	1179.8	1116.7	1052.5	1275.6

Panel C: Cost calculations

	Stockholm & Scania sample	Total healthcare costs (30 days)	Cumulative costs	Cumulative costs (30 days)
Initially to nurse	-11.2	-9.70	-6.52	-21.0
	(18.4)	(16.9)	(16.5)	(25.8)
Full controls	✓	✓	✓	✓
Observations	262,144	495,555	495,555	495,555
Baseline mean	1203.8	1854.9	1476.1	2648.3

Panel D: Alternative markdown assumptions

	Alternative markdown $\theta = 0.10$	Alternative markdown $\theta = 0.20$	Alternative $\theta = 0.30$
Initially to nurse	-17.8	-12.4	-7.04
	(13.1)	(13.1)	(13.1)
Full controls	✓	✓	✓
Observations	495,555	495,555	495,555
Baseline mean	1254.8	1204.8	1154.8

Panel E: External validity

	Weighted to match in-person primary care
Initially to nurse	28.8
	(56.0)
Full controls	✓
Observations	495,555
Baseline mean	1179.8

	Pre Covid-19 period	Pre Covid-19 + interlude period	First and second Covid-19 wave	2 months before Covid-19 period	2 months after Covid-19 period
Initially to nurse	88.3***	27.6	-40.8**	83.7***	70.1***
	(21.1)	(17.2)	(20.7)	(25.9)	(17.9)
Full controls	✓	✓	✓	✓	✓
Observations	212,372	308,293	186,158	157,342	284,470
Baseline mean	1127.4	1156.7	1219.5	1108.9	1144.7

Notes: This table presents regression estimates from alternative model specifications assessing robustness. We show the effects of initial case assignment to a nurse rather than directly to a doctor, on the outcome variable total healthcare costs. Each specification includes either baseline controls (symptoms fully interacted with demographics and with prior provider use, and region), additional controls (health risk and socioeconomic variables), or patient fixed effects as indicated in the table. All regressions control for login time based on date-by-3 hour fixed effects. Each cell shows estimates for *Initially to nurse* from the Ordinary Least Squares (OLS) or the Instrumental Variables (IV) specification using *Log median doctor wait time* as an instrument and estimated via Two-Stage Least Squares (2SLS), following Equations 7 and 19. The baseline mean represents the average of the dependent variable for cases directly assigned to a doctor. The first-stage K-P F-statistic refers to the Kleibergen-Paap F-statistic. Robust standard errors are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table A9. Percentage change in aggregate outcomes and inputs from counterfactuals relative to the actual assignment

	$\frac{\Delta\text{Healthcare costs}}{\text{Healthcare costs}}$	$\frac{\Delta\text{Healthcare costs}}{\text{Firm costs}}$	$\frac{\Delta\text{Taxpayer payments}}{\text{Taxpayer payments}}$	$\frac{\Delta\text{Firm net revenue}}{\text{Firm net revenue}}$	$\frac{\Delta\text{Cases assigned to nurses}}{\text{Cases assigned to nurses}}$
	(1)	(2)	(3)	(4)	(5)
A: Percentage changes in outcome from switching from actual assignment to assignments to non-optimized assignments					
1) Cases assigned uniformly to nurse/doctor proportional to clinician shares in *prior* 15 minutes	0.46%	1.66%	0.35%	−1.45%	−0.73%
2) Cases assigned uniformly to nurse/doctor proportional to clinician shares in *fixed* 15-minute windows	0.45%	1.60%	0.33%	−1.53%	0.00%
3) Cases assigned uniformly to nurse/doctor proportional to clinician shares in *prior* 3 hours	0.47%	1.69%	0.43%	−0.15%	−5.66%
4) Cases assigned uniformly to nurse/doctor proportional to clinician shares in *fixed* 3-hour windows	0.47%	1.68%	0.34%	−1.63%	0.00%
5) All cases are initially assigned to nurses	0.34%	1.23%	−5.15%	−89.84%	339.94%
6) All cases are initially assigned to doctors	0.70%	2.51%	2.13%	24.12%	−100.00%
B: Percentage changes in outcome from switching from actual assignment to assignments that minimize total healthcare costs					
1) Unconstrained minimization of total healthcare costs	−4.05%	−14.50%	−6.62%	−46.21%	193.19%
2) Constrained minimization of total healthcare costs (Nurses constant within 3-hour windows)	−2.72%	−9.71%	−2.63%	−1.37%	0.00%
3) Constrained minimization of total healthcare costs (Nurses constant within 15-minute windows)	−2.16%	−7.72%	−2.07%	−0.72%	0.00%

Note: This table shows the percentage change in aggregate outcomes from counterfactual assignments relative to the actual assignment. To avoid randomization, the counterfactuals in this table assign the costs associated with "fractional clinicians": q × the predicted cost of a nurse meeting plus (1-q) × the predicted cost of a doctor meeting, where q is in the interval [0,1]. Only the assignments in Panel A lines 2 and 4 and Panel B lines 2 and 3 hold total clincian assignments fixed as in reality; the rest entail using more (or fewer) nurses as indicated in the final column.

Table A10. Description of key case characteristics

Variable	Description	Data source
Baseline case characteristics		
Age	Set of indicator variables for patient age, defined for each integer year from 2 to 70, with one combined indicator for ages 70 and above.	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).	Provider
Gender	Indicates whether the patient is female.	Provider
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.	Provider
COVID-19 periods	Set of indicator variables classifying COVID-19 spread into four periods: pre-COVID (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)	Public Health Agency of Sweden
Symptom category	Set of indicator variables for a case's main symptom, as indicated by the patient before the consultation.	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.	SCB
Last visit 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.	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.	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.	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.	Provider

Table A10. Description of key case characteristics (Continued)

Variables	Descriptions	Data source
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.	Provider
Any ED (7 days)	Indicates any emergency department visit within 7 days of the initial consultation.	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.	NBHW
Any hospitalization (7 days)	Indicates any hospitalization within 7 days of the initial consultation, excluding ICD codes in category O related to childbirth.	NBHW
Any urgent care (7 days)	Indicates any urgent care center visit within 7 days of the initial consultation.	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.	NBHW
Any specialist care (7 days)	Indicates any specialist outpatient care (excluding ED) within 7 days of the initial consultation.	NBHW
Any unbooked revisit (7 days)	Indicates whether the patient has an unscheduled return visit with the same symptom category within 7 days.	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.	Provider
Death (3 years)	Indicates whether the patient passed away within 3 years of the initial consultation	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.	Provider
Income decrease $>$ 20 pct. (cal. month after)	Indicator for whether the patient's income fell by at least 20% relative to their average income over the previous three months, conditional on that the patient has a stable income (received income for x months etc look at the data)	SCB
Informative diagnosis (7 days)	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) within 7 days.	Provider
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.	Provider
Rating: top score (7 days)	Indicates whether the patient gave a top satisfaction rating (5 out of 5) for any consultation within the firm, within 7 days of the initial consultation. Many patients do not give a rating (51.18%) . These nonresponses are coded as zeroes in this variable.	Provider
Zero income (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 et al., 2019).	SCB
Outcomes: Cost measures		
Total healthcare costs (7 days)	The total costs of healthcare provision within 7 days after the initial assignment, calculated as total taxpayer costs 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	NBHW
Total taxpayer costs	The taxpayer's incurred costs of total healthcare provision within 7 days of the initial assignment, covering both within-firm and downstream spending. These include costs at the firm on the calendar day of the initial consultation and the cost of any occurrence of each downstream event (an additional within-firm visit, emergency department visit, specialist or urgent care visit, hospitalization, or prescription within 7 days).	NBHW
Taxpayer payments to the firm (7 days)	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	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).	Provider
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 costs of the initial and potential follow-up consultations, multiplied by θ, where θ denotes the firm's markdown.	Provider
Firm costs	The firm's costs of providing care on the calendar day of the initial assignment, defined as the sum of the cost of the nurse consultation and any subsequent doctor consultation.	Provider

Table A10. Description of key case characteristics (Continued)

Variables	Descriptions	Data source
Additional case characteristics		
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 (further educ. $<$ 3 years); secondary education $\geq$ 3 years (further educ. $\geq$ 3 years). Educational degrees are reported for patients $\geq$ age 25.	SCB
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.	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.	Provider
Any comorbidity	Indicates an Elixhauser comorbidity index of 1 or higher, 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.	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.	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.	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.	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.	NBHW
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.	NBHW
Any benefits	Indicates any receipt of social security benefits in 2018. Benefits are reported for patients $\geq$ age 20 in Nov 2018.	SCB
Income above median	Indicates whether the patient's annual income in Nov 2018 is greater than the in-sample median of non-missing incomes (294,700 SEK, approximately 28,000 USD). Income is reported for patients aged 20 or older.	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, income, or benefits) contain missing values.	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.	NBHW
No show (7 days)	Indicates whether any consultation had a technical error including appointments scheduled through the knowledge hierarchy.	Provider
No show (initial day)	Indicates whether the initial consultation was recorded as a "no show".	Provider
Technical error (7 days)	Indicates whether any consultation was recorded as a "no show" including appointments scheduled through the knowledge hierarchy.	Provider
Tehcnical error (initial day)	Indicates whether the initial consultation had a technical error.	Provider

Table A11. 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: 5 Nov 2024
Nurse	275	2020	Recommendation on remuneration for digital healthcare services to healthcare providers.	Sveriges Kommuner och Regioner (SKR)[2] Last access: 5 Nov 2024
External primary care consultation				
Physical consultation	1452.35	2019/2020	Weighted average cost: 0.396×614 SEK$+0.604 \times$ 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[1] Last access: 5 Nov 2024 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: medicine, pulmonology, infectious diseases, dermatology, urology, orthopedics, ophthalmology, otolaryngology, pediatrics, and gynecology. We consider specialties for acute and routine specialty services for an average one-time cost and exclude, for example, oncology and psychiatry which likely involve ongoing therapies.	Own calculation, based on information from Norra sjukvårdsregionförbundet (NRF)[3] Last access: 5 Nov 2024
Emergency department visit	3991.5	2019/2020	Average costs in Southern Sweden, $(3,963 + 4,020) * 0.5 = 3,991.5$.	Södra Sjukvårdsregionen[4] Last access: 5 Nov 2024
Urgent care center visit	2002	2019/2020 information from 2024	Using the information provided—"Without an EU card, you must pay the full cost of the care yourself. [...] An appointment with a doctor at a vårdcentralen (healthcare centre) or a visit to the närakuten (urgent care centre) costs SEK 2,093."—we base our analysis on 2019/2020 costs, assuming that urgent care visits have the same cost as regular primary care visits.	Region Stockholm (1177)[5] Last access: 5 Nov 2024
Hospitalization	28487.5	2020	There are four cost components of hospitalization costs: the one-time admission fee (Intagning, vårdavdelning), admission costs (including doctor consultations)(Intagning, läkarinsats), daily medical care costs (including doctor consultations) (Läkarinsats per vårddag), and daily nursing care costs (Omvårdnadsdag). The average length of stay is three days, so costs (excluding initial admission-related expenses) are multiplied by three. Since care intensity varies across units, the final cost is calculated by taking the median total cost for a three-day hospitalization across all units.	Own calculation, based on information from Södra Sjukvårdsregionen[6] Last access: 5 Nov 2024

[1] Full source: https://www.vardanalys.se/rapporter/besok-via-natet/
[2] Full source: https://skr.se/download/18.32563d7d1784aa279ece294c/1618741364556/11-2020-WEBB-Rek-om-ersattning-for-digitala-vardtjanster.pdf
[3] Full source: https://www.norrasjukvardsregionforbundet.se/halso-och-sjukvard/avtal-och-priser/arkiv/
[4] Full source: https://sodrasjukvardsregionen.se/verksamhet/avtal-priser/regionala-priser-och-ersattningar-foregaende-ar/
[5] Full source: https://www.1177.se/en/Stockholm/other-languages/other-languages/soka-vard/hitta-ratt-vard-nara-dig—andra-sprak—stockholms-lan/
[6] Full source: https://sodrasjukvardsregionen.se/verksamhet/regionala-priser-och-ersattningar/regionala-priser-och-ersattningar-foregaende-ar/

Table A12. Monotonicity subgroup test

	Prior hospitalization		Prior ED visit		Prior specialist visit	
	Any	None	Any	None	Any	None
Log median doctor wait time	0.057***	0.063***	0.063***	0.061***	0.060***	0.065***
	(0.0024)	(0.0011)	(0.0018)	(0.0013)	(0.0013)	(0.0017)
Baseline characteristics	✓	✓	✓	✓	✓	✓
Observations	92,343	397,756	164,120	326,142	312,264	177,973
K-P F-statistic	559	3,077	1,210	2,412	2,196	1,419
	Prior urgent care		Comorbidity		Gender	
	Any	None	Any	None	Female	Male
Log median doctor wait time	0.062***	0.062***	0.061***	0.062***	0.058***	0.069***
	(0.0021)	(0.0012)	(0.0023)	(0.0011)	(0.0013)	(0.0017)
Baseline characteristics	✓	✓	✓	✓	✓	✓
Observations	115,592	374,565	102,135	388,003	307,509	182,738
K-P F-statistic	841	2,774	694	2,924	2,068	1,575
	Income		Migration background		Further education	
	> median	≤ median	Any	None	Any	None
Log median doctor wait time	0.060***	0.062***	0.068***	0.060***	0.062***	0.061***
	(0.0018)	(0.0013)	(0.0021)	(0.0012)	(0.0012)	(0.0020)
Baseline characteristics	✓	✓	✓	✓	✓	✓
Observations	159,066	331,116	123,050	367,183	351,347	138,868
K-P F-statistic	1,157	2,470	1,025	2,596	2,681	943

Note: This table reports the results of Equation 19, which represents the first stage of the IV regression, across various subsamples, including the baseline case characteristics as control variables. The instrument is the winsorized natural logarithm of the median wait time in the up to 10 prior doctor consultations within 60 minutes (*Log median doctor wait time*). The baseline case characteristics include the symptom categories, demographics, region, along with login date-by-3 hour fixed effects. *Further educ.* refers to any post-secondary education reported for patients aged 25 or older, while income is reported for patients over the age of 20. The subsamples split by further education and median income exclude patients with missing or undefined information. Descriptions of all variables are provided in Appendix Table A10. The baseline mean refers to the mean of the variable *Initial assignment to nurse*. The First-stage K-P F-statistic refers to the Kleibergen-Paap F-statistic. Robust standard errors are shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

C 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 considerations. 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 Landsting SKL). 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*),[12] where nurses are lower cost than doctors, and thus represent the lower level of care when clinically appropriate (Socialdepartementet 2016). Accordingly, the firm is only paid for one consultation with the highest-level clinician type, while it is not paid for additional same-day consultations.[13]

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 Landsting SKL). The firm uses an algorithmic assessment before a full clinical evaluation by a human for this purpose. In practice, some cases are always to nurses (where nurses are clearly ex ante effective), some always to doctors (where nurses clearly cannot resolve the case), and the remaining cases are placed in an overlap group where either clinician type could be suitable for the initial assignment.

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 the national inspection bodies such as in Sweden and the UK, 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 and other clinicians in order to 'prioritize doctor consultations to where they make the most difference' (Kry International AB 2024).

Both clinician types differ in training. Doctors typically complete an 11-semester medical degree and an 18-month internship to obtain their license (Allmäntjänstgöring, AT), 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 do not.[14]

[12] Similar principles exist in other countries, such as Norway, the Netherlands, and England (NHS England n.d.; Ministry of Health, Welfare and Sport 2018; Royal Norwegian Ministry of Health and Care Services 2006).

[13] This rule has been confirmed by firm representatives and by regional payment guidelines (Region Kronoberg 2025; Region Skane 2023).

[14] In our sample, about 10% of nurses hold a specialty qualification, compared with more than 75% of doctors who

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 treat more complex conditions. Hourly wages in Nordic countries are relatively compressed (Mogstad et al. 2025), but cross-country comparisons are difficult due to differences in employment arrangements; in Sweden in 2020, average annual salaries were about 38,700 SEK for nurses and 79,200 SEK[15] for doctors (Statistics Sweden SCB). Sweden faces shortages of healthcare workers, particularly among specialized nurses and specialized doctors (Socialstyrelsen 2022). The nurse-to-doctor ratio is about 2.5:1, similar to other OECD systems (Table 1).

D Data Appendix

D.1 Data sources

The primary data is supplied by a large healthcare firm in Sweden. Their data contains 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 purpose of the consultation, the clinician type, the contact method, and primary diagnoses recorded in the form of the ICD-10 codes, the standardized system of disease classification. 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 us to link patients' consultations to additional individual-level information from Statistics Sweden (SCB). We also observe patients' healthcare utilization by linking their unique identifiers from the consultation records to information from the National Board of Health and Welfare (Socialstyrelsen, NBHW). In addition, Region Scania and Region Stockholm provide us with data on all primary care visits (not only within the firm we study but also for other providers) in Scania and Stockholm for patients who are registered in Scania or Stockholm.

Below, we describe our data sources in more detail.

D.2 Primary Care Data

We observe all primary care visits, including visits outside the firm Scania is a region in southern Sweden with about 13% of the country's population, and Stockholm is a the densely populated capital region with about 20% of Sweden's population. Data from Region Scania covers the period from 2013 to 2020, and data from Region Stockholm covers the years 2013 to 2023. These primary care data are linked to our other data sources through the unique patient identifiers. For primary care visits in Scania, we observe the date of the contact, clinician type, contact type, healthcare clinic, and main diagnoses in the form of the ICD-10 code. For primary care visits in Stockholm, we observe the date of the contact, the care branch, the pseudonymized identifier of the healthcare clinic,

are in or have completed specialty training.

[15] 1 SEK = 0.1087 USD in 2020 (Sveriges Riksbank 2020).

and main ICD-10 diagnoses. We cleaned symptom names by merging identical symptoms for children and adults, as well as symptoms with slightly different names but referring to the same condition, including differences in abbreviations and capitalization, and translated all Swedish symptom names into English.

D.3 Demographic and Socioeconomic Data

Statistics Sweden (SCB) provides us with basic information about patients' immigration background. In addition, we obtain information about patients from three key datasets from SCB: 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' civil status and municipality of residence from 2013 to 2018. LISA provides socio-economic and demographic data for patients, including income, rehabilitation benefits, employment status, paid sickness, and income-related benefits, for the years 2013 to 2020, with additional income and employment data available for 2019 to 2022. Yearly data at LISA is typically updated in November of each year. Finally, UREG offers data on patients' educational backgrounds for the years 2013 to 2020.

We use demographic and socioeconomic micro-data about patients from Statistics Sweden, specifically from LISA, RTB and UREG. The dataset provides data at the patient-year level, but information is collected at different points during the year: Income and employment status are measured in October each year, educational attainment is measured at the next year's spring semester, immigration background does not change over time, and the remaining variables (such as civil status, registered region) are measured at the end of each year.

We use the 2018 values for all control variables to ensure they are recorded before the earliest consultation starts. When there are missing values for education, Missing education values in 2018 are imputed using a backward fill: we first use 2017 values, and if those are unavailable, we use 2016 values.

D.4 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. 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 outside of primary care is 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 including 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 to the national insurance system and any additional patient costs), ATC codes, and a prescriber category (such as primary care, specialist care, school nursery etc.).

The prescription data cover all patients in our sample. Using the unique clinic identifiers and information on primary care prescriptions allow us to infer visits to primary care centers outside of the provider, which is missing for patients outside Stockholm and Scania.

D.5 Details of primary care and downstream costs

Cost estimates are obtained from the sources presented in the table A11. 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 Stockholm and 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, pulmonology, infectious diseases, dermatology, urology, orthopedics, ophthalmology, otolaryngology, pediatrics, and gynecology; the estimate covers routine and acute specialty visits and excludes fields like oncology and psychiatry that typically involve ongoing therapy (Norra sjukvårdsregionförbundet 2018). Hospitalization costs are computed as the sum of median average costs across four components: the one-time admission fee, admission costs, daily medical care, and daily nursing care (daily care types are multiplied by the sample average three-day stay). Detailed costs of hospitalization is provided by the Southern Healthcare Region Södra sjukvårdsregionen (2020), and admission and medical care costs include doctor consultations.

D.6 Matching consultations to their booked visits

Here we describe how we 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 A10 presents definitions of the case characteristics used in our main analysis.

E Model Appendix

Here we specialize the general framework of Section 4.1 to our setting. The *experts* are doctors and the *frontline workers* are nurses, so 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 $Q_t^{\text{ov}} \geq 0$ denote the number of doctor-group and overlap-group cases, respectively, to be assigned in period t.[16]

Doctors are first assigned to doctor-group cases. Define

$$\tilde{E}_t \;\equiv\; E_t - \min(E_t,\, Q_t^D) \;\geq\; 0 \tag{9}$$

as the number of doctors remaining after doctor-group cases have been assigned; any unserved doctor-group cases carry over to the next period. The $\tilde{E}_t$ remaining doctors and F_t nurses are then available for overlap-group cases. Let

$$M_t^{\text{ov}} \;\equiv\; \min(\tilde{E}_t + F_t,\, Q_t^{\text{ov}}) \tag{10}$$

denote the number of overlap-group cases assigned in period t; unassigned overlap-group cases carry over. Note that M_t^{ov} plays the role of M_t from Section 4.1, restricted to the overlap group. We assume escalated cases are processed in later slots outside the main queue, so $\tilde{E}_t$ in (9) is calculated before any escalated cases are added back to the 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 in both the direct-to-doctor route and in the escalated branch of the knowledge hierarchy it is the doctor who resolves the case, this rule implies

$$r_{F,E} \;=\; r_E. \tag{11}$$

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

We parameterize revenues using a common wedge $\theta \in (0,1)$ between administratively set prices and the firm's production costs:

$$r_E \;=\; \frac{\kappa_E}{1-\theta}, \qquad r_F \;=\; \frac{\kappa_F}{1-\theta}, \tag{12}$$

so that each price covers the corresponding production cost with a proportional margin $\theta/(1-\theta)$. Together, (11) and (12) give the full revenue vector

$$(r_E,\, r_F,\, r_{F,E}) \;=\; \left(\frac{\kappa_E}{1-\theta},\, \frac{\kappa_F}{1-\theta},\, \frac{\kappa_E}{1-\theta}\right), \tag{13}$$

with $r_{F,E} = r_E > r_F$ since $\kappa_E > \kappa_F$.

Unconstrained assignment. Substituting (13) into (3), the net revenue advantage of the knowledge hierarchy simplifies to

$$\Delta(\tau) \;=\; p_\tau \cdot \frac{\kappa_F - \theta\kappa_E}{1-\theta} - \kappa_F. \tag{14}$$

Since $\kappa_F < \kappa_E$ and $p_\tau \le 1$, we have $\Delta(\tau) \le (\kappa_F - \theta\kappa_E)/(1-\theta) - \kappa_F = -\theta\kappa_E/(1-\theta) < 0$ for all τ.[17] By Result 1, the unconstrained firm therefore assigns all overlap-group cases directly to doctors.

Constrained assignment. When doctor availability binds ($M_t^{\text{ov}} > \tilde{E}_t$), the firm must assign some overlap cases to nurses. Evaluating condition (3) under (13) gives

$$\frac{\partial \Delta(\tau)}{\partial p_\tau} \;=\; r_F - r_{F,E} + \kappa_E \;=\; \frac{\kappa_F - \theta\kappa_E}{1-\theta}, \tag{15}$$

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 our setting (see Section 4.3). By Result 3, 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.

Two alternative (hypothetical) revenue structures. We consider two departures from (13) to illustrate how different pay structures affect incentives.

Fixed payment per resolved case. Suppose the firm receives a flat fee f per resolved case irrespective of clinician type:

$$(r_E,\, r_F,\, r_{F,E}) \;=\; (f,\, f,\, f). \tag{16}$$

Then $\partial\Delta(\tau)/\partial p_\tau = f - f + \kappa_E = \kappa_E > 0$, so condition (3) holds unconditionally and the constrained firm assigns higher-p_τ cases to nurses.

Flat payment per consultation. Suppose the firm is paid ρ per *consultation* regardless of clinician type. Because escalating an unresolved case to a doctor generates two billable consultations (one

[17] More precisely, $\Delta(\tau) < 0$ for all $p_\tau \in (0,1)$ because $\kappa_F - \theta\kappa_E < \kappa_F(1-\theta)$ iff $\theta\kappa_F < \theta\kappa_E$, which holds since $\kappa_E > \kappa_F$.

with the nurse, one with the doctor), while a directly resolved case generates only one, the revenue vector is

$$(r_E,\, r_F,\, r_{F,E}) \;=\; (\rho,\, \rho,\, 2\rho). \tag{17}$$

Then $\partial\Delta(\tau)/\partial p_\tau = \rho - 2\rho + \kappa_E = \kappa_E - \rho$. If $\kappa_E < \rho$, condition (3) 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 pay structure therefore incentivizes the firm to *maximize* rather than minimize production costs when constrained, which may rationalize governments' avoidance of per-consultation fees.

Social planner's problem. The social planner's objective specializes equation (6) in a natural way. 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 total production costs:

$$\Delta^{SC}(\tau) \equiv \mathrm{SC}_F(\tau) - \mathrm{SC}_E(\tau) \;=\; [\kappa_F + (1-p_\tau)\kappa_E + \mathbb{E}[C_{F,\tau}]] - [\kappa_E + \mathbb{E}[C_{E,\tau}]], \tag{18}$$

where $C_{F,\tau}$ and $C_{E,\tau}$ are the downstream cost random variables defined in Section 4.1. The planner assigns type-τ cases to the knowledge hierarchy whenever (18) is negative, and to doctors otherwise. When clinician availability is constrained, the planner ranks symptoms by (18) and fills nurse capacity from the top of this ranking, exactly as in Section 4.1. Because 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 4.1.

F Additional robustness

F.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. In the IV strategy, our first stage takes the following form:

$$Nurse_i = \gamma_i \text{Doctor Congestion}_i + X_{1i}\lambda_{1i} + X_{2i}\lambda_{2i} + \nu_i, \tag{19}$$

where 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 Equations 7 and 19 with robust standard errors.

In the IV setup, δ in Equation 7 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. First, the congestion instrument must be relevant by shifting the initial assignment. Figure A2 shows that the probability of initial nurse assignment increases with doctor congestion. Appendix Table 3 confirms this relationship conditional on the full set of controls (Column 3) and reports the first-stage estimates from Equation 19. The first-stage coefficient is 0.056 and precisely estimated, with an F-statistic of 3,725, confirming strong instrument relevance.

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_{1i} 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 1 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 A11 shows that, for every symptom, the share of cases initially assigned to nurses weakly increases with congestion. We further assess monotonicity by examining the relationship between congestion and assignment across subsamples (Bhuller et al. 2016; Dobbie et al. 2018). Appendix Table A12 reports first-stage estimates by health risk, demographics, and socioeconomic status. Across all subsamples, coefficients are positive, statistically significant, and similar in magnitude, providing no evidence of monotonicity violations.

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 A5), 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.

A potential concern is if congestion directly affects outcomes. Appendix Table A5 shows that 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, waits even in the highest congestion quartile remain below 45 minutes, which is unlikely to meaningfully affect care quality or costs, especially relative to physical primary care, where over 25% of patients wait more than 24 hours for an initial doctor appointment.

F.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.