Organizing Expertise with Externalities: Evidence from Primary Care

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

[1] *University of Zurich,* [2] *Stockholm School of Economics; University of Copenhagen,* [3] *Stockholm University,* [4] *London School of Economics*

March 27, 2026
PRELIMINARY - PLEASE DO NOT POST OR CIRCULATE

Abstract

Many modern organizations use knowledge hierarchies, where frontline workers handle incoming tasks and pass those they cannot resolve to experts. But the implications of this organization of expertise are not well-understood, especially when they cause downstream externalities. We study a knowledge hierarchy (KH) using 500,000 patient cases from an online primary care firm, linked to outcomes and costs across Sweden's healthcare system. We use variation in assignment arising from temporary doctor congestion, controlling for date-by-three-hour windows and flexible case characteristics. Within the KH, nurses resolve 70% of cases and pass the remainder to doctors. Relative to doctor-only care, the KH modestly affects lower-stakes outcomes (e.g., primary care follow-up) but has no adverse effects on patient satisfaction or higher-stakes outcomes, including hospitalizations, earnings losses, and mortality. Although the KH does not change the mean systemwide (firm plus downstream) production costs, this masks substantial heterogeneity. Compared to using only doctors, the savings from the firm's selective use of the KH equal 2.3% of online primary care costs. However, reassigning doctors and nurses within 15-minute windows based on comparative advantage in systemwide production costs would save 6.7% of online primary care costs.

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

This paper was previously presented as "Does Division of Labor Increase Productivity? Evidence from Primary Care". 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" and SSE's House of Sustainable Society. 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, the division of labor often centers on organizing expertise using *knowledge hierarchies*, in which frontline workers handle simpler tasks and pass harder tasks to experts only when needed (Garicano 2000). Despite the prevalence of knowledge hierarchies, there is little empirical evidence on their productivity effects, the type of tasks assigned to them, or whether their use is efficient. This is especially true when there are downstream externalities, such as the shifting of tasks to experts in other organizations.

Empirically evaluating knowledge hierarchies relative to a natural benchmark – 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 a setting, discussed below, which allows us to overcome the empirical challenges and identify the effects of a knowledge hierarchy, including its heterogeneous effects across task types and its downstream externalities.

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

To guide the empirical analysis, we develop a parsimonious model of task assignment between a knowledge hierarchy and a direct-to-expert route. The model isolates key tradeoffs that firms face when experts and frontline workers differ in their costs, their probabilities of resolving different tasks, and the revenues that such resolutions generate. We study how revenue structures affect assignment and highlight the role of downstream externalities.

Understanding how knowledge hierarchies are used and that this implies is especially important in sectors where expertise is scarce, demand is rising, and services are publicly financed. The healthcare sector, which we study, exemplifies these forces: aging populations, clinician shortages (Maier and Aiken 2016), and rapid technological change are reshaping how expertise is produced and deployed (Dahlstrand 2026; Dahlstrand, Le Nestour, and Michaels 2024; Bronsoler, Doyle, and Reenen 2022; Rajpurkar et al. 2022). In such settings, delegation is not merely an internal managerial choice – it can influence quality, costs, and who ultimately bears the costs. 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 – possibly including taxpayers, e.g., in Medicare, Medicaid, and many European health systems. We quantify these productivity

and externality effects, bringing organizational design directly into the domain of public policy.

Our specific 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. We leverage an algorithmic assignment of patient cases to routes. Algorithmic assignment is particularly well-suited to our identification strategy: unlike human triage, it relies on a fixed and observable set of inputs, leaving little room for unobservable characteristics to drive selection into routes. Conditional on these inputs, residual variation in assignment is plausibly exogenous.

We apply our above-mentioned model to study the firm's algorithm, which assigns cases either directly to doctors or initially to nurses (the KH). The firm maximizes its net revenues, 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 clinician use constant), it does not minimize systemwide production costs. This is because the firm is not given a contract that internalizes the downstream costs, which are dependent on the route assignment – a problem common to many contractual environments. In our model, we show that efficient assignment of tasks to routes follows the comparative advantage of the KH in systemwide production 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 includes measures of a wide range of 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. We link these cases 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 measure 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 interactions of case characteristics observed by the algorithm. We show that, conditional on these controls, case characteristics that the algorithm cannot observe are almost exactly balanced. Our main findings are also robust to using a measure of short-term doctor congestion as an instrumental variable. This congestion only causes short delays – more than 90% of cases are handled within an hour and more than 99.9% within a day. In primary care settings such as ours, such delays are unlikely to materially affect health outcomes. This sets us apart from the hospital emergency department settings in Costantini (2025), Godøy et al. (2024), or Gruber, Hoe, and Stoye (2023).

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. The 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. While the KH modestly increases the rate of external primary care visits and emergency department visits, it does not significantly change the incidence of hospitalizations. We also find no adverse effects of the KH on major earnings losses in the calendar month following the initial consultation, nor on three-year mortality rates. Together, these results suggest that on average, the KH does not materially affect patient health.

We next assess the KH's effects on costs, both within the firm and in downstream healthcare services used within seven days. We estimate that the KH reduces taxpayer payments to the firm (equivalent to firm gross revenues) under current payment rules, where the firm is only paid for the clinician who resolves a case but incurs costs for all clinicians involved, which lowers the firm's revenue net of labor input (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 systemwide production 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, but on average, this cost saving is offset by increased costs higher up in the distribution. This substantial heterogeneity suggests that the assignment of tasks to the KH or directly to doctors is an important lever for healthcare productivity.

Turning to the economic drivers of case assignment to routes, we find that when the firm is unconstrained – that is, there is no doctor congestion – it assigns almost all cases directly to doctors. When short-term doctor congestion arises, the key characteristics that determine assignment are the patients' symptoms: symptoms with a higher relative net revenue advantage under the KH (cases that nurses are likely to resolve themselves) are more likely to be assigned to nurses. Consistent with the model, this assignment pattern also reduces taxpayer payments to the firm, conditional on keeping clinician use fixed. Compared to a counterfactual scenario in which only the direct-to-doctor route is used, the firm's selective use of the KH reduces systemwide production costs by 0.7% (2.3% 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 systemwide production costs, while respecting short-run clinician capacity constraints. Reallocating cases on this basis within 15-minute windows without changing actual nurse or doctor use reduces systemwide production costs by 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 systemwide production costs. Moreover, it suggests that the KH itself is a useful organizational tool to reduce systemwide production costs when applied to the most suitable tasks. Implementing it for the right tasks – those 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 seen in in-person primary care; and that our main results are similar when we reweight by demographics to represent in-person care users, or focus solely on the period before Covid-19. Although these relevance checks do not guarantee that our point estimates generalize exactly to other settings, 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 setting reflects a broader problem of task allocation across skill levels when processing at different tiers generates downstream spillovers outside the decision-maker's objective. For example, in customer support, delegation may affect retention and platform demand; in infrastructure maintenance, network reliability and congestion; and in financial screening, portfolio risk borne by investors or taxpayers. In such settings, privately optimal delegation need not coincide with socially efficient assignment.

Our paper contributes to the literature by providing 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 knowledge hierarchies (Garicano 2000) in three ways: in our empirical focus; in the cost of knowledge hierarchies that we emphasize – frontline workers must attempt a task before determining whether they can resolve it, rather than incurring communication costs; and in our consideration of downstream externalities.

Our study also differs from research that compares occupations head-to-head, most notably Chan and Chen (2022). Rather than estimating relative productivity when two clinician types are direct substitutes, we study a knowledge hierarchy in which nurses attempt cases first and selectively pass some to doctors. We focus not only on absolute advantage, but also on comparative advantage and on its role in the selective assignment of task types between routes. We further differ in our setting of primary care rather than the emergency department, and in analyzing 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, Mahoney, and Udalova 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, Kilby, and Marks 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 A.

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

[1] During our sample period, annual out-of-pocket spending for outpatient visits was capped at roughly 1,150 SEK, or about USD 125 (1 SEK = 0.1087 USD in 2020, Sveriges Riksbank 2020). Online primary care visits carried a copay of 100–200 SEK (about USD 11–22) for adults and were free for minors and the elderly; these payments counted toward the annual cap. Taxpayer fee-for-service payments to healthcare providers are reduced by the amount of the copay if a copay is paid.

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 (Kry International AB 2024). 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 levels intended to reflect production costs (Sveriges Kommuner och Landsting (SKL) 2019).

The firm's task assignment is influenced by a notion of '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 (we refer to this as the 'resolving clinician' rule). As a result, the firm always assigns certain cases to nurses, others always to doctors, and the remainder ('*overlap-group cases*') to either clinician type. Cases of a given type are processed as 'first-come-first-served' (FCFS), reflecting expectations of transparency and predictability and limiting long or arbitrary waits. Although these constraints reflect the Swedish environment, similar considerations are common in many other settings.

Similar revenue structures arise in other knowledge-hierarchy settings. For example, in IT and customer support, the level of incident resolution may result in non-additive payments across layers, and in professional services such as law and consulting, higher-skill work is often billed at higher rates, but contracts may be similarly non-additive.

2.2 Cases at the firm

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

Our study focuses on *overlap cases* after spring 2019 from patients who were not registered with the firm; this means that the firm was not these patients' in-person primary care provider, but an additional venue for care. During the period of our study, assignment to a nurse or a doctor was

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 remains 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.

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

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 routine care and patient communication, whereas doctors undergo at least seven years of formal and on-the-job medical 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 with some differences noted below.

Clinicians use a digital interface that shows the consultations assigned to them, but information on the firm's total current congestion is less easily observable than in most in-person healthcare settings. Upon assignment, clinicians observe patients' demographics, reported symptoms, and medical histories from prior interactions with the firm. Doctor and nurse consultations may result in home-care advice, suggestions about over-the-counter medicines, and/or recommendations about further care (e.g., to visit the ED). Doctor consultations may also result in prescriptions, longer-term sickness absence certificates, and/or referrals to specialists outside the firm.[2] In this setting, nurses go beyond triage – they do not merely prioritize or route patients, but often complete the episode of platform-based care and give health advice.

If a nurse cannot resolve a case, the nurse passes it to a doctor by scheduling a follow-up consultation for a later time. The follow-up consultation specifies that a doctor is required, but does not name an individual doctor. The nurse also completes a handoff note summarizing symptoms, the reason for passing the case, and any relevant clinical information. The mean total time spent on each case by the initial clinician is around 10-12 minutes (with a standard deviation of around 7 minutes), irrespective of clinician type and whether the case was resolved.

[2]Doctors and 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.

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 B.

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 the age of 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 incomplete consultations in Appendix Section D.2 and address potentially selective attrition in robustness checks in Section 6.3.

3.2 Variables

We summarize the key variables used in our analysis. Additional details are provided in Appendix B, where we also provide the precise variable definitions.

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 (X_1). 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 before 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 in which patients have not selected 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 algorithm's assignment rules, we interact symptoms with other variables that the algorithm observes: the patient's age, gender, and prior cases at the firm, including whether the patient's most recent observed case at the firm (if any) was assigned to a doctor and whether the patient has received a prescription written by a doctor at the firm. We also include patients' aggregated region of residence, measured 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 patient registers the case 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 short-term congestion that has no correlation with predicted healthcare costs.

Additional controls (X_2). 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 earnings, welfare 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 doctor (or another doctor consultation in the case of the direct-to-doctor route), case resolution, patient satisfaction, and diagnostic informativeness; (ii) downstream healthcare use within seven days after the initial consultation date, including any new prescription and use of external primary, specialist, urgent (often out-of-hour centers), emergency, and hospital care; and (iii) longer-run consequences, capturing earnings declines and mortality to capture adverse events for patients outside their direct healthcare utilization.

Cost outcomes. We also examine how the knowledge hierarchy affects taxpayer payments to the firm and as a whole, firm revenues, and systemwide production 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. We also measure downstream costs — the costs of out-of-firm care borne by taxpayers — counting up to one instance of each downstream event within seven days after the initial consultation. Downstream costs are

calculated using service-specific cost estimates from Appendix Table A12. In robustness checks, we vary the assumptions on production costs and we also include full aggregation of all 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 — most cases involve very short waits, with over 60% below 20 minutes, while a small fraction experience relatively longer waits: fewer than 10% wait more than 60 minutes (Subfigure A2a). Moreover, 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. While this table shows the raw data assignment, in Section 5, we discuss balance, accounting for the baseline control variables in our identification strategy.

The analysis sample spans a broad range of demographic characteristics, but with a somewhat different mix than 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.

4 Model

This section introduces a model whose main role is to organize our empirical analysis and clarify mechanisms that determine the use of knowledge hierarchies. We begin with a general model of a firm that faces a stream of heterogeneous tasks (e.g., patients, customers, orders, or applications) that require processing. We focus on short-run task assignment, abstracting from staffing and hiring decisions. Tasks can be assigned either directly to experts or to frontline workers, who resolve a subset and pass the remainder to experts (the knowledge hierarchy, KH). The model highlights how firms determine task allocation; how revenue structures (e.g., differential pay for expert versus frontline worker resolution) affect whether profit maximization entails cost minimization; and how downstream externalities can lead to task misallocation. We then apply this model to our setting,

where tasks are patient cases, frontline workers are nurses and experts are doctors, and the firm is paid by taxpayers. Finally, we discuss the model's implications for our empirical analysis.

Our model emphasizes the cost of frontline workers' attempts to resolve tasks even when these attempts fail. This differs from the canonical model of KH (Garicano 2000), which emphasizes communication frictions as the key cost to using KH. We note that communication costs have fallen dramatically in recent years (see, e.g., Bloom et al. 2014), and in many settings, including ours, they are likely less important than attempt costs. We also differ by examining when the firm's revenue structure shifts it away from cost minimization, and by analyzing 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.[3] When a frontline worker cannot resolve a task, they pass it to an expert; we assume experts resolve all tasks assigned to or passed 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.[4] 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 firm net revenues (gross revenues minus variable costs), ignoring fixed costs that do not vary with task assignment. For a task of type $\tau \in \mathcal{T}$, the firm's expected net revenues under the KH and the expert route are, respectively:

$$\begin{aligned} \Pi_F(\tau) &\equiv p_\tau (r_F - \kappa_F) + (1 - p_\tau)(r_{F,E} - \kappa_F - \kappa_E), \\ \Pi_E(\tau) &\equiv r_E - \kappa_E. \end{aligned}$$

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

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

Since p_τ enters $\Pi_F(\tau)$ only as a mixing weight between two fixed payoffs, and all revenue and cost

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

[4] We discussed in Section 2.2 that task processing durations vary little across worker types in our setting. Therefore, we could equivalently model the number of slots for each worker type in each period.

parameters are constants by assumption, $\Delta(\tau)$ is linear in p_τ with constant slope, given by:

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

Intuitively, if $\frac{\partial \Delta(\tau)}{\partial p_\tau} > 0$, and the firm has to choose how to assign two tasks between a frontline worker and an expert, it will assign to the frontline worker the easier task (with higher p_τ). Here and throughout the model, we assume that when indifferent, the firm breaks ties randomly.

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

Weak first-come-first-served (FCFS) constraint. The firm must (weakly) process tasks in the order they arrive when workers are available. This allows the firm to batch-process the remaining tasks in the queue that arrived earliest (subject to worker availability), and push the remaining tasks to the next period.[5]

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

Solving the firm's problem. Let M_t denote the number of tasks, taken from the front of the queue, to be assigned in period t, where $M_t \leq E_t + F_t$. Any remaining tasks are moved to the front of the queue for period $t+1$. The firm is *unconstrained* if the number of tasks with $\Delta(\tau) > 0$ does not exceed F_t and the number of tasks with $\Delta(\tau) < 0$ does not exceed E_t; otherwise, assignment is *constrained*. Whether the firm's constrained assignment also minimizes costs depends on its net revenue structure.

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)**.** *When staffing constraints bind, the firm allocates scarce worker types to the tasks for which they yield the largest net revenue advantage. The firm ranks tasks by $\Delta(\tau)$, assigns frontline workers to the highest-ranked tasks with $\Delta(\tau) > 0$ up to capacity F_t, assigns experts to the lowest-ranked tasks with $\Delta(\tau) < 0$ up to capacity E_t, and allocates any remaining available workers to any remaining tasks.*

Result 3 (Firm production cost minimization)**.** *If $\frac{\partial \Delta(\tau)}{\partial p_\tau} > 0$, then $\Delta(\tau)$ increases in task ease p_τ. Under staffing constraints, the firm therefore assigns available frontline workers to the tasks with the highest resolution probabilities (i.e., the easiest tasks). In this case, profit maximization coincides with firm production cost minimization.*

[5] Some firms may apply FCFS within task groups. We discuss such a scenario in Section 4.2.

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

The social planner's problem. Denote downstream (out-of-firm) costs for a type-τ as the random variable $C_{F,\tau}$ (where $\mathbb{E}[C_{F,\tau}] < \infty$) if the case is initially assigned to a frontline worker (the knowledge hierarchy, KH) and $C_{E,\tau}$ (where $\mathbb{E}[C_{E,\tau}] < \infty$) if directly assigned to an expert.[6] We consider a social planner who minimizes expected total production costs, which are the sum of firm production costs and downstream costs. For a type-τ case, expected total production costs are $SC_F(\tau) \equiv \kappa_F + (1 - p_\tau)\kappa_E + \mathbb{E}[C_{F,\tau}]$ under the KH, and $SC_E(\tau) \equiv \kappa_E + \mathbb{E}[C_{E,\tau}]$ under direct-to-expert assignment. The social planner's assignment rule follows the difference in these costs:

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

When staff availability is unconstrained, the social planner assigns type-τ tasks to the KH if and only if $\Delta^{SC}(\tau)$ is negative – i.e., it assigns according to the KH's absolute advantage. When staff availability is constrained, the planner instead ranks task types by Equation 3, and prioritizes assignment to the KH for tasks with the lowest $\Delta^{SC}(\tau)$, subject to the constraints. Since the firm does not internalize downstream costs, the social planner's assignment may be different from the firm's.

4.2 The problem of the primary care firm

In Appendix C we characterize the problem of the primary care firm we study, and the associated social planner problem, and we provide an overview here. 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.[7] We focus on a setting in which production is paid for by taxpayers. The firm is compensated only for the clinician type that resolves the case, and we parameterize revenues

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

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

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.

4.3 Implications for the empirical analysis

Parameter values. We set $r_F = 275$ and $r_E = r_{F,E} = 500$, since the firm is only paid for the 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 the cases into types. We begin our empirical analysis with a flexible definition of types and later focus on symptoms.

Since we do not have data on the firm's production 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, which sets the administrative product prices, we see little role for markups. Instead, we focus on labor market power, and assume $\theta = 0.25$, which is close to the middle of the range implied by the literature on monopsony markdowns (e.g. Azar and Marinescu 2024). In Section 6.3 we assess the robustness of our findings to varying θ, and allow for different wedges by clinician type.

Empirical predictions. Our model yields implications for the empirical analysis, in which 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, types fully determine potential outcomes for firm production costs and downstream costs, so controlling for case-type fixed effects suffices for OLS identification. In practice, we face two concerns: first, we do not know how types are defined in the firm's assignment algorithm, and second, even for cases of a given type, potential outcomes in downstream costs for a type may vary by seasons, days of the week, times of the day, and other transitory factors (e.g., pandemics). To address the first concern, we control flexibly for interactions of the patients' symptoms, demographics, and medical histories within the provider – the key factors that the firm's assignment algorithm observes. To address the second, we control flexibly for time windows within a calendar day.

Prediction 1a: First-come-first-served (FCFS). The model assumes that within each case group (doctor group, nurse group, and overlap group), 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. Appendix C shows that in our setting $\Delta(\tau) < 0$ for all τ, so by Result 1, the unconstrained firm assigns all overlap-group cases directly to doctors. For overlap-group cases, temporary doctor shortages shift the firm's problem from the unconstrained situation (when all cases are assigned to doctors) to the constrained situation, where some are assigned to nurses. Thus, 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. If, in a congested situation, the cases immediately preceding (overlap-group) case i in the queue are doctor-group cases, case i may be assigned to a nurse. We test this prediction, but do not use queue composition as an instrument, since it only affects assignment during congested situations, and is therefore weaker than the congestion instrument.

Prediction 2: Mean effects of the knowledge hierarchy. Using our identification strategy (Prediction 1), we estimate the effects of the KH on care quality and costs for overlap cases. Since $\Delta(\tau) < 0$ for all τ, the model predicts that KH reduces firm net revenues and taxpayer-to-firm payments. Effects on healthcare outcomes and downstream costs (and thus on total taxpayer payments and systemwide production costs) are theoretically ambiguous and are estimated empirically.

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

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

Prediction 5: Counterfactuals. Because the firm does not observe (nor internalize) downstream costs, its case assignment does not necessarily minimize the systemwide production costs of producing care. In other words, the firm may exert a downstream cost externality by allocating cases to clinicians who reduce firm production costs but increase downstream care costs. Our analysis of the social planner's objectives motivates a counterfactual case assignment that minimizes expected systemwide production costs 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. Prediction 1 motivates our identifying assumption that, conditional on the (short) time window and the case type, the initial assignment of each overlap case to a nurse (versus a

doctor) is random. To translate this into an estimation strategy, we need to make assumptions about both the time windows and the types, which we do by estimating OLS specifications of the form:

$$Y_i = \delta_i Nurse_i + X_{1i}\beta_{1i} + X_{2i}\beta_{2i} + \epsilon_i, \tag{4}$$

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_{1i} includes predetermined case characteristics likely used by the assignment algorithm (date-by–3-hour windows of patients' log in time; and symptom indicators interacted with indicators for the patient's age, gender, aggregated region, whether they have any past prescription from the firm, and whether their last case at the firm – if any – was with a doctor). This specification reflects a pragmatic tradeoff: ensuring that the time windows are tight and the variables used to determine "types" are flexible, while retaining sufficient degrees of freedom. X_{2i} includes other predetermined characteristics (health risk and socioeconomic background), which, to the best of our knowledge, the algorithm cannot observe (see the description of 'Additional controls' in Section 3.2). δ_i captures the (potentially heterogeneous) effect of initial nurse assignment – the knowledge hierarchy. Since treatments are not naturally clustered, we report robust standard errors (Abadie and Cattaneo 2018) on the error term (ϵ_i).

To support this identification strategy, we present here two pieces of evidence. First, Figure 1 examines balance in initial case assignment across additional health risk and socioeconomic characteristics X_{2i}, conditional on baseline case characteristics X_{1i}. Assignment is balanced for most variables, and even when it is not, the differences are economically negligible. We summarize these into a measure that predicts systemwide production costs, and we find an economically small relationship, for which we cannot reject the null of balance (coef.: $-3.22 \cdot 10^{-6}$, s.e.: $2 \cdot 10^{-6}$, p-value: 0.11). The predicted healthcare costs are strongly correlated with actual systemwide production costs (coef.: 0.881, s.e.: 0.022).[8] Second, Table 3 reports the explanatory power of case characteristics and exogenous shifters for initial nurse assignment. Baseline characteristics (X_1) alone account for a substantial share of the variation (adjusted $R^2 = 0.195$, Column 1). This contrasts with using only the additional controls (X_2), which have almost no explanatory power or additional explanatory power (Columns 2 and 3). Adding doctor congestion increases the explanatory power (adjusted $R^2 = 0.200$, Column 4), but doctor-group status in prior cases does not (Column 5).

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

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

[8] Moreover, we conduct two one-sided t-tests (TOSTs) for each additional case characteristic, conditional on the baseline controls. We 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).

cases (Columns 3 and 4), consistent with the model's assumption regarding temporary doctor shortages, while nurse-group cases are substantially delayed (Column 4), consistent with lower urgency and processing outside the main queue.

Short-term doctor congestion. We measure doctor congestion using *log median doctor wait time* of recent patients, defined above in Section 3.2 and further discussed in Section D.1 in the Appendix. Table 3 and Figure A2 test the model's Prediction 1b: that doctor congestion increases the propensity for initial nurse assignment, and show consistent evidence that there is a significant increase in nurse assignment when there is more doctor congestion. Appendix Figure 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, attrition (measured by no-shows and technical errors), or very short consultations.[9] 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 (Table 3, Column 4), we focus on doctor congestion as our instrumental variable.

6 Results

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

6.1 Effects of the knowledge hierarchy on patient outcomes

Table 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 B.

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 within the firm, unbooked (re)visits 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

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

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 (Table 4), or by 6.4 percentage points (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 recurrent 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.[10]

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 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 earnings and mortality. If anything, KH reduces the probability of earnings declines from primary employment in the calendar month following the initial consultation, relative to average earnings in the preceding three months, for patients with stable prior employment. KH reduces the incidence of earnings declines of 20% or more by 2.1 percentage points from a baseline of 22%, and reduces the probability of zero earnings by 0.3 percentage points from a baseline of 5%. We estimate a precise zero effect on three-year mortality.[11]

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 and

[10]The unconditional new prescription rate for cases passed to doctors in the KH, 42%, is comparable to that among cases assigned directly to doctors. 18% of cases that nurses resolve (within the firm) result in new prescriptions, likely reflecting prescriptions in downstream care.

[11]Observed earnings reductions reflect sick leave, as employer-paid earnings fall to 80% for the first 14 days and to zero thereafter, at which point the Swedish Social Insurance Agency pays sickness benefits. Employees can self-certify sickness for the first seven days, but longer sick leaves need to be signed off by doctors; thus, the slightly lower rate of earnings losses after KH may partly reflect lower access to doctors (Försäkringskassan 2025).

on costs including downstream care. Our cost measures are constructed using taxpayer payments for the categories listed in Appendix Table A12 and the model assumptions described in Section 4.

Taxpayer payments to the firm and firm net revenue. The KH substantially reduces taxpayer payments to the firm (which equal firm gross revenue). Over a seven-day window (when including up to one additional within-firm visit), payments in the KH route are 149 SEK lower, a reduction of $\approx 28\%$ 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 firm production cost effects. As predicted by the model, these taxpayer savings are offset by losses to the firm, since initial nurse assignment reduces firm net revenue by 108 SEK, relative to a baseline of 125 SEK under doctor assignment.

Firm production costs, total taxpayer payments, and systemwide production costs. Within-firm, the KH reduces production costs by 40 SEK, more than 10% relative to a baseline mean of 375 SEK. To calculate total taxpayer payments and systemwide production costs, we consider payments and costs on the day of the initial consultation and count at most one instance of each out-of-firm downstream event within seven days, assuming that downstream costs are all borne by taxpayers. Taxpayer payments reflect public healthcare spending, while systemwide production costs net out transfers between taxpayers and the firm. The KH reduces mean total taxpayer payments by 117 SEK, or about 9% relative to the baseline mean of 1,304.8 SEK. In contrast, mean systemwide production 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 systemwide production costs of care.

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

Subfigure 2a presents the CDFs of taxpayer payments to the firm 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 systemwide production 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 systemwide production costs.

Assessing identification. Our baseline OLS specification conditions on a rich set of predetermined case characteristics likely used by the assignment algorithm, as well as on 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 A9 shows that estimated effects on systemwide production costs are mostly unchanged when we (i) control only for baseline case characteristics, (ii) use finer-grained time fixed effects (date-by-1-hour windows), and (iii) explicitly control for doctor congestion by including congestion-quartile indicators. To address selection on unobserved patient traits, such as patient assertiveness, we also 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 any remaining concerns about unobserved selection using two approaches. First, in Panel A of Appendix Table A9 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 systemwide production costs, –35.6 to 76.5 SEK, includes zero and remains small. Second, Appendix Tables A7 and A8 complement these checks with an instrumental variable (IV) strategy based on Prediction 1b using log median doctor wait time in recent prior cases as an instrument (Kleibergen-Paap F-statistic: 3,670). This strategy does not require our controls to capture the exact functional form of the assignment mechanism and therefore provides a useful check 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 systemwide

production costs remain small and statistically insignificant. Quantitatively, most IV estimates are similar to OLS as a proportion of the baseline mean.[12] 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 D.2 discusses this and the data we have on attrition in more detail. Panel B of Appendix Table A9 addresses this concern using two complementary approaches: bounding the maximum bias, and probing whether our estimates differ in subsamples in which incentives or opportunities for selection 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 sample, we use administrative case-type-specific dropout rates from a later period and calculate the Lee bounds conditional on case types rather than our baseline controls, X_1. Differential attrition is small in that data, with a 2.68 percentage-point higher dropout rate on average for cases assigned to nurses. In our sample, the attrition we can measure consists of no-shows and technical errors, and the pattern there is the opposite: a 0.3 percentage point higher attrition in the direct-to-doctor route. Nonetheless, since healthcare costs exhibit fat right tails, the Lee bounds will be more conservative. 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 systemwide production costs. In addition, exploiting a change in the assignment process around mid-2019, when clinician type was assigned only shortly before the consultation, we test for differences in the treatment effect across periods (April-June versus July-September 2019) by interacting the nurse indicator with a second-period indicator.[13] 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 costs. Appendix Table A9 examines the robustness of our mean and distributional results under alternative assumptions and specifications.

[12] For some patient outcomes, the baseline means are very low, which can yield imprecise point estimates using the IV strategy. This, combined with the fact that the OLS and IV estimates are overall very similar, is a reason to use the OLS as our main identification strategy.

[13] 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.

Alternative cost calculations and markdown assumptions. In our main specification, we calculate 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 A9 reports mean total healthcare cost estimates under alternative cost calculation schemes: restricting the sample to cases in Stockholm and Scania, for whom downstream care including all external primary care is fully observed; and using aggregate costs, defined as the summed costs for all downstream events within 7 or 30 days. These alternative cost calculations do not affect our main conclusions.

In addition, Appendix Figure A7 shows the robustness of systemwide production costs to varying and differential markdowns by clinician types. Our main specifications assume a common wedge of $\theta = 0.25$ between the price and the firm's production costs of doctor or nurse consultations, consistent with existing monopsony markdown estimates implying $\theta \in [0.13, 0.33]$ (Section 3.2). Given the administrative price-setting procedure, it is likely that the wedges do not vary much by clinician type.[14] 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 $\approx 5\%$, reflecting that firm production costs constitute only a small share of systemwide production costs and most costs accrue in downstream care.

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 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 A9 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 D of Appendix Table A9 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.

Effects on the cost distribution. Our baseline estimates imply a small and statistically insignificant mean effect of the KH on systemwide production costs, but the KH meaningfully reshapes the

[14]The prices are set by policymakers based on cost calculations of the labor costs of each clinician group and the firm's other operating costs (Sveriges Kommuner och Landsting (SKL) 2019).

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 Task assignment

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

7.1 Task assignment by the firm

Prediction 4 is that overlap-group cases, for which KH has a *relative net revenue advantage* (higher p_τ) are more likely to be assigned to nurses, and that these cases entail lower *relative taxpayer payments to the firm*, partially aligning incentives between the firm and the taxpayers (holding staffing costs), but ignoring downstream costs. Both patterns are driven by constrained doctor availability.

Estimation. To study this – and how task assignment varies for other related outcomes – we measure the KH's relative advantage over the direct-to-doctor route for each symptom using four outcome measures: (i) Systemwide production costs; (ii) Firm net revenues; (iii) Taxpayer payments to the firm (i.e., firm gross revenues); and (iv) Total taxpayer payments. We begin by modifying Equation 4 to allow the effect of initial nurse assignment to vary by symptom:

$$Y_{i,k(i)} = \delta_k \, Nurse_i \times Symptom_{k(i)} + \tilde{X}_{1i}\beta_1 + X_{2i}\beta_2 + \varepsilon_i, \tag{5}$$

where $Symptom_{k(i)}$ denotes a vector of indicators for symptoms k for each case i 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, and $\tilde{X}_{1i}$ is X_{1i} excluding the symptoms interaction terms, as we already interact symptom indicators with $Nurse_i$ to obtain the symptom-specific KH effects.

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, $\sum_i \hat{Y}_{i,k}(n) / \sum \hat{Y}_{i,k}(d)$.

Finally, we relate the firm's average KH assignment probability to the KH's relative advantage for each symptom 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). 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 the same-day drop-in consultations.

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 systemwide production 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.

Figure 4 examines how the alignment between the firm's incentives and taxpayer objectives varies with doctor congestion. Regressions of initial nurse assignment propensity on relative net revenue advantage by doctor congestion quartile reveal significantly steeper slopes at higher congestion levels, again in line with the model. [15] For total taxpayer payments, we observe a similar but weaker pattern. Taken together, the evidence in this section shows how the lack of information on downstream costs (and incentives to internalize them) leads to task allocations that do not minimize systemwide production costs.

7.2 Counterfactual assignment of tasks

This section compares the firm's observed task assignment to two sets of counterfactuals: (i) non-optimized assignments that ignore comparative advantage across case types; and (ii) assignments that maximize the social planner's objective with or without staffing constraints, following absolute or comparative advantage (Prediction 5).

Characterizing the social planner's objective. For each task type, the social planner minimizes expression 3, which we estimate using Equation 5, with systemwide production costs as the dependent variable. We interpret the KH route as having an *absolute advantage* in producing care for symptom k if $\sum_i \hat{Y}_{i,k}(n)/\sum \hat{Y}_{i,k}(d) < 1$. The *comparative advantage* of the KH route is determined by how low $\sum_i \hat{Y}_{i,k}(n)/\sum \hat{Y}_{i,k}(d)$ is relative to the cost ratio for other symptoms. Figure 5 illustrates heterogeneity in this cost ratio across symptoms. Subfigure 5a ranks symptoms by this ratio, reflecting the KH's comparative advantage, in the spirit of Dornbusch, Fischer, and Samuelson (1977). Subfigure 5b reports the absolute effects of assignment to KH, and its breakdown into effects

[15] For the lowest congestion quartile, around 90% of all case types are sent to doctors, which also agrees with the model, since absolute firm net revenues are higher under direct doctor assignment for all cases.

on firm production costs and downstream costs. For most symptoms, downstream costs outweigh firm costs.

To provide intuition, we discuss key drivers of downstream costs. For symptoms where direct-to-doctor assignment has the largest comparative advantage (right-hand side of Figure 5), we find patterns consistent with direct-to-doctor assignment reducing the tail risk of high-cost downstream care through faster or more certain access to antibiotic treatment (e.g., beta-lactam antibacterials), which are commonly prescribed after online visits with these symptoms. For example, for sore throat and urinary tract infection, KH assignment increases the probability of ED visits and hospitalizations within seven days (hospitalization effect for sore throat: 0.0003, 0.0024, and for UTI: 0.0176, 0.0024. ED effect for sore throat: -0.0008, 0.0012, and for UTI: 0.0053, 0.0014), with common diagnoses reflecting complications from bacterial infections (e.g., peritonsillar abscess, acute tonsillitis, and acute tubulo-interstitial nephritis).

In contrast, for symptoms where downstream costs are higher under direct-to-doctor assignment (left-hand side of Figure 5), KH assignment may reduce costs by limiting ineffective prescribing. This is particularly relevant when prescription drugs may provide limited benefit or have harmful side effects, such as when symptoms are diffuse (e.g., diarrhea or vomiting) or treatment knowledge is evolving (e.g., Covid-19). In such cases, care advice from a nurse, or combined guidance from a nurse and doctor, may be more effective.

Defining counterfactual case assignments. Using our estimates of expected systemwide production costs, we define several counterfactual assignment rules. The counterfactual assignments we consider are all deterministic (given our estimates above). For example, if in time t, there are E_t doctors and F_t nurses to assign to $E_t + F_t$ cases with the same symptom, we assign each the baseline mean cost (under direct-to-doctor assignment) plus $E_t/(E_t + F_t)$ times the estimated KH effect.

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

We then consider counterfactuals that minimize systemwide production costs. In the counterfactual that maximizes the social planner's objective without constraints, we ignore staff availability constraints, and instead assume that clinicians can be assigned without affecting wages or prices. This allows assignment by *absolute advantage*: all cases of a symptom are assigned to the KH if its expected total healthcare cost is lower, and to doctors otherwise – that is, the cost ratio in Subfigure 5a is below one.

In counterfactuals with a constrained social planner, we respect short-run clinician availability within fixed time windows (15 minutes; 3 hours), which results in assignments according to *comparative advantage*. Within each time window, we fix the total number doctors and nurses, and reassign cases

between symptoms to minimize systemwide production costs, according to the ranking of symptoms by their cost ratio shown in Subfigure 5a. These counterfactuals assign cases to the available nurse capacity, starting with those symptoms that yield the largest cost savings, and moving down the ranking until all nurses are assigned.

Results. Table 6 summarizes the implications of the different counterfactuals for various outcomes. Panel A considers the non-optimized assignments that ignore the KH's heterogeneous cost effects across symptoms. Allocations 1)-4) assign cases proportionally to clinician availability. These assignments increase systemwide production costs between 0.55-0.58%, equivalent to about a 1.9% increase in firm production costs. Assignment 5), which uses only the KH, increases systemwide production costs by 0.79% (equivalent to 2.6% in firm production costs). Finally, using only doctors increases systemwide production costs by 0.71% (equivalent to 2.3% increase in firm production costs) – in other words, compared to using only doctors, the firm's actual (selective) use of KH generates meaningful savings.

Panel B implements the social planner's objective. Counterfactual 1) shows that in the unconstrained optimum the use of the KH route increases by about 170%. This reduces systemwide production costs by 4.07%, but also involves sizeable redistribution, as it reduces the firm's net revenues by almost 40%. Because this counterfactual assumes no change in factor prices, it can be interpreted as an upper bound on the savings achievable through changes in staffing — for instance, by expanding nurse training.

Instead, counterfactuals 2) and 3) respect actual clinician use within 3-hour or 15-minute windows, which still yield significant savings, with minimal implications for the firm's revenues.[16] The reassignment within 15-minute windows reduces systemwide production costs by about 2.1%, which is equivalently to a 6.7% savings in firm production costs.

Appendix Table A10 reports robustness tests on the main counterfactual result from Table 6. First, bootstrapped standard errors indicate that the estimated reductions—2.1% in systemwide production costs and 6.7% in firm production costs— are precise. Second and third, we replace the indicator-based measure of systemwide production costs with aggregated costs over both the standard 7 days and over 30 days. The estimated savings in systemwide production remain comparable at 2.2% and 1.6%, respectively, while the corresponding savings relative to firm production costs increase to 7.8% and 8.9%. Fourth, using only cases from Stockholm and Scania yields estimates similar to those from the full sample. Finally, we split the sample, estimating the effects using odd calendar days and constructing the counterfactual using even calendar days, and we find comparable savings of 2.3% (systemwide production costs) and 7.5% (firm production costs).

This potential to make substantial cost savings with existing staff arises from a lack of information and inadequate incentives for the firm. A simple corrective is to issue guidelines that steer the allocation of cases toward lower systemwide production costs. It is common for healthcare payers to issue guidelines to prioritize cost-effective care, such as those mentioned in Section 2, or the UK NICE guidelines. In the firm we study, such guidelines could be easily implemented by adjusting the

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

algorithm that assigns cases, without further administrative burden, and with only trivial losses for the firm. In the longer term, staffing or hiring can change, as could the supply of nurses and doctors. The study of such changes is left for future research.

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 higher-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 systemwide production 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. 2018. "Econometric Methods for Program Evaluation". In: *Annual Review of Economics* 10 (1): 465–503. DOI: 10.1146/annurev-economics-080217-053402.

Acemoglu, Daron and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings". In: *Handbook of Labor Economics*. Vol. 4. Elsevier, pp. 1043–1171. DOI: 10.1016/S0169-7218(11)02410-5.

Ala, Ali and Feng Chen. 2022. "Appointment Scheduling Problem in Complexity Systems of the Healthcare Services: A Comprehensive Review". In: *Journal of Healthcare Engineering* 2022, 5819813. DOI: 10.1155/2022/5819813.

Alexander, Diane and Molly Schnell. 2019. "Just what the nurse practitioner ordered: Independent prescriptive authority and population mental health". In: *Journal of Health Economics* 66, 145–162. DOI: 10.1016/j.jhealeco.2019.04.004.

Azar, José and Ioana Marinescu. 2024. "Chapter 10 - Monopsony power in the labor market". In: *Handbook of Labor Economics*. Ed. by Christian Dustmann and Thomas Lemieux. Vol. 5. Elsevier, pp. 761–827. DOI: 10.1016/bs.heslab.2024.11.007.

Becker, Gary S. and Kevin M. Murphy. 1992. "The Division of Labor, Coordination Costs, and Knowledge". In: *The Quarterly Journal of Economics* 107 (4): 1137–1160. DOI: 10.2307/2118383.

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

Bloom, Nicholas, Luis Garicano, Raffaella Sadun, and John Van Reenen. 2014. "The Distinct Effects of Information Technology and Communication Technology on Firm Organization". In: *Management Science* 60 (12): 2859–2885. DOI: 10.1287/mnsc.2014.2013.

Bronsoler, Ari, Joseph Doyle, and John Van Reenen. 2022. "The Impact of Health Information and Communication Technology on Clinical Quality, Productivity, and Workers". In: *Annual Review of Economics* 14 (Volume 14, 2022): 23–46. DOI: 10.1146/annurev-economics-080921-101909.

Buehler, Aidan, Joshua D. Gottlieb, Jeffrey Hicks, Lisa Laun, Maarten Palme, Maria Polyakova, Victoria Udalova, and Maria Ventura. 2026. "International Comparison of Physician Incomes". In: *American Economic Association: Papers and Proceedings* 116.

Caliendo, Mion, Opromolla & Rossi-Hansberg. 2020. "Productivity and Organisation in Portuguese firms". In: *Journal of Political Economy* 128 (11): 4211–4257. DOI: 10.1086/710533.

Caliendo, Monte & Rossi-Hansberg. 2015. "The Anatomy of French Production Hierarchies". In: *Journal of Political Economy* 123 (4): 809–852. DOI: 10.1086/681641.

Cattaneo, Matias D., Richard K. Crump, Max H. Farrell, and Yingjie Feng. 2024. "On Binscatter". In: *American Economic Review* 114 (5): 1488–1514. DOI: 10.1257/aer.20221576.

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

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

Chernozhukov, V., I. Fernandez-Val, and A. Galichon. 2009. "Improving point and interval estimators of monotone functions by rearrangement". In: *Biometrika* 96 (3): 559–575. DOI: 10.1093/biomet/asp030.

Costantini, Sydney. 2025. "How Do Mental Health Treatment Delays Impact Long-Term Mortality?". In: *American Economic Review* 115 (5): 1672–1707. DOI: 10.1257/aer.20240226.

Dahlstrand, Amanda. 2026. "Defying Distance? The Provision of Medical Services in the Digital Age". In: *Working Paper*. [Link].

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

Dillender, Marcus, Anthony T. Lo Sasso, Brian J. Phelan, and Michael Richards. 2024. "Occupational Licensing and the Healthcare Labor Market". In: *Journal of Human Resources*, 0722–12450R2. DOI: 10.3368/jhr.0722-12450R2.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang. 2018. "The Effects of Pre-Trial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges". In: *American Economic Review* 108 (2): 201–240. DOI: 10.1257/aer.20161503.

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

European Observatory on Health Systems and Policies. n.d.(a). *Health System Summaries*. [Link].

European Observatory on Health Systems and Policies. n.d.(b). *Health Systems Reviews*. [Link].

Folkhälsomyndigheten. 2020. *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.

Försäkringskassan. 2025. *Sjuklön: När behöver medarbetaren läkarintyg?* Försäkringskassan. [Link].

Garicano, Luis. 2000. "Hierarchies and the Organization of Knowledge in Production". In: *Journal of Political Economy* 108 (5): 874–904. DOI: 10.1086/317671.

Godøy, Anna, Venke F. Haaland, Ingrid Huitfeldt, and Mark Votruba. 2024. "Hospital Queues, Patient Health, and Labor Supply". In: *American Economic Journal: Economic Policy* 16 (2): 150–81. DOI: 10.1257/pol.20210399.

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

Gruber, Jonathan, Thomas P. Hoe, and George Stoye. 2023. "Saving Lives by Tying Hands: The Unexpected Effects of Constraining Health Care Providers". In: *The Review of Economics and Statistics* 105 (1): 1–19. DOI: 10.1162/rest_a_01044. eprint: https://direct.mit.edu/rest/article-pdf/105/1/1/2066656/rest_a_01044.pdf.

Guo, Jiapei, Angela E. Kilby, and Mindy S. Marks. 2024. "The impact of scope-of-practice restrictions on access to medical care". In: *Journal of Health Economics* 94, 102844. DOI: 10.1016/j.jhealeco.2023.102844.

Hartman, Erin and F. Daniel Hidalgo. 2018. "An Equivalence Approach to Balance and Placebo Tests". In: *American Journal of Political Science* 62 (4): 1000–1013. DOI: 10.1111/ajps.12387.

Jones, Benjamin F.. 2009. "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?". In: *The Review of Economic Studies* 76 (1): 283–317. DOI: 10.1111/j.1467-937X.2008.00531.x.

Katz, Lawrence F. and Kevin M. Murphy. 1992. "Changes in Relative Wages, 1963–1987: Supply and Demand Factors". In: *The Quarterly Journal of Economics* 107 (1): 35–78. DOI: 10.2307/2118323.

Kleiner, Morris M.. 2016. "Battling over Jobs: Occupational Licensing in Health Care". In: *American Economic Review* 106 (5): 165–170. DOI: 10.1257/aer.p20161000.

Kry International AB. 2024. *Annual Report and Consolidated Accounts*. [Link].

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

Maier, Claudia B. and Linda H. Aiken. 2016. "Task shifting from physicians to nurses in primary care in 39 countries: a cross-country comparative study". In: *European Journal of Public Health* 26 (6): 927–934. DOI: 10.1093/eurpub/ckw098.

McMichael, Benjamin J.. 2023. "Supply-side health policy: The impact of scope-of-practice laws on mortality". In: *Journal of Public Economics* 222, 104901. DOI: 10.1016/j.jpubeco.2023.104901.

Ministry of Health, Welfare and Sport. 2018. *The Right Care in the Right Place*. Tech. rep. The Hague: Ministry of Health, Welfare and Sport (Netherlands). [Link].

Mogstad, Magne, Kjell G. Salvanes, and Gaute Torsvik. 2025. "Income Equality in the Nordic Countries: Myths, Facts, and Lessons". In: *Journal of Economic Literature* 63 (3): 791–839. DOI: 10.1257/jel.20251636.

Myndigheten för vård- och omsorgsanalys. 2022. *Besök via nätet*. [Link].

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

NHS England. n.d. *Principle 3 – Right care, right time, right place for carers*. [Link].

Norra sjukvårdsregionförbundet. 2018. *Price Lists of Northern Healthcare Region*. https://www.norrasjukvardsregionfo
och-sjukvard/avtal-och-priser/arkiv/.

OECD. 2021a. *Health at a Glance*. [Link].

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

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

OECD. 2025b. *Health Statistics Data Explorer: Physicians*. [Link].

OECD. 2025c. *OECD Health Statistics Data Explorer: Practising nurses*. [Link].

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence". In: *Journal of Business & Economic Statistics* 37 (2): 187–204. DOI: 10.1080/07350015.2016.1227711.

Rajpurkar, Pranav, Emma Chen, Oishi Banerjee, and Eric J. Topol. 2022. "AI in health and medicine". In: *Nature Medicine* 28 (1): 31–38. DOI: 10.1038/s41591-021-01614-0.

Region Kronoberg. 2025. *Riktlinje för registrering av vårdkontakter*. [Link].

Region Skane. 2023. *Regionala riktlinjer för registrering och klassificering i Region Skåne*. [Link].

Royal Norwegian Ministry of Health and Care Services. 2006. *National Health Plan for Norway (2007–2010)*. Tech. rep. Oslo, Norway: Royal Norwegian Ministry of Health and Care Services. [Link].

Saez, Emmanuel, Benjamin Schoefer, and David Seim. 2019. "Payroll Taxes, Firm Behavior, and Rent Sharing: Evidence from a Young Workers' Tax Cut in Sweden". In: *American Economic Review* 109 (5): 1717–1763. DOI: 10.1257/aer.20171937.

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

Socialdepartementet. 2016. *Effektiv vård: Slutbetänkande av En nationell samordnare för effektivare resursutnyttjande inom hälso- och sjukvården*. SOU 2016:2. Stockholm, Sweden: Statens offentliga utredningar. [Link].

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

Socialstyrelsen. 2022. *Bedömning av tillgång och efterfrågan på legitimerad personal i hälso- och sjukvård samt tandvård*. [Link].

Södra sjukvårdsregionen. 2020. *Reports About Healthcare Costs*. https://sodrasjukvardsregionen.se/verksamhet/avtal-priser/regionala-priser-och-ersattningar-foregaende-ar/.

Statistics Sweden (SCB). 2020. *Wages and salaries by occupation*. Statistics Sweden Statistical Database. [Link].

Sveriges Kommuner och Landsting (SKL). 2019. *Meddelande från styrelsen - Utvecklingen av digitala vårdtjänster*. [Link].

Sveriges Kommuner och Regioner. 2024. *Ekonomi- och verksamhetsstatistik inom hälso- och sjukvården*. Sveriges Kommuner och Regioner (SKR). [Link].

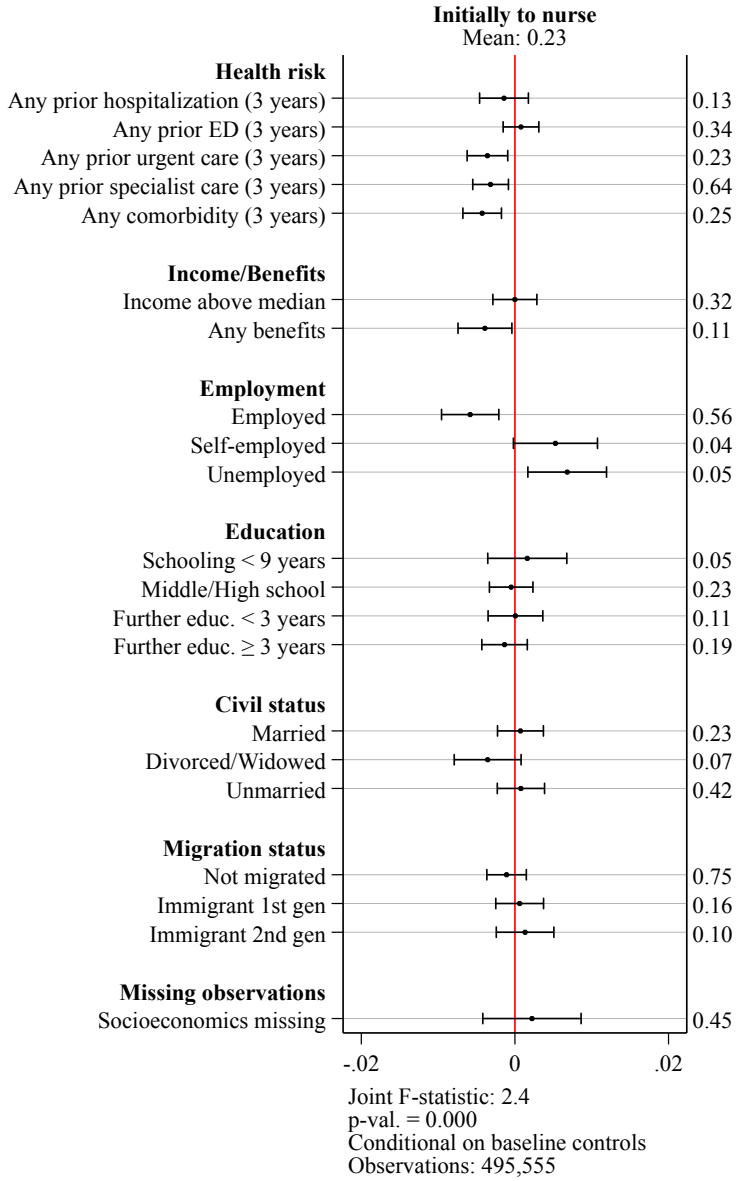
Sveriges Riksbank. 2020. *Exchange rates – SEK to USD (daily rates 2020)*. Sveriges Riksbank statistics. [Link].

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

Traczynski, Jeffrey and Victoria Udalova. 2018. "Nurse practitioner independence, health care utilization, and health outcomes". In: *Journal of Health Economics* 58, 90–109. DOI: 10.1016/j.jhealeco.2018.01.001.

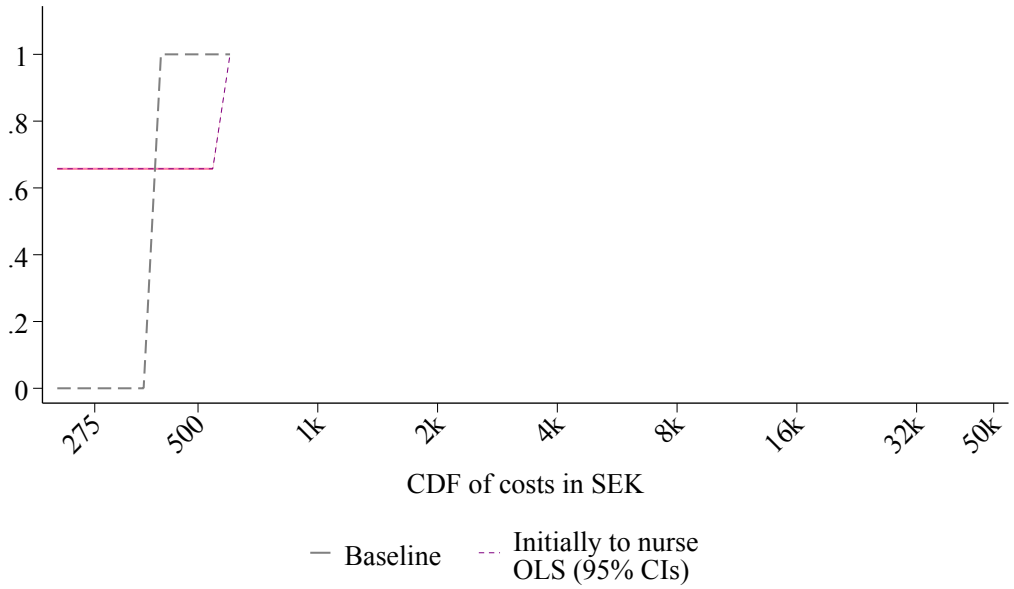
Main Figures

Figure 1. Balance of the treatment in case characteristics

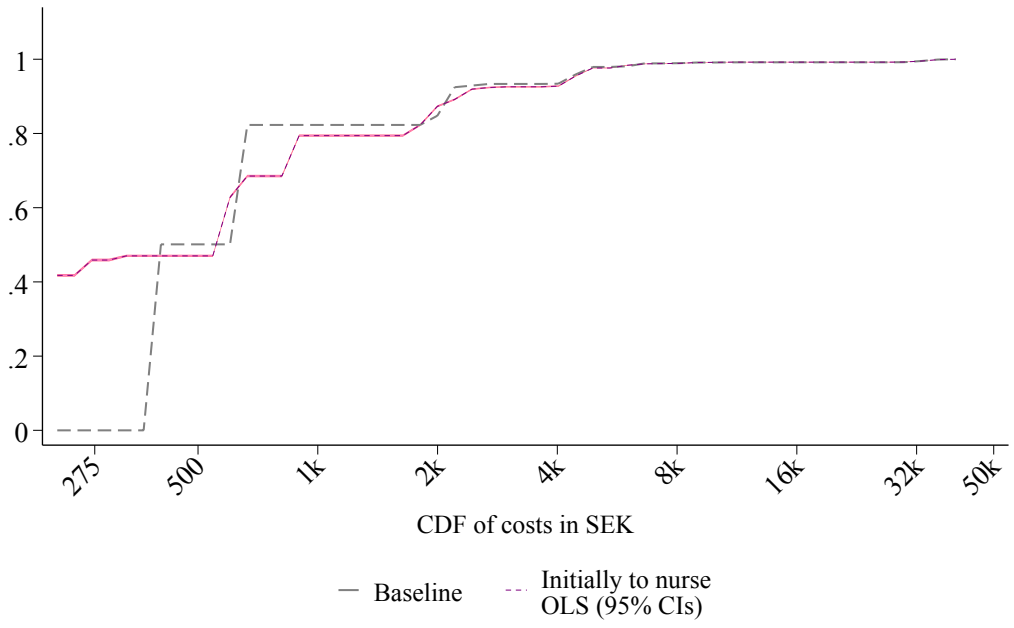


Notes: This figure presents balance tests for the treatment variable (*Nurse*). Each row shows a separate regression of the treatment on the characteristic listed on the vertical axis (each from X_2), conditional on all baseline controls (X_1): login time fixed effects (date-by-3-hour), main symptom categories, patient demographics, and a demographics × prior firm use interaction term. Horizontal lines represent 95% confidence intervals based on robust standard errors. The F-statistic is a joint F-test of the instrument on all additional case characteristics (X_2), conditional on all baseline controls (X_1). The numbers on the right show sample means for each characteristic.

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



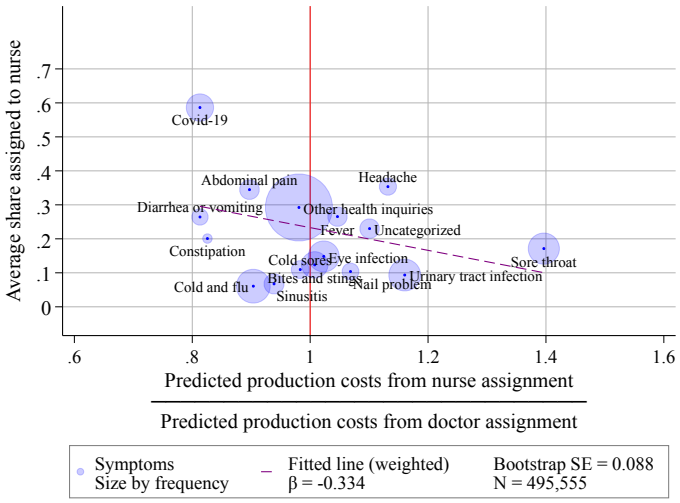
(a) Distribution of firm production costs



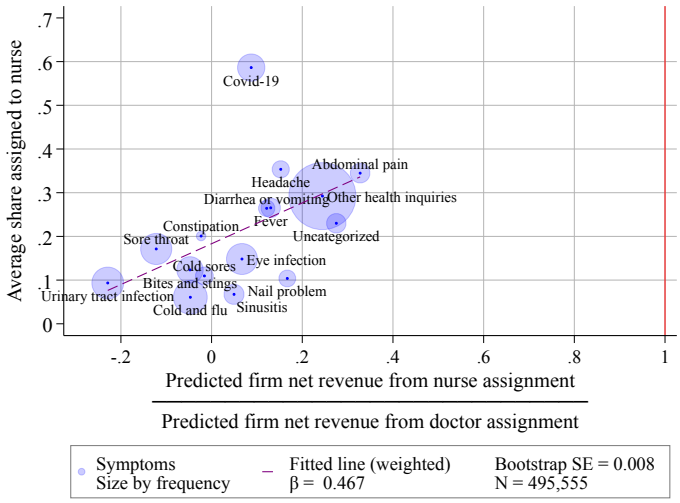
(b) Distribution of systemwide production costs

Notes: These figures present the estimated probability that firm costs and systemwide production costs fall below the threshold shown on the x-axis (shown on a logarithmic scale). The y-axis shows estimated probabilities from OLS regressions based on Equation 4, where the dependent variable is an indicator for costs exceeding ln(cost) thresholds at 0.1 intervals, with the full set of controls (X_1 and X_2). The shaded areas show 95% confidence intervals. CDF monotonicity is ensured following Chernozhukov, Fernandez-Val, and Galichon (2009).

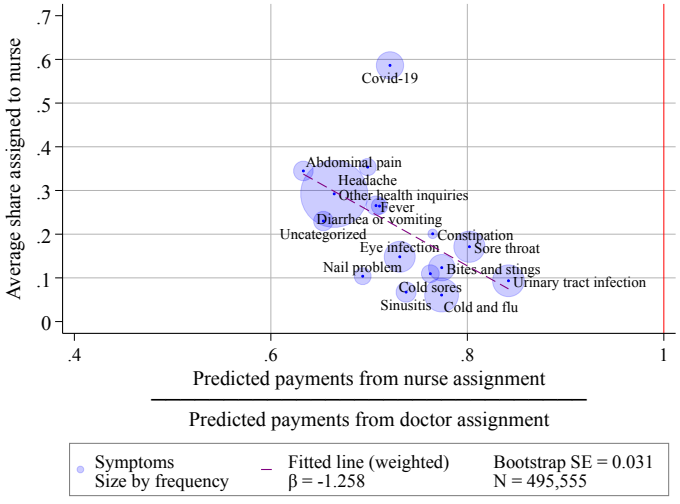
Figure 3. Relative revenue and cost advantage across case types



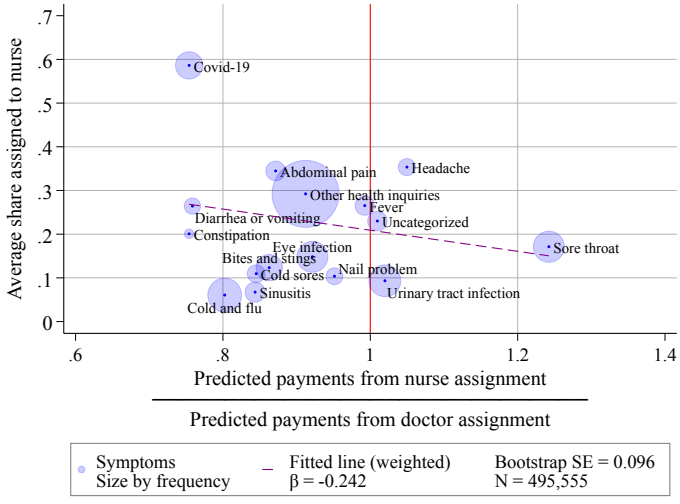
(a) Systemwide production costs



(b) Firm net revenue



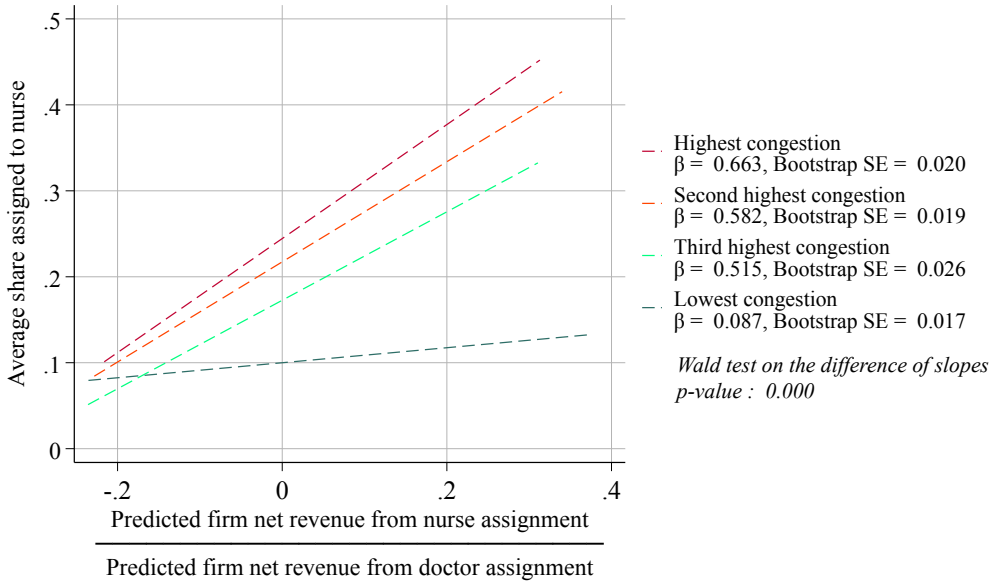
(c) Taxpayer payments to the firm



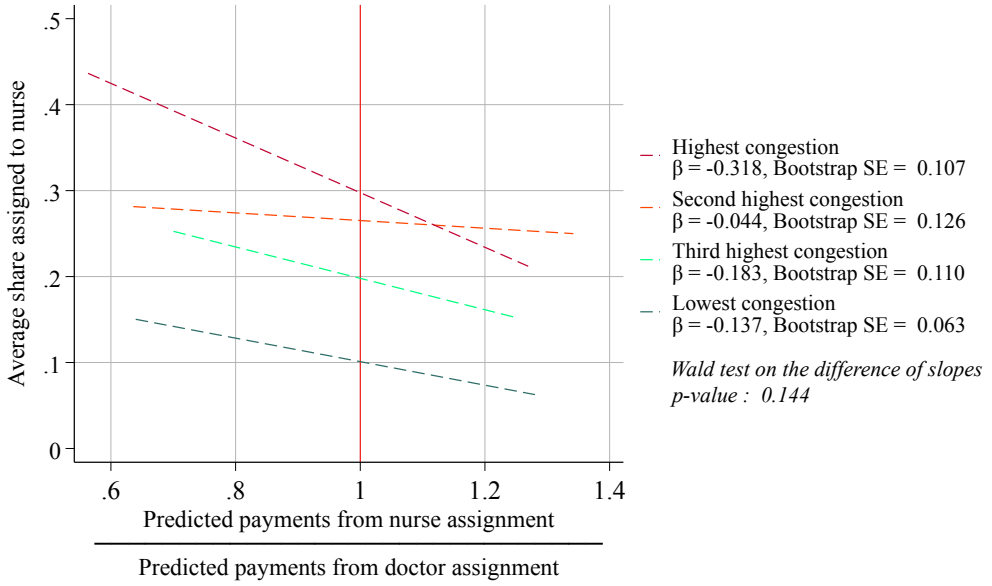
(d) Total taxpayer payments

Notes: Each panel shows the firm's case assignment by the relative revenue or cost advantage of initial nurse versus initial doctor assignment across symptom groups. The y-axis plots the predicted probability of initial nurse assignment from the first-stage regression (Appendix Equation 16). The x-axis plots the relative revenue or cost advantage under initial nurse versus initial doctor assignment, derived from a modified version of Equation 4 interacting symptom groups with *Nurse*. Panel 3a has on the x-axis the relative advantage in systemwide production costs; Panel 3b the relative advantage in firm net revenue; Panel 3c the relative advantage in taxpayer payments to the firm; Panel 3d the relative advantage in total taxpayer payments. Relative advantages are constructed by summing predicted outcomes within each symptom group and taking ratios. Circle sizes reflect symptom frequencies. Fitted lines are weighted OLS regressions. Standard errors are from 100 bootstrap replications (resampled at the consultation level). The vertical red line indicates equal costs or revenues. All regressions use the analysis sample with full controls (X_1 and X_2).

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



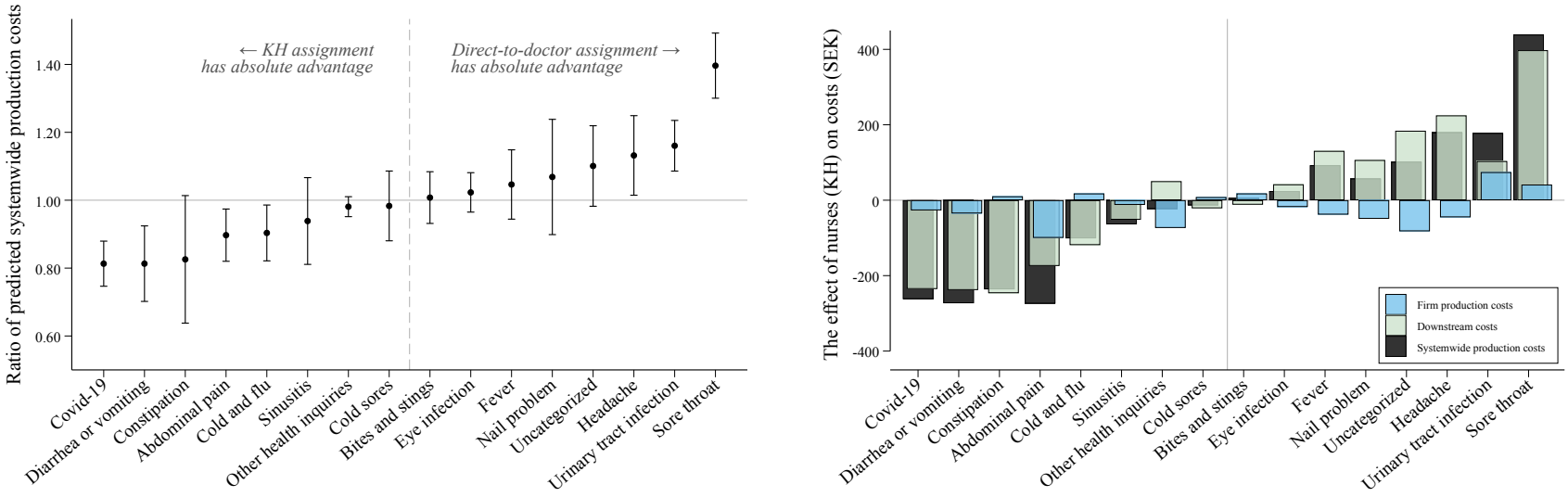
(a) Relative firm net revenue advantage



(b) Relative total taxpayer payment advantage

Notes: These figures show the relative revenue and cost advantage of 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 (Appendix Equation 16). The x-axis plots the relative revenue or cost advantage, derived from a modified version of Equation 4 interacting symptom groups with *Nurse*. Relative advantages are constructed by summing predicted outcomes within each symptom group and taking ratios. Dashed lines are weighted OLS regression fits (weighted by symptom frequency). Standard errors are from 100 bootstrap replications (resampled at the consultation level).

Figure 5. KH effect and average predicted cost by symptoms



(a) Comparative advantage

(b) Breakdown of systemwide production costs

Notes: These figures show heterogeneous effects of the KH across symptom groups. The left panel shows predicted systemwide production costs for initial nurse and initial doctor assignment for each symptom group, derived from Equation 5 interacting symptom groups with *Nurse*; predicted costs are averaged within each symptom group. For the first eight symptom groups, predicted costs are lower under initial nurse assignment (nurse absolute advantage); the vertical line marks where this advantage reverses. The right panel decomposes systemwide production costs into downstream costs and firm costs using the same specification.

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 OECD (2025a), National Center for Health Statistics (2023), and Centers for Medicare & Medicaid Services (CMS) (2025). Row (4) uses data from OECD (2021a). Row (5) uses data from OECD (2025c) and OECD (2025b). Row (6) uses data from countries' governmental websites. Rows (7) and (9) use data from OECD (2021b). Row (8) uses data 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
Symptoms							
Abdominal pain	0.030	[0.17]	0.046	[0.21]	0.025	[0.16]	0.000
Cold and flu	0.089	[0.28]	0.024	[0.15]	0.11	[0.31]	0.000
Cold sores	0.025	[0.16]	0.012	[0.11]	0.029	[0.17]	0.000
Constipation	0.0070	[0.083]	0.0062	[0.078]	0.0072	[0.085]	0.000
Covid-19	0.058	[0.23]	0.15	[0.36]	0.031	[0.17]	0.000
Diarrhea or vomiting	0.021	[0.14]	0.024	[0.15]	0.020	[0.14]	0.000
Eye infection	0.075	[0.26]	0.049	[0.22]	0.083	[0.28]	0.000
Fever	0.029	[0.17]	0.034	[0.18]	0.028	[0.16]	0.000
Headache	0.023	[0.15]	0.036	[0.19]	0.019	[0.14]	0.000
Nail problem	0.022	[0.15]	0.010	[0.10]	0.026	[0.16]	0.000
Other health inquiries	0.35	[0.48]	0.45	[0.50]	0.32	[0.47]	0.000
Bites and stings	0.054	[0.23]	0.030	[0.17]	0.062	[0.24]	0.000
Sinusitis	0.031	[0.17]	0.0093	[0.096]	0.038	[0.19]	0.000
Sore throat	0.076	[0.26]	0.057	[0.23]	0.081	[0.27]	0.000
Uncategorized	0.030	[0.17]	0.030	[0.17]	0.029	[0.17]	0.419
Urinary tract infection	0.079	[0.27]	0.032	[0.18]	0.092	[0.29]	0.000
Demographics							
Female	0.63	[0.48]	0.59	[0.49]	0.64	[0.48]	0.000
Patient age	29.2	[16.5]	29.3	[16.6]	29.2	[16.5]	0.569
West Sweden	0.20	[0.40]	0.20	[0.40]	0.20	[0.40]	0.066
Stockholm	0.44	[0.50]	0.45	[0.50]	0.44	[0.50]	0.000
Middle Sweden	0.18	[0.39]	0.18	[0.39]	0.19	[0.39]	0.001
South Sweden	0.095	[0.29]	0.093	[0.29]	0.095	[0.29]	0.015
Norrland	0.034	[0.18]	0.031	[0.17]	0.035	[0.18]	0.000
Småland + the islands	0.042	[0.20]	0.042	[0.20]	0.042	[0.20]	0.254
Health risk factors							
Any prior hospitalization (3 years)	0.13	[0.34]	0.14	[0.34]	0.13	[0.34]	0.443
Any prior ED (3 years)	0.34	[0.47]	0.35	[0.48]	0.34	[0.47]	0.000
Any prior urgent care (3 years)	0.23	[0.42]	0.22	[0.41]	0.23	[0.42]	0.000
Any prior specialist care (3 years)	0.64	[0.48]	0.63	[0.48]	0.64	[0.48]	0.000
Any comorbidity (3 years)	0.25	[0.43]	0.25	[0.43]	0.25	[0.43]	0.006
Socioeconomic characteristics							
Income above median	0.32	[0.47]	0.31	[0.46]	0.33	[0.47]	0.000
Any benefits	0.11	[0.31]	0.11	[0.31]	0.11	[0.32]	0.000
Schooling $<$ 9 years	0.049	[0.21]	0.052	[0.22]	0.048	[0.21]	0.000
Middle/High school	0.23	[0.42]	0.23	[0.42]	0.24	[0.42]	0.002
Further educ. $<$ 3 years	0.11	[0.31]	0.10	[0.31]	0.11	[0.31]	0.002
Further educ. $\geq$ 3 years	0.19	[0.39]	0.18	[0.39]	0.20	[0.40]	0.000
Employed	0.56	[0.50]	0.55	[0.50]	0.57	[0.50]	0.000
Self-employed	0.041	[0.20]	0.042	[0.20]	0.041	[0.20]	0.577
Unemployed	0.050	[0.22]	0.052	[0.22]	0.049	[0.22]	0.000
Married	0.23	[0.42]	0.22	[0.41]	0.23	[0.42]	0.000
Divorced/Widowed	0.074	[0.26]	0.075	[0.26]	0.074	[0.26]	0.802
Unmarried	0.42	[0.49]	0.42	[0.49]	0.42	[0.49]	0.161
Not migrated	0.75	[0.43]	0.73	[0.45]	0.75	[0.43]	0.000
Immigrant 1st gen	0.16	[0.36]	0.17	[0.38]	0.15	[0.36]	0.000
Immigrant 2nd gen	0.096	[0.29]	0.100	[0.30]	0.095	[0.29]	0.000
Socioeconomics missing	0.45	[0.50]	0.47	[0.50]	0.45	[0.50]	0.000
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 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)	(5)
X_1	✓		✓	✓	✓
X_2		✓	✓	✓	✓
Log median doctor wait time				✓	✓
Doctor-group status in prior cases					✓
R-squared	0.216	0.000	0.216	0.221	0.221
Adjusted R-squared	0.195	0.000	0.195	0.200	0.200
Observations	495,555	495,555	495,555	495,555	495,505

Notes: This table reports the predictive power from regressions of the treatment (*Initial assignment to nurse*) on case characteristics and exogenous assignment shifters. Baseline characteristics (X_1) and additional controls (X_2) are described in the main text. The exogenous assignment shifters include doctor congestion (*Log median doctor wait time*) and indicators for the doctor-group status of the previous 50 cases with either doctor-group or overlap-group symptoms. K-P F-statistics are 3,726.36 for column (2), 3,725.22 for column (3), and 3,670.24 for column (4). Robust standard errors are in parentheses.

Table 4. Effects of the knowledge hierarchy on patient outcomes

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)
X_1 and X_2	✓	✓	✓	✓
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 new prescription (7 days)	Any external PCP consultation (7days)	Any urgent care (7 days)	Any specialist care (7 days)
Initially to nurse	-0.10***	0.022***	0.0033***	0.0031***
	(0.0017)	(0.0013)	(0.00074)	(0.00051)
X_1 and X_2	✓	✓	✓	✓
Observations	495,555	495,555	495,555	495,555
Baseline mean	0.44	0.11	0.035	0.016

	Any ED (7 days)	Any hospitalization (7 days)
Initially to nurse	0.0049***	-0.00049
	(0.00075)	(0.00038)
X_1 and X_2	✓	✓
Observations	495,555	495,555
Baseline mean	0.030	0.0082

Panel C: Long-term consequences

	Earnings decrease > 20 pct. (cal. month after)	Zero earnings (cal. month after)	Death (3 years)
Initially to nurse	-0.021***	-0.0031**	-0.000012
	(0.0025)	(0.0013)	(0.00018)
X_1 and X_2	✓	✓	✓
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. Outcomes are defined as follows: "Any clinician-booked doctor visit" 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 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) to their initial consultation; "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) at the initial consultation; "Any new prescription" 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; "Earnings decrease" indicates whether the patient's earnings fell by at least 20% relative to their average earnings over the previous three months; "Zero earnings" 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 earnings measures are defined for a subsample with stable earnings of employed patients earning more than 3,533 SEK (2010 values) per month in each of the three prior months, based on an annual threshold of 42,400 SEK (Saez, Schoefer, and Seim 2019). The columns in each subpanel report the OLS estimate for *Nurse* based on Equation 4. All regressions control for baseline characteristics (X_1) and additional controls (X_2) as described in the main text. The baseline mean is the average of the dependent variable for cases directly assigned to a doctor. Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5. Effects of the knowledge hierarchy on costs

	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)
X_1 and X_2	✓	✓	✓
Observations	495,555	495,555	495,555
Baseline mean	500	544.6	125
	Firm production costs	Total taxpayer payments	Systemwide production costs
Initially to nurse	-40.3***	-117.4***	-9.72
	(0.55)	(13.2)	(13.2)
X_1 and X_2	✓	✓	✓
Observations	495,555	495,555	495,555
Baseline mean	375	1304.8	1179.8

Note: This table presents regression results for the effects of initial case assignment to a nurse (the *knowledge hierarchy*) rather than directly to a doctor. Outcomes are defined as follows: "Taxpayer payments to the firm (within-day)" are the taxpayer's incurred costs for within-day consultations at the firm; "Taxpayer payments to the firm (7 days)" are the taxpayer's incurred costs for consultations at the firm within 7 days of the initial assignment; "Firm net revenue (within-day)" is the firm's net revenue on the initial consultation day, defined as revenues minus costs of provision; "Total taxpayer payments" are the taxpayer's incurred costs of total healthcare provision within 7 days of the initial assignment, including within-firm and downstream spending; "Systemwide production costs" are the costs of healthcare provision within 7 days after the initial assignment, defined as total taxpayer payments minus the firm's net revenues. The column in each subpanel reports the OLS estimate for *Nurse* based on Equation 4. All regressions control for baseline characteristics (X_1) and additional controls (X_2) as described in the main text. The baseline mean is the average of the dependent variable for cases directly assigned to a doctor. Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

	$\frac{\Delta \text{ Systemwide production costs}}{\text{Systemwide production costs}}$	$\frac{\Delta \text{ Systemwide production costs}}{\text{Firm production costs}}$	$\frac{\Delta \text{ Total taxpayer payments}}{\text{Total taxpayer payments}}$	$\frac{\Delta \text{ Firm net revenue}}{\text{Firm net revenue}}$	$\frac{\Delta \text{ Cases to nurses}}{\text{Cases to nurses}}$
	(1)	(2)	(3)	(4)	(5)
A: Switching from actual assignment to non-optimized assignments					
1) Cases assigned uniformly to nurse/doctor proportional to clinician shares in *prior* 15 minutes	0.58%	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: Switching from actual assignment to assignments that minimize Systemwide production costs					
1) Unconstrained minimization of systemwide production costs	−4.07%	−13.32%	−6.81%	−39.43%	169.26%
2) Constrained minimization of systemwide production costs (Nurses constant within 3-hour windows)	−2.55%	−8.36%	−2.41%	−0.71%	0.00%
3) Constrained minimization of systemwide production costs (Nurses constant within 15-minute windows)	−2.05%	−6.71%	−1.91%	−0.28%	0.00%

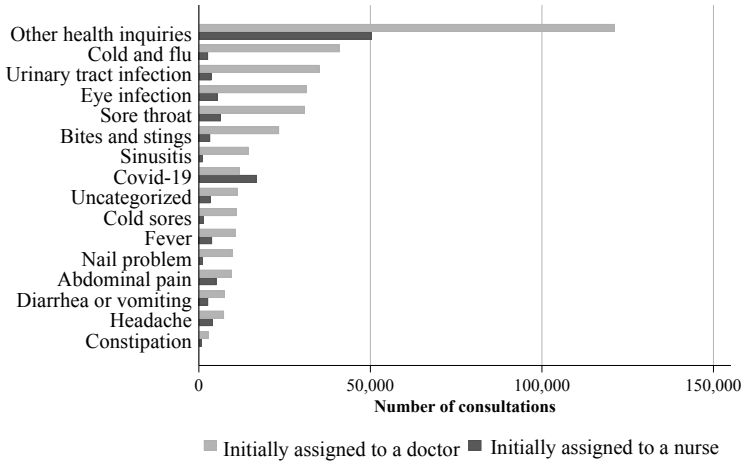
Note: This table shows the percentage change in outcomes from counterfactual assignments relative to the actual assignment. To avoid randomization, the counterfactuals assign the costs associated with "fractional clinicians": $q\times$ the predicted cost of a nurse meeting plus $(1-q)$ times the predicted cost of a doctor meeting, where $q \in [0,1]$. Only the assignments in Panel A rows 2 and 4 and Panel B rows 2 and 3 hold total clinician assignments fixed as in reality; the remaining counterfactuals entail using more or fewer nurses as indicated in the final column. The number of actual nurse assignments is 111,625.

Appendix

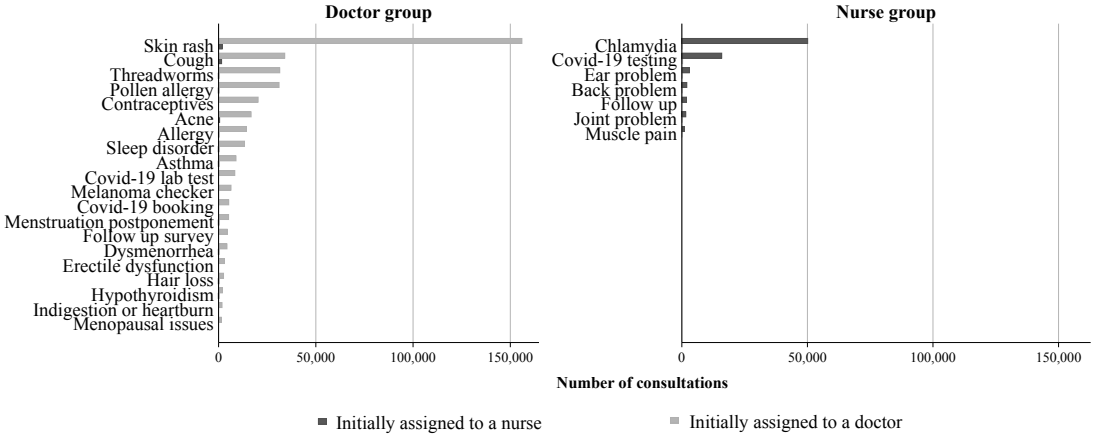
Does Division of Labor Increase Productivity? Evidence from Primary Care

Appendix Figures

Figure A1. Case groups



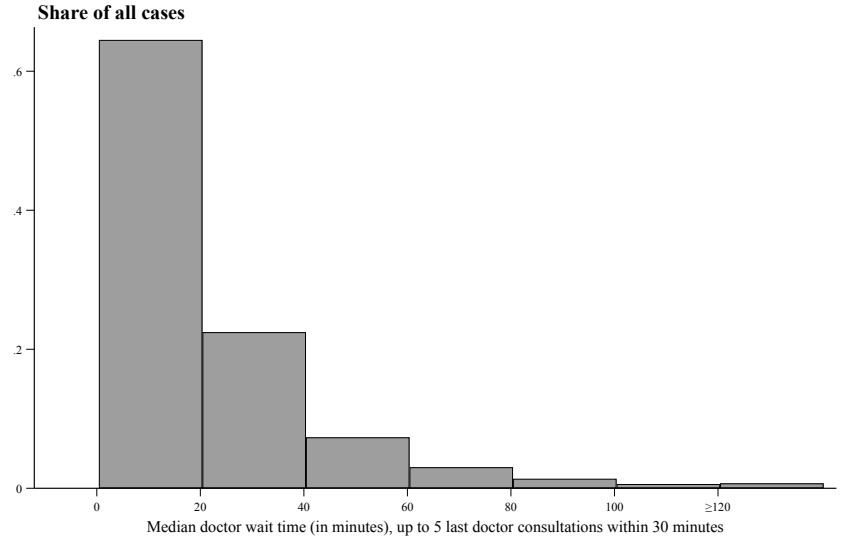
(a) Overlap group



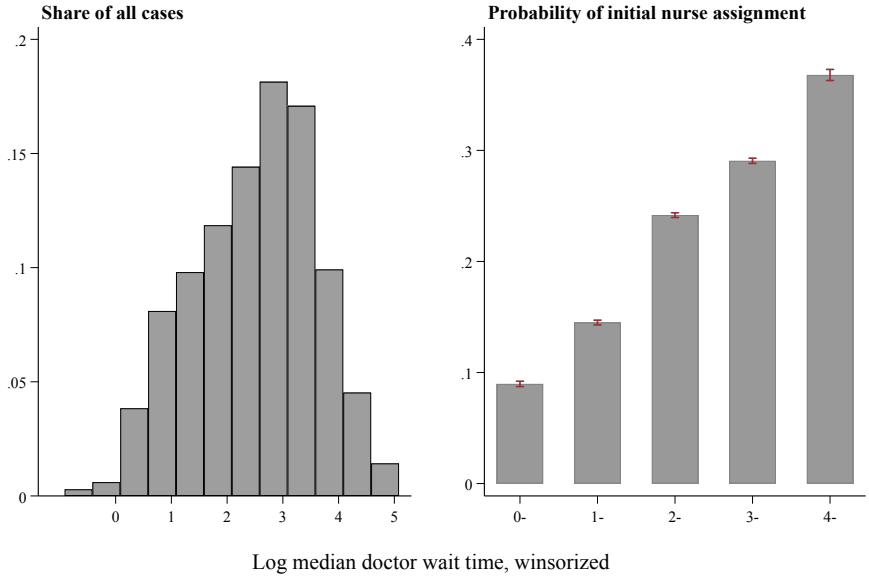
(b) Doctor group and nurse group

Notes: Panel (a) shows the self-reported symptom categories for patients in the analysis sample. Panel (b) shows the patient symptom categories by group assignment: overlap group (analysis sample), doctor group, and nurse group.

Figure A2. Congestion measure



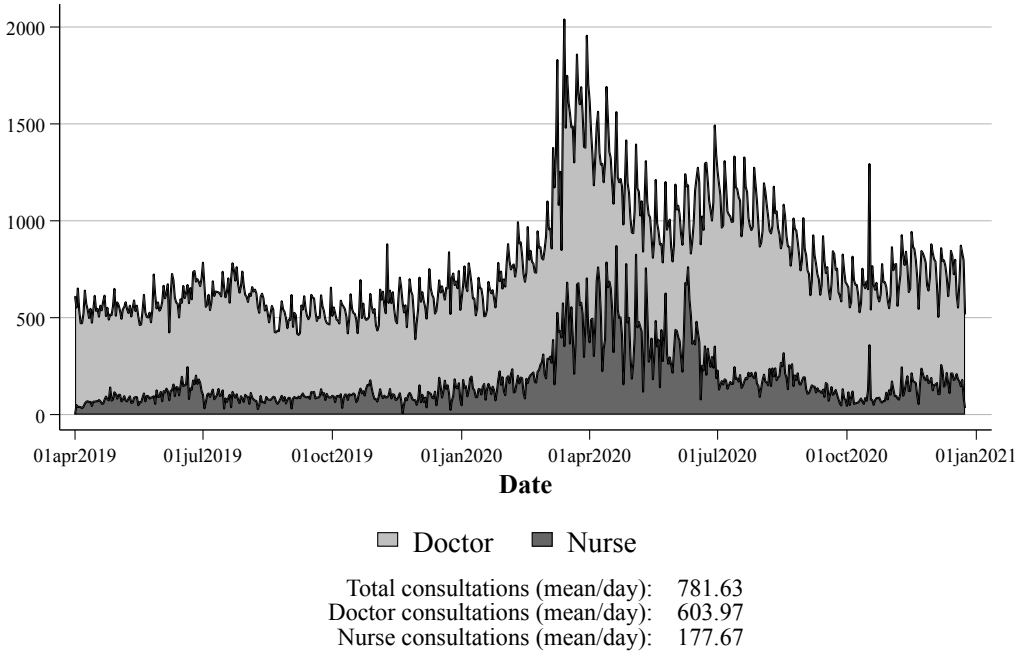
(a) Doctor wait time before transformation



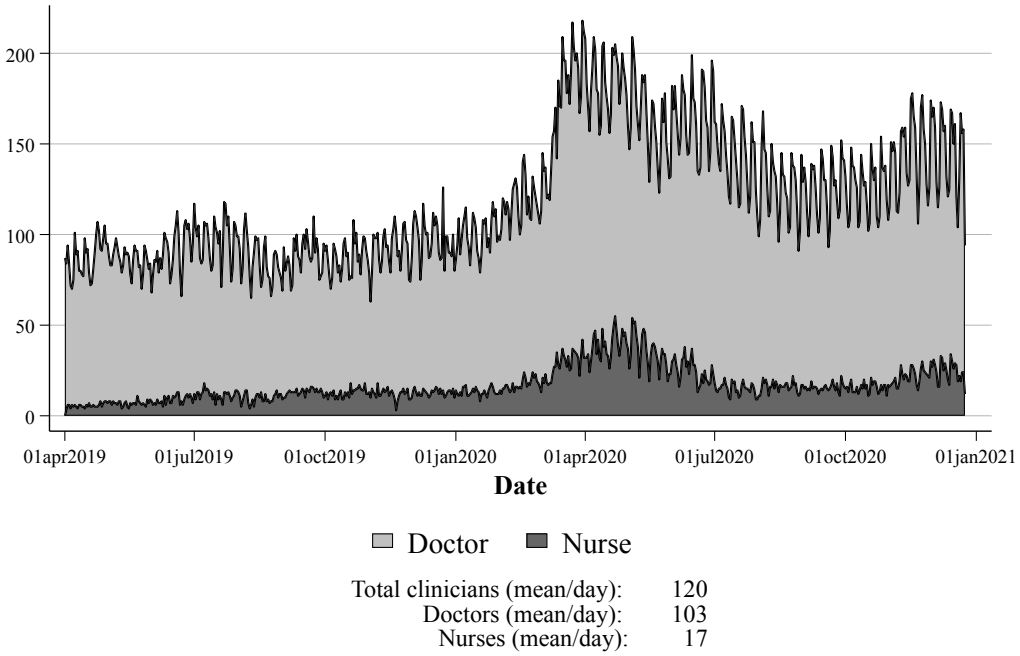
(b) Congestion measure and initial nurse assignment

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

Figure A3. Number of consultations and clinicians per calendar day



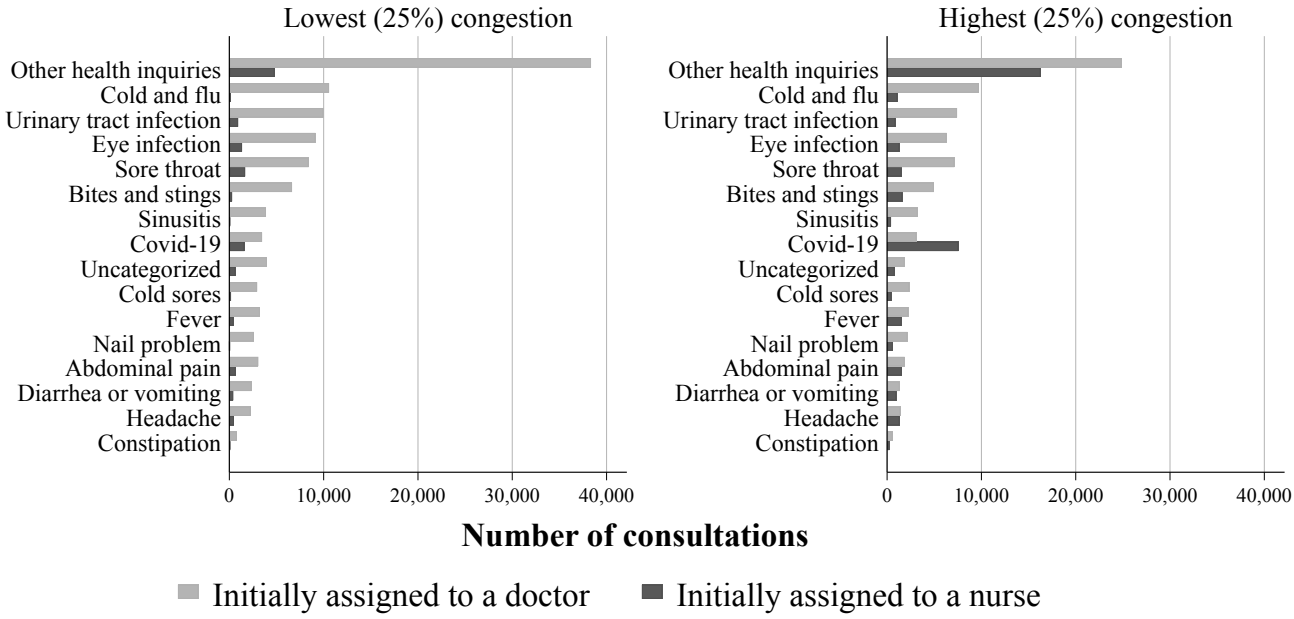
(a) Number of consultations



(b) Number of clinicians

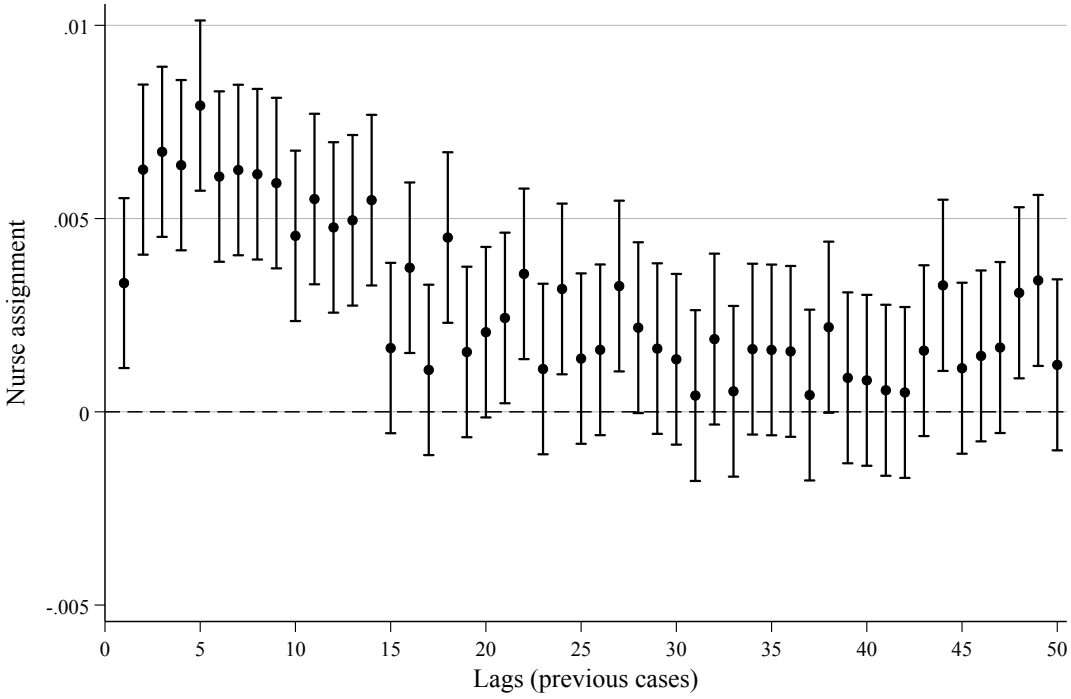
Notes: Panel (a) shows the number of consultations per calendar day for nurses and doctors in the main analysis sample. Panel (b) shows the number of active clinicians per calendar day for nurses and doctors in the analysis sample.

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



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

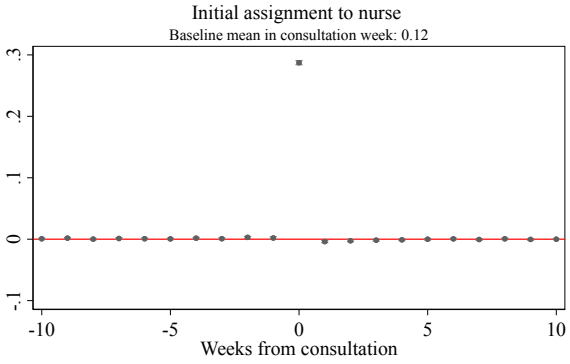
Figure A5. Effects of lagged doctor group case on case assignment



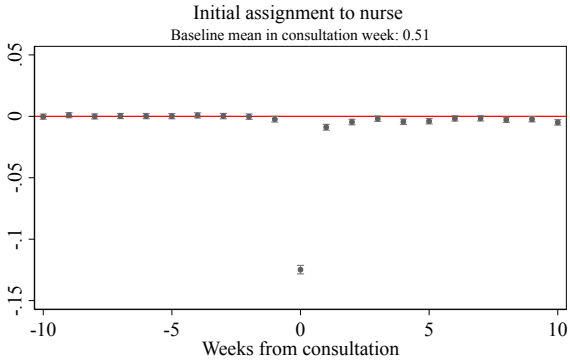
Notes: This figure reports estimates from regressions of initial nurse assignment on indicators for lagged doctor-group cases, ordered by patient login time. Each point corresponds to the coefficient from a separate OLS regression using the case at lag n. All regressions include date-by-3-hour fixed effects and use the analysis sample of overlap-group cases. Vertical lines denote 95% confidence intervals.

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

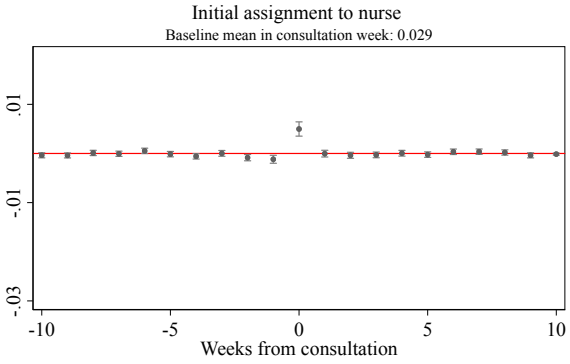
(a) Any consultation at the provider



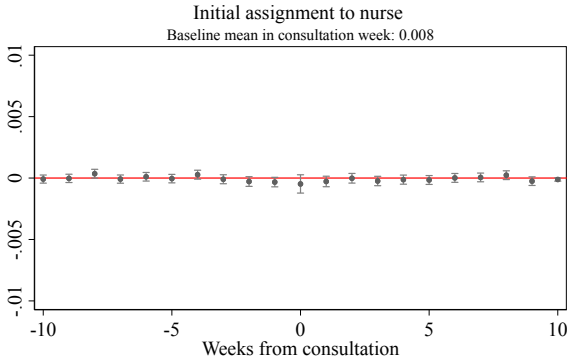
(b) Any prescription



(c) Any emergency department visit

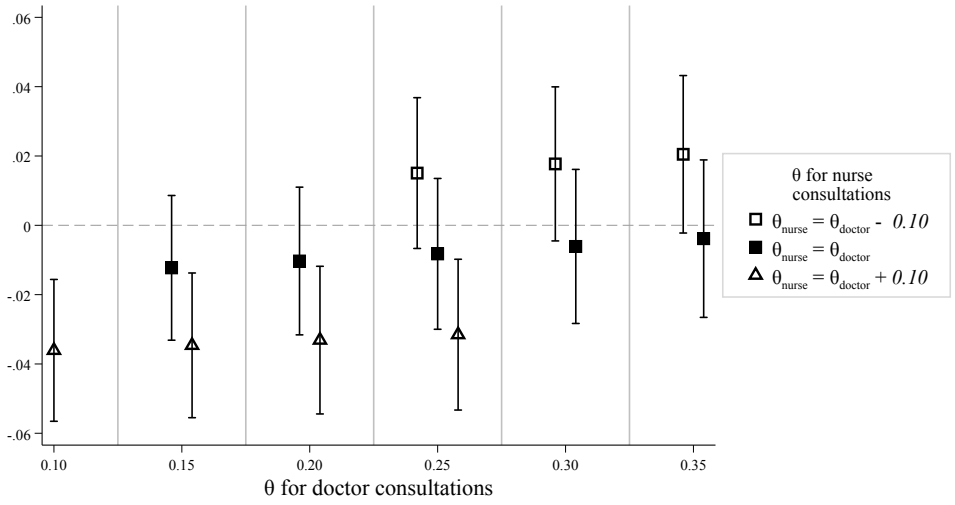


(d) Any hospitalization



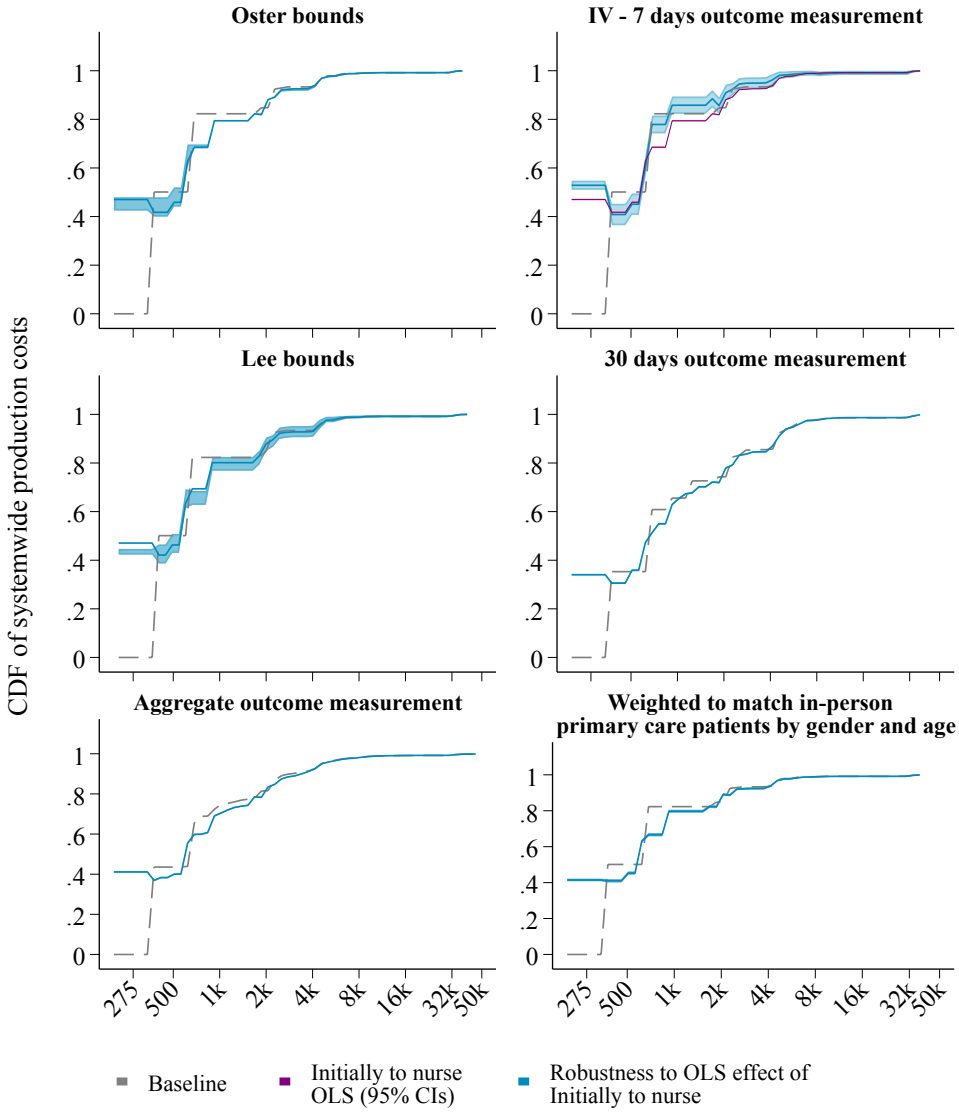
Notes: These figures show the estimated effect of initial assignment to a nurse (the *knowledge hierarchy*) on outcomes in the 10 weeks before and after the initial consultation. "Week 0" marks the week starting with the initial consultation date. Each estimate is from a separate OLS regression of a lagged or lead outcome on *Nurse*, based on Equation 4 with the full set of controls and date-by-3-hour fixed effects. The baseline mean is the average of the week-0 outcome for cases directly assigned to a doctor.

Figure A7. Systemwide production costs when markdowns θ vary between nurses and doctors



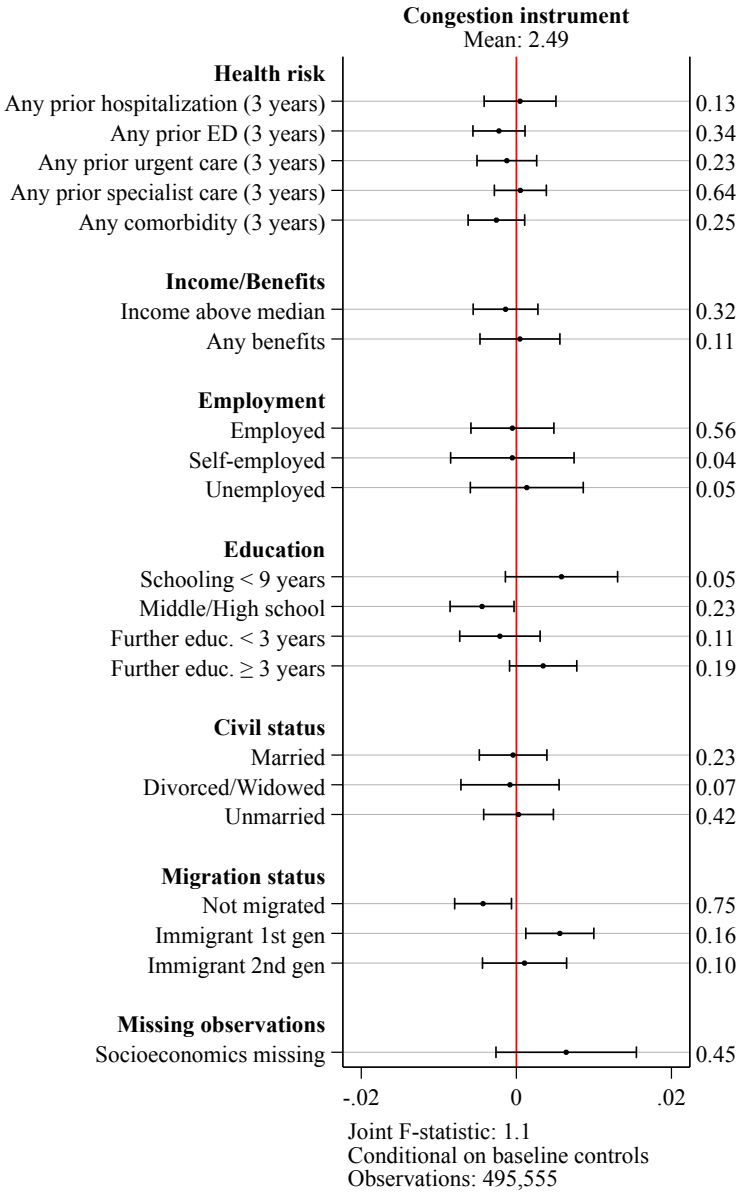
Notes: This figure shows the estimated proportional effect of initial nurse assignment on systemwide production costs across combinations of clinician-specific markdowns. The x-axis shows the doctor markdown and the y-axis shows the nurse markdown. Standard errors are in parentheses. The full set of X_1 and X_2 is included.

Figure A8. Robustness on the distribution of systemwide production costs



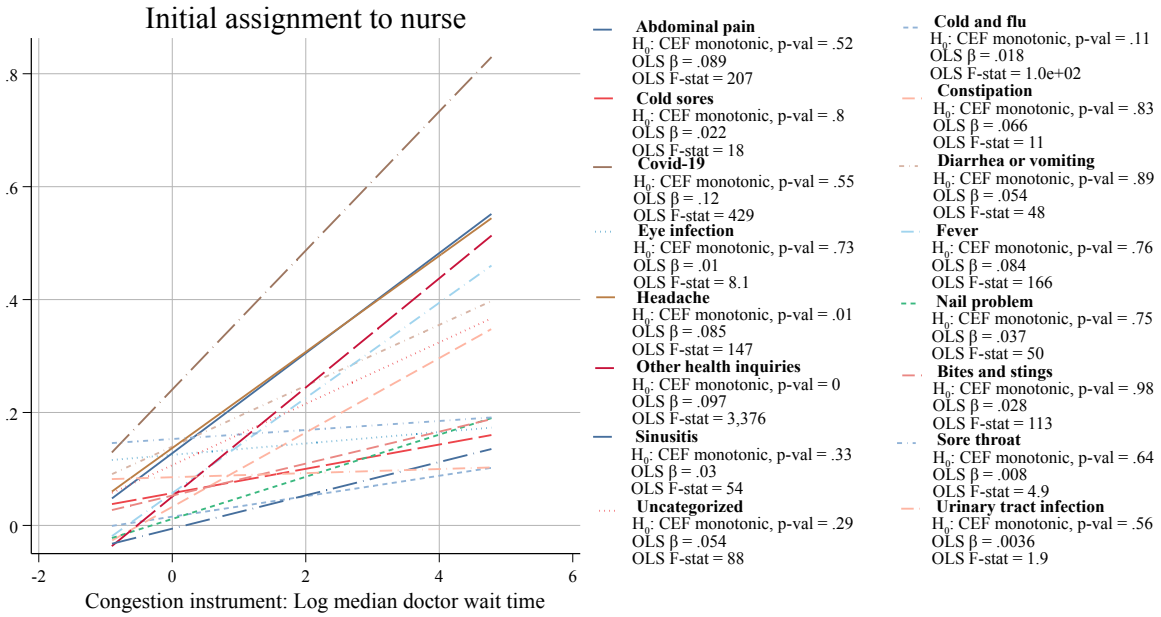
Notes: These panels provide robustness checks for the CDF estimates in Figure 2b. The top-left panel shows Oster bounds following Oster (2019). The grey dashed line reports the baseline specification, while the blue line shows the bias-adjusted estimates assuming the maximum R^2 is 1.3 times the controlled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

Figure A9. Balance of the congestion instrument in case characteristics



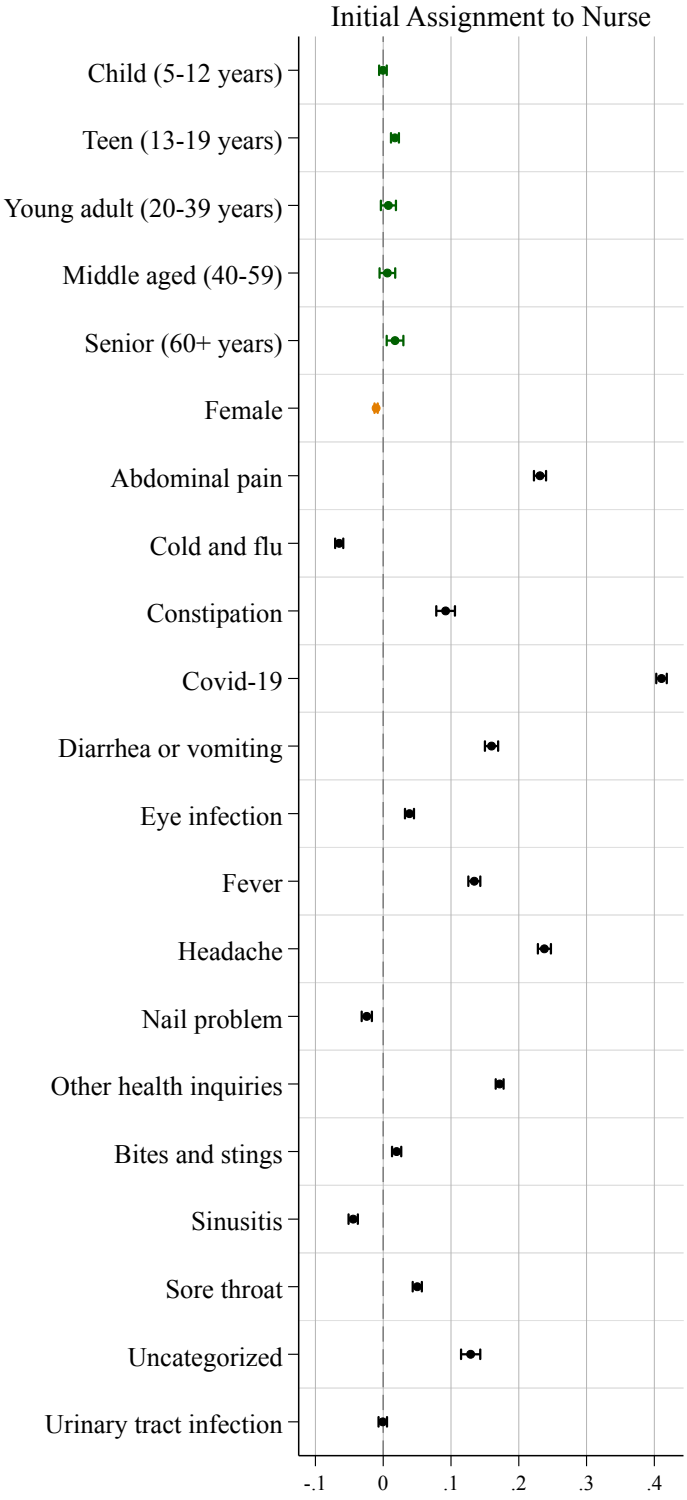
Notes: This figure presents balance tests for the congestion instrument (*Log median doctor wait time*). Each row reports a separate regression of the instrument on the characteristic listed on the vertical axis (each from X_2), conditional on all baseline controls (X_1): login time fixed effects (date-by-3-hour), main symptom categories, patient demographics, and a demographics × prior firm use interaction term. Horizontal lines represent 95% confidence intervals based on robust standard errors. The F-statistic is a joint F-test of the instrument on all additional case characteristics (X_2), conditional on all baseline controls (X_1). The numbers on the right show sample means for each characteristic.

Figure A10. Instrument monotonicity test



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

Figure A11. Nurse assignment heterogeneity



Notes: This figure shows heterogeneity in nurse assignment. The coefficients are estimated from a regression of nurse assignment on X_2 and a modified version of X_1. In this modified version, the interaction terms in X_1 exclude age, female, and symptom-group indicators, and the age dummies are replaced with age-group indicators. Horizontal lines show 95% confidence intervals based on robust standard errors.

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 (rather than a doctor) and the total number of consultations. Each row introduces an additional 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 overlap-group nurse consultations took place 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. Restriction 6 excludes consultations where the same patient consulted within the prior two hours. Restriction 7 excludes consultations with missing baseline characteristics (login time, patient age, gender, region) and defines the *broad drop-in sample*. From this sample, we define three subgroups 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 have fewer than 1,000 observations in the broad drop-in sample. The *analysis sample* consists of the 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	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	Pain, not elsewhere classified	2.69	3.08	43.7	25.4
N39	Other disorders of urinary system	2.59	0.52	19.0	13.6
J01	Acute sinusitis	2.35	0.52	4.23	4.71
Z53	Persons encountering health services for specific procedures, not carried out	2.69	0.033	16.2	1.80

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

Table A3. Comparison of sample characteristics to the Swedish population

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

Notes: This table compares average case characteristics in the analysis sample (N = 495,555) to the Swedish population and the Scania in-person primary care visit sample. Column (1) reports means for our sample; Column (2) reports means for Sweden in 2019; Column (3) reports means for non-acute in-person primary care visits in the Scania region. "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. "Any migration" is an indicator for individuals born outside Sweden (first-generation migrants) or whose parents were both born outside Sweden (second-generation migrants). "Annual earnings" includes annual earnings from wages and self-employment and is reported for individuals strictly over age 20.

Table A4. Exit rank on login 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 (X_1 and X_2). Exit rank is defined as rank sorted by meeting start time; login rank is rank sorted by patient login time. The third and fourth columns include indicators for deterministic nurse-group and doctor-group tasks. Robust standard errors are in parentheses.

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 below one minute	0.047	0.051	-0.0038	0.00	123,871	124,071
	(0.21)	(0.22)				
Consultation 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 consultation characteristics during periods of low and high doctor congestion. Low (high) congestion is defined as the lower (upper) quartile of the congestion instrument (*Log median doctor wait time*). Columns 1 and 2 show means and standard deviations (in parentheses). Columns 3 and 4 report the difference in means and its p-value from a t-test. Columns 5 and 6 show the number of observations in each group.

Table A6. Effects of the knowledge hierarchy on additional outcomes

	Top score rating (7 days)	No show (initial)
Initially to nurse	0.047***	0.00043**
	(0.0019)	(0.00014)
X_1 and X_2	✓	✓
Observations	495,555	495,555
Baseline mean	0.42	0.0010

	Technical error (initial)	Any external PCP consultation (7 days), Stockholm & Scania Sample
Initially to nurse	-0.0024***	0.030***
	(0.00026)	(0.0018)
X_1 and X_2	✓	✓
Observations	495,555	262,144
Baseline mean	0.0066	0.11

	Informative diagnosis (7 days)
Initially to nurse	-0.064***
	(0.0016)
X_1 and X_2	✓
Observations	495,555
Baseline mean	0.80

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 dependent variable. "Top score rating (7 days)" indicates a top satisfaction rating (5 out of 5) for any consultation within 7 days of the initial consultation; "Informative diagnosis (7 days)" indicates a non-missing diagnosis excluding ICD-10 chapters R and Z within 7 days; "Any external PCP consultation (7 days)" indicates a primary care visit outside the firm within 7 days, observed directly in Stockholm and Scania. "No show" and "Technical error" indicate whether the initial consultation was recorded as a no-show or a technical error. The columns report OLS estimates for *Nurse* based on Equation 4. All regressions control for baseline characteristics (X_1) and additional controls (X_2) as described in the main text. The baseline mean is the average of the dependent variable for cases directly assigned to a doctor. Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7. Effects of the knowledge hierarchy on patient outcomes – IV estimates

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***	-0.0079	0.012***	-0.0012	-0.14***	-0.064***
	(0.0015)	(0.010)	(0.00094)	(0.011)	(0.0019)	(0.023)	(0.0017)	(0.019)
X_1 and X_2	✓	✓	✓	✓	✓	✓	✓	✓
Observations	495,555	490,046	495,555	490,046	495,555	495,555	495,555	495,555
First-stage K-P F-statistic		3,639		3,639		3,670		3,670
Baseline mean	0.014	0.014	0.058	0.055	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***	-0.024	-0.10***	-0.029	0.0033***	-0.010	0.0031***	0.00046
	(0.0013)	(0.015)	(0.0017)	(0.022)	(0.00074)	(0.0089)	(0.00051)	(0.0061)
X_1 and X_2	✓	✓	✓	✓	✓	✓	✓	✓
Observations	495,555	490,046	495,555	490,046	495,555	490,046	495,555	490,046
First-stage K-P F-statistic		3,639		3,639		3,639		3,639
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***	-0.0077	-0.00049	-0.00019
	(0.00075)	(0.0086)	(0.00038)	(0.0045)
X_1 and X_2	✓	✓	✓	✓
Observations	495,555	490,046	495,555	490,046
First-stage K-P F-statistic		3,639		3,639
Baseline mean	0.030	0.030	0.0082	0.0083

Panel C: Long-term consequences

	Earnings decrease > 20pct. (cal. month after)		Zero earnings (cal. month after)		Death (3 years)	
	OLS	IV	OLS	IV	OLS	IV
Initially to nurse	-0.021***	0.0061	-0.0031**	0.0080	-0.000012	-0.0047**
	(0.0025)	(0.031)	(0.0013)	(0.017)	(0.00018)	(0.0022)
X_1 and X_2	✓	✓	✓	✓	✓	✓
Observations	214,674	212,906	214,674	212,906	495,555	490,046
First-stage K-P F-statistic		1,449		1,449		3,639
Baseline mean	0.22	0.22	0.050	0.050	0.0021	0.0020

Note: This table presents OLS and instrumental variable estimates for the effects of initial case assignment to a nurse (the *knowledge hierarchy*) rather than directly to a doctor. Each subpanel corresponds to a different dependent variable. Outcomes are defined as in Table A11. The left column in each subpanel presents the OLS estimate for *Nurse* based on Equation 4. The right column provides the IV estimate using *Log median doctor wait time* as an instrument for *Nurse*, estimated via Two-Stage Least Squares (2SLS). *Log median doctor wait time* is defined as the winsorized natural logarithm of the median wait time among the up to 5 prior doctor consultations within 30 minutes. All regressions control for baseline characteristics (X_1) and additional controls (X_2) as described in the main text. The baseline mean is the average of the dependent variable for cases directly assigned to a doctor. The K-P F-statistic is the Kleibergen-Paap F-statistic. Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A8. Effects of the knowledge hierarchy on cost outcomes – IV estimates

	Taxpayer payments to the firm (7 days)		Taxpayer payments to the firm (within-day)		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)
X_1 and X_2	✓	✓	✓	✓	✓	✓
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
	Firm production costs		Total taxpayer payments		Systemwide production costs	
	OLS	IV	OLS	IV	OLS	IV
Initially to nurse	-40.3***	-62.9***	-117.4***	-242.8	-9.72	-144.2
	(0.55)	(3.38)	(13.2)	(154.2)	(13.2)	(154.2)
X_1 and X_2	✓	✓	✓	✓	✓	✓
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	375	375	1304.8	1304.8	1179.8	1179.8

Note: This table presents OLS and instrumental variable estimates for the effects of initial case assignment to a nurse (the *knowledge hierarchy*) rather than directly to a doctor. Each subpanel corresponds to a different dependent variable. Outcomes are defined as in Table A11. The left column in each subpanel presents the OLS estimate for *Nurse* based on Equation 4. The right column provides the IV estimate using *Log median doctor wait time* as an instrument for *Nurse*, estimated via Two-Stage Least Squares (2SLS). *Log median doctor wait time* is defined as the winsorized natural logarithm of the median wait time among the up to 5 prior doctor consultations within 30 minutes. All regressions control for baseline characteristics (X_1) and additional controls (X_2) as described in the main text. The baseline mean is the average of the dependent variable for cases directly assigned to a doctor. The K-P F-statistic is the Kleibergen-Paap F-statistic. Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A9. Robustness checks: Systemwide production costs

Panel A: Alternative controls and unobserved selection

	Only baseline controls	Full controls with date-by-1 hour windows	Full controls with symptom controls x 3-hours windows x weekday x quarter	Full controls including congestion quartiles
Initially to nurse	-11.3	-5.68	-3.25	-9.43
	(13.1)	(13.3)	(13.2)	(13.1)
X_1	✓			
Date-by-1 hour FEs		✓		
Congestion quartiles				✓
X_1 and X_2		✓	✓	✓
Observations	495,555	494,819	495,485	495,555
Baseline mean	1179.8	1179.0	1179.8	1179.8

	Full controls including patient fixed effects	Oster bounds
Initially to nurse	4.90	-9.72
	(23.2)	(13.1)
X_1 and X_2	✓	✓
Patient FEs	✓	
Observations	228,694	495,555
Bias-adjusted β		-35.6
Uncontrolled β		76.5
Baseline mean	1189.2	1179.8

Panel B: Differential attrition

	Lee bounds	Cases with past nurse consultations	Difference between first period (April-June) and second period (July-September)
Initially to nurse × second period			62.1
			(65.0)
Initially to nurse		-43.5	
		(39.4)	
Lower bound	-256.1		
Upper bound	246.0		
Symptom controls			✓
X_1 and X_2	✓	✓	
Observations	495,555	52,107	104,137
Baseline mean	1179.8	1275.6	1086.0

Panel C: Cost calculations

	Stockholm & Scania sample	Aggregate costs (7 days)	Aggregate costs (30 days)
Initially to nurse	-11.2	-6.52	-21.0
	(18.4)	(16.5)	(25.8)
X_1 and X_2	✓	✓	✓
Observations	262,144	495,555	495,555
Baseline mean	1203.8	1476.1	2648.3

Panel D: External validity

	Weighted to match in-person primary care	Pre Covid-19 period	Pre Covid-19 + interlude period
Initially to nurse	28.8	80.5***	18.3
	(56.0)	(22.0)	(17.8)
X_1 and X_2	✓	✓	✓
Observations	495,555	200,779	296,647
Baseline mean	1179.8		

	First and second Covid-19 wave	2 months before Covid-19 period	2 months after Covid-19 period
Initially to nurse	-25.1	83.7***	70.1***
	(19.9)	(25.9)	(17.9)
X_1 and X_2	✓	✓	✓
Observations	197,806	157,342	284,470
Baseline mean		1108.9	1144.7

Notes: This table presents robustness checks for the effect of initial case assignment to a nurse on systemwide production costs. Each cell shows the OLS estimate for *Initially to nurse*, under alternative model specifications. Baseline controls (X_1) include symptom categories fully interacted with demographics and with prior provider use, and region. Additional controls (X_2) include health risk and socioeconomic variables. Patient fixed effects are included where indicated. All regressions use date-by-3-hour fixed effects for login time. The IV uses *Log median doctor wait time* as an instrument, estimated via 2SLS following Equations 4 and 16. The baseline mean is the average of the dependent variable for cases directly assigned to a doctor. The K-P F-statistic is the Kleibergen-Paap F-statistic. Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A10. Robustness: Percentage change in systemwide production costs under constrained minimisation, holding nurses constant within 15-minute windows

	$\frac{\Delta \text{Systemwide production costs}}{\text{Systemwide production costs}}$	$\frac{\Delta \text{Systemwide production costs}}{\text{Firm production costs}}$
1) Main specification	−2.05% (0.21)	−6.71% (0.68)
3) Aggregate costs (7 days)	−2.17%	−7.76%
2) Aggregate costs (30 days)	−1.57%	−8.88%
4) Sample restricted to Stockholm and Scania	−2.03%	−6.80%
5) Split sample using even and odd days	−2.30%	−7.52%

Note: This table shows the percentage change in outcomes from counterfactual assignments relative to the actual assignment for different systemwide production cost specifications. Standard errors are bootstrapped 93 times and reported only for the main specification due to computational constraints. The two columns differ in their denominator: Column 1 normalises the change in systemwide production costs by total systemwide production costs, while Column 2 normalises by firm production costs. Row 1 (main specification) uses 7-day systemwide costs for Column 1 and initial-day firm production costs for Column 2. Rows 2 and 3 use aggregated costs for both columns over 30-day and 7-day horizons, respectively. Row 4 restricts the sample to patients in the Stockholm and Scania regions. Row 5 estimates the comparative advantage of symptoms using odd calendar days and applies the resulting counterfactual assignment to even calendar days, providing an out-of-sample assignment rule.

Table A11. Description of key case characteristics

Variable	Description	Data source
Baseline case characteristics (X_1)		
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
Female	Indicates whether the patient is female.	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 (Statistics Sweden)
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
Symptom category	Set of indicator variables for a case's main symptom, as indicated by the patient before the consultation. The full list of symptoms is shown in the Table 2.	Provider
Last visit within the firm was with a doctor	Indicates whether the patient's most recent consultation with the firm, prior to the initial consultation, was initially assigned to a doctor; coded 0 if no prior consultation exists.	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 (National Board of Health and Welfare)
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
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, excluding visits with ICD codes in category O related to childbirth.	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, excluding visits with ICD codes in category O related to childbirth.	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, excluding visits with ICD codes in category O related to childbirth.	NBHW
Any unbooked revisit (7 days)	Indicates whether the patient has an unscheduled return visit with the same symptom category within 7 days.	Provider

Table A11. Description of key case characteristics (Continued)

Variables	Descriptions	Data source
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
Earnings decrease $>$ 20 pct. (cal. month after)	Indicator for whether the patient's earnings fell by at least 20% relative to their average earning over the previous three months, conditional on that the patient has a stable income (received earnings 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 earning (cal. month after)	Indicates whether the patient had no earnings from their main income source in the calendar month following the initial consultation. Defined for a subsample with stable earnings consisting of employed patients who earn more than 3,533 SEK (2010 values) in each of the three months prior, based on an annual threshold of 42,400 SEK (Saez et al., 2019).	SCB
Outcomes: Cost measures		
Systemwide production costs	The total costs of healthcare provision within 7 days after the initial assignment, calculated as total taxpayer payment minus the firm's net revenues created using the indicator approach. 30 days in parenthesis refer to the version that counts within 30 days healthcare provisions	NBHW
Aggregate costs	The systemwide production costs incurred within 7 days after the initial assignment, calculated as total costs minus the firm's net revenues created using all the healthcare events within 7 days, rather than the indicator approach.	NBHW
Total taxpayer payments	The total taxpayer cost of healthcare services incurred within 7 days of the initial consultation. This includes costs at the firm on the calendar day of the initial consultation and the cost of up to one occurrence of each downstream event within 7 days, including an emergency department visit, specialist visit, urgent care visit, hospitalization, or prescription.	NBHW
Taxpayer payments to the firm	The taxpayer's incurred costs for consultations at the firm, including up to one consultation within-day (per reimbursement rules) and up to one additional follow-up consultation within 7 days of the initial assignment (per the indicator approach discussed in 6.2). Within the week, two consultations may occur if both a nurse and a doctor are involved; otherwise, at most one consultation is included.	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
Downstream costs	The taxpayer's incurred costs of total healthcare provision within 7 days of the initial assignment, covering only downstream, out of firm spending. These include costs of up to one occurrence of each downstream event emergency department visit, specialist or urgent care visit, hospitalization, or prescription within 7 days).	NBHW
Firm net revenue	The firm's net revenue from this case on the calendar day of the initial assignment, defined as gross revenues minus costs of provision. Gross revenues are earned for up to one consultation per day (per reimbursement rules). However, costs would be incurred for all consultations. Costs are measured as within-day taxpayer payments of the initial and potential follow-up consultations, multiplied by θ, where θ denotes the firm's markdown.	Provider
Firm production costs	Consultations that are not reimbursed are treated as costs to the firm. Costs are measured as within-day taxpayer costs of the initial consultations, multiplied by (1 - θ), where θ denotes the firm's markdown.	Provider

Table A11. Description of key case characteristics (Continued)

Variables	Descriptions	Data source
Additional case characteristics (X_2)		
Education level	Set of indicator variables for the highest educational degree reported in Nov 2018, as one of the following categories: schooling below 9 years; middle/high school; secondary education $<$ 3 years; secondary education $\geq$ 3 years. Educational degrees are reported for patients $\geq$ age 25.	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
Civil status	Set of indicator variables for the patient's civil status reported in Nov 2018, as one of the following categories: married; divorced or widowed; unmarried. Civil status is reported for patients $>$ age 18.	SCB
Any comorbidity	Indicates an Elixhauser comorbidity index of 1 or higher pre-2019, as diagnosed within specialty or hospital care within the 3 years prior to the initial consultation, excluding visits in the 2 days immediately preceding it. The Elixhauser list of comorbidities is adapted from 4-digit to 3-digit ICD codes to fit the number of digits in the reporting from Socialstyrelsen.	NBHW
Any prior ED	Indicates any ED visits within the 3 years prior to the initial consultation, excluding visits in the 2 days immediately preceding it. Excludes visits with ICD codes in category O related to childbirth.	NBHW
Any prior hospitalization	Indicates any inpatient hospitalization within the 3 years prior to the initial consultation, excluding visits in the 2 days immediately preceding it. Excludes visits with ICD codes in category O related to childbirth.	NBHW
Any prior specialist care	Indicator for any specialist visits within the 3 years prior to the initial consultation, excluding visits in the 2 days immediately preceding it. Excludes visits with ICD codes in category O related to childbirth.	NBHW
Any prior urgent care	Indicates any urgent care center visit within the 3 years prior to the initial consultation, excluding visits in the 2 days immediately preceding it. Excludes visits with ICD codes in category O related to childbirth.	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 receipt of at least one social insurance or housing related benefit in 2018, including youth disability insurance, disability insurance, sickness benefits, rehabilitation benefits, or housing allowance. Benefits are reported for individuals aged 20 or older in November 2018.	SCB
Earnings above median	Indicates whether the patient's annual earning in Nov 2018 is greater than the in-sample median of non-missing earnings (294,700 SEK, approximately 28,000 USD). earning 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, earnings, 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	Indicates whether the initial consultation was recorded as a "no show".	Provider
Technical error	Indicates whether the initial consultation had a technical error.	Provider
Case characteristics used in robustness		
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)	The Public Health Agency of Sweden

Table A12. 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 A13. 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)
X_1	✓	✓	✓	✓	✓	✓
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)
X_1	✓	✓	✓	✓	✓	✓
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
	Earnings		Migration background		Further education	
	$>$ median	$\leq$ 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)
X_1	✓	✓	✓	✓	✓	✓
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 first-stage results of Appendix Equation 16 across various subsamples. The instrument is the winsorized natural logarithm of the median wait time in the up to 5 prior doctor consultations within 30 minutes (*Log median doctor wait time*). *Further educ.* refers to any post-secondary education reported for patients aged 25 or older; income is reported for patients over the age of 20. Subsamples split by further education and median income exclude patients with missing or undefined information. Variable definitions are provided in Appendix Table A11. All regressions control for (X_1) as described in the main text. The baseline mean is the average of *Initial assignment to nurse*. The K-P F-statistic is the Kleibergen-Paap F-statistic. Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A Additional institutional details

Cost-effectiveness. The firm's choice of a clinician type (such as nurse or doctor) for a case is guided by cost-effectiveness 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) 2019). Cost-effectiveness is formalized as a requirement to deliver care at the 'lowest effective care level' (referred to as the *LEON* principle, based on the Swedish *Lägsta Effektiva Omhändertagande-Nivå*),[17] 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.[18]

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) 2019). 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 in addition to doctors, in order to 'prioritize doctor consultations to where they make the most difference' (Kry International AB 2024).

Training and scope of practice. The clinician types differ in training in a similar way to other countries' doctors and nurses. Doctors typically complete an 11-semester medical degree, followed by an 18-month internship (Allmäntjänstgöring, AT) to obtain their license, and five or more years of residency training to become specialists (Specialiseringstjänstgöring, ST), amounting to seven or more years of formal and on-the-job training. Nurses complete a three-year degree for their license and may, after at least one year of experience, pursue a 1.5-year specialization (such as district

[17] 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).

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

nursing, midwifery, or anesthesia), although most nurses do not.[19] Nurses and doctors also differ in their scope of practice. Nurses typically handle patient communication, routine check-ups, and triage, whereas only doctors can prescribe most medications, issue specialist referrals, and issue longer sick notes.

Wages and labor markets. In Sweden in 2020, average monthly salaries were about 38,700 SEK for nurses and 79,200 SEK for doctors (Statistics Sweden (SCB) 2020).[20] While others have noted that doctor incomes in Sweden are less concentrated at the top of the income distribution than in other high-income countries (Buehler et al. 2026), this is not specific to doctors but reflects an overall compression of the wage distribution in Nordic countries (Mogstad, Salvanes, and Torsvik 2025). Buehler et al. (2026) also notes that relatively low doctor wages are consistent with monopsony markdowns, as we also model. Sweden faces shortages of healthcare workers, particularly among specialized nurses and specialized doctors (Socialstyrelsen 2022). The nurse-to-doctor ratio is about 2.5 nurses to one doctor, which is close to the OECD average ratio (Table 1).

B Data Appendix

B.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.

B.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

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

[20] 1 SEK = 0.1087 USD in 2020 (Sveriges Riksbank 2020).

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. We define a diagnosis as *informative* if the International Classification of Diseases (ICD) code is neither missing nor a symptom or health status update (R00-R99, Z00-Z99). Diagnoses are rarely missing in our analysis sample (0.32% for initial consultations), as the provider is required to provide a diagnosis for reimbursement by the public payer. In physical primary care in Scania, 52.59% of diagnoses from nurse or doctor consultations are informative. For primary care visits in Stockholm, we observe the date of the contact, the care branch, the pseudonymized identifier of the healthcare clinic, 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.

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

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.

B.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. We include both prescriptions from the firm and external prescriptions. To exclude prescriptions for chronic conditions, we consider only drugs on the chemical subgroup level (Anatomical Therapeutic Chemical level 4) that patients have not yet purchased at pharmacies within the 6 months prior to the initial consultation. Mortality data, available for 2019 to 2023, include the date and cause of death. Outpatient and inpatient data, spanning 2013 to 2023, include visit and admission dates, ICD-10 diagnostic codes, and, for outpatient data, additional information on contact types.

Healthcare data 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.

B.5 Details of primary care and downstream costs

Cost estimates are obtained from the sources presented in the Appendix Table A12. 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.

B.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 A11 presents definitions of the case characteristics used in our main analysis.

C Model Appendix

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

Queue structure. In each period t, the queue consists of two groups: *doctor-group* cases, for which only direct doctor assignment is suitable, and *overlap-group* cases, for which either clinician type may be assigned initially. Let $Q_t^D \geq 0$ and $Q_t^{\text{OV}} \geq 0$ denote the number of doctor-group and overlap-group cases, respectively, to be assigned in period t.[21]

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

Doctors are first assigned to doctor-group cases. Define

$$\tilde{E}_t \;\equiv\; E_t - \min(E_t,\, Q_t^D) \;\geq\; 0 \tag{6}$$

as the number of doctors remaining after doctor-group cases have been assigned in period t; any unserved doctor-group cases carry over to period $t+1$. The $\tilde{E}_t$ remaining doctors and F_t nurses are then available for overlap-group cases. Let

$$M_t^{\mathrm{OV}} \;\equiv\; \min(\tilde{E}_t + F_t,\, Q_t^{\mathrm{OV}}) \tag{7}$$

denote the number of overlap-group cases assigned in period t; unassigned overlap-group cases carry over to period $t+1$. Note that M_t^{OV} corresponds to M_t from Section 4.1, restricted to overlap-group cases. We assume that cases passed from nurses to doctors are processed in later slots outside the main queue, so $\tilde{E}_t$ in (6) is calculated before such cases are added back to a doctor's workload.

Revenue structure. The firm's per-case revenues are administratively set and follow a *resolving-clinician* payment rule: the firm is compensated only for the clinician type that resolves the case. Because it is the doctor who resolves the case in both the direct-to-doctor route and in the knowledge hierarchy route when a case is passed from a nurse to a doctor, this rule implies

$$r_{F,E} \;=\; r_E. \tag{8}$$

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

$$r_E \;=\; \frac{\kappa_E}{1-\theta}, \qquad r_F \;=\; \frac{\kappa_F}{1-\theta}, \tag{9}$$

so that each payment to the firm covers its production cost with a proportional margin $\theta/(1-\theta)$. Together, Equations (8) and (9) give the full revenue vector

$$(r_E,\, r_F,\, r_{F,E}) \;=\; \left(\frac{\kappa_E}{1-\theta},\, \frac{\kappa_F}{1-\theta},\, \frac{\kappa_E}{1-\theta}\right), \tag{10}$$

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

Unconstrained assignment. Substituting the firm's revenue structure (10) into Equation (1), the net revenue advantage of the knowledge hierarchy simplifies to

$$\Delta(\tau) \;=\; p_\tau \cdot \frac{\kappa_F - \theta\kappa_E}{1-\theta} - \kappa_F. \tag{11}$$

Since costs are such that $\kappa_E > \kappa_F$ and $\Delta(\tau)$ is linear in p_τ with bounds

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

we see that $\Delta(\tau) < 0$ for all τ. 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 the firm's revenue structure (10) gives

$$\frac{\partial \Delta(\tau)}{\partial p_\tau} \;=\; r_F - r_{F,E} + \kappa_E \;=\; \frac{\kappa_F - \theta\kappa_E}{1-\theta}, \tag{12}$$

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 the firm's revenue structure (10) in our setting to illustrate how different pay structures affect firm incentives.

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

$$(r_E,\, r_F,\, r_{F,E}) \;=\; (f,\, f,\, f). \tag{13}$$

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 passing on an unresolved case to a doctor generates revenues from two consultations (one with the nurse, one with the doctor), while a directly resolved case generates only one paid consultation, the revenue vector is

$$(r_E,\, r_F,\, r_{F,E}) \;=\; (\rho,\, \rho,\, 2\rho). \tag{14}$$

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

total production costs of healthcare:

$$\Delta^{SC}(\tau) \equiv \text{SC}_F(\tau) - \text{SC}_E(\tau) \;=\; [\kappa_F + (1-p_\tau)\kappa_E + \mathbb{E}[C_{F,\tau}]] - [\kappa_E + \mathbb{E}[C_{E,\tau}]], \tag{15}$$

where $C_{F,\tau}$ and $C_{E,\tau}$ are the downstream cost random variables defined in Section 4.1.[22] The planner assigns type-τ cases to the knowledge hierarchy whenever (15) is negative, and to doctors otherwise. When clinician availability is constrained, the planner ranks case types by (15) 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.

D Additional robustness

D.1 Alternative identification: Instrumental variable strategy

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

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

$$Nurse_i = \gamma_i \text{Doctor Congestion}_i + X_{1i}\lambda_{1i} + X_{2i}\lambda_{2i} + \nu_i, \tag{16}$$

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 Equation 4 and Appendix Equation 16 with robust standard errors.

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

Validity of the congestion instrument. 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

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

set of controls (Column 3) and reports the first-stage estimates from Appendix Equation 16. 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 A10 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, Goldin, and Yang 2018). Appendix Table A13 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, even in the highest congestion quartile, the mean wait time is below 45 minutes. This is unlikely to meaningfully affect care quality or costs, especially when compared to physical primary care, where over 25% of patients wait more than 24 hours for an initial doctor appointment.

D.2 Attrition

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

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