Physician workload and treatment choice: the case of primary care*

Ity Shurtz[†1], Alon Eizenberg[‡2], Adi Alkalay[3], and Amnon Lahad[3,4]

[1]Department of Economics, Ben-Gurion University of the Negev
[2]Department of Economics, The Hebrew University of Jerusalem, and CEPR
[3]Clalit Health Services
[4]School of public health, The Hebrew University of Jerusalem

January 2022

Abstract

Primary care is a notable example of a service industry where capacity-constrained suppliers face fluctuating demand levels. To meet this challenge, physicians trade off their time with patients with other inputs such as lab tests and referrals. We study this tradeoff using administrative data from a large Israeli HMO where the absence of colleagues generates exogenous variation in physician workload. We motivate and estimate a range of specifications, from a classic exclusion restriction within a linear model to non-parametric, partially-identified models. The results suggest that diagnostic inputs are unlikely to properly compensate for a decline in time spent with patients.

*First public version: February 2018. We are grateful for helpful conversations and comments from Tobias Kline, Jianjing Lin, Charles Manski, Fernanda Marquez-Padilla, and John Pepper. All errors are our own. We have also benefitted from comments by seminar participants at Ben Gurion, Carnegie Mellon, Mannheim, Tel-Aviv, and Tilburg, and conference participants at the Israeli IO Day 2017, iHEA Boston 2017 Congress, Barcelona GSE Summer Forum, Policy Evaluation in Health 2017, the 7th Annual Conference of the American Society of Health Economists, and the IIOC 2018. Nadav Perlov provided excellent research assistance. Financial support from the Maurice Falk Institute for Economic Research in Israel and from the Israel National Institute for Health Policy Research is gratefully acknowledged.

[†]shurtz@bgu.ac.il
[‡]alon.eizenberg@mail.huji.ac.il

1 Introduction

In many industries, firms' capacity is either fixed or very costly to modify while demand fluctuates between high and low states. Airlines, hotels and car rental agencies address this challenge via price adjustments. Prices then efficiently reflect the varying shadow cost of the capacity constraint. Professional service providers, in contrast, often refrain from price adjustments, and opt to adjust service quality instead (Chatain and Eizenberg (2015)). This article studies the nature of such adjustments in healthcare, focusing on the primary care context.

Primary care services play an important role in delivering public health while curbing health care costs (Starfield et al. (2005)). Importantly, providers and policy makers have been looking for remedies to the "primary care crunch," i.e., the shortage in primary care physicians in the United States and elsewhere. The capacity-quality tradeoff has been identified as a crucial issue in this context, as noted in Anand et al. (2011):

> "(a) major difficulty in improving productivity in such customer-intensive services is the sensitivity of...service quality... to the speed of service: as the service speed increases, the quality of service inevitably declines...*(p)rimary health-care practice in the United States epitomizes this problem*" (*italics* added by the authors).

The nature of this phenomenon, however, is far from obvious, owing in part to the complexity of the decisions made by primary care physicians.[1] It is not easy to determine how primary care physicians trade off one key input — their time with patients — with other inputs such as the prescription of medication, tests or referrals to specialists. A number of questions arise: does the shadow cost of physician capacity manifest itself in poor treatment of routine issues? Is it reflected in a limited ability to address a wide scope of long-term health issues? Does the tightening physician capacity result in more intense prescription of antibiotics or painkillers? Do diagnostic inputs such as lab tests substitute for face-time with patients?

We address these questions using detailed visit-level administrative data from eleven clinics of a large Israeli HMO during 2011-2014.[2] Fluctuating demand states interact with physician capacity to determine the number of patients seen per hour, or, equivalently, the average visit length — a commonly used indicator of physician workload. We empirically examine the causal effect of workload on physicians' decisions.

We interpret this causal effect within an analytical framework where physicians produce patient health using multiple inputs: time with patients, diagnostic tests, medication prescriptions, and referrals to the Emergency Room (ER) or to specialists. Physicians choose the optimal allocation of those inputs taking into account input costs. Physician workload is linked

[1] As noted in Scott (2000), "GPs make many different types of decisions that influence the amount, type and location of care received by patients. These include decisions to refer to a specialist or other health professionals, prescribe medication, arrange follow-up, and order tests."

[2] The issue is as relevant in Israel as in the United States. In 2011, for example, a physician strike resulted in a decision to allow primary care physicians to see five rather than six patients per hour.

in this framework to the opportunity cost of physician face-time. The causal effect of a tightening time constraint is therefore related to the substitutability or complementarity between physician time and other inputs.

In addition to providing context to our research questions, the analytical framework also helps guide our empirical strategies, aimed at identifying the causal effect of workload. The identification challenge stems from the presence of unobserved factors that simultaneously affect both the outcome of interest, say, a binary indicator for the prescription of medication, and the physician's workload.

The route cause of this endogeneity challenge is a potential relationship between workload and the underlying distribution of medical conditions presented to the physician on a given day. The analytical framework describes various channels that generate this relationship, from local infections that affect both the number of visiting patients and their medical needs, to self-selection of patients with acute conditions to seek treatment when the expected face-time with the physician is low.

We address this challenge using the absence of fellow physicians at the clinic as a source of exogenous variation in workload: physicians attend to patients of the absent colleague in addition to their regular patients. To avoid confounding the effect of workload with the effect of treating unfamiliar patients, we only analyze the physicians' decisions with respect to their regular patients.

The analytical framework clarifies what assumptions on the underlying behavior of patients and physicians are needed to justify this exclusion restriction. It also articulates assumptions that are compatible with weaker restrictions, namely, a Monotone Instrumental Variable (MIV, Manski and Pepper (2000)) approach that allows the instrument to shift the response function. We therefore estimate the effect of physician workload employing a range of techniques, from standard 2SLS to the nonparametric estimation of bounds of the Average Treatment Effect (ATE), while clarifying the different sets of assumptions that provide credence to each of these methods in our context.

Results. One may expect physicians to be more conservative and administer higher levels of medical treatment when workload is higher, but we find very limited evidence for such behavior. Our 2SLS regressions imply that physician workload has no significant impact on referrals to the ER or on the prescription of painkillers. Some specifications do provide mixed evidence that higher workload is associated with an increase in prescription of antibiotics. By and large, physicians do not appear to adjust the intensity of medical treatment in response to heightened workload levels.

We do, however, find that workload has a significant effect on the use of diagnostic inputs. The 2SLS regressions suggest that diagnostic inputs serve as complements to physician time: a one minute decrease in average visit length causes a 9 percent decrease in referrals to specialists, a 3.8 percent decrease in referrals to lab tests, and a statistically insignificant decrease in referrals to imaging. We also find that high workload causes physicians to limit the scope of

issues they address during a visit, as captured by the number of recorded diagnoses.

We investigate the robustness of the complementarity between diagnostic tools and physician face-time by relaxing the linear structure and the strict exclusion restriction. Instead, we turn to nonparametric analysis, and consider the weaker MIV restriction, while gaining additional identifying power via a Monotone Treatment Selection (MTS) restriction. The resulting estimated 95 percent confidence interval suggests that high workload may increase the probability of using diagnostics by at most 1 percent, but may reduce it by as much as 44 percent. Taken together with the 2SLS analysis, the results therefore rule out that physician time and diagnostic inputs are substitutes in an economically-meaningful fashion, while leaving a considerable scope for complementarity between time with patients and the use of diagnostics.

We further examine whether patients compensate for a physician's limited attention on a high-workload day by returning to the clinic on another day. The answer appears negative: we find no evidence that workload increases the likelihood of subsequent visits.

Policy relevance. The results contribute several insights to our understanding of the "primary care crunch." Given the considerable public interest in this phenomenon, it is important to understand its nature: what precisely is lost as the speed of primary care service rises?

The results suggest that the shadow cost of physician capacity does not include direct and indirect costs of intensified prescriptions or ER referrals.[3] We do find, however, that physicians record fewer diagnoses and make less use of diagnostic inputs when facing a tightening time constraint. This points to the conclusion that what is lost due to tighter physician capacity is the ability to follow up on a wide scope of medical issues.

Discussing non-urgent matters and following up on them is an important aspect of preventive care, a staple of the primary care apparatus. The results therefore suggest that a key component of the shadow cost of physician capacity may be the lower provision of preventive care.[4]

The findings inform an additional aspect of the tight capacity of primary care physicians. It is sometimes argued that investment in health care infrastructure and technology can alleviate the strain that the shortage in physicians places on the system (Bodenheimer and Smith, 2013). At least in the context of diagnostic technology such as lab tests, our results suggest that such investments may fall short in achieving these goals: physicians are not found to use such diagnostic tools to substitute for their face-time with patients.

On the other hand, the results suggest that providers could partially compensate for the tightening physician capacity by investing in IT infrastructure that specifically enhances the ability to monitor long-term issues. For example, artificial intelligence could be used to suggest to physicians what routine tests and follow-up care to consider for each patient, and information systems could help physicians communicate such issues to patients remotely and efficiently, e.g., via text messages or brief electronic encounters.[5]

[3] Indirect costs refer, for example, to negative externalities from over-prescription of antibiotics or painkillers.

[4] Our analysis remains silent on the question of what is the "optimal" level of preventive care.

[5] A related issue is the growth of urgent care centers, where patients walk in to address an acute medical

Finally, we note that these findings would be of limited importance if providers have a good understanding of their physicians' response to the capacity constraint, *and* trade off their investment in expanded capacity versus investment along other margins in a socially-optimal fashion. Neither one of these conditions, however, is likely to be met.

The complex set of tools at the physicians' disposal suggests that providers are not likely to have a clear sense of the channels via which physicians respond to workload. And even if providers were perfectly informed regarding these mechanisms, and optimally solved their private cost-minimizing problem, they could still fail to internalize the social costs associated with physician workload. As taxpayers end up shouldering some of the costs associated with poorer long-run health outcomes, a misalignment of incentives is to be expected.

Literature. A growing literature studies the role of supply side factors in driving the well documented variation in health care utilization (Finkelstein et al., 2016; Currie et al., 2016). Recent work has focused on financial incentives (Clemens and Gottlieb, 2014; Ho and Pakes, 2014), liability concerns (Currie and MacLeod, 2008; Frakes, 2013), group practice size (Gaynor et al., 2004) and the degree of team work among physicians (Chan, 2016).

While we do not formally measure the quality of care, our work connects with a literature that examines the workload-quality tradeoff in various work environments. Perdikaki et al. (2012) study the relationship between store traffic, labor, and sales performance and find that the conversion of incoming traffic into sales declines with shoppers' traffic. Conversely, Tan and Netessine (2014) study restaurants and find that higher workload is associated with higher effort, higher sales and lower labor costs. Chatain and Eizenberg (2015) study a legal service provider and find that service quality is increasing in the available capacity of relevant personnel.

Health economists have paid particular attention to the role of workload in hospitals. Studying an emergency unit, Batt and Terwiesch (2012) find that workload induces a service slowdown and that providers adjust their clinical behavior to accelerate the service. Kc and Terwiesch (2009) show that the system load increases the service rate and results in lower quality of care. Kim et al. (2016) find that congestion has a significant impact on admission decisions and patient outcomes in intensive care units. Powell et al. (2012) find that physician workload reduces the share of "severe" patients and, consequently, hospital reimbursement.

We study the effect of workload in the context of primary care. Unlike the ICU or ER contexts, one cannot use a simple indicator, such as the mortality rate, to measure service quality.[6] Given this difficulty, we do not attempt to measure service quality, and instead identify the concrete channels via which workload affects primary care physician's decisions.

problem. Chang et al. (2015) attribute this growth, in part, to lengthy wait times for primary care appointments. Our results suggest primary care physicians are capable of addressing acute issues with their regular patients even under a tight time constraint. The centers employ family medicine physicians (as well as emergency medicine experts) and therefore compete, to some extent, over the same limited supply of such physicians.

[6]The health economics literature has attempted to measure quality in a variety of fashions to overcome this issue. Some examples include Fleitas (2018) who uses test scores from graduate school medical specialty admissions to measure physician quality, and Brunt et al. (2018) who measure the quality of treatment using data from the Medicare's Physician Compare Quality Reporting System.

Some recent work that is related to our analysis include Neprash (2016) and Freedman et al. (2021) who examine the impact of schedule disruptions on physician actions. Our approach differs from that pursued in these articles in three ways. First, we explicitly study the impact of the overall daily level of workload on treatment choices at the individual visit level — as opposed to studying the impact of visit-level schedule disruptions on visit-level outcomes. Our approach is therefore complementary to that of these other articles in that it focuses on the impact of tightening capacity constraints. Second, our identification strategy is different, relying on absence of colleagues as a shifter of the physician's daily workload level. Finally, we employ both point- and partial-identification strategies, allowing us to account for possible violations of our exclusion restriction.

Despite these differences, and the fact that neither the outcomes nor the treatments considered by these articles coincide exactly with our definitions, important aspects of their findings are in line with ours. Freedman et al. (2021) report that increased time pressure causes physicians to limit the number of topics with which they deal during a visit and reduce recommended preventive care. Neprash (2016) shows that in response to schedule disruptions, physicians shorten the visit's duration, perform fewer procedures, increase referrals of a new patient to a specialist and prescribe more antibiotics and painkillers. Some differences also obtain: for example, we do not find that workload significantly increases the likelihood of a return visit.

A growing empirical literature places nonparametric bounds on the Average Treatment Effect following Manski and Pepper (2000) in various areas including health economics (Gerfin and Schellhorn, 2006; Bhattacharya et al., 2012), the economics of education (Gonzalez, 2005; De Haan, 2011; De Haan and Leuven, 2016), public economics (Kreider et al., 2012), and network effects in social media (Shriver et al., 2013). An alternative approach for placing bounds on treatment effects is provided in Mogstad et al. (2018). Interest in partial identification strategies in applied work has been on the rise, as surveyed by Ho and Rosen (2015).

The remainder of the article is structured as follows. Section 2 describes the data. Section 3 describes our simple analytical framework, and the empirical strategy motivated by it. Section 4 reports our results, and section 5 concludes.

2 Data and environment

We use detailed administrative data that cover all primary care visits in eleven clinics, all in the Jerusalem area, of Clalit Health Services — the largest of four HMOs that deliver most of the country's primary care — in the period 2011-2014.[7] Primary care physicians are (typically) directly employed by this HMO and patients are enrolled with a regular primary care physician. They schedule ten minute appointments, online or by phone, at the fixed price of zero. Appointments are scheduled until the schedule is full. Patients with urgent medical

[7]The Israeli primary care system offers universal coverage and is largely publicly-funded. Israeli residents can freely choose an HMO, and HMOs compete, in part, by striving to improve the quality of care.

issues may drop-in or call, asking to see a physician even if the schedule is full. The clinic's office would then try to fit them in, so long as it is feasible. The number of patients seen by physicians at the clinic therefore fluctuates between high and low demand states.

Physician work times at the clinic are fixed and wiggle room around them is strongly limited. Discussions with physicians suggest that they tend to stick to their regular hours rather than exceed them. One of the reasons for this is that the clinic often closes down at the end of the shift. As a consequence, a rather fixed physician time capacity interacts with fluctuating levels of demand, absent an ability to regulate demand via price adjustments. This gives rise to our research question: how do physicians trade off their time with patients with other inputs at their disposal?

Normally, a primary care visit is scheduled with the patient's regular physician. If patients need urgent care, outside their physician's office hours or in her absence, they are typically referred to another physician at the clinic. We restrict our sample to visits in which physicians encounter their regular patients, on weekdays when they see at least twelve of those. We thus focus on physicians' interaction with regular patients in a typical day.[8]

At the visit level we observe the patient's regular physician identity, the identity of the attending physician, and the visit actual start time — as opposed to the visit's scheduled start time. Visit length is calculated as the elapsed time between the visit start time, and the start time of the subsequent visit. Also observed are patient characteristics: gender, age, country of origin and the presence of any of 113 chronic conditions. Importantly, a detailed description of the visit's outcomes is recorded including diagnoses, prescriptions, and referrals to specialists, laboratory tests, and imaging.

The final sample contains 823,349 office visits made by 78,959 patients to 93 physicians. Table 1 provides descriptive statistics regarding these face-to-face visits. The mean patient age is 47.6, and 58 percent of visits are by women. Thirty percent of the visiting patients are smokers and 26 percent are obese. Hypertension, hyperlipidemia and ischemic heart disease characterizes 34, 45 and 15 percent of visiting patients, respectively. A substantial mass of office visits is, therefore, generated by patients exhibiting chronic medical conditions.

Office visits last 11.56 minutes on average. Diagnostic tools are used quite frequently: a referral to a specialist, to imaging or to a lab test is administered in 14, 8 and 20 percent of visits, respectively. Visits may also result in more immediate treatment outcomes: patients are referred to the emergency room, prescribed antibiotics, or prescribed painkillers in 1 percent, 10 percent, and 5 percent of visits, respectively.

Physician workload. Our empirical exercise focuses on the relationship between a visit-level outcome (e.g., an indicator for a referral to a lab test) and a measure of the physician's daily level of workload. A common measure of workload in the primary care setting is the number of patients seen per hour or, equivalently, the average visit length (see e.g. Hobbs et al.

[8]In Israel, weekdays are Sunday through Thursday, thus we drop Friday and Saturday visits. The data include about 1,090,000 weekday office visits.

(2016)). We therefore define the main explanatory variable *workload* as the daily average of a physician's visit lengths.[9] For example, the workload measure equals twelve minutes per-patient for a physician who works for two hours, during which she sees ten patients.

Figure 1 displays the distribution of the workload measure, featuring considerable variation around the mean of 11.5 minutes per patient.[10] The 10th percentile of the workload distribution is 7.6 minutes, and more than doubles to 16 minutes at the 90th percentile.

Absent colleagues. The task of studying the relationship between visit-level outcomes and the daily workload measures is fraught with endogeneity and simultaneity challenges: workload may be shifted by unobserved factors that also affect the distribution of medical conditions presented at the clinic on a given day. To address this challenge we exploit exogenous variation in physician workload generated by the absence of colleagues at the clinic.

When a colleague is absent, her patients are referred to other physicians at the clinic, affecting their workload. Physicians have fixed shifts at the clinic, and we build on this regularity to define a physician as *absent* on a given day if two conditions are satisfied. First, the physician treats zero of her (positive number of) patients, namely, the physician is not present at the clinic, yet some of her patients do arrive to seek treatment. Second, the physician has worked (and has seen at least 5 of her patients) in the two weeks before and after the relevant day, on the same weekday. The observation therefore pertains to a weekday in which the physician normally works at the clinic.

Having defined physicians' days of absence, we calculate, for each physician who is present at the clinic on a given day, a proxy for the added workload brought about by absent colleagues: the share of the absent colleague's patients out of the total number of patients seen by the physician on that day. For example, a physician who saw 7 regular patients and 3 patients of an absent colleague obtains a 3/10=0.3 value for this proxy variable.[11] We refer to this proxy as the share of an absent physician's patients, and use it to instrument for physician workload. We also define an extensive margin instrumental variable: an indicator taking the value one for physicians who see any patients of absent colleague on the given day, and zero otherwise.

3 Empirical strategy

We seek to identify the causal effect of physician workload on visit-level outcomes such as the use of diagnostics, or the prescription of medication. To address this challenge, we begin in section 3.1 by introducing a simple analytical framework describing the economics of physician workload.

The analytical framework serves two purposes. First, the causal effects of physician workload

[9] The length of the last visit of the day cannot be calculated, and is omitted from the calculation of *workload*.

[10] As the 99th percentile of the distribution is 21.5, about 2000 observations with a visit length in excess of 25 minutes were excluded from this figure for illustrative purposes.

[11] We exclude patients who are not regular patients of either the physician or her missing colleague.

are captured within this framework via comparative statics. This helps us provide an economic interpretation for the objects we end up estimating. Specifically, these objects capture the degree of substitutability or complementarity of physician's time with other inputs within a production function framework. Second, the framework helps motivate our econometric strategies.

To that end, section 3.2 next describes the basic econometric framework and presents alternative econometric restrictions. Those include a standard exclusion restriction, but also weaker restrictions that do not require full exclusion to hold, such as the Monotone Instrumental Variable (MIV) restriction from Manski and Pepper (2000). We then present a mapping between these alternative econometric restrictions, and various assumptions on the Data Generating Process (DGP) within the analytical framework described above. This clarifies what assumptions on the DGP must hold to justify the use of the various estimators we employ.

Finally, section 3.3 describes some practical implementation aspects of the empirical strategies. In particular, the implementation of the strategies involving partial identification is nonparametric in nature, and we describe it in detail.

3.1 The economics of physician workload: an analytical framework

Indexing days of work at the clinic by ℓ, we assume that the physician's day unfolds as follows:

1. $N_\ell \geq 0$ of the physician's regular patients report to the clinic. Also (potentially) reporting are $A_\ell \geq 0$ patients of an absent colleague. The expected average visit length at this point is therefore $C_\ell/(N_\ell + A_\ell)$, where $C_\ell > 0$ is the physician's daily time capacity.

2. Some of the physician's regular patients are deterred by her expected workload and leave. The number of such patients is denoted N_ℓ^D where $0 \leq N_\ell^D < N_\ell$.[12] The average visit length is updated to $C_\ell/(N_\ell + A_\ell - N_\ell^D)$.

3. The physician meets with each of the $(N_\ell + A_\ell - N_\ell^D)$ patients and makes clinical decisions.

Both the potential, and the actual numbers of patients to be seen by the physician on day ℓ are therefore determined once at the beginning of the day. In reality, this process may involve dynamic features with patients showing up, calling in or giving up their slot during the day. Nonetheless, the simple timeline captures the key features of the setup: a fixed capacity of physician time, fluctuating demand levels, and the self-selection of patients into seeing the physician on a given day.

The number of the physician's regular patients showing up on day ℓ (before attrition), and the number of an absent colleague's patients are stochastically determined: $N_\ell = \overline{N} + \phi_\ell$, $A_\ell = \overline{A} + \eta_\ell$, where $\overline{N}$ and $\overline{A}$ are long-run averages of N_ℓ and A_ℓ, respectively, and (ϕ_ℓ, η_ℓ) are

[12] Patients of the absent colleague are assumed not to be deterred, as they likely seek urgent attention.

shocks. Assumptions on potential correlations of these shocks with the distribution of medical conditions presented by the physician's regular patients has a central role in our analysis.

Patient types. We index the physician's regular patients who show up on day ℓ by $i = 1, ..., N_\ell$. Each such patient is characterized by a type, $\theta_{i\ell} \in [\underline{\theta}, \overline{\theta}]$, capturing the positive utility she would gain by seeing the physician today. The patient's type summarizes both the importance of the medical issue in question, and its urgency.[13] Let $\Theta_\ell = \{\theta_{1\ell}, \theta_{2\ell}, ..., \theta_{N_\ell \ell}\}$ denote the set of types for all the physician's own-patients who show up on day ℓ (before attrition), and $\overline{\Theta}_\ell$ denote the mean of these values.

Correlations. Several important scenarios can be captured via assumptions on the correlation between the mean patient type $\overline{\Theta}_\ell$ and the shocks that determine patient volume. Concretely, we consider the possibility that the mean type $\overline{\Theta}_\ell$ is positively correlated with the shock to the volume of the physician's regular patients who show up, ϕ_ℓ. Such positive correlation captures the possibility that local infections increase both the volume of patients seen at the clinic, and the intensity with which they require inputs such as medication and diagnostic tools. This possibility is a key driver of endogeneity in our empirical analysis.

Another important possibility is that local infections also affect the health of colleagues. In this case, $\overline{\Theta}_\ell$ is positively correlated with the shock to the volume of an absent colleague's patients, η_ℓ. This scenario poses a threat to our exclusion restriction. We study below the implication of such correlations for our empirical strategies.

Patient attrition. The number of deterred patients, N_ℓ^D, is endogenously determined. A regular patient i incurs a cost for seeing her physician today, given by $g(f_\ell)$, where f_ℓ is the expected face-time with the physician, and $g(\cdot)$ is a function satisfying $g' < 0$, i.e., patient utility is penalized by the physician's workload.[14] The pre-attrition expected face-time satisfies $f_\ell = C_\ell/(N_\ell + A_\ell)$ as described above.

Patient i therefore obtains a utility of $\theta_{i\ell} - g(f_\ell)$ if she sees the physician today, and a normalized utility of zero if she gives up her slot, capturing the option value of rescheduling. It follows that a cutoff patient type characterizes the attrition, i.e., deterred patients present relatively mild medical issues:

$$N_\ell^D = \#\Big\{ i \in \{1, ..., N_\ell\} : \theta_{i\ell} < g(f_\ell) \Big\}. \tag{1}$$

This mechanism introduces another unobserved channel via which the distribution of medical conditions presented by the physician's visiting patients could be related to her workload. As described below, this poses yet another identification challenge with respect to the causal effects of workload.

[13] For example, $\theta_{i\ell}$ could be low for patients with severe conditions if those are chronic rather than urgent.

[14] This penalty captures both an expected decline in the quality of care, and the likelihood of a long wait time, that are characteristic of attempting to see the physician on a high-workload day.

The physician's problem. Patient i's wellbeing is a function of three inputs: the face-time spent with her physician, denoted F_i, treatment (e.g., prescribing medication) denoted T_i, and diagnostic tools (e.g., a referral to a specialist) denoted D_i. The wellbeing function, which depends also on the patient type θ, is therefore $W(F_i, T_i, D_i; \theta)$.

The physician produces patient wellbeing taking into account the input costs. She therefore determines the optimal allocation (F_i^*, T_i^*, D_i^*) for each patient i to maximize:

$$W(F_i, T_i, D_i; \theta) - c_{F,\ell} \cdot F_i - c_D \cdot D_i - c_T \cdot T_i, \tag{2}$$

where $c_{F,\ell}, c_D$ and c_T are the physician's costs of utilizing time, diagnostics and treatment resources, respectively. The physician's time constraint is implicitly reflected in the shadow cost $c_{F,\ell}$: an additional minute spent with patient i implies one minute less spent with other patients today.[15] The other costs, c_T and c_D, can be viewed as reflecting the HMO's monetary costs of prescribing medication or referring to specialists, translated into constraints faced by the physician. Unlike the cost of face-time $c_{F,\ell}$, these other costs are not modeled as varying across days and so are not indexed by ℓ.

Comparative statics and their empirical relevance. In our empirical analysis we proxy for the cost of face-time $c_{F,\ell}$ using the daily average visit length $C_\ell/(N_\ell + A_\ell - N_\ell^D)$. Workload is declining in average visit length as shorter average visit length implies higher workload. Thus, $c_{F,\ell}$ is the negative of daily average visit length and is defined as our workload measure. This is the actual workload, taking into account patient attrition. To clarify: we do not observe patient attrition, but we do observe all visits actually taking place. We compute the average visit length by simply averaging the observed visit lengths at the physician-clinic-day level.

The goal of the empirical analysis is to identify the sign and magnitude of two objects:

$$\frac{\partial D_i^*}{\partial c_F}, \ \frac{\partial T_i^*}{\partial c_F}.$$

That is, we study the causal effect of a tightening time constraint on the physician's use of diagnostic tools and treatment inputs. Each of those derivatives are positive (negative) if and only if the time spent with a patient is a substitute (complement) to the relevant input.[16]

The simple model outlined above provides an economic interpretation for these causal effects. The model clarifies that a positive (negative) effect implies substitution (complementarity) between the relevant resource and physician time. The framework is also helpful in determining appropriate methods for the consistent estimation of those objects as discussed

[15] A more complicated model would have the physician maximize the (weighted) sum of patients' wellbeing measures subject to a time constraint. Deciding how much time to spend with patient i would then require predictions regarding the medical issues of patients seen later in the day. In our simplified setup the maximization is solved separately with respect to each patient, and those problems are linked via the shadow cost $c_{F,\ell}$.

[16] In the sense of the Hotelling/Lau gross elasticity of substitution (HLES), see e.g., Stern (2011).

next.

3.2 Econometric framework

Consider the following regression model:

$$y = \alpha + \beta \cdot workload + x \cdot \gamma + \psi, \tag{3}$$

where y is a binary visit-level outcome (e.g., an indicator for a referral to a lab test), and *workload* is the negative of the daily average of a physician's visit lengths. Per the analytical framework above, it is equal to minus $C_\ell/(N_\ell + A_\ell - N_\ell^D)$. The regression also contains x, a rich set of controls, and ψ, an error term.

The parameter β is the derivative of the outcome—the probability with which diagnostic or treatment resources are deployed—with respect to workload. This is the effect of tightening the time constraint on the physician's behavior ($\frac{\partial D_i^*}{\partial c_F}$ or $\frac{\partial T_i^*}{\partial c_F}$ in the analytical framework). Since the variable we use in the analysis—the daily average of a physician's visit lengths—is the negative of the physician's workload, the sign of β is negative (positive) if the relevant visit-level outcome y increases (decreases) with workload.

This regression involves considerable identification challenges. Those originate in several channels via which physician workload may be correlated with unobserved shifters of patients' medical conditions, as described by the analytical framework presented above. We next elaborate on an array of econometric restrictions with which we tackle the identification challenge.

We discuss, in turn, four such econometric restrictions: *exogenous workload*, *monotone Treatment Selection (MTS)*, *Instrumental Variable (IV)*, and *Monotone Instrumental Variable (MIV)*. While all restrictions are discussed intuitively here in terms of the linear regression model in (3), the implementation of some of them is nonparametric. We elaborate on the nonparametric implementation in section 3.3 below.

1. *Exogenous workload*. This restriction requires workload to be uncorrelated with the error term ψ in (3) enabling consistent estimation of β via OLS. Our analytical framework, however, contained two features that render this restriction unlikely to hold, as they suggest a likely positive correlation between *workload* and ψ.

First, one may expect the day-ℓ mean patient type, $\overline{\Theta}_\ell$, to be positively correlated with the shocks that determine patient volume ϕ_ℓ due to local infections that generate higher than usual physician workload and a sicker than usual patient pool. Second, high workload may cause some patients to give up their appointment. The analytical framework showed that such patient attrition (denoted there by $N_\ell^D > 0$) will be concentrated among patients with mild conditions. Within the econometric model, this suggests — again — positive correlation between the *workload* measure and the error term ψ.

2. *Monotone treatment selection (MTS).* In light of the above mechanisms, this econometric restriction allows *workload* to be correlated with the econometric error term ψ in (3), but restricts this correlation to be positive. Again, a sicker population of visiting patients would be associated with higher realized workload either due to an omitted variable issue ($COV(\phi_\ell, \overline{\Theta}_\ell) > 0$), via attrition of the physician's original patients presenting mild conditions ($N_\ell^D > 0$), or both. Such a sicker population also has a higher probability of requiring various treatment and diagnostic inputs regardless of the physician's workload level, resulting in larger values of the econometric error term ψ.

The *MTS* restriction provides partially-identifying information on the causal effect of physicians' workload by exploiting this positive correlation. In practice, we implement this restriction nonparametrically rather than via the parametric regression model in (3), as discussed below.

3. *Instrumental variable (IV).* This restriction draws on the absence of colleagues at the clinic as a source of exogenous variation in the physician's workload. It imposes an exclusion restriction: the instrumental variable (i.e., the share of patients seen today that are generated by the absence of a colleague) is allowed to affect clinical decisions only via its effect on the *workload* variable. This restriction imposes a correlation of zero between the instrument and the error term ψ in (3).

We provide below evidence that absences are strongly correlated with the endogenous workload variable, ruling out weak IV concerns. The exclusion restriction, however, is not testable and may be challenged by two channels captured by the analytical framework.

First, a local infection may affect both patient health outcomes and the absence of colleagues. This possibility is modeled in the analytical framework as a positive correlation between absent colleague's patients volume η_ℓ and the mean patient type $\overline{\Theta}_\ell$. Second, a colleague's absence will cause longer wait times, prompting patients with mild conditions to give up their slot, thus affecting the unobserved true distribution of patient health conditions presented to the physician via the attrition channel ($N_\ell^D > 0$).

Those challenges to the exclusion restriction, nonetheless, motivate an additional econometric restriction: Monotone Instrumental Variable.

4. *Monotone instrumental variable (MIV).* The *MIV* restriction relaxes the exogeneity of the instrument: it allows for its correlation with the error term ψ in (3), but restricts the sign of the correlation to be positive.

Unlike the full exclusion embedded in the *IV* restriction, the weaker *MIV* is consistent with the two channels discussed above. It is consistent with a positive correlation between an absent colleague's patients volume η_ℓ and the mean patient type $\overline{\Theta}_\ell$, and with the possibility that the extra workload generated by a colleague's absence prompts attrition by patients with mild conditions.

Both of these channels suggest that the instrument is positively correlated with the error term in (3) in the sense of being associated with a sicker pool of regular patients and hence with a higher probability of administering medical treatment and employment of diagnostic tools.

As in the case of *MTS*, the *MIV* restriction shall also be taken to data within a nonparametric framework, and was discussed here in terms of a parametric model for expositional purposes only.

Connecting the econometric restrictions to the analytical framework: a summary. We have thus far described various econometric restrictions, and their relationship to the economics of physician workload as captured in the analytical framework outlined in section 3.1. Table 2 summarizes this discussion. It charts a map between assumptions regarding the underlying DGP, and the econometric strategies that can be justified based on such assumptions.

The left column of Table 2 lists assumptions regarding the DGP, focusing on the relationship between the shocks ϕ_ℓ, η_ℓ that introduce variation in the physician's daily workload, and the daily set of medical conditions presented by this physician's regular patients, Θ_ℓ. The other columns list *econometric restrictions* that, under the corresponding assumptions on the DGP, deliver consistent estimation of the effects of interest. The table lists a total of seven cases involving different assumptions on the DGP. In cases (1)-(5), the patient attrition mechanism is shut down, i.e. $N_\ell^D \equiv 0$. In cases 6-7, attrition is allowed.

In case 1, the errors that shift the daily number of regular patients, N_ℓ, and the number of patients contributed by an absent colleague, A_ℓ, are completely independent from the set of medical conditions presented by the physician's regular patients, Θ_ℓ. The empirical workload measure $C_\ell/(N_\ell + A_\ell)$ is therefore exogenously determined. As a consequence, OLS estimation will be consistent. All other restrictions (*IV, MIV, MTS*) are also valid in this case.

Case 1 provides a useful, though implausible benchmark. In particular, it rules out the possibility that local infections increase physician workload while also changing the distribution of medical conditions presented to the physician. Case 2 allows for this possibility by assuming positive correlation between ϕ_ℓ and $\overline{\Theta}_\ell$. Higher physician workload may now be associated with the regular patients seen by the physician being sicker, increasing the likelihood of prescribing medication or using diagnostic inputs. While the OLS estimator is no longer consistent, an estimator derived from the *MTS* restriction remains consistent in this case. Furthermore, since $\eta_\ell \perp \Theta_\ell$ continues to hold, the instrument remains exogenous and so the *MTS* restriction may be combined with either the *IV* or the *MIV* restrictions.

Case 3 weakens the assumptions further: not only does it allow ϕ_ℓ to be correlated with Θ_ℓ, but it allows the correlation to be either positive or negative. The *MTS* restriction is then no longer valid, but since we maintain that $\eta_\ell \perp \Theta_\ell$, the *IV* and *MIV* restrictions still are.

Case 4 asserts that both shocks, ϕ_ℓ and η_ℓ, are positively correlated with the mean patient type. It allows a local infection to affect not only the health of the physician's regular patients, but to also result in a higher incidence of absent colleagues. The *IV* restriction becomes invalid, but both *MTS* and *MIV* continue to hold. Case 5 weakens the assumptions of case 4 by leaving the correlation of η_ℓ and Θ_ℓ unrestricted. This invalidates *MTS*, but *MIV* still holds.

Cases 6 and 7 introduce patient attrition via equation (1) of our analytical framework.

This creates additional identification challenges: even if the volume of the physician's regular patients who initially show up is independent of their health status, high physician workload would deter the ones exhibiting mild conditions, once again generating positive correlation between the physician's workload and the econometric error in (3). The *Exogenous workload* restriction therefore fails, and so does the *IV* restriction: the absence of colleagues increases the physician's workload, but now, via the attrition mechanism, it is also associated with a sicker group of the physician's regular patients with whom she meets.

In both cases 6 and 7, however, we maintain that the absence of colleagues is non-negatively correlated with sicker regular patients to begin with. The deterrence mechanism simply reinforces a non-negative correlation between the absence of colleagues and the error term in (3), validating the *MIV* restriction. In case 6, we also assume $COV(\phi_\ell, \overline{\Theta}_\ell) \geq 0$, and so the *MTS* restriction also holds. In case 7, this latter correlation is unrestricted, so that, regardless of attrition, *MTS* fails and only the *MIV* restriction remains valid.

To recap the exercise, we have sketched a theoretical framework with the purpose of guiding our empirical work. This model connected our estimated objects to the substitutability or complementarity of physician time and other inputs such as diagnostic tools and prescriptions. Furthermore, it motivated a range of econometric restrictions that deliver consistent estimates given a range of alternative assumptions on the underlying DGP.

The analytical framework was kept simple on purpose. For example, we do not model the precise mechanism that generates physician absences, instead we model them as stemming from a simple stochastic process. Nonetheless, by discussing correlations of that process with other factors, we are able to capture a variety of mechanisms and their implications for our identification strategies. For example, we do not formally discuss whether the absence of colleagues is known in advance. If it is known in advance, and still many patients of the absent physician show up, it could indicate an infection affecting the local community, resulting in a positive correlation between η and $\overline{\Theta}_\ell$, motivating the *MIV* estimator. If it is not known in advance, it could indicate that the physician herself (or her children) are sick, again potentially indicating a local infection and once again suggesting the use of *MIV*.

Finally, a concern arises from the possibility for non-random assignment of an absent colleague's patients among non-missing physicians. For example, it is possible that the clinic manager uses her familiarity with physicians to divert extra patients towards those physicians who are likely to perform well under pressure.[17] We address this concern by including physician fixed effects in the 2SLS analysis, and also report, in Appendix A.6, results using a clinic-level version of this instrument: namely, an indicator taking the value one given any absence at the clinic, and zero otherwise. Reassuringly, this generates qualitatively similar results.

[17] An intricate set of incentives affects the decision of the clinic manager to allocate such additional patients among physicians. Physicians receive a monetary compensation of about 7 dollars per such visit. Some physicians may be more reluctant to accept extra patients, or have more bargaining power than others.

3.3 Practical implementation of the econometric restrictions

Under the *exogenous workload* restriction, we estimate the linear regression model in (3) via OLS. Implementation of the *IV* restriction applies the standard 2sls estimator to the same linear regression, resting on the premise that the instrument is uncorrelated with the error ψ and yet correlated with the endogenous *workload* variable.

We further estimate models relying on the *MIV* restriction, and on a combination of that restriction with *MTS*. This is performed within a nonparametric framework following Manski and Pepper (2000, MP2000). We next provide notation for that framework and formally introduce the *MTS* and *MIV* within that nonparametric framework following MP2000.

A nonparametric regression model. The population of interest contains a set $j \in \mathcal{J}$ of individuals characterized by heterogeneous *response functions* $y_j(\cdot) : T \to Y$, where $t \in T$ are discrete treatments, and $y \in Y$ are discrete outcomes.

In our application the population $\mathcal{J}$ contains all visits (patient-physician-day triplets). The outcome space is binary, i.e., $Y = \{0, 1\}$. For example, a patient is either referred to some diagnostic procedure, or not. We define a binary workload treatment, so $T = \{0, 1\}$. A value of $t = 1$ implies that the physician's workload exceeds the 75th percentile of the physician-specific workload distribution, and $t = 0$ implies workload below or at the 75th percentile.[18]

The observed variables are (x, z, y), where x_j is a covariate vector for visit j. The variable $z_j \in T$ is the *realized treatment*. That is, it takes the value 1 for visits that take place on a day when the physician is *observed* to experience high workload, and zero otherwise. Finally, $y_j = y_j(z_j)$ is the observed outcome.

The covariate vector is $x = (w, \nu) \in \mathcal{X} = W \times V$, where $\nu \in V$ is our instrument: the daily share of patients seen by the physician that is generated by the absence of a colleague. In the baseline linear model, the instrument is treated as a continuous variable. In the nonparametric analysis, we treat the space of instrument values V as discrete.[19] Our object of interest is the distribution $\mathcal{P}[y(\cdot)]$ of response functions, or, rather, its conditional version $\mathcal{P}[y(\cdot)|w]$. Given the binary outcome, the Average Treatment Effect is:

$$ATE(1, 0|w) = P[y(1)|w] - P[y(0)|w] \tag{4}$$

We estimate (bounds on) the treatment effect of physician workload on the probability with which the physician uses specific inputs. Relying on our theoretical framework, we interpret a negative (positive) ATE as reflecting complementarity (substitution) between physician time and the input in question. We consider identification of this treatment effect follows from the *IV, MIV* and *MTS* restrictions. Those restrictions were discussed intuitively within the

[18] In the baseline linear model the treatment t is a continuous variable: the physician's workload, measured by the daily average visit length.

[19] We divide it into 10 bins ranging from 0 to 0.4, with a fixed width of 0.04. Days in which more than 40 percent of the patients seen by the physician are due to an absent colleague are rare and not likely to be representative, and are dropped from the analysis.

parametric regression model in (3) in the previous section. We now formally define those restrictions within the nonparametric framework.

Assumption 1 ***IV:***

$$E\big[y(t)|w,\nu=u'\big]=E\big[y(t)|w,\nu=u\big]\ \forall t\in T, w\in W,(u,u')\in V\times V$$

In words, the *IV* assumption does not allow the instrument to affect the response function. It can only affect the outcome via its effect on the treatment t, delivering the classic exclusion restriction. In contrast, the weaker Monotone Instrumental Variable assumption allows the response function to be shifted by the instrument in a pre-specified direction:

Assumption 2 ***MIV:***

$$E\big[y(t)|w,\nu=u_2\big]\geq E\big[y(t)|w,\nu=u_1\big]\ \forall t\in T, w\in W,(u_1,u_2)\in V\times V\ \textit{such that}\ u_2\geq u_1$$

In our context, the *MIV* restriction allows the absence of a colleague to be associated with higher use of an input (e.g., a diagnostic input) *conditional on the physician's workload level*. Finally, the Monotone Treatment Selection (*MTS*) restriction allows the realized treatment z to serve as an additional monotone instrumental variable:

Assumption 3 ***MTS:***

$$E\big[y(t)|w,\nu=u,z=1\big]\geq E\big[y(t)|w,\nu=u,z=0\big]\ \forall t\in T, w\in W,u\in V$$

MTS implies that the response function shifts upward, at any treatment level (either $t=1$ or $t=0$), if the *realized* treatment is that of high workload ($z=1$).

Appendix B provides the complete derivation of the nonparametric estimators employing the *IV*, *MIV*, and a combined *MIV-MTS* restrictions, and the final form of these estimators. With the exception of *MIV-MTS*, the derivations follow Manski and Pepper (2000) exactly.[20]

As that appendix shows, these restrictions result in *set-identification* of the Average Treatment Effect of physician workload on her use of various inputs. Implementing this approach therefore allows us to relax the strict exclusion restriction, yet results in a loss of point identification.

[20]In contrast to Manski and Pepper (2000), the MTS restriction employed here conditions on the instrument ν (the share of patients seen by the physician that are contributed by an absent colleague), in addition to the covariate vector w. Manski and Pepper (2000) do not use the MTS restriction in conjunction with an additional instrumental variable, and so the conditioning on ν does not arise there. Our approach is consistent with that of De Haan and Leuven (2016) who also combine *MTS* with an additional MIV assumption.

4 Results

We present the results in three stages. First, section 4.1 describes the "first stage": the source of variation employed in the instrumental variable approach. Section 4.2 proceeds by presenting OLS and 2SLS estimates of the linear regression model in (3). The OLS analysis is presented as a benchmark only, as it is only justified under very strong assumptions (case 1 of Table 2). The 2SLS results are justified under the weaker assumptions of cases 2 and 3. Those weaker assumptions still require a full exclusion restriction, and, in particular, cannot be justified if the patient attrition mechanism is present.

We present below evidence that the attrition mechanism is not likely to be quantitatively important. Just the same, with or without attrition, full exclusion could still be violated if the extra workload caused by an absent colleague is correlated with the distribution of medical conditions presented by the physician's regular patients (per cases 4 through 7).

In section 4.3 we present nonparametric bounds under the weaker *MIV* restriction which holds in all seven cases summarized in Table 2. We tighten these bounds via the *MTS* restriction to obtain the *MIV-MTS* estimator. As discussed above, the *MTS* is justified if the shocks to the volume of the physician's regular visiting patients are non-negatively correlated with the acuteness of their medical conditions. This non-negative correlation is allowed under cases 1, 2, 4 and 6 in Table 2.

Finally, in section 4.4 we take stock of what we have learned from these various analyses regarding the questions of interest.

4.1 The first stage

Figure 2 provides a graphical illustration of the relationship between the instrument (the share of absent colleagues' patients out of the physician's total daily patient count) and the physician's daily workload measure (average visit length). The figure classifies the instrument into 0.025 percentage point bins and the workload measure is calculated for each bin. We regress the workload measure on the instrument, and use the solid line to display the relationship predicted by this regression. The figure shows that a ten percentage points increase in the value of the instrument is associated with a decrease of about one minute in the physician's average visit length — a sizable effect, given that the average visit length is about 11.6 minutes.[21]

To further illustrate the effect of absent colleagues on the physician's daily workload we aggregate the data to the physician-day level and perform an event study analysis. Let D_{st} be an indicator that takes the value one when at least one physician is absent from clinic s at time t, and zero otherwise. Suppose for example that a physician was absent on January 5th, 2013 at clinic 5, then $D_{5,1/5/2013} = 1$. Next, define τ_{st}, the *event relative time*, as the number

[21] This association does not arise from cross sectional differences between physicians. In Figure A.1 we show nonparametric estimates of the first stage using dummy variables for the same bins of our instrument as in Figure 2 and controlling for physician and time fixed effects. The results are almost identical to Figure 2.

of days that elapsed since the absence. In this example $\tau_{5,1/5/2013} = 0$, $\tau_{5,1/4/2013} = -1$ and $\tau_{5,1/7/2013} = 2$. Indexing physicians by j we estimate the following regression:

$$workload_{jst} = \alpha + \nu_j + \nu_t + \gamma_1 \cdot \tau_{-k} + ... + \gamma_{k+1} \cdot \tau_0 + \gamma_{k+2} \cdot \tau_1 + ... + \gamma_{2k+1} \cdot \tau_{k-1} + \epsilon_{jst}, \quad (5)$$

where ν_j are physician fixed effects, and ν_t contains year-month and day of the week fixed effects. The objects of interest are the coefficients on the τ variables.

Our hypothesis is that the coefficient on τ_0 should be negative (recalling that seeing a higher number of patients implies a *lower* average visit length), while the coefficients on days before and after the absence should be zero. Figure 3 displays the estimates of these indicators, ranging from τ_{-7} to τ_6, along with 95 percent confidence intervals. The pattern confirms our hypothesis: in the days before and after the event, the effect of an absence is not statistically different than zero. On the day of the event, the average visit length drops by about a third of a minute, and this effect is statistically significant.

The first stage is formally estimated via the following regression, indexing visits by i:

$$workload_{jsti} = \alpha + \nu_j + \nu_t + sa_{jst} \cdot \beta_1 + x_{jsti} \cdot \beta_2 + \epsilon_{jsti}, \quad (6)$$

where again ν_j and ν_t capture fixed effects for physician, year-month, and day of the week. The variable sa_{jst} is the instrument: the share of an absent physician's patients out of physician j's total count of patients at clinic s on day t. We denote this instrument by $IV1$, and additionally define an extensive margin instrument, IV2, taking the value 1 if $sa_{jst} > 0$, and zero otherwise. The vector x contains visit and patient level characteristics, and ϵ is an error term.[22]

Table 3 displays the first-stage results. Column 1 of panel (a) shows that $IV1$ has a negative and statistically-significant effect on workload with a point estimate of about -4.8. A ten percentage points increase in the instrument is therefore associated with a decrease of 0.48 minutes in average visit length. The hypothesis that the instrument may be excluded from the model is rejected at 99.9% significance, alleviating weak instruments concerns.

Recall that the exclusion restriction may be violated due to potential interaction between the absence of colleagues on a given day, and the composition of medical conditions displayed by the physician's regular patients visiting her on that day. To explore this possibility, the regression reported in column (2) of Panel (a) controls for patient characteristics, and for visits of an administrative nature. If the instrument affects the type of visits or the patient pool, the estimates in this specification would be different from those in column (1). The results demonstrate that adding those patient and visit level controls leaves the first stage results

[22]The patient level characteristics we use are: age, age squared, a gender dummy, and 113 indicators for chronic conditions. We additionally include dummy variables for visits which primary goal is: issue a medical certificate, prescription renewal, filling out forms, or an administrative visit.

virtually unchanged, mitigating such concerns.

Yet another threat to the exclusion restriction is the "deterrence" mechanism: the attrition of patients with mild conditions when their physician's workload is high. To gauge the likely importance of this issue, we examine whether the number of regular patients seen by a physician is lower on a day when a colleague is absent. We run an event study analysis, similar in nature to the one performed above. The dependent variable here is the number of the physician's regular patients seen per hour.[23] Figure 4 shows that there appears to be no change in the number of patients a physician sees per hour on days on which a colleague is absent.

Taken together, these results lend support to the exclusion restriction reflected in cases 1-3 of our conceptual framework (Table 2), in that they support the assumption that $N_\ell^D = 0$ as well as that $COV(\eta_\ell, \Theta_\ell) = 0$, and, therefore, our 2SLS analysis. Nonetheless, we report below results based on weaker restrictions such as the *MIV*.

Finally, panel (b) of Table 3 reports the first stage performance of the extensive-margin instrument, IV2, revealing exactly the same patterns demonstrated for IV1. Column (1) shows a point estimate of -0.63, implying that seeing any of the absent physician's patients results in a decrease of 0.63 minutes in average visit length. Similarly as in the case of IV1, adding patient characteristics and visit level controls in column (2) does not change the estimates.

4.2 Linear regression results: OLS and 2SLS

To address our research questions concerning the potential substitutability or complementarity of physician time with other inputs at her disposal we estimate the following version of the model from equation (3):

$$y_{jsti} = \alpha + \nu_j + \nu_t + \beta_1 \cdot workload_{jst} + x_{jsti} \cdot \beta_2 + \epsilon_{jsti}, \tag{7}$$

where, again, j, s, t, i index physicians, clinics, time and visits, respectively. The dependent variable y_{jsti} is an indicator capturing an outcome of interest, e.g., referring a patient to a specialist. Our main explanatory variable is $workload_{jst}$, measuring physician j's average visit length at clinic s, time t. We use IV1 and IV2 to instrument for this endogenous variable in 2SLS regressions.

We analyze the effect of workload on physician choices in face-to-face encounters with patients focusing on two sets of outcomes. The first set of outcomes involves the use of diagnostic inputs. Here, the dependent variables are indicators for the following visit-level outcomes: a referral to a specialist, a referral to imaging (x-ray, ultrasound, CT or MRI) and a referral to lab tests (e.g. blood or urine tests). The second set involves treatment outcomes: indicators for a referral to the emergency room, for a prescription of antibiotics, and for a prescription of

[23] We use patients per-hour, rather than simply counting patients per-day, to account for the fact physicians' shift-lengths may vary across physicians and over time.

painkillers. As the dependent variable is binary and the effects of interest concern probabilities, we multiply the estimates by one hundred.

The effect of workload on diagnostic outcomes. It is not *a-priori* clear whether physician time and diagnostic outcomes are substitutes or complements. When workload is higher, physicians may be able to substitute diagnostic procedures for conversation and physical examination. On the other hand, a tightening time constraint may limit the scope of the medical issues that can be addressed during the visit, resulting in the use of fewer diagnostic procedures. This motivates empirical work of the sign and magnitude of this effect, providing insights into the nature of the shadow cost of physician capacity.

Table 4 displays the diagnostic outcomes analysis. All specifications include year-month and day of the week fixed effects, mitigating the possibility that an omitted factor such as weather conditions affects both the absence of colleagues at the clinic, and the composition of patient medical conditions on a given day. Furthermore, conditional on these fixed effects, it is more likely that workload generated by absent colleagues is uncorrelated with the medical conditions presented by own-patients, supporting the exclusion restriction.[24]

We also control for physician fixed effects, so that we rely on within-physician variation and avoid basing our causal inference on differences in practice style among non-absent physicians. The literature suggests that controlling for such effects is important: in particular, Doyle et al. (2010) show that physicians trained in lower-ranked programs tend to order more diagnostic tests. Our goal is to identify how workload affects the utilization of such tests controlling for any other source of variation.

Panel (A) studies the overall utilization of diagnostic inputs: the dependent variable takes the value 1 if *any* diagnostic input is used, and zero otherwise. The OLS estimate of the effect of the daily average visit length, the workload measure, is reported in column (1) to equal 0.48, indicating that time with patients and the use of diagnostic inputs are complements.[25] Adding patient characteristics (including fixed effects for specific patient chronic conditions) in column (2) decreases the estimates slightly to 0.45.

Columns (3)-(6) implement the IV restriction to address the endogeneity of the workload measure using both the intensive and extensive margin versions of the instrument. The sign of the coefficient on workload continues to be positive in all specifications, reconfirming that the utilization of diagnostics is an increasing function of the time with patients. Using IV1, the estimated coefficient of 1.64 on the average visit length is considerably larger than the OLS coefficients, and is barely affected by adding patient characteristics in column (4). Using IV2 generates even larger coefficients of 1.86 in column (5), which is again not sensitive to controlling for characteristics in column (6).

This is in line with the MTS restriction, whereby high workload days are associated with

[24] In the analytical framework of section 3.2, this potential lack of correlation was described by $cov(\eta_\ell, \Theta_\ell) = 0$.

[25] Recall that since workload is declining in average visit length, a positive β implies that the visit-level outcome decreases with workload.

sicker patients that require more diagnostics, offsetting some of the effect found in the IV analysis.

To assess the quantitative implications, recall that diagnostic inputs are used in 35 percent of visits. The IV1 results therefore imply that a 1 minute decrease in average visit length causes a 4.6 percent decrease in the probability of utilization of diagnostic inputs.

Additional panels of Table 4 examine the impact of workload on the utilization rate of each diagnostic input separately. Panel (B) displays results concerning referrals to specialists. The results are qualitatively similar as those in Panel (A). Both OLS and IV coefficients on the average visit length are positive, and controlling for patient characteristics has a very minimal effect on these estimates. Given the mean of the dependent variable, 0.14, the results using IV1 imply that a 1 minute decrease in average visit length causes a 9 percent decrease in the probability of a referral to a specialist.

Panel (C) repeats the analysis where the outcome is a referral to a lab test, yielding the exact same pattern of results as above. Here too, the OLS estimates are positive, with the IV estimates being even larger. Given that 20 percent of visits result in a referral to a lab test, the results indicate that a 1 minute decrease in average visit length causes a 3.8 percent decrease in the probability of such a referral. Finally, panel (D) reports results concerning referrals to imaging. The OLS estimates in columns (1) and (2) are positive with point estimate of 0.22 and 0.2 respectively. The IV1 and IV2 estimates are also positive, albeit statistically insignificant.

Taken together, these results indicate that a tightening time constraint causes physicians to reduce their utilization of diagnostic inputs. This pattern is not driven by any particular diagnostic input but rather holds for each input separately (except that the results for imaging are statistically insignificant).

To interpret these results, we first note that issuing a referral entails an administrative and professional burden: the physician has to conclude that the referral is warranted, and to type into her computer a detailed note in this regard. A tightening time constraint interferes with the ability to perform this task. Perhaps more importantly, more time with patients allows the physician to explore a broad scope of issues.

Given ample face-time, the physician may take an overall look at the patient's health, suggest routine checkups, and ask about issues beyond the medical complaint that brought the patient into the clinic. Consider, for example, an elderly patient who scheduled the appointment on account of a sore elbow. Given a tight schedule, the physician may quickly attend to the elbow and avoid discussing any other matters. With additional time, the physician can ask about symptoms such as fatigue, low appetite or memory loss, and suggest appropriate tests and examination by specialists. The empirical findings reported above suggest that this preventive care aspect may be restricted as the time constraint tightens.

To further assess this possibility, we examine the effect of workload on the number of diagnoses coded by the physician within the visit. In panel (E) of Table 4 we repeat the same

regression analysis, but with the simple count of coded diagnoses as the dependent variable.[26] Both OLS and IV estimates are positive and statistically significant, indicating that higher workload is associated with a smaller number of diagnoses per visit. These results further support the notion that high workload results in a reduction in the scope of medical issues being addressed within a visit.[27]

The effect of workload on the choice of medical treatment. Having analyzed the effect of workload on the use of diagnostic inputs, we next examine its effect on medical treatment intensity. One may hypothesize that higher workload would drive physicians to be more conservative and provide more treatment: substitute office time and examination with the prescription of antibiotics or painkillers, or referrals to the emergency room. The data, however, lends little support for this hypothesis.

We examine this issue in Table 5, beginning in Panel (A) where the dependent variable is an indicator taking the value 1 if *any* of the medical treatments we consider were used, and zero otherwise. The OLS estimate in column (1) is positive but when we add patient characteristics in column (2), it becomes small and statistically insignificant. The instrumental variable estimates in columns (3)-(6) are all negative. The estimates with IV1 are statistically insignificant with point estimates of about -0.5 percentage points. With IV2 the estimates are larger and statistically significant with point estimates of about -0.8 percentage points, reflecting an increase of about 5 percent in the probability of receiving treatment given a 1 minute decrease in the average visit length.

Unpacking this aggregate effect to consider specific treatment outcomes, Panel (B) reports the impact of the average visit length on the probability of a referral to the emergency room. The OLS estimates in column (1) and (2) are positive and significant. However, the instrumental variable estimates in columns (3)-(6) are all very small and statistically insignificant. Estimates for specifications in which an indicator for the prescription of painkillers serve as the dependent variable are reported in Panel (C). Again, the OLS estimates are positive and significant, while the IV estimates are negative and statistically insignificant.

In Panel (D), the dependent variable is an indicator for the prescription of antibiotics. The OLS estimate in column (1) is -0.03 percentage points, and is statistically insignificant. Controlling for patient characteristics, the estimate, shown in column (2), is statistically significant at -0.04 percentage points. The estimates using IV1, shown in column (3)-(4), are -0.32 yet they are statistically insignificant. Using the second instrument, IV2, the estimate in column (5) is -0.51 and controlling for patient characteristics in column (6) the result becomes statistically significant with a point estimate of -0.55. As the probability of receiving antibiotics is on average 10 percent, This result implies that a 1 minute decrease in the daily mean visit length increases the visit-level probability of receiving antibiotics by 5 percent.

Overall, these results provide little evidence for a tendency to increase the intensity of

[26] Each visit has at least one such diagnosis, but physicians may code additional ones.

[27] Section A.3 of the appendix explores potential heterogeneity across patient populations.

treatment in response to higher physician workload. There appears to be no effect on the incidence of referrals to the emergency room or on the prescription of painkillers, yet there is some (mixed) evidence that increased workload tends to increase the use of antibiotics.[28]

The effect of workload on subsequent encounters and other interactions. We also use our baseline 2sls analysis to examine two additional questions. First: to the extent that a patient has not been able to address important issues during a visit due to physician workload, she may return for a subsequent visit to discuss these additional issues. If this is the case, the system may be said to "correct itself" with respect to some of the limitations brought about by physician workload. As Appendix A.1 shows, however, we find little evidence for an increase in the likelihood of a subsequent visit driven by current-visit physician workload.

Second, we examine how physician workload affects other dimensions of patient-physician interactions: those involving phone calls and physician response to online patient queries. As reported in Appendix A.2, we find that both these activities are reduced by physician workload.

Takeaways from the linear regressions. The 2sls analyses, motivated by the *IV* restriction, indicate that the intensity of treatment is not significantly affected by workload. More intense prescription of medication or painkillers, or referrals to the ER, are neither substitutes nor complements to the primary care physician's time with patients. Instead, physicians respond to a tightening time constraint along other dimensions. They respond less to online queries and perform fewer phone calls with patients.

More importantly, they make fewer diagnoses per visit, and reduce the use of diagnostic tests or referrals to specialists. Diagnostic inputs are therefore complements, rather than substitutes, to physician time. Appendix A.5 provides reduced-form regressions (i.e., where the explanatory variable is the instrument) that further support these findings.

In the next section we go beyond the full exclusion restriction to consider nonparametric bounds estimators relaying on different restrictions, namely, the weaker *MIV* restriction, and the *MTS* restriction. We focus this analysis on the role of diagnostic inputs, since these are the inputs for which the baseline linear analysis above provided the strongest evidence for a workload effect. Put differently, we wish to examine the robustness of our primary conclusion — that diagnostic inputs are complements to physician time — using alternative restrictions.

4.3 Physician workload and diagnostic inputs: nonparametric bounds

We next follow the nonparametric approach of section 3.3 to place bounds on the Average Treatment Effect of physician workload on the probability of employing diagnostic inputs, relying on set-identifying econometric restrictions. The derivation of the set estimators' formulae is provided in Appendix B.1.

Figure 5 provides a graphic illustration of these bounds, while Table 6 provides the estimates. Panel (a) of Figure 5 displays bounds on the probability of using any diagnostic input. The

[28] See Section A.4 of the appendix for analysis of heterogeneity across patient populations.

vertical red and blue lines show the bounds under low-workload and high-workload, respectively. The figure provides these bounds for four sets of econometric restrictions: *IV, MIV* and the *IV-MTS* and *MIV-MTS* combinations.[29] Panel (b) of the figure displays the corresponding bounds on the ATE — the average effect of a switch from low to high workload on the probability of using diagnostic inputs, shown in equation (B4) of Appendix B.1.

Starting with the IV restriction, presented in the most-left part of the figure, the upper bound on the estimated probability given low workload lies above the upper bound given high workload, and the same holds for the respective lower bounds. Nonetheless, it is not possible to sign the treatment effect. This can be seen in Panel (a) of the figure because the intervals on the probability of using diagnostics given high and low workload overlap, and in Panel (b), showing that the estimated interval for the ATE contains the value zero.

The same information is displayed numerically in the first row of Table 6. The estimated set for the ATE is $[-.321, .138]$, with a 95 percent confidence interval of $[-.351, .174]$.[30] The confidence intervals in the IV case are displayed graphically in Panel (b) of Figure 5 and reported numerically in Columns (7)-(8) of Table 6.

Moving to the right, Figure 5 next presents bounds generated under the weaker *MIV* restriction, where the estimator is again provided in Appendix B.1 (equation B6). The same pattern is observed: the intervals on the probability of using diagnostics given low workload are higher than those given under high workload — but they overlap. Therefore, once again, it is not possible to sign the ATE. This is hardly surprising, given that the *MIV* restriction is weaker than the *IV* restriction employed above.[31]

Next, Figure 5 employs the *MTS* restriction in conjunction with the *IV* and the *MIV* restrictions, yielding tighter upper bounds on the probability of using diagnostics given high workload ($t = 1$). As shown in panel (a) of the figure, the estimated intervals for the probability of using diagnostic inputs given low and high workload no longer overlap. Indeed, Table 6 shows that the estimated intervals for the ATE of workload on the probability of using diagnostic inputs now contain only negative values, in both the *IV-MTS* and the *MIV-MTS* cases. For example, under *MIV-MTS*, this interval is $[-.432, -.025]$.

Consistent with the results from the linear models in section 4.2, the *MIV-MTS* bounds

[29] In Appendix B.2 we repeat the same analysis for subsamples that correspond to different conditioning covariates (denoted by w in equation (B3)) matching the heterogeneity specification in the linear analysis reported in Appendix Table A.5.

[30] We follow the common practice (e.g., de Haan and Leuven 2016, Kreider et al. 2012) of relying on Imbens and Manski (2004) to derive the 95% confidence intervals. Those are obtained using 100 bootstrap replications. As our sample size is much larger than that commonly used in such applications, we do not perform the correction for finite sample bias prescribed in these articles. We check the sensitivity of the results to the number of bootstrap replications for the full sample case and find that the results are quite stable between 50 and a 1000 replications, as shown in Appendix Table B.1.

[31] Comparing the IV to the MIV results in Table 6 shows that the MIV lower bound is always smaller than the IV lower bound, as should be expected given that the MIV restriction is weaker. At the same time, the two assumptions deliver the exact same upper bounds. This happens because the smallest "no-assumptions" upper bound is obtained at the largest value of the instrument. A detailed explanation is provided in Appendix B.1.

imply that physician time and diagnostic inputs are complements. As workload intensifies, the probability of using diagnostics declines by 2.5 to 43.2 percent. The 95 percent confidence intervals for this ATE, however, do contain the value of zero: under *MIV-MTS* the confidence interval is $[-.440, .010]$. One cannot, therefore, reject the hypothesis that the average effect of workload on the probability of using diagnostic tools is zero. This is shown graphically in panel (b) of Figure 5: the confidence intervals slightly cross the horizontal red line marking zero.

Nevertheless, the *MIV-MTS* analysis is informative. The ATE is bounded between a substantial negative value, and a very small positive value. The results imply that diagnostics could be very strong *complements* to physician time: switching from low to high workload may reduce the probability of using diagnostics by as much as 44 percent. At the same time, diagnostics cannot be strong *substitutes* for time: at the upper bound, a switch to high workload could increase the probability of using diagnostics by at most 1 percent.[32]

The *MIV-MTS* analysis therefore provides evidence against the possibility that physician time and the use of diagnostic tools are substitutes, while leaving a substantial scope for complementarity. We next summarize the overall message that emerges from the range of estimators employed in this study regarding the questions of interest.

4.4 Taking stock: takeaways from different estimation strategies regarding diagnostic inputs and physician time

Our baseline analysis, employing linearity and a strict exclusion restriction, provided a negative and statistically significant point estimate for the effect of workload on the probability of using diagnostic tools, suggesting that physician time and the employment of such tools are complements. This estimated effect could be interpreted either as a homogeneous treatment effect pertaining to all units, or as a Local Average Treatment Effect.

The nonparametric bounds analysis employing the combined *MIV-MTS* restrictions supplements the baseline analysis by pursuing different assumptions, and by estimating a different object. It does not impose a linear functional form, and relies on an *MIV* assumption which is weaker than the strict exclusion restriction used in the baseline analysis. It is therefore valid under a much more broad set of assumptions on the underlying DGP, as summarized in Table 2. It also adds the *MTS* assumption which we view as a natural reflection of the very endogeneity issue we tackle, and was also formalized within our analytical framework. In terms of the estimated object, rather than providing a point estimate of a LATE, the *MIV-MTS* delivers bounds on the ATE.

The estimated *MIV-MTS* bounds reinforce the results from our baseline linear analysis: they effectively rule out the possibility that diagnostic tools serve as substitutes to physician time in a quantitatively important fashion, while leaving a substantial scope for complementarity.

[32] In Appendix B.2 we repeat the analysis within patient subsamples, obtaining similar results.

Across these analyses, we therefore obtain the robust finding that diagnostic tools do not serve as a substitute for physician time. The baseline linear results go further and unequivocally establish that those tools are complements to physician time. The sections above provided very detailed arguments in favor of the exclusion restriction employed in our baseline analysis, and so *we view the complementarity result as robust.* But even if one were to discredit this exclusion restriction and place more trust in the *MIV-MTS* setup, we would still learn the valuable information that the ATE of workload on the probability of using diagnostics is not likely to be positive at an economically meaningful level, and that the data do not rule out the possibility that this effect is strongly negative.

By presenting both the linear *IV* results and the nonparametric *MIV-MTS* results we consider a range of identifying restrictions. In section 3, we connected these different restrictions to different underlying assumptions regarding the DGP. We therefore learn what the data robustly tell us about the key questions of interest under a broad range of alternative assumptions.

The economic implications of these results were discussed in the introduction. We learn that the shadow cost of a tight physician capacity is not reflected in more intense treatment of short-run issues, but rather in a reduction of the scope of issues discussed with the patient, and the extent of follow up on long-term issues. We, therefore, obtain a multi-faceted picture of the intricate role played by physician workload in determining clinical courses of action, and, ultimately, in the administration of primary care.

5 Conclusions

In this study we examine the effect of workload on physician behavior, focusing on two primary questions. First, we asked how does a tightening time constraint affect the intensity of medical treatment such as the prescription of painkillers or antibiotics, or referrals to the Emergency Room. We find that such an effect is negligent, suggesting that more intense use of such inputs should not be considered as a component of the shadow cost of physician capacity.

Second, we ask whether the use of diagnostic inputs can serve as a substitute for physician time. If that were the case, policy makers may be able to invest in such diagnostic tools as a possible remedy to the "primary care crunch," i.e., the chronic short capacity of primary care physicians and their time with patients.

We find that the answer is negative: across a range of econometric strategies, we find that diagnostic inputs are either complements to time with patients — or, at least, cannot be said to substitute for it in an economically meaningful fashion. These results are robust across a variety of economic assumptions on the underlying Data Generating Process, motivating a variety of econometric techniques.

The results leave a strong scope for the possibility that, as time with patients decreases, physicians discuss fewer medical issues with patients, and make a lesser use of referrals to tests

and specialists. The shadow cost of physician capacity may, therefore, very well include a reduction in the delivery of preventive care, an important mission of primary care services.

Our identification strategy relies on the absence of colleagues as a source of exogenous variation in workload. This setting corresponds conceptually to a very simple "comparative statics" analysis examining the effect of a temporary increase in workload on physicians' behavior holding other parameters of the environment constant. While we believe that the exclusion restriction is reasonable and have provided support for this claim, we have also used an analytical framework to discuss scenarios where it could fail.

In those cases, other econometric restrictions deliver consistent estimation of the effects of interest. Our resulting employment of a variety of econometric techniques — from standard 2SLS to nonparametric bounds estimation of the effects of interest — allows us to explore how different assumptions combine with data to produce robust conclusions.

Some open questions remain: while workload affects physician behavior, the ultimate impact on patient well-being and the efficiency of the system has yet to be completely understood. Furthermore, our analysis examines temporary increases in workload that allow for inter-temporal substitution of tasks. The effect of permanent increases in workload however may have a stronger impact on patient well-being.

References

Krishnan S Anand, M Fazil Pac, and Senthil Veeraraghavan. Quality-speed conundrum: Trade-offs in customer-intensive services. *Management Science*, 57(1):40–56, 2011.

Robert J Batt and Christian Terwiesch. Doctors under load: An empirical study of state-dependent service times in emergency care. *Working Paper, The Wharton School*, 1, 2012.

Jay Bhattacharya, Azeem M Shaikh, and Edward Vytlacil. Treatment effect bounds: An application to swan–ganz catheterization. *Journal of Econometrics*, 168(2):223–243, 2012.

Thomas S Bodenheimer and Mark D Smith. Primary care: proposed solutions to the physician shortage without training more physicians. *Health Affairs*, 32(11):1881–1886, 2013.

Christopher S Brunt, John Bowblis, and Joshua R Hendrickson. Physician competition and quality of care: Empirical evidence from medicare's physician quality reporting system. Technical report, Mimeo, 2018.

David C Chan. Teamwork and moral hazard: evidence from the emergency department. *Journal of Political Economy*, 124(3):734–770, 2016.

Olivier Chatain and Alon Eizenberg. Demand fluctuations, capacity constraints and repeated interaction: An empirical analysis of service quality adjustments. Technical report, CEPR, DP10545, 2015.

Jeffrey Clemens and Joshua D Gottlieb. Do physicians' financial incentives affect medical treatment and patient health? *American Economic Review*, 104(4):1320–49, 2014.

Janet Currie and W Bentley MacLeod. First do no harm? tort reform and birth outcomes. *The Quarterly Journal of Economics*, 123(2):795–830, 2008.

Janet Currie, W Bentley MacLeod, and Jessica Van Parys. Provider practice style and patient health outcomes: the case of heart attacks. *Journal of health economics*, 47:64–80, 2016.

Monique De Haan. The effect of parents' schooling on child's schooling: a nonparametric bounds analysis. *Journal of Labor Economics*, 29(4):859–892, 2011.

Monique De Haan and Edwin Leuven. Head start and the distribution of long term education and labor market outcomes. *CESifo Working Paper No. 5870*, 2016.

Joseph J.Jr. Doyle, Steven Ewer, and Todd H. Wagner. Returns to physician human capital: Evidence from patients randomized to physician teams. *Journal of Health Economics*, 29(6): 866–882, 2010.

Amy Finkelstein, Matthew Gentzkow, and Heidi Williams. Sources of geographic variation in health care: Evidence from patient migration. *The quarterly journal of economics*, 131(4): 1681–1726, 2016.

Sebastián Fleitas. Who benefits when inertia is reduced? competition, quality and returns to skill in health care markets. Technical report, Mimeo, 2018.

Michael Frakes. The impact of medical liability standards on regional variations in physician behavior: evidence from the adoption of national-standard rules. *American Economic Review*, 103(1):257–76, 2013.

Seth Freedman, Ezra Golberstein, Tsan-Yao Huang, David Satin, and Laura Barrie Smith. Docs with their eyes on the clock? the effect of time pressures on primary care productivity. *Journal of Health Economics*, 77, 2021.

Martin Gaynor, James B Rebitzer, and Lowell J Taylor. Physician incentives in health maintenance organizations. *Journal of Political Economy*, 112(4):915–931, 2004.

Michael Gerfin and Martin Schellhorn. Nonparametric bounds on the effect of deductibles in health care insurance on doctor visits–swiss evidence. *Health economics*, 15(9):1011–1020, 2006.

Libertad Gonzalez. Nonparametric bounds on the returns to language skills. *Journal of Applied Econometrics*, 20(6):771–795, 2005.

Kate Ho and Ariel Pakes. Physician payment reform and hospital referrals. *American Economic Review*, 104(5):200–205, 2014.

Kate Ho and Adam M. Rosen. Partial identification in applied research: Benefits and challenges. *NBER Working paper 21641*, 2015.

FD Richard Hobbs, Clare Bankhead, Toqir Mukhtar, Sarah Stevens, Rafael Perera-Salazar, Tim Holt, Chris Salisbury, et al. Clinical workload in uk primary care: a retrospective analysis of 100 million consultations in england, 2007–14. *The Lancet*, 387(10035):2323–2330, 2016.

Guido W Imbens and Charles F Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Diwas S Kc and Christian Terwiesch. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):1486–1498, 2009.

Song-Hee Kim, Carri W Chan, Marcelo Olivares, and Gabriel J Escobar. Association among icu congestion, icu admission decision, and patient outcomes. *Critical Care Medicine*, 44(10): 1814–1821, 2016.

Brent Kreider, John V Pepper, Craig Gundersen, and Dean Jolliffe. Identifying the effects of snap (food stamps) on child health outcomes when participation is endogenous and misreported. *Journal of the American Statistical Association*, 107(499):958–975, 2012.

Charles F Manski and John V Pepper. Monotone instrumental variables: with an application to the returns to schooling. *Econometrica*, 68(4):997–1010, 2000.

Magne Mogstad, Andres Santos, and Alexander Torgovitsky. Using instrumental variables for inference about policy relevant treatment effects. *Econometrica (forthcoming)*, 2018.

Hannah T Neprash. Better late than never? physician response to schedule disruptions. Technical report, Mimeo, 2016.

Olga Perdikaki, Saravanan Kesavan, and Jayashankar M Swaminathan. Effect of traffic on sales and conversion rates of retail stores. *Manufacturing & Service Operations Management*, 14(1):145–162, 2012.

Adam Powell, Sergei Savin, and Nicos Savva. Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient. *Manufacturing & Service Operations Management*, 14(4):512–528, 2012.

Anthony Scott. Chapter 22 economics of general practice. In *Handbook of Health Economics*, volume 1, Part B, pages 1175 – 1200. Elsevier, 2000.

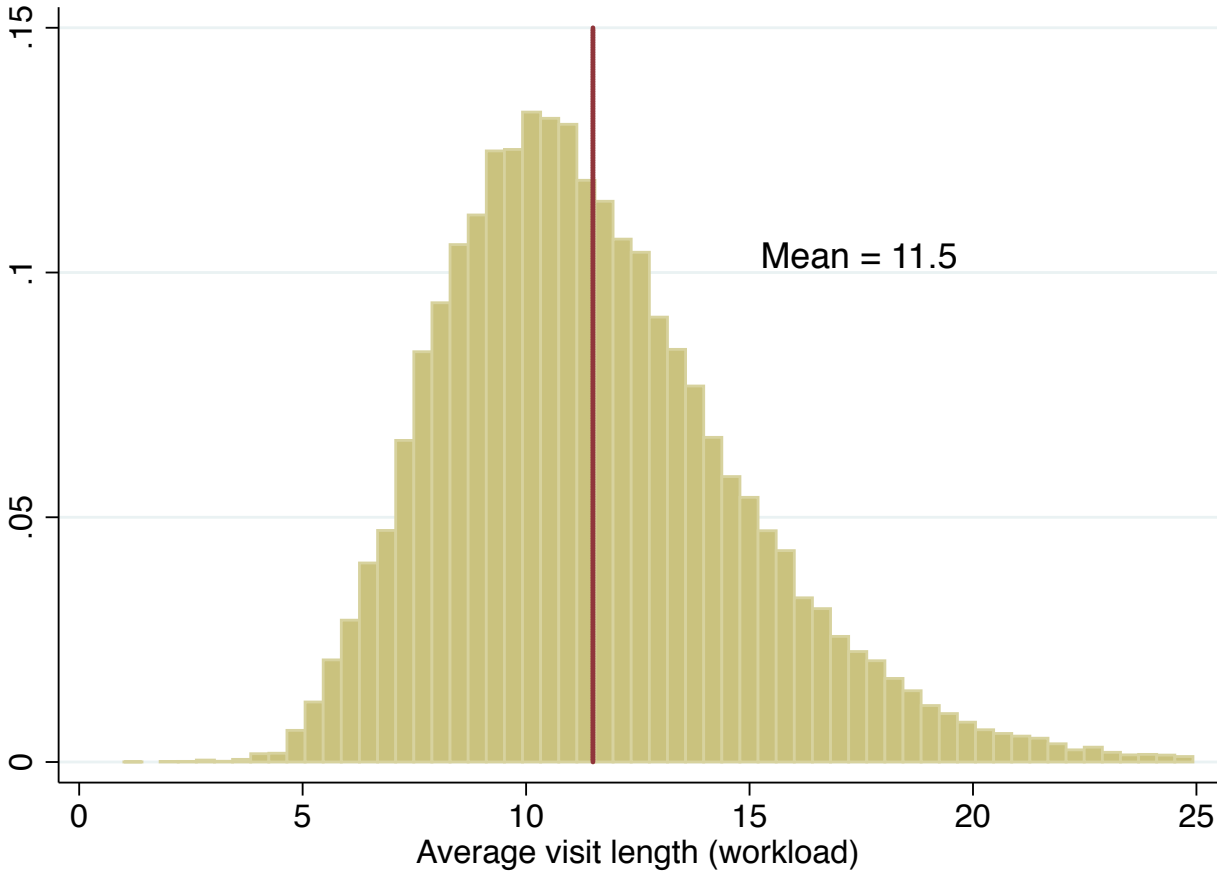
Scott K Shriver, Harikesh S Nair, and Reto Hofstetter. Social ties and user-generated content: Evidence from an online social network. *Management Science*, 59(6):1425–1443, 2013.

Barbara Starfield, Leiyu Shi, and James Macinko. Contribution of primary care to health systems and health. *Milbank quarterly*, 83(3):457–502, 2005.

David I Stern. Elasticities of substitution and complementarity. *Journal of Productivity Analysis*, 36(1):79–89, 2011.

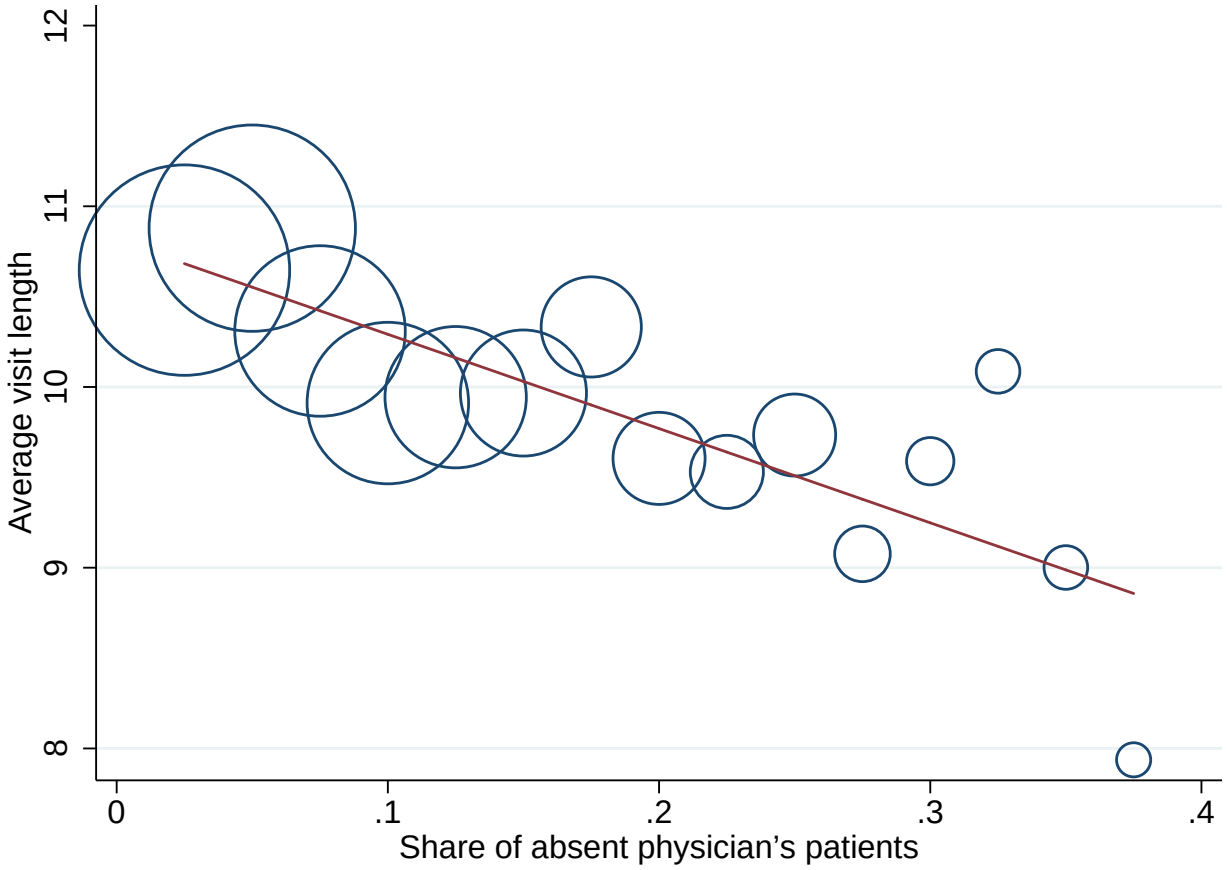
Tom Fangyun Tan and Serguei Netessine. When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science*, 60(6):1574–1593, 2014.

Figure 1: Distribution of average visit length (workload)



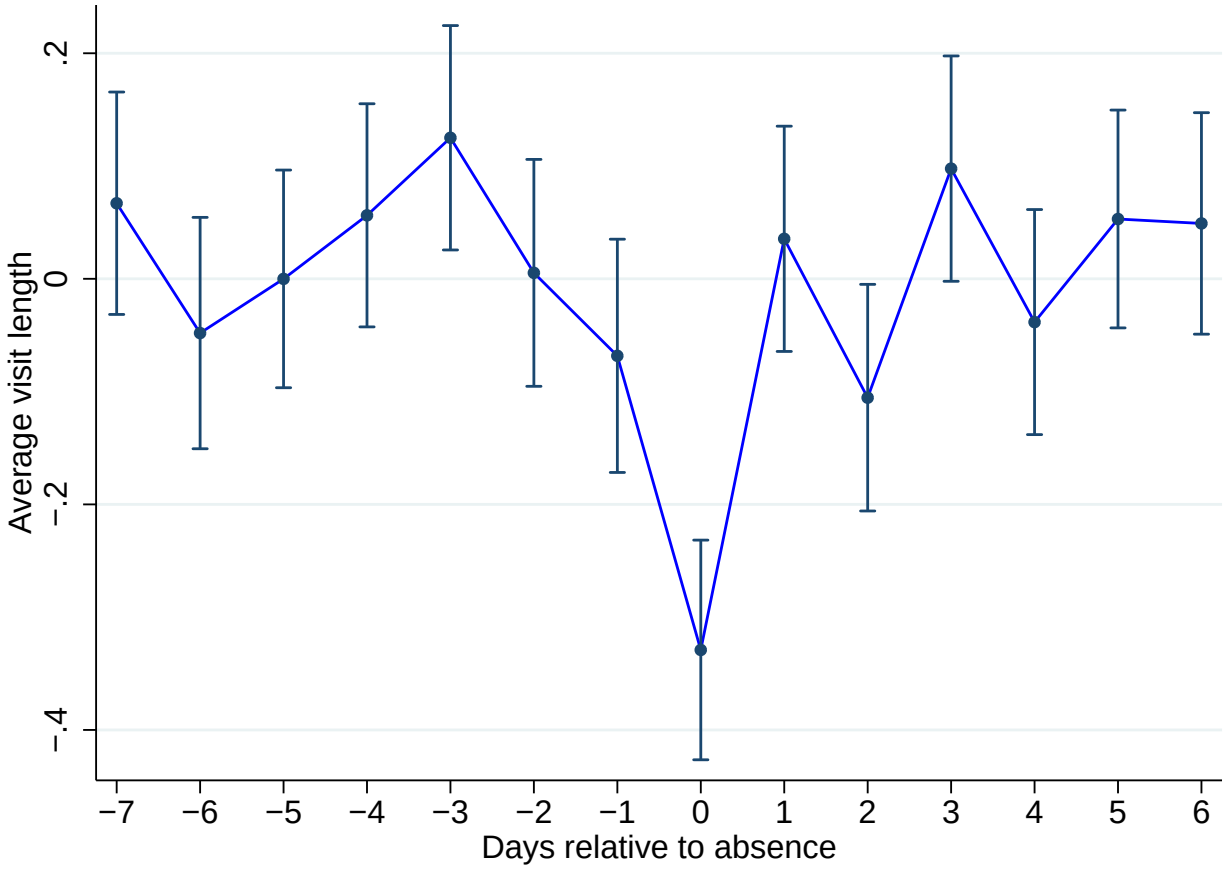
Note: The figure plots the average visit length–our workload measure. The (red) vertical line denotes the workload measure's mean. As the 99^{th} percentile of the distribution is 21.5, about 2000 observations pertaining to visits in excess of 25 minutes were excluded for illustrative purposes.

Figure 2: Share of missing physician's patients and workload



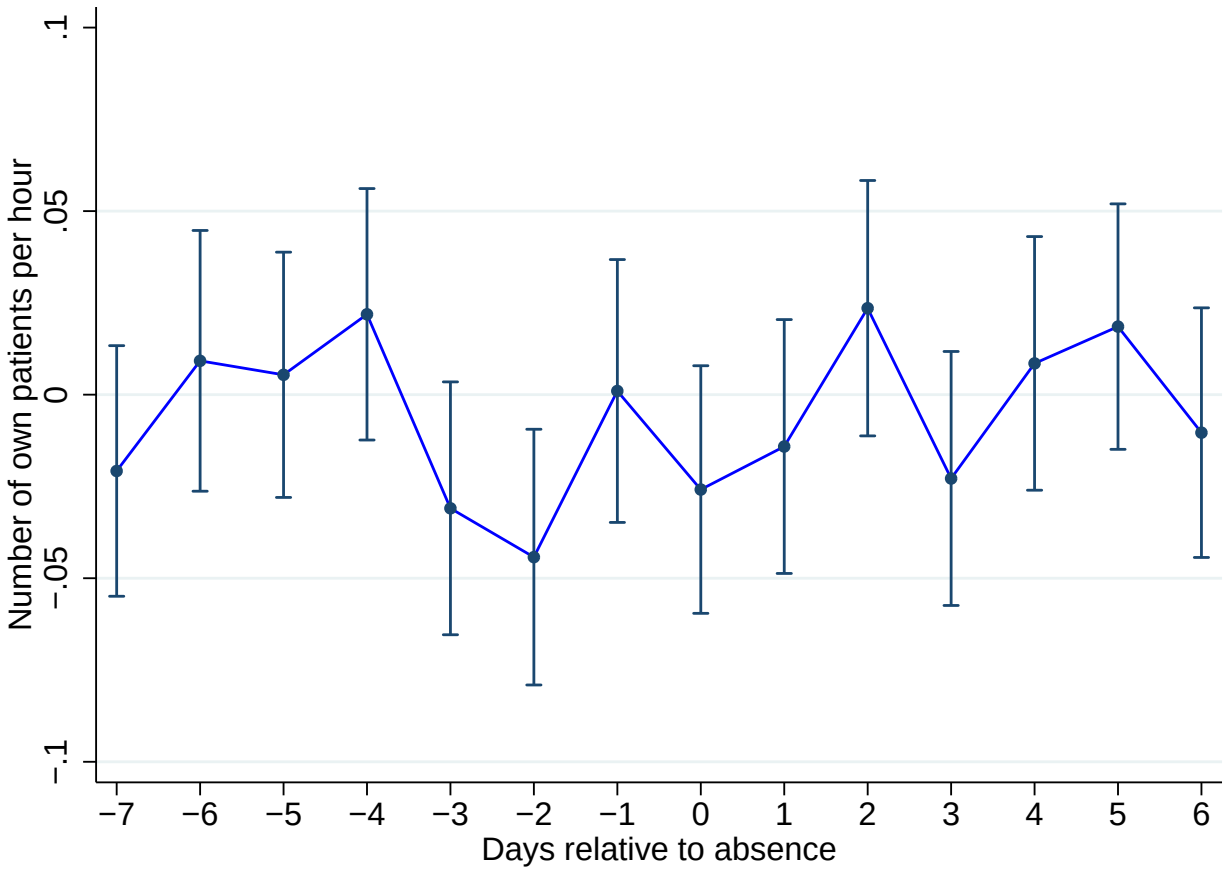
Note: The figure plots means of average visit length (workload) in 0.025 percentage point bins of the share of the missing physician's patients. The superimposed line is the predicted relation between absence and workload.

Figure 3: Workload around days of a colleague's absence



Note: The figure plots the coefficients and standard errors from the event study model described in Equation (5). The dependent variable is average visit length (workload).

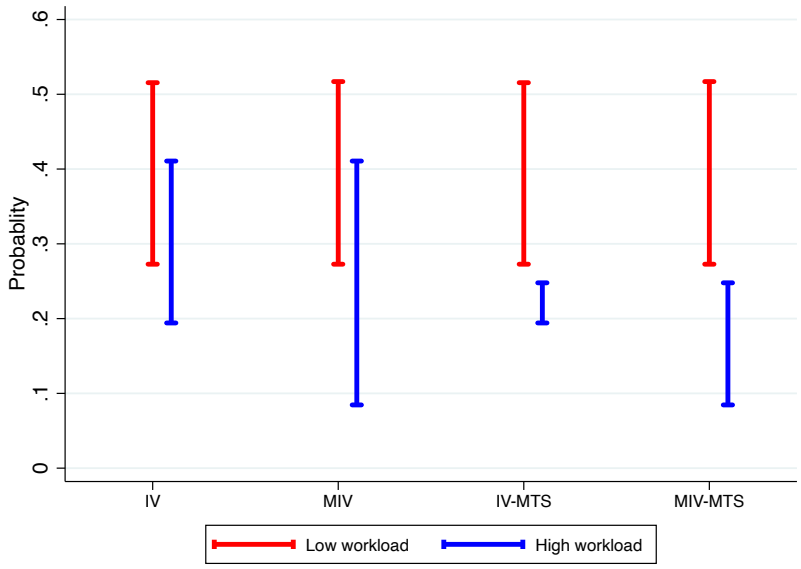
Figure 4: Number of own patients per hour around days of a colleague's absence



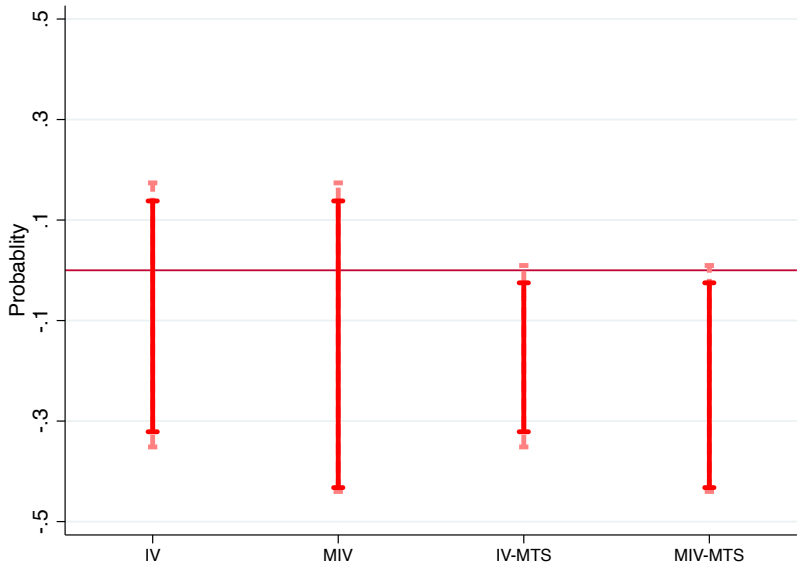
Note: The figure plots the coefficients and standard errors from the event study model akin to the model in Equation (5). The dependent variable is the number of a physician's own patients per hour.

Figure 5: Bounds on the effect of workload on the utilization of diagnostic inputs

(a) Low and high workload



(b) ATE



Note: Estimates from the bounds analysis. Panel (a) presents estimated sets for the probability of using diagnostics under different workload levels. Panel (b) presents estimated sets and 95 percent Confidence Intervals for the ATE of a switch from low to high workload on this probability.

Table 1: Summary statistics of visits data

	Mean (1)	SD (2)
Patient characteristics		
Mean age	47.60	26.68
Share women	0.58	0.49
Share born in Israel	0.61	0.49
Share smokers	0.30	0.46
Share obese	0.26	0.44
Share hypertension	0.34	0.47
Share hyperlipidemia	0.45	0.50
Share ischemic heart disease	0.15	0.36
Office visits characteristics		
Visit length	11.56	9.31
Share referral to specialist	0.14	0.35
Share referral to imaging	0.08	0.27
Share referral to lab tests	0.20	0.40
Share referrals to ER	0.01	0.11
Share Painkiller	0.05	0.21
Share antibiotics	0.10	0.30
Number of patients	78,959	
Number of physicians	93	
Observations	823,349	

Notes: The table includes Sunday-Thursday face-to-face visits in the clinics used in this study in the period 2011-2014 (see text).

Table 2: Assumptions about DGP and estimators

Assumptions on DGP	Econometric restrictions			
	Exogenous workload	MTS	IV	MIV
No attrition: $N_\ell^D \equiv 0$				
1. $\phi_\ell \perp \Theta_\ell, \eta_\ell \perp \Theta_\ell$	✓	✓	✓	✓
2. $COV(\phi_\ell, \overline{\Theta}_\ell) \geq 0, \eta_\ell \perp \Theta_\ell$		✓	✓	✓
3. $\phi_\ell \not\perp \Theta_\ell, \eta_\ell \perp \Theta_\ell$			✓	✓
4. $COV(\phi_\ell, \overline{\Theta}_\ell) \geq 0, COV(\eta_\ell, \overline{\Theta}_\ell) \geq 0$		✓		✓
5. $\phi_\ell \not\perp \Theta_\ell, COV(\eta_\ell, \overline{\Theta}_\ell) \geq 0$				✓
Attrition allowed: $N_\ell^D \geq 0$				
6. $COV(\phi_\ell, \overline{\Theta}_\ell) \geq 0, COV(\eta_\ell, \overline{\Theta}_\ell) \geq 0$		✓		✓
7. $\phi_\ell \not\perp \Theta_\ell, COV(\eta_\ell, \overline{\Theta}_\ell) \geq 0$				✓

Notes: The assumptions on the DGP use the notation of the theoretical framework in Section 3.1: ϕ_ℓ and η_ℓ are the stochastic elements affecting the number of the physician's regular patients, N_ℓ (before attrition), and the number of an absent colleague's patients, A_ℓ, showing up on day ℓ at the clinic, respectively. Θ_ℓ is the mean patient type. N_ℓ^D is the number of deterred patients, which is endogenously determined. The econometric restrictions are defined in sections 3.2 and 3.3.

Table 3: The effect of absences on workload

	(1)	(2)
A. IV 1		
Share of absent physician's patients of all patients	-4.82**	-4.82**
	(0.26)	(0.26)
F-statistic	356	357
p-value	0.000	0.000
B. IV 2		
Seeing absent physician's patients	-0.63**	-0.62**
	(0.05)	(0.05)
F-statistic	188	188
p-value	0.000	0.000
Year-month, day & physician FE	Yes	Yes
Patient age, gender & condition controls	No	Yes
Observations	823,349	823,349

Notes: All columns report estimates of effect of absence on workload, as per Equation (6). The Year-month fixed effects consist of a dummy variable for each of the calendar months in our data. Patient age, gender & condition controls include: age, age squared, a gender dummy, and 113 indicators for chronic conditions. Additionally, we include dummy variables for visits for which the main reason is: issue a medical certificate, prescription renewal, filling out forms, and an administrative visit. Standard errors clustered by physician-day are reported in parentheses. One or two asterisks indicate significance at 5% or 1%, respectively.

Table 4: The effect of workload on utilization of diagnostic inputs

	OLS		IV 1		IV 2	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Dependent Variable: all diagnostic inputs (mean =0.36)						
Average visit length	0.48**	0.45**	1.64**	1.62**	1.86**	1.84**
	(0.02)	(0.02)	(0.36)	(0.36)	(0.44)	(0.44)
B. Dependent Variable: referral to a specialist (mean =0.14)						
Mean visit length	0.33**	0.30**	1.13**	1.11**	1.34**	1.32**
	(0.02)	(0.02)	(0.25)	(0.25)	(0.32)	(0.32)
C. Dependent Variable: referral to a lab test (mean =0.20)						
Average visit length	0.16**	0.16**	0.76*	0.77*	0.89*	0.93**
	(0.02)	(0.02)	(0.31)	(0.31)	(0.36)	(0.36)
D. Dependent Variable: referral to imaging (mean =0.08)						
Mean visit length	0.22**	0.20**	0.11	0.09	0.15	0.12
	(0.01)	(0.01)	(0.18)	(0.18)	(0.22)	(0.22)
E. Dependent Variable: number of diagnoses codes (mean =1.56)						
Average visit length	0.011**	0.010**	0.015*	0.015*	0.027**	0.026**
	(0.000)	(0.000)	(0.006)	(0.006)	(0.007)	(0.007)
Year-month, day & physician FE	Yes	Yes	Yes	Yes	Yes	Yes
Patient age, gender & condition controls	No	Yes	No	Yes	No	Yes
Observations	823,349	823,349	823,349	823,349	823,349	823,349

Notes: Panels (A), (B), (C) and (D) of this table report estimates of effect of workload on the probability of use of any of the diagnostic inputs, referral to a specialist, referral to a lab test and referral to imaging, respectively, as per Equation (7). Panel (E) reports the dianoses number estimates. The Year-month fixed effects consist of a dummy variable for each of the calendar months in our data. Patient age, gender & condition controls include: age, age squared, a gender dummy, and 113 indicators for chronic conditions. Additionally, we include dummy variables for visits for which the main reason is: issue a medical certificate, prescription renewal, filling out forms, and an administrative visit. Standard errors clustered by physician-day are reported in parentheses. One or two asterisks indicate significance at 5% or 1%, respectively.

Table 5: The effect of workload on treatment decision

	OLS		IV 1		IV 2	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Dependent Variable: all treatments (mean =0.15)						
Average visit length	0.04*	0.01	-0.49	-0.52	-0.81*	-0.88**
	(0.02)	(0.02)	(0.27)	(0.27)	(0.32)	(0.32)
B. Dependent Variable: referral to the emergency room (mean =0.01)						
Mean visit length	0.05**	0.04**	0.00	-0.00	-0.07	-0.07
	(0.01)	(0.01)	(0.07)	(0.08)	(0.09)	(0.09)
C. Dependent Variable: prescription of pain killers (mean =0.05)						
Mean visit length	0.03**	0.02*	-0.17	-0.17	-0.29	-0.29
	(0.01)	(0.01)	(0.15)	(0.14)	(0.18)	(0.18)
D. Dependent Variable: prescription of antibiotics (mean =0.10)						
Mean visit length	-0.03	-0.04**	-0.32	-0.32	-0.51	-0.55*
	(0.02)	(0.01)	(0.23)	(0.23)	(0.26)	(0.26)
Year-month, day & physician FE	Yes	Yes	Yes	Yes	Yes	Yes
Patient age, gender & condition controls	No	Yes	No	Yes	No	Yes
Observations	823,349	823,349	823,349	823,349	823,349	823,349

Notes: Panels (A), (B), (C) and (D) of this table report estimates of effect of workload on the probability of any of the treatments, referral to the emergency room, prescription of pain killers and prescription of antibiotics, respectively, as per Equation (7). The Year-month fixed effects consist of a dummy variable for each of the calendar months in our data. Patient age, gender & condition controls include: age, age squared, a gender dummy, and 113 indicators for chronic conditions. Additionally, we include dummy variables for visits for which the main reason is: issue a medical certificate, prescription renewal, filling out forms, and an administrative visit. Standard errors clustered by physician-day are reported in parentheses. One or two asterisks indicate significance at 5% or 1%, respectively.

Table 6: The effect of workload on the use of diagnostic inputs: bounds results

	Low workload		High workload		ATE (Low to High)		ATE CI		Obs.
	(Lower)	(Upper)	(Lower)	(Upper)	(Lower)	(Upper)	(Lower)	(Upper)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
IV	0.273	0.516	0.194	0.411	-0.321	0.138	-0.351	0.174	822,416
MIV	0.273	0.517	0.085	0.411	-0.432	0.138	-0.440	0.174	822,416
IV-MTS	0.273	0.516	0.194	0.248	-0.321	-0.025	-0.351	0.010	822,416
MIV-MTS	0.273	0.517	0.085	0.248	-0.432	-0.025	-0.440	0.010	822,416

Notes: The table reports the estimates from the bounds analysis under the IV, MIV, IV-MTS and MIV-MTS assumptions, respectively. The ATE is defined in (4) and pertains to the average change in the probability with which diagnostic inputs are used following a switch from a low to a high workload state. Columns (7) and (8) report the lower and upper endpoints of 95 percent confidence intervals computed following Imbens and Manski (2004). See text.