Physician workload and treatment choice: the case of primary care

16 November 2017, 11:15-12:30, bld. 72, room 465

Ity Shurtz (HUJI)

Abstract: We examine how primary care physicians' treatment choices respond to physician workload, using detailed administrative data from eleven clinics of a large Israeli HMO. We use absences of colleagues at the clinic as a source of an exogenous increase in the physician's workload. Using a standard homogeneous-effects linear model, we find that physician time and utilization of diagnostic inputs are complements: during face-to-face visits, a one minute decrease in average (daily) visit length causes a 9 percent decrease in referrals to specialists, and a 3.8 percent decrease in referrals to lab tests. We find much smaller effects on the choice of treatment prescribed during the visit: our results imply no significant impact of workload on referrals to the emergency room, or on the prescription of painkillers, though there is some evidence that higher workload causes an increased prescription of antibiotics. Finally, when physicians experience higher workload they decrease the amount of non face-to-face encounters with patients. Our results are robust to relaxing the linearity and homogeneous-effects assumptions: following Manski and Pepper (2000), we compute nonparametric bounds on the Average Treatment Effects, resulting in qualitatively similar findings. Relaxing the exogeneity assumption of the instrument following a Monotone Instrumental Variable approach also results in similar conclusions. Our analysis provides important lessons to insurers and policy makers alike, as they reveal the channels via which practitioners respond to increased pressure brought about by limited capacity (the \primary care crunch"). In particular, we confirm that increased workload impairs primary care clinicians' ability to deliver preventive care, one of the key aspects of managed care health systems.

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

Adi Alkalay
Clalit Health Services

Alon Eizenberg†
Department of Economics
Hebrew University of Jerusalem

Amnon Lahad
School of public health
Hebrew University of Jerusalem

Ity Shurtz‡
Department of Economics
Hebrew University of Jerusalem

August 2017

Preliminary and Incomplete

Abstract

We examine how primary care physicians' treatment choices respond to physician workload, using detailed administrative data from eleven clinics of a large Israeli HMO. We use absences of colleagues at the clinic as a source of an exogenous increase in the physician's workload. Using a standard homogeneous-effects linear model, we find that physician time and utilization of diagnostic inputs are complements: during face-to-face visits, a one minute decrease in average (daily) visit length causes a 9 percent decrease in referrals to specialists, and a 3.8 percent decrease in referrals to lab tests. We find much smaller effects on the choice of treatment prescribed during the visit: our results imply no significant impact of workload on referrals to the emergency room, or on the prescription of painkillers, though there is some evidence that higher workload causes an increased prescription of antibiotics. Finally, when physicians experience higher workload they decrease the amount of non face-to-face encounters with patients. Our results are robust to relaxing the linearity and homogeneous-effects assumptions: following Manski and Pepper (2000), we compute nonparametric bounds on the Average Treatment Effects, resulting in qualitatively similar findings. Relaxing the exogeneity assumption of the instrument following a Monotone Instrumental Variable approach also results in similar conclusions. Our analysis provides important lessons to insurers and policy makers alike, as they reveal the channels via which practitioners respond to increased pressure brought about by limited capacity (the "primary care crunch"). In particular, we confirm that increased workload impairs primary care clinicians' ability to deliver preventive care, one of the key aspects of managed care health systems.

*We are grateful to Charles Manski for helpful conversations, and to participants at the Israeli IO Day 2017, iHEA Boston 2017 Congress and Barcelona GSE Summer Forum, Policy Evaluation in Health 2017 for helpful comments. Nadav Perlov provided excellent research assistance. Financial support from the Maurice Falk Institute for Economic Research in Israel is also gratefully acknowledged. All errors are our own.

†alon.eizenberg@mail.huji.ac.il
‡ity.shurtz@huji.ac.il

1 Introduction

In many industries and economic contexts, capacity is fixed or modifiable at a high cost while demand tends to be uncertain. In industries such as airlines, hotels and car rentals, a common practice is to adjust prices to reflect the varying shadow cost of capacity, i.e., the cost of serving the marginal consumer. By contrast, in professional services industries such as banking, legal services and healthcare, firms tend to refrain from price adjustments. Keeping prices largely stable, these firms may respond to demand fluctuations by adjusting other margins, such as service quality. In this study we examine whether, and in what sense, is such behavior present in healthcare, with a specific focus on primary care providers.

Primary care is a particularly interesting setup in this context because better primary care provision is thought to be associated with improved population health and lower health care costs (Starfield et al. (2005)).[1] In fact, primary care is perhaps the most important context in which the capacity-quality tradeoff arises because the stakes that are involved, namely the potential harm to patient health and well-being, may be very high relative to other professional services industries.[2]

In the setting of primary care, the capacity constraint and fluctuations in demand determine the number of patients a physician sees within a given amount of time, i.e. the physician's workload. Thus, we examine how primary care physicians respond to workload. We use a unique detailed administrative data from eleven clinics of a large Israeli HMO during 2011-2014 to study this issue. A commonly used measure of workload, in the primary care setting, is the number of patients a physician sees in a given amount of time or equivalently, the average visit length. Based on this notion, the measure of workload we use in this study is the physician's average daily length of office visits. In addition to visit length, the data also record other visit-level information that captures the physician's actions. In particular, we observe whether the patient is referred to a specialist, whether any medication is prescribed, or whether

[1]Scott (2000) notes that "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]Anand et al. (2011) for example notes that "A major difficulty in improving productivity in such customer-intensive services is the sensitivity of the service quality provided to the speed of service: as the service speed increases, the quality of service inevitably declines... Primary health-care practice in the United States epitomizes this problem"

a referral to an Emergency Room (ER) has been issued.

Primary care physicians play an important role in the healthcare system. First, they serve as gatekeepers for the HMO, regulating the treatments and referrals that patients receive to avoid unnecessary and inefficient treatment. Second, they oversee preventive care action, referring patients to tests and taking other action to detect and solve health issues before they deteriorate. Measuring the quality of primary care from data is therefore quite complex, as the link from physician choices to patient outcomes such as well being or mortality rates is far from obvious. For this reason, we study the impact of workload on a number of clearly identifiable visit-level outcomes. In particular, we study the impact of workload on the utilization of a variety of elements in the physician's toolcase. Those can be broadly categorized into three types: the use of diagnostics, the choice of treatment, and the utilization of non face-to-face interaction with patients.

A basic challenge to the identification of a causal effect of workload on these outcomes is endogeneity. As we clarify in more detail below, this may stem from both measurement error in the workload variable, and from the presence of unobserved factors that can simultaneously affect both the outcome, and the physician's workload. We address these challenges via an instrumental variable approach: absences of fellow physicians at the clinic are used as a source of an exogenous increase in the physician's workload. In practice, our instrument is the fraction of the total count of patients that visit a physician on a given day that is attributed to an absent colleague's patients. We complement this intensive margin approach with an extensive margin alternative, using an indicator for a colleague's absence as an instrument.

While our analysis uses the presence of other colleague's patients as a shifter of a physician's workload, we only analyze the physician's decisions with respect to her regular patients, excluding other patients. We do this to avoid confounding the effect of workload with the effect of treating unfamiliar patients. In the sections below, we discuss potential pitfalls to our identification strategy, and the manner with which we use the rich structure of our data to address them. One threat to identification is that seasonal effects, such as a flu epidemic, may affect both the absence of colleagues and patients' health. The use of time fixed effects helps us mitigate this concern. Another concern is that the allocation of an absent colleague's patients among the non-absent colleagues may not be random, i.e., that those “extra patients” are referred by the clinic’s manager to physicians that are better able to handle workload, or that have less bargaining power within the clinic. In a homogeneous-effect framework, our inclusion in the model of physician fixed effects ameliorates such problems. In a heterogeneous-effects framework, however, this will still result in identifying the “effect on the treated,” a common issue in instrumental variables models. Nonetheless, our bounds approach, discussed below, admits such heterogenous effects. Finally, another concern is that the absence of colleagues may affect the distribution of patient types that visit the physician on a given day. This may happen if, observing that wait time is longer than usual, patients with less urgent medical concerns give up their slot and decide to try seeing the doctor another day, implying that on days when colleagues are missing, the physician’s own patients represent, on average, more severe cases. We tackle this possibility via two channels: first, we demonstrate using descriptive statistics that no evidence is found for such an effect. Second, to the extent that this effect is still present, it is addressed via a Monotone Instrumental Variable approach as we explain below.

Results. We begin our analysis with a standard homogeneous-effects linear model estimated via Two Stage Least Squares regressions where the dependent variable is an indicator for a particular outcome (e.g., referral to a specialist) and the right-hand side variables include our main variable of interest, the daily average visit length, and a rich set of controls. We begin by considering face-to-face interactions, i.e., visits that involve the presence of the patient at the clinic, as opposed to interaction via phone or responses to online queries. In this context, we examine two aspects of physician behavior: utilization of diagnostic inputs, and the choice of treatment. Overall, we find that utilization of diagnostic inputs during an office visit tends to decrease with workload, namely, diagnostic inputs are complements rather than substitutes to physician time. Particularly, 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 an insignificant decrease in referrals to imaging. These results suggest that when physicians experience high workload they tend to limit the scope of issues they address during a single visit. Our results also speak to heterogeneity: the effects on patients of age 60 and above are stronger than those on younger patients.

With respect to the choice of treatment, one may hypothesize that physicians would be more conservative when workload is higher, and increase treatment intensity. We find only limited evidence that such a response arises in practice. We find no significant impact of workload on referrals to the emergency room. We also do not find evidence that workload affects the prescription of painkillers. We do find some evidence that higher workload is associated with an increase in prescription of antibiotics: a one minute decrease in average visit length increases the prescription of antibiotics by about 5%.

We proceed by analyzing the impact of workload on the likelihood of subsequent visits. We find some evidence that workload increases the likelihood of subsequent visits. However, the results are not statistically significant in most specifications and the magnitude of this effect appears to be quite small.

We next examine the effect of workload on non face-to-face encounters, which include patients' online queries and phone calls. We find that when physicians experience higher workload they decrease the amount of non face-to-face encounters with patients: a one minute reduction in average visit length decreases the amount of response to queries by 4 percent, reflecting an elasticity of 0.45. The number of phone calls physicians make with patients also decreases when workload is higher. A one minute reduction in average visit length causes a 10 percent decrease in the number of phone calls with patients (an elasticity of 1.1). The impact of workload on non face-to-face encounters is therefore quite strong.

We next relax the homogeneous-effect and linearity assumptions by computing nonparametric bounds on the Average Treatment Effect (ATE) following Manski (1990) and Manski and Pepper (2000). Here we retain the IV assumption but allow the effect of workload to vary across units of observation (i.e., visits), while also not imposing a linear structure. Preliminary results from this approach reaffirm those reported above for the linear model, and are again more pronounced for visits involving elderly patients.

Finally, we relax the assumptions further by not requiring that our instrument, a measure of the extra workload imposed by a missing colleague, be exogenous. Instead, we follow Manski and Pepper (ibid.) Monotone Instrumental Variable (MIV) approach and allow the instrument to affect the response not only via its impact on the treatment variable (i.e. the workload) in a pre-specified direction. We use this assumption to allow higher levels of the instrument to correspond to higher levels of the response function. This takes care of the possibility that the absence of a colleague implies a higher incidence of severe own-physician patient conditions described above (e.g., because less urgent cases choose to give up their office visit). This more conservative approach delivers bounds that appear to confirm the previous conclusions, albeit in a weaker fashion. While the ATE is not bounded away from zero under this weaker assumption, it still holds that the MIV upper (lower) bound on the probability of using diagnostics given low workload is above the upper (lower) bound on this probability given high workload. For elderly patients, the estimated intervals on the probability of utilizing diagnostic tools barely overlap, suggesting that the analysis only marginally falls short of signing the effect. The literature often combines the MIV assumption with a Monotone Treatment Response (MTR) assumption that further tightens the bounds. Taking this approach in our case will once again bound the ATE away from zero (in progress). In short, the MIV analysis provides further support to our main findings.

Policy implications. Our results reveal a multi-faceted picture of primary care physicians' response to workload. Such physicians have multiple margins along which they may respond to an increase in the number of patients they need to examine. Our results imply that physicians largely avoid changing the course of treatment on account of such pressure. They do not appear to prescribe more antibiotics or painkillers in a consistent fashion as they become busier, whereas one may have worried that they would use such prescriptions as a substitute to more thorough examination of the patient. Referrals to the emergency room also appear largely insensitive to the physician's workload. On the other hand, physicians do restrict the usage of diagnostic tools (referrals to tests and to specialists), as well as the amount of non face-to-face interaction with patients, as their time resources becomes more strained. The impact on referrals to specialists and to other tests is of particular importance: an early and accurate diagnosis of medical problems is considered one of the main benefits to managed care systems such as HMOs, and such referrals are very important in that context.

The implications of our results for policy considerations merits some discussion. After all, if providers are aware of their physicians' response to the capacity constraint, and optimally trade off their investment in increased capacity (i.e., the hire of additional physicians) versus their investment along other margins (i.e., referrals and diagnostics), then our results may have little policy implications. Both issues, however, are far from obvious. First, providers may not necessarily be aware of the cost, in terms of service quality, of their limited capacity. Our results may inform providers of this implied cost, noting that failure to properly refer patients to routine and nonroutine checkups and diagnostics may result in short-run savings, but also in long-run costs associated with poorer health. Second, even if providers optimally solve their own cost-minimizing problem, they may fail to fully internalize the social costs associated with patient health and well-being. In a setup where taxpayers, and not the providers, end up shouldering some of the costs associated with poorer long-run health outcomes, such misalignment of incentives is possible. Finally, our analysis is related to the so-called "primary-care crunch" — the shortage in primary care physicians in the United States. It is often argued that shortage of physicians induces increased workload, and lower quality healthcare.

Our analysis speaks to these policy issues. It shows that workload indeed has a substantial impact on physician behavior. Under higher workload, physicians appear to change their behavior both by reducing the number of non face-to-face encounters and changing their practice style during face-to-face encounters with patients. Thus, these results support a view often heard in the public discourse that physicians' high workload and physician shortage disrupt the delivery of healthcare and presumably lead to lower quality of care. Our analysis reveals the channels via which such effects occur, paying attention to the possibility that physicians may be able offset some of the harm to patients well-being by making optimal choices under the workload constraint. In this sense, our analysis also leaves some open questions: while workload affects physician behavior, the ultimate impact on patient well-being and the efficiency of the system has yet to be completely understood. Finally, it is worth noting that our analysis examines temporary increases in workload which allow intertemporal substitution of tasks. The effect of permanent increases in workload however may have a stronger impact on patient well-being.

Related literature. One strand of literature that is closely related to this study looks at crowding in the healthcare system and its effects on delivery of treatment. Using operational data from a hospital emergency department, Batt and Terwiesch (2012) find that workload induced service slowdown and that care 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 a reduction in the quality of care. Kim et al. (2016) study admission to intensive care units (ICU) and find that ICU congestion can have a significant impact on ICU admission decisions and patient outcomes. Powell et al. (2012) find that physician workload reduces that share of "severe" patients and consequently hospital reimbursement.

Another strand of literature that is related to this paper studies the impact of workload on worker productivity. Tan and Netessine (2014) use data from restaurants to examine to effect of workload, defined as the number of tables handled, on performance that is measured by sales and meal duration. Surprisingly, they find that workload is associated with higher sales effort in a manner that may lead to higher sales and to lower labor costs. Perdikaki et al. (2012) study the relationship between store traffic, labor, and sales performance. They find that the conversion of incoming traffic into sales declines with shopper's traffic. Chatain and Eizenberg (2015) study a legal service provider and find that service quality is an increasing function of the amount of available resources.

While there is a growing literature on the issue of workload-quality trade off in various work environments, this study and the unique setting it builds upon, is the first, to the best of our knowledge, to address this issue comprehensively in the context of primary care physicians. As the primary care environment is perhaps the most important setting in which this issue arises (see Anand et al. (2011)), this study contributes to filling an important gap in the understanding of this issue. This study also complements the growing literature that documents similar issues in the hospital environment. In terms of methods, our paper joins a growing empirical literature that places nonparametric bounds on the Average Treatment Effect under various sets of assumptions following Manski and Pepper (2000). Examples of such applications include health economics Gerfin and Schellhorn (2006), the economics of education (Gonzalez (2005), De Haan (2011)), public economics (Kreider et al. (2012)), and online network effects within social media (Shriver et al. (2013)).

The remainder of the paper is structured as follows. Section 2 describes the data, section 3 describes the empirical strategy, and section 4 reports our baseline findings. Section 5 presents our nonparametric bounds results, while section 6 concludes.

2 Data and Variables

We use a detailed administrative database that covers all the primary care visits in eleven clinics in the Jerusalem area of Clalit Health Services — the largest of four HMO's that provide the vast majority of health insurance in the country and deliver most of its primary care — in the period 2011-2014.

In Clalit Health Services, patients are enrolled with a primary care physician. Normally, a primary care visit is scheduled with the regular physician. However, there are exceptions to this routine. If patients need urgent care, outside of their physician's office hours or when their physician is absent, they are typically referred to one of their physician's colleagues at the clinic. In our analysis we restrict the sample to visits in which physicians see their regular patients in days in which they see at least twelve of their regular patients. By doing so we aim to capture the behavior of physicians treating only their regular patients in a typical day.

The data include information about visit characteristics such as visit time, visit length, and the patient's regular physician identity. They also include patient characteristics such as gender, age, country of origin and chronic conditions. Finally, a detailed description of the visit is recorded including diagnoses, prescriptions, referrals, laboratory tests, imaging and so on.

Table 1 provides descriptive statistics. The sample includes 825,660 visits by 80,084 patients. The number of physicians in the sample is 93. With respect to patient characteristics, the mean patient age is 47.6, 58 percent of the patient visits are by women, and most patients are native Israeli. Thirty percent of the patients are smokers and 26 percent are obese. Hypertension characterizes 34 percent of the visiting patients, almost 45 percent of them have hyperlipidemia, while 15 percent suffer from ischemic heart disease.

Office visits last 11.56 minutes on average. Fourteen percent of the visits result in a referral to a specialist, 8 percent result in referrals to imaging and 20 percent result in a referral to lab tests. This indicates that diagnostics are an important part of the regular treatment provided by primary care physicians. Patients are referred to the emergency room in one in every hundred visits. Antibiotics are prescribed in one in ten visits and pain killers are prescribed in one in twenty visits.

3 Empirical strategy

3.1 Workload

A common measure of workload in the primary care setting is the number of patients a physician sees per-hour or, equivalently, the average visit length (see e.g. Hobbs et al. (2016)). This is the measure of workload we use here. Thus, we define the main explanatory variable $workload$ as the daily average of a physician's visit lengths. For example, a physician that had an overall office visit time of two hours and, within that time, saw ten patients has a workload of twelve minutes per-patient. An issue that arises with respect to this measure is measurement error, as we discuss next.

3.2 Identification

Consider the following empirical model of the relationship between workload and physician behavior

$$y = \alpha + \beta \cdot workload + x \cdot \gamma + \epsilon \tag{1}$$

where y is an outcome and x is a rich set of controls.

Analyzing this model using OLS may provide biased estimates. Intuitively, our workload measure is based on the number of patients the physician sees per hour and ignores possible random shocks to workload, particulary, workload that arises because of patient characteristics such as a age, chronic conditions and so on. Thus, arguably the measure of workload that we use falls into a measurement error framework where the econometrician observes workload with some noise.

$$\widetilde{workload} = workload + u \tag{2}$$

In such cases attenuation bias may arise and the estimates of $\hat{\beta}$ may be biased towards zero.

Additionally, while the data are highly detailed and contain much of the relevant information that underlies the realization of the outcomes we study, omitted variables may still be correlated with our measure of workload. For example, a local infection may increase the number of patients the physician sees per hour and also the probability of prescribing antibiotics. We address both the omitted variable issue and the measurement error issue via an instrumental variable approach, as described in detail in the next section.

3.3 The IV approach

To identify the causal effect of workload on physician behavior we use the absence of colleagues at the clinic as an exogenous source of variation in a physician's workload. In Clalit Health Services, when a colleague is absent, her patients are referred to other physicians at the clinic. The absent physician's patients increase the workload of physicians who are present at the clinic. We use this source of variation in workload to identify its effects on physician behavior.

Our implementation of this approach builds on the orderliness of weekdays of work at the clinic. Physicians have fixed days and hours during the week in which they schedule appointments and see patients. We exploit this regularity and define an absence as a day in which two conditions are satisfied. First, on this day, a physician treats zero of her (positive number of) patients, namely, the physician is not present at the clinic in that day, although 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. This condition ensures that it is one of the physician's routine days. Having defined the physicians' days of absence from the clinic, we calculate, for each physician that was present at the clinic during that day, a proxy for the added workload brought about by the colleague's absence. Our proxy is the share of the missing colleague's patients out of the total patients seen by the physician on that day, where this total includes both the physician's regular patients, and the missing colleague's patients seen by this physician.[3] We refer to this proxy as the share of the missing physician's patients hereinafter. We use the share of the missing physician's patients as an instrumental variable for physician workload. We also define an additional instrumental variable, reflecting an "extensive margin" approach: an indicator taking the value 1 for physicians who see any missing colleagues' patients on the given day, and zero otherwise.

Threats to identification. Our identification strategy may run into some pitfalls that are summarized as follows. First, in periods that are prone to disease (e.g., the winter when the flu is more prevalent), the disease may affect the distribution of patient conditions that arrive at the clinic, while *also* leading to higher rates of physician absence. To minimize the potential biases from such issues we include a rich set of time-period controls that should ameliorate seasonal effects.

Second, it may be that the allocation of a missing colleague's patients among the non-missing physicians is not random. Some physicians may be more reluctant to see their colleague's patients than others, or may have more "bargaining power" that allows them to divert the additional workload towards others. It could also be that the clinic manager knows the extent to which the various physicians handle workload successfully, and diverts patients, as much as possible, towards those physicians who are likely to perform well under pressure. To the extent that such unobserved physician characteristics (bargaining power, ability to handle pressure) are correlated with therapeutic style (e.g., the general intensity with which a physician is prone to refer to specialists, or to prescribe painkillers), our instrument may not be valid. At the same time, as long as these physician characteristics are fixed over time, our inclusion of physician fixed effects should largely diminish such concerns.

Finally, there may be another channel via which the distribution of the physician's own patients is otherwise affected by the absence of a colleague. Imagine, for example, that these patients observe that wait times are worse than usual. Such patients who do not display

[3] We exclude patients that are neither the physician's regular patients nor the missing physician's regular patients.

acute medical conditions may then give up their appointment.[4] Our data analysis does not indicate much scope for such possibilities. In particular, we find that neither the number of the physician's own patients, nor their characteristics (notably, age) are affected by the instrument. We present this evidence in the next section, where we also display evidence regarding the "first stage" correlation between the instrument and our endogenous variable, workload. We also note that, while we do not believe that this mechanism plays an important role, our MIV strategy, presented in Section 5 below, accounts for it.

4 Baseline results: the homogenous-effect linear model

4.1 The first stage

We first illustrate graphically the source of variation we use in our instrumental variable approach. Figure 1 depicts the relationship between our instrument (the share of absent physician patients) and our measure of workload. To do so, we classify the share of the absent physician patients into bins of 0.2 percentage points. For each bin, we calculate the average of our measure of workload — the average visit length. We also regress our workload measure on the share of the absent physician patients, and use the solid line to display the relationship predicted by this regression. As the figure shows, an increase of ten percentage points in the share of the absent physician patients is associated with a decrease of about one minute in average visit length.

To further illustrate the effect of days with absences on physician workload at the clinic, we aggregate the data at the physician-day level and analyze an event study model. We 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 in clinic 5, then $D_{5,1/5/2013} = 1$. Next, define τ_{st}, the *event relative time*, as the number of days that elapsed since the absence. Thus, in our example $\tau_{5,1/5/2013} = 0$, $\tau_{5,1/4/2013} = -1$ and

[4] This scenario may involve patients who arrive and leave without being examined, or patients who are advised on the phone that rescheduling may be more attractive than showing up for their appointment on account of unusual workload at the clinic.

$\tau_{5,1/7/2013} = 2$. Letting physicians be indexed by j, we analyze a statistical model of the form:

$$
\begin{aligned}
(3) \qquad workload_{jst} =& \alpha + \nu_j + \nu_t + \\
& \gamma_1 \cdot \tau_{-k} + ... + \gamma_{k+1} \cdot \tau_0 + \gamma_{k+2} \cdot \tau_1 + ... + \gamma_{2k+1} \cdot \tau_{k-1} + \epsilon_{jst}
\end{aligned}
$$

where ν_j is a vector of physician fixed effects, and ν_t is a set of dummy variables for year-month combinations, and for the day of the week. The variables $\tau_{-k} - \tau_{k-1}$, the objects of interest, are indicators that capture the effect of the event on workload. Specifically, our hypothesis is that in the periods before the absence, the effect of these indicators is not significantly different from zero. At the time of the event, the effect should be negative, and after the event the effect should again not be significantly different from zero.

Figure 2 displays the estimates of these indicators, ranging from τ_{-7} to τ_6. The pattern is consistent with the above hypothesis: in the seven days before the event, the effect of the event is insignificantly different form zero. At the time of the event, the average visit length at the clinic drops by about a third of a minute. In the days after the event, average visit lengths are not significantly different from zero.

After illustrating the relationship between absences and physician workload we turn to estimating it formally. Indexing visits by i, the first stage regression is:

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

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 share of an absent physician's patients out of physician j's total count of patients at clinic s on day t. In the analysis below we denote this instrument by $IV1$. We also use a second, discrete version of the instrument which takes the value 1 if $sa_{jst} > 0$, and zero otherwise. We denote this instrument $IV2$ and we report the estimates using both instruments. The vector x is a set of visit-level characteristics including patient characteristics.[5]

[5]The patient level characteristics we use are: age, gender and chronic conditions. We additionally include dummy variables for visits for which the main reason is: issue a medical certificate, prescription renewal, filling out forms, and an administrative visit.

Table 2 displays the first stage results of the two instrumental variables we use in the analysis. As column 1 of panel (a) of the table shows, our first instrument, $IV1$, has a negative effect on workload with a point estimate of about -4.8, implying that an increase of ten percentage points in the share of an absent physician's patients is associated with a decrease of 0.48 minutes in average visit length.

As discussed above, one threat to our instrumental variable approach is a potential interaction between the absence of colleagues, and the composition of the physician's own patient pool. For example, if the less urgent cases are deterred by the physician's workload and decide to return on a different day, our identification assumption could be violated since the instrument would then affect the outcome we measure not just via its effect on $workload$. To assess this potential selection issue, we examine the sensitivity of the first stage regressions to patient characteristics. We also account for visits of an administrative nature. If the instrument affects the type of visits or the patient pool, the first stage estimates in this specification would be different from the previous estimates. As column 2 of panel (a) shows, adding patient and visit level controls virtually does not change the first stage results. These results are consistent with our identification assumption, namely they support the assumption that the instrument does not affect the composition of patients.

To further examine if a colleague's absence, and the resulting physician workload, creates deterrence to the physician's own patients, we examine if the number of the physicians' own patients is affected by a colleague's absence (in the spirit of McCrary (2008)). Concretely, we examine whether the number of patients in the clinic decreases on a day when a colleague is absent. If this is the case, we may worry that the patients that decide to reschedule their appointment are systematically different from those who choose to stay in the clinic, in the sense of presenting a less urgent medical problem. To implement the examination, we run an event study analysis, similar in nature to the analysis described in Equation (3) (and that was displayed in Figure 2). The dependent variable here is the number of the physician's own patients per hour. Figure 3 displays the results of this analysis. As the figure shows, there appears to be no change in the number of patients a physician sees per hour on days on which a colleague is absent. Namely, the number of visits of the regular patients of the present physician, on a day of a colleague's absence, is not statistically different from this number in other days. This result further alleviates the concerns that absence affects the outcomes we measure in channels other than through its effect on workload.

Finally, we turn to the first stage performance of our alternative instrument, IV2. This is addressed in panel (b), where 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. As in the case of the first instrument, adding patient characteristics and visit level controls in column 3 does not change the estimates.

4.2 The effect of workload on physician behavior

We now turn to our main research question: what is the impact of workload on physician behavior? To that end, we estimate the following version of the model we laid out in equation (1):

$$y_{jsti} = \alpha + \nu_j + \nu_t + \beta_1 \cdot workload_{jst} + x_{jsti} \cdot \beta_2 + \epsilon_{jsti} \tag{5}$$

were y_{jsti} is an outcome of interest, e.g. an indicator for referring a patient to a specialist or an indicator for referring a patient to the emergency room. We estimate this model for different outcomes following the discussion in the introduction. As described above, we use IV1 and IV2 to instrument for the main explanatory variable of interest, $workload_{jst}$.

4.3 The effect of workload on physician behavior during face-to-face visits

We turn to analyze the effect of workload on physician behavior in face-to-face encounters with patients. We focus on two sets of outcomes. The first set is diagnostic outcomes: indicators for a referral to a specialist, a referral to imaging (such as x-ray, ultrasound, CT or MRI) and a referral to lab tests (such as blood or urine test). The second is treatment outcomes: indicators for a referral to the emergency room, for the prescription of antibiotics, and for the prescription of pain killers.[6]

4.3.1 The effect of workload on diagnostic outcomes

It is not *a-priori* clear whether workload and diagnostic outcomes are substitutes or complements. On the one hand, when workload is higher, physicians may be able to substitute face-to-face time and physical examination with diagnostic procedures. On the other hand, under a tightening time constraint, physicians may limit the scope of the medical issues they address during the visit, and therefore may use fewer diagnostic procedures. The sign of the effect, as well as its magnitude, are a matter for empirical examination.

Table 3 displays the results of the diagnostic outcomes analysis. All specifications include Year-month, day and physician fixed effects. As explained above, the time fixed effects are helpful in addressing 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. The physician fixed effects, for their part, help guard against the possibility that the allocation of an absent colleague's patients among the non-absent physicians is non-random.

We begin by analyzing the overall utilization of diagnostic inputs by considering an indicator dependent variable, taking the value 1 if any of the diagnostic inputs categories we consider are used, and zero otherwise. Panel (A) of the table shows the results of this specification. The OLS estimate of the effect of the daily average visit length, our measure of workload, is reported in column (1) to equal 0.48. The estimate indicates that workload is negatively correlated with utilization of diagnostic inputs. Adding patient characteristics in column (2) decreases the estimates slightly to 0.45. The instrumental variable estimate in column (3) is positive and much larger with a point estimate of 1.62. This estimate is not affected by adding patient characteristics, as column (4) reports. The results are somewhat larger, with a point estimate of 1.85, with IV2 and, as column (6) reports, and they are also not sensitive to the inclusion of patient characteristics. Given that, on average, diagnostic inputs are used in 35 percent of the visits, the results from IV1 imply that a 1 minute decrease in average visit length causes a 4.6 percent decrease in the probability of utilization of diagnostic inputs. Next, we

[6] All the models in this section are linear probability models. The estimates in this section are all multiplied by a hundred.

examine each of the diagnostic inputs separately.

Panel (B) displays results concerning referrals to specialists. The results are qualitatively similar as those in Panel (A). The OLS estimate in column (1) is 0.33, indicating that workload is negatively correlated with referrals to specialists. In column (2), with patient characteristics, the estimates decrease slightly to 0.3. The instrumental variable estimate in columns (3) is again much larger with a point estimate of 1.13. Column (4) again shows that this estimate is not affected by adding patient characteristics. The results using the second instrument are again somewhat larger. Since the mean of the dependent variable here is 0.14, those latter results 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 with the dependent variable being an indicator for a referral to lab test results. Here too, the OLS estimates in columns (1) and (2) are positive with point estimate of 0.16. The IV estimates using the first instrument (column (3)-(4)) remain positive and they are much larger with point estimates of 0.76. The estimates that use IV2 are yet larger with point estimates of about 0.9. Since on average 20 percent of the visits result in a lab referral, the results indicate that a 1 minute decrease in average visit length causes a 3.8 percent decrease in the probability of a referral to a lab test. 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 IV estimates using IV1 (columns (3)-(4)) are positive, but are not statistically significant. The estimates that use IV2 are also statistically insignificant.

Our results suggest that, under a tightening time constraint, physicians avoid referring patients to specialists. This may reflect the administrative and professional burden of generating a referral: it requires the physician to write a detailed note to the specialist explaining the reasons and background for the referral. An exception is the case of referrals to imaging (panel (D)) that do not appear to be sensitive to workload. This may reflect the possibility that the utilization of imaging tends to be associated with acute conditions that receive attention regardless of workload, or that they involve a lower administrative burden.

Overall, the results in this section indicate that diagnostic inputs serve as complements, rather than substitutes, to the physician's time. It therefore appears that, under high workload, physicians tend to address fewer medical issues during a visit.

Heterogeneity. In Table 4, we explore heterogeneity in the effects of workload on diagnostic inputs. We focus on two margins of heterogeneity. The first is the patient's age. We split the sample to two subsamples, one with patients that are older than 60 and another with patients that are 60 or younger. For each subsample we analyze the same regression model as in Table 3.

We report the results of this analysis in columns (1)-(6) of Table 4. Panel (A) of the table shows the results for the overall utilization of diagnostic inputs. The point estimates for IV1, in columns (2) and (5) are 2.04 and 1.39 for patients over 60 and patients aged 60 or less, respectively. With averages of 0.33 and 0.39, these estimates indicate a 6% and a 3.6% decrease in utilization for the older and younger patients, respectively. This difference remains apparent, although it is attenuated, with IV2. Panel (B) shows a similar and even more pronounced pattern in referrals to specialists. Namely, that the effect is stronger among older patients. The referral to lab tests analysis in Panel (C) is also quite similar, although the difference between the two age groups is only apparent with IV1. The referrals to imaging results, shown in Panel (D) remain insignificant.

The second dimension of heterogeneity we explore is patient condition. To get at this, we take advantage of the fact that we observe, for every patient, detailed individual level characteristics as well as the number of visits that the patient makes in each time period. Using this information, we aim to attribute a "utilization score" to each patient in the data based on their personal characteristics. To do so, we analyze the data at the year patient level. We regress the number of visits per year against the set of patient level characteristics. We attribute to each patient in the data their predicted number of visits per period. The interpretation of the "utilization score" is intuitively, the number of visits per time period that a patient with these characteristics would make, on average, in a given time period. We then split our sample to high utilization–visits by patients above the median score, and low utilization–visits by patients with a below median score. We describe the process of creating the utilization score in detail in Appendix B.

The results of this analysis are displayed in columns (7)-(12) of Table 4. In Panel (A), the results for the overall utilization of diagnostic inputs, the point estimates for IV1, reported in columns (8) and (11) are 1.86 and 1.45 indicating a 5.4% and a 3.9% decrease in utilization for high utilization and low utilization patients, respectively. This difference arises also with IV2. Panel (B), referrals to specialists, shows a similar pattern. Namely, that the effect is stronger among high utilization patients. Panel (C) also shows a stronger decrease in referrals to lab tests among high utilization patients. The referrals to imaging results, shown in Panel (D) are insignificant.

4.3.2 The effect of workload on the choice of treatment

In this section we analyze the relationship between workload and the choice of treatment. Here, there appears to be a plausible hypothesis about the sign of the effect. Under higher workload, physicians may tend to be more conservative and provide more treatment. Namely, substitute office time and examination with treatment such as the prescription of antibiotics or painkillers, or referrals to the emergency room.

We examine this issue in Table 5. We first analyze the overall utilization of treatment. Analogously to the previous section, we use an indicator that takes the value 1 if any of the treatments we consider was used, and zero otherwise. Panel (A) shows the results of this specification, with columns (1) and (2) showing the OLS estimates without, and with accounting for patients characteristics, respectively. The estimate in Column (1) is positive but when we add patient characteristics it becomes small and statistically insignificant. The instrumental variable estimates in columns (3)-(6) are all negative. The estimates with IV1 are 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, that reflects an increase of about 5 percent in the probability of receiving treatment.

Unpacking this aggregate effect to consider specific treatment outcomes, Panel (B) reports estimates of the relationship between workload and 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 pain killers serve as the dependent variable are reported in panel (C). Again, the OLS estimates in column (1) and (2) are positive and significant, while the IV estimates in columns (3)-(6) are all negative and statistically insignificant.

Finally, Panel (D) displays the results regarding the prescription of antibiotics. The OLS estimate in column (1) is -0.03 percentage points, and is statistically insignificant. With patient characteristics the estimate, shown in column (2), is statistically significant at -0.04 percentage points. The IV estimates using the first instrument, shown in column (3)-(4), are -0.032 yet they are statistically insignificant. Using the second instrument (recall this is an indicator for whether the physician sees any patients of an absent colleague) the estimate in column (5) is -0.51 and adding 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 appointment length, increases the probability of receiving antibiotics by 5 percent.

Overall, these results indicate that there is only limited evidence of a tendency to increase the utilization in treatment options under higher 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 utilization of antibiotics.

Heterogeneity. In Table 6, we explore heterogeneity in the effect of workload on treatment along the two dimensions we described above, age and patient condition. We report the results from the analysis by age in columns (1)-(6) of Table 6. Panel (A) of the table shows the results for the overall utilization of treatment. The point estimates for IV1, in columns (2) and (5) are -0.89 and -0.3 for patients over 60 and patients aged 60 or less, respectively. With IV2 the corresponding point estimates are -1.25 and -0.57. Overall the effect on treatment appears to be present only for older patients, although the estimates are only statistically significant for IV2. The referral to the emergency room results in Panel (B) are statistically insignificant in all IV specifications. In Panel (C) we report the results for the prescription of pain killers as the dependent variable. The estimates using IV1 and IV2 are -0.65 and (a significant) -0.72 for older patients and 0.06 and 0 for the younger patients. I.e. there appears to be an increase in prescription of pain killer only in older patients. The results of prescription of antibiotics, reported in Panel (D) are all statistically insignificant. Overall, there appears to be little evidence for an effect of workload on treatment, and where present, the effect seems to be relevant for older patients.

The results of the analysis by patient condition are displayed in columns (7)-(12) of Table 6. In Panel (A), the results for the overall utilization of treatment appear to be similar for the two groups. The point estimates for IV1, in columns (8) and (11) are -0.41 and -0.58, for high utilization and low utilization patients, respectively, both statistically insignificant. The estimates for IV2, reported in columns (9) and (12) are larger, -0.76 and (a significant) -0.98. The results for referrals to the emergency room, reported in Panel (B), show no statistically significant IV results. In Panel (C) the results for the prescription of pain killers are statistically insignificant. The results of prescription of antibiotics, reported in Panel (D) are also all statistically insignificant.

4.4 The effect of workload on subsequent face-to-face encounters

We proceed by examining the impact of workload on subsequent face-to-face encounters. We create four indicator variables, each takes the value one if, after the visit, the patient arrives at the clinic again within one of four time windows of 15, 30, 60 and 90 days. Table 7 reports the results. Column (1) of the table reports the OLS regression results with no patient level controls. The estimate for the 15 days window, reported in panel (A) is a statistically significant 0.11 percentage points (the mean is 0.38), indicating a positive correlation between average visit length and subsequent visits. Adding patient levels controls in column (2), the result decreases to 0.09 percentage points. The OLS results for the other three time windows, that we report in panels (B)-(D), are all quite similar, positive and significant. The point estimates for IV1 without and with patient level controls are reported in columns (3) and (4) of Table 7, respectively. The estimates for the 15 days window, reported in panel (A) are -0.44 and -0.49 - both are statistically insignificant. The estimates in panel (B), the 30 day window, are also negative with insignificant point estimates of -0.5 and -0.56 percentage points. The 60 day window results are a marginally significant -0.59 percentage points, without patient controls and a significant -0.65 percentage points with patient controls. In the 90 days window the results are an insignificant -0.5 percentage points without patient level controls and they become significant with point estimate of -0.56 percentage points with patient level controls. The point estimates for IV2 (columns (5) and (6)) are quite similar - they are all negative yet statistically insignificant.

Overall, all the IV estimates for all time windows are negative and quite small, and most of them are statistically insignificant, indicating that patients who meet the physician when workload is high may be slightly more likely to schedule a subsequent visit.

Heterogeneity. Here too, we explore heterogeneity in the effects of workload along the two dimensions we described above, age and patient condition. We report the results in Table 8. The results from the analysis by age are reported in columns (1)-(6) of the table. The OLS results for patients that are over 60 and patients that are aged 60 or less, are reported in columns (1) and (4) respectively. The OLS estimates for patients that are over 60 in all time windows are positive while those of patients that are aged 60 or less are roughly zero. The IV estimates for patients that are over 60 are negative and statistically insignificant in all time windows. In the group of patients that are aged 60 or less all the estimates are insignificant and they appear to be smaller in absolute terms. Overall, the small tendency to increase the likelihood of a subsequent visit appears to be more prevalent among patients over 60.

The results of the analysis by patient condition are displayed in columns (7)-(12) of the table. Over all, there is no apparent evidence of a difference across patient condition. In the narrower windows of 15 and 30 days, there seems to be a stronger negative effect in the group of high utilization patients. However this difference in not apparent in the wider windows of 60 and 90 days.

4.5 The effect of workload on non face-to-face encounters

We complete our baseline results by considering how workload affects the number of non face-to-face physician patient encounters. Particularly, we look at two outcomes: responses to patients' online queries, and phone calls made with patients. As we analyze daily averages, we analyze the data in this section at the physician-day level.

Patients can make an online query to their physician using the Clalit Health Services website. The vast majority of those queries are prescription renewals and administrative requests. Response time is four working days and the Clalit Health Services website explicitly notes that queries are not for emergency cases. Physicians respond to those queries during the work day. This is an interesting outcome to explore since, unlike face-to-face office visits, physicians have a large degree of freedom to choose the timing in which they respond to those queries. Thus, this is a "cheap" channel, in terms of patient well being, that is available to physicians to manage their time. Our hypothesis is, therefore, that higher workload would result in a decreased number of replies to queries that physicians make during the day. Apart from online queries, another channel of communication between physicians and patients is phone calls. The extent to which a physician returns phone calls is another channel via which she may manage her time under the workload constraint. Therefore, we expect a reduced number of phone calls between physicians and their patients when workload is higher.

The analysis of the effect of workload on these two outcomes is summarized in Table 9. Panel (A) of the table displays the online queries results. The OLS estimate in column (1) is negative, indicating that workload is positively correlated with the number of responses to queries. The IV estimates in columns (2) and (3) are positive with magnitudes of 0.07 and 0.08, respectively. With a mean of 1.76 queries per hour, these estimates imply that a decrease in 1 minute in average visit length decreases responses to queries by 4 percent. Since the mean appointment length is 11.5, the estimates reflect an elasticity of 0.45.

Panel (B) of the table summarizes the results for phone calls. The OLS estimate in column (1) is, again, negative, while the IV estimates in columns (2) and (3) are again positive, with a level of 0.03. The mean number of phone calls per hour is 0.31 and thus, these estimates imply that a decrease in 1 minute in average visit length decreases responses to phone calls by 10 percent, reflecting an elasticity of 1.1. Consistent with our expectation, higher workload is associated with a decrease in response to online queries and in the number of phone calls with patients.

Overall, the results in this section indicate that the choice of treatment is not significantly affected by workload. Taken together, the results above imply that physicians prefer adjusting their choices on other channels to cope with higher workload. Namely, they respond less to online queries and perform fewer phone call with patients, and make a lesser use of diagnostic tests or referrals to specialists. The implication is that the shadow cost of physician capacity is not an oversubscription of medication, but rather a poorer long-term management of patient health via a reduced amount of diagnostic tests and thorough examination of medical problems by specialists.

5 Bounds analysis

The baseline analysis in the previous section entailed two key, implicit assumptions: a linear functional form, and a homogeneous treatment effect across units. In other words, it was assumed that increased workload has the same effect on all observations. Of note, it is possible to interpret the Two Stage Least Squares results reported above as indicators of Local Average treatment Effects (LATE, Imbens and Angrist 1994) within a model that allows for heterogeneity. This interpretation requires additional assumptions. In particular, it is necessary to assume that no physician responds to the absence of a colleague by spending *more* time with her own patients, on a daily average basis.

An alternative approach to introducing heterogeneity to the response function is to go beyond the linear model and explore nonparametric bounds on the average treatment effect (ATE) of workload on physician behavior. We do so following the framework from Manski and Pepper (2000). We next explain how to derive such results under an IV assumption akin to the one used above. Then, we explore the possibility of imposing a weaker assumption, Monotone Instrumental Variable (MIV), on the relationship between our instrument and the response function. As explained below, this assumption has a quite natural intuition in our setting.

5.1 Theory and Assumptions

Adopting the framework and notation from Manski and Pepper (2000), our setup can be described as follows. The population of interest contains a set $j \in \mathcal{J}$ of individual units (in our case, patient visits). Each individual is characterized by a *response function* $y_j(\cdot) : T \to Y$, where $t \in T$ are discrete treatments, and $y \in Y$ are discrete outcomes.

More concretely, in our application, the set $\mathcal{J}$ contains all visits (triplets of patient, physician and day). Our outcome space is binary, i.e., $Y = \{0, 1\}$: for example, a patient is either referred to a specialist, or not. Our original modeling of the treatment t is the level of physician workload on the relevant day, measured by the time spent with each patient on average. This is a continuous measure, but to keep the framework as transparent as possible, we define the treatment, too, as a binary variable, so $T = \{0, 1\}$. A value $t = 1$ implies that the physician experiences workload above a certain threshold (e.g., the 80th percentile of the physician-specific workload distribution), while $t = 0$ implies values below or at that threshold.

The observables in this framework are (x, z, y), where x_j is a covariate vector for appointment j. The variable $z_j \in T$ is the *realized treatment*. That is, it takes the value 1 for appointments that take place on a day when the physician is *observed* to experience higher-than-normal workload, and zero otherwise. Finally, $y_j = y_j(z_j)$ is the observed outcome. our object of interest is the distribution $\mathcal{P}[y(\cdot)]$ of response functions, or, its conditional version $\mathcal{P}[y(\cdot)|x]$. Specifically, as we focus on a binary outcome the ATE is defined by:

$$ATE(1, 0|x) = P[y(1)|x] - P[y(0)|x] \tag{6}$$

The covariate vector can be written as $x = (w, \nu) \in \mathcal{X} = W \times V$, where $\nu \in V$ is our instrument. In our application, recall that this is the share of patients seen by the physician that are attributed to the absence of a colleague. We treat the space of instrument values V as discrete, dividing it into 20 bins ranging from 0 to 0.4. This choice reflects our view that cases where more than 40 percent of the patients seen on a given day are not the physician's regular patients, but rather the patients of a missing colleague, are both rare, and probably extreme.

We emphasize two different assumptions that may be employed to characterize the relationship between the instrument, the treatment (i.e., the workload level), and the outcome: an IV assumption, and an MIV assumption. The Instrumental Variable assumption is the following:

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, this assumption does not allow the instrument to affect the response function,i.e., the outcomes that would be observed given various treatments. It can, therefore, only affect the outcome via its effect on the treatment t.

Alternatively, the Monotone Instrumental Variable assumption allows the response function to depend on the instrument, but in a pre-specified direction:

Assumption 2 ***MIV:***

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

Let us explain the content of this assumption in the context of a concrete outcome, say, the prescription of antibiotics. This assumption makes it possible for the absence of a colleague to be associated with higher probabilities of prescribing antibiotics conditional on any physician's workload level. This may be the case if, as discussed earlier, an absent colleague creates longer wait times that cause patients with non-acute medical complaints to give up their appointment. In this case, it is possible that the absence not only affects the workload, but increases the probability of prescribing antibiotics *conditional on the workload level.* As discussed above, we find little evidence for such mechanisms in the data. Nonetheless, it is interesting to note that the MIV assumption allows for consistent estimation even is such a mechanism is, in fact, in place.

5.1.1 Bounds under IV assumption

To derive bounds on the ATE under the IV assumption, the Law of Iterated Expectations can be used as follows:

$$E\big[y(t)|w,\nu=u\big] = E\big[y|w,\nu=u,z=t\big]\cdot P(z=t|w,\nu=u)$$
$$+ E\big[y(t)|w,\nu=u,z\neq t\big]\cdot P(z\neq t|w,\nu=u)$$

Recalling that the joint distribution of (y,w,ν,z) is given, the only unknown on the RHS of this expression is the "counterfactual" outcome under the treatment that was not assigned, $E\big[y(t)|w,\nu=u,z\neq t\big]$. But, since y can only be 0 or 1, we can bound this unknown terms by 0 from below, and by 1 from above. This results in the following bounds:

$$\underline{b}(w,u,t) \leq E\big[y(t)|w,\nu=u\big] \leq \overline{b}(w,u,t),\ \text{with} \tag{7}$$

$$\underline{b}(w,u,t) \equiv E\big[y|w,\nu=u,z=t\big]\cdot P(z=t|w,\nu=u)$$
$$\overline{b}(w,u,t) \equiv E\big[y|w,\nu=u,z=t\big]\cdot P(z=t|w,\nu=u) + P(z\neq t|w,\nu=u)$$

So far we have not used the IV assumption. Under this assumption, $E\big[y(t)|w,\nu=u\big] = E\big[y(t)|w\big]\ \forall u\in\mathcal{V}$. This implies that each value of the instrument generates bounds on the same quantity of interest $E\big[y(t)|w\big]$, and we can obtain the tightest upper (lower) bounds on it by sweeping over the instrument values to obtain the smallest (largest) values. In other words, we can bound $E\big[y(t)|w\big]$ from above and below by quantities that are known given the joint distribution of the observables:

$$\max_{u\in\mathcal{V}} \underline{b}(w,u,t) \leq E\big[y(t)|w\big] \leq \min_{u\in\mathcal{V}} \overline{b}(w,u,t) \tag{8}$$

Condition (8) defines a lower bound, $LB^t \equiv \max_{u\in\mathcal{V}} \underline{b}(w,u,t)$, and an upper bound, $UB^t \equiv$

$\max_{u \in \mathcal{V}} \underline{b}(w,u,t)$, on the average value of the response function, evaluated at a specific treatment level t. Simply put, this latter quantity is the probability of the binary outcome (e.g, referral to a specialist) taking place given a specific workload level. Recalling that $t = 0, 1$ values correspond to low and high workload levels, respectively, these results provide bounds on the ATE:

$$[LB^1 - UB^0, UB^1 - LB^0] \tag{9}$$

Note that LB^1, UB^0, UB^1, and LB^0 are easily estimated using nonparametric methods. A confidence interval on the ATE can then be computed following, say, Kreider et al. (2012) (in progress).

5.1.2 Bounds under MIV

By the weaker MIV assumption, for any $u_1 \leq u \leq u_2$, we have:

$$E\big[y(t)|w, \nu = u_1\big] \leq E\big[y(t)|w, \nu = u\big] \leq E\big[y(t)|w, \nu = u_2\big]$$

Combining with (7), we obtain for any $u_1 \leq u \leq u_2$:

$$\underline{b}(w, u_1, t) \leq E\big[y(t)|w, \nu = u\big] \leq \overline{b}(w, u_2, t)$$

Sweeping over all values (u_1, u_2) that satisfy the inequalities, we can thus obtain sharp bounds on $E\big[y(t)|w, \nu = u\big]$:

$$\sup_{u_1 \leq u} \underline{b}(w, u_1, t) \leq E\big[y(t)|w, \nu = u\big] \leq \inf_{u_2 \geq u} \overline{b}(w, u_2, t) \tag{10}$$

But we are interested in bounds on $E\big[y(t)|w\big]$. To eliminate the conditioning on the instrument ν, and exploiting the fact that it only takes the values 0 and 1 in our application, we can write, again by the Law of Iterated Expectations:

$$E\big[y(t)|w\big] = \sum_{u\in\mathcal{V}} E\big[y(t)|w, \nu = u\big] \cdot Pr(\nu = u)$$

By substituting the bounds from (10), we obtain the following:

$$\sum_{u\in\mathcal{V}} \left[Pr(\nu = u) \sup_{u_1\leq u} \underline{b}(w, u_1, t)\right] \leq E\big[y(t)|w\big] \leq \sum_{u\in\mathcal{V}} \left[Pr(\nu = u) \inf_{u_2\geq u} \overline{b}(w, u_2, t)\right] \tag{11}$$

Equation (11) provides bounds on our object of interest under the MIV assumption. Bounding the ATE follows in the same fashion as described above for the IV case.

5.2 Nonparametric Bounds Estimates

In this section we report the bound estimates. We focus attention on the main finding reported in our baseline analysis: the negative effect of workload on the utilization of diagnostic inputs. Figures 4 provides the graphic illustration of the bounds under IV assumption and Panel (A) of Table 10 provides the estimates.

Panel (a) of the figure displays the bounds on the probability of the outcome, i.e., the utilization of any diagnostic input. The vertical red lines and blue lines show the bounds under low workload and high workload, respectively. The six pairs of red and blue bounds in the figure each correspond to a different subsample, as indicated on the x-axis. The subsamples match the heterogeneity specification in the regression analysis to allow comparisons between the regression and bounds results. Panel (b) of the figure displays the bounds on the ATE–moving from low to high workload. Formally, they correspond to different conditioning covariates, denoted by w in the derivations above.

At first glance, the general impression is that, indeed, moving from low to high workload decreases utilization of diagnostic inputs, as the regression analysis suggested: in all subsamples, the upper bound on the estimated probability given low workload lies above the upper bound given high workload, and the same holds for the lower bound. Nonetheless, it is not always possible to sign the treatment effect, as the intervals often overlap. More specifically, let us examine the results for each subsample moving from left to right. The first subsample is, in fact, the full sample, representing patients of all ages. In this case, one cannot sign the effect of workload. This can be seen in Panel (a) because the two bounds overlap, or in Panel (b) as the bounds on the ATE contain the value of zero. The numbers in Table 10 confirm this impression: the lower bound for low workload is 0.254 while the upper bound for High workload is 0.295.

We note that the results presented here are estimated bounds computed using the derivations in the previous subsection. Making statistical statements regarding the significance of the sign of the effects requires confidence intervals on these estimated bounds (in progress).

Moving to the second subsample, those aged 60 or less, the bounds are broader with a larger overlap, indicating that the effect of workload, once again, cannot be signed. By contrast, the bounds on the third subsample, patients aged more than sixty, are much narrower and it is apparent that they do not overlap. The figures in Table 10 confirm this. The lower bound under low workload is 0.381 and the upper bound under high workload is 0.267. Therefore, in this case it is possible to conclude that the workload has a negative effect on utilization of any diagnostic inputs, as the ATE bounds in Panel (b) show. The ATE in the case is bounded between $[-0.114, -0.247]$. In the low utilization subsample, it is evident that the bounds do not provide a sharp prediction. In the case of the high utilization patients, however, it is possible to conclude that the effect of workload is negative and bounded between $[-0.019, 0.250]$ as Table 10 indicates. Finally, the subsample of patients who are both older and characterized by high utilization also clearly indicates a negative effect of workload with quite tight bounds at $[-0.098, 0.236]$.

Figures 5 provides the graphic illustration of the bounds under MIV assumption and Panel (B) of Table 10 provides the corresponding estimates. In this case, under all subsamples, it is not possible to sign the ATE. Nonetheless, the MIV results do seem to be largely consistent with our previous findings. For the subsamples of patients of age above 60, and those who are both above 60 and high utilization patients, the estimated intervals under low and high workload barely overlap. The relevant literature often combines the conservative MIV assumption with a Monotone Treatment Response assumption, further tightening the bounds. In this case, at least for these two subsamples, it seems likely that this approach (in progress) would allow us to sign the treatment effect, just as we did in the IV case above.

Taken together, the bounds results presented in this section largely support the findings of our baseline analysis that relied on a homogeneous-effect linear model: the effect of workload on the utilization of diagnostic inputs emerges as negative, particularly so for older and higher-utilization patients.

6 Conclusions

In this study we examine the effect of workload on physician behavior. We find that when workload is higher physicians respond less to online patient queries and make less phone calls with patients. Physicians' utilization of diagnostic inputs decreases, mostly, referrals to specialists and to lab tests. Finally, increased workload does not affect the tendency of physicians to refer patients to the emergency room or to prescribe pain killers. However, workload is associated, to some extent, with prescribing more antibiotics.

These results show that at a higher workload physicians change their practice style and the treatment they provide. These changes should be accounted for when optimal workload and size of workforce in healthcare are set.

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.

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

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

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.

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.

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. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

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

Justin McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2):698–714, 2008.

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.

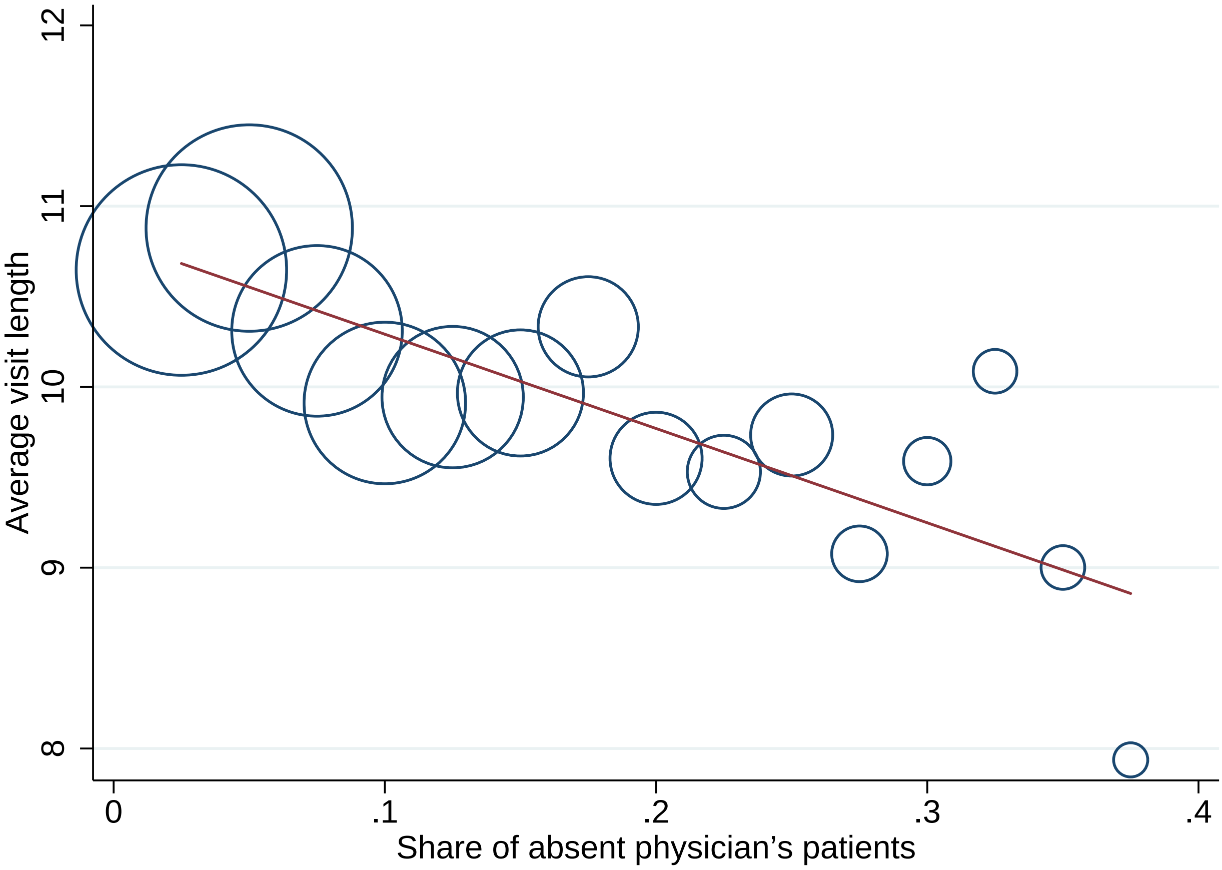
Anthony Scott. Chapter 22 economics of general practice. volume 1, Part B of *Handbook of Health Economics*, pages 1175 – 1200. Elsevier, 2000. doi: http://dx.doi.org/10.1016/S1574-0064(00)80035-9. URL `http://www.sciencedirect.com/science/article/pii/S1574006400800359`.

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.

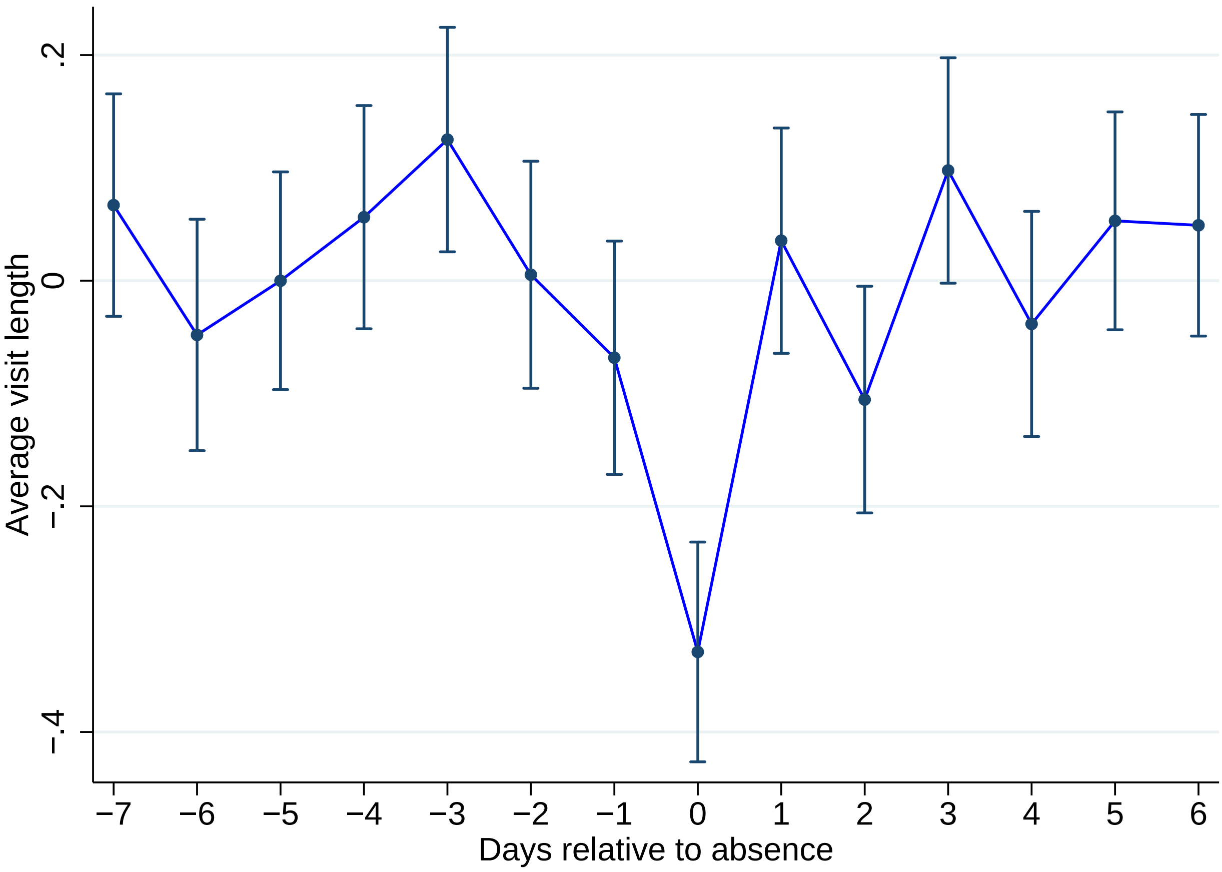
Tom Fangyun Tan and Serguei Netessine. When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science*, 60(6):1574–1593, 2014.

Figure 1: Share of missing physician's patients and workload



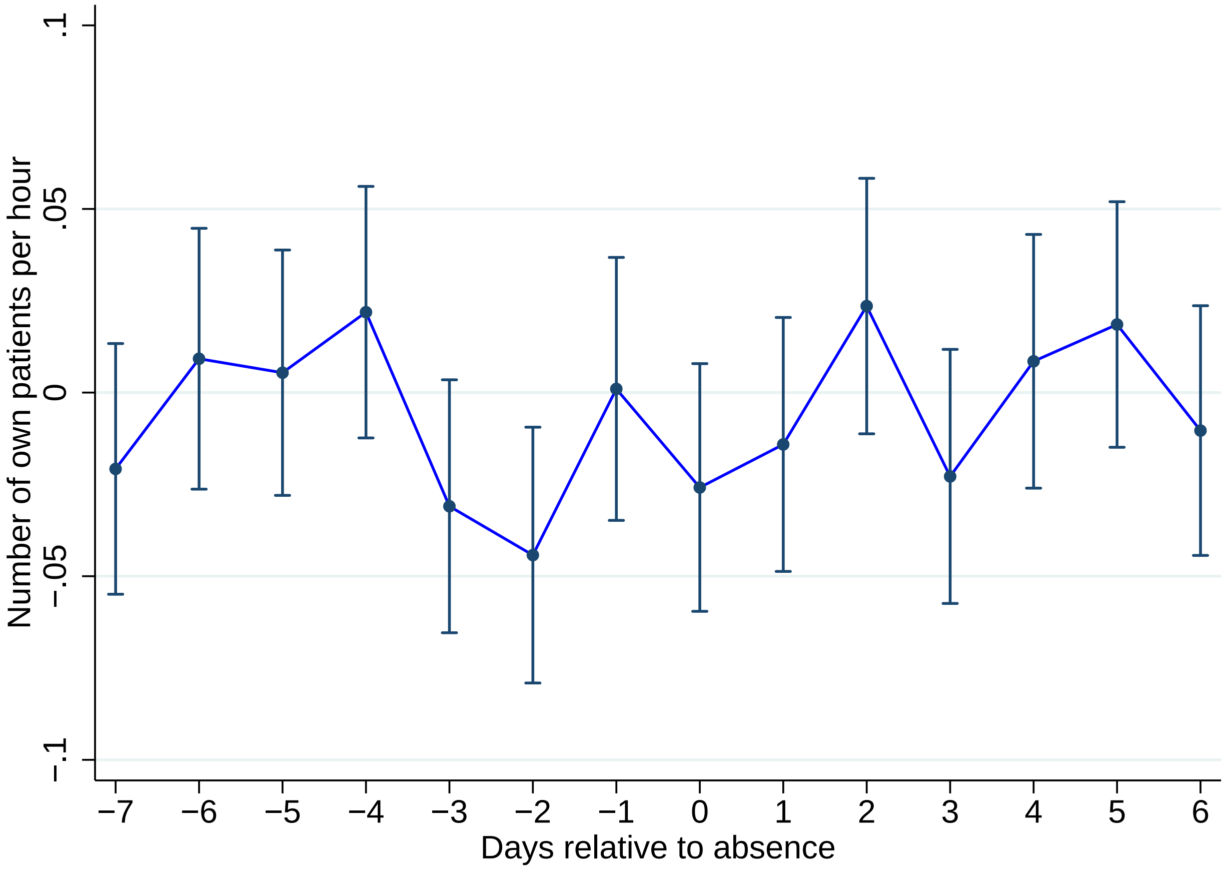
Note: The figure plots the mean visit length for bins of share of the missing physician's patients. The superimposed line is the predicted relation between absence and workload.

Figure 2: 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 3. The dependent variable is mean visit length.

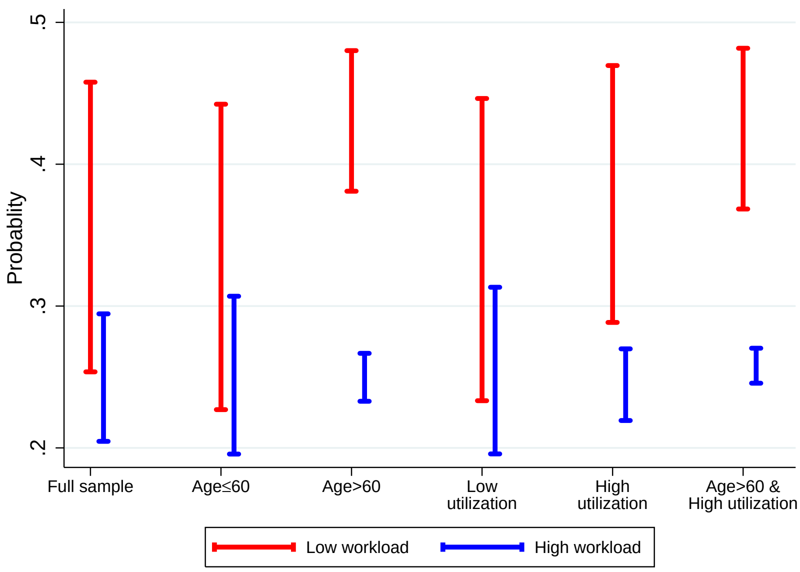
Figure 3: Number of own patients per hour around days of a colleague's absence



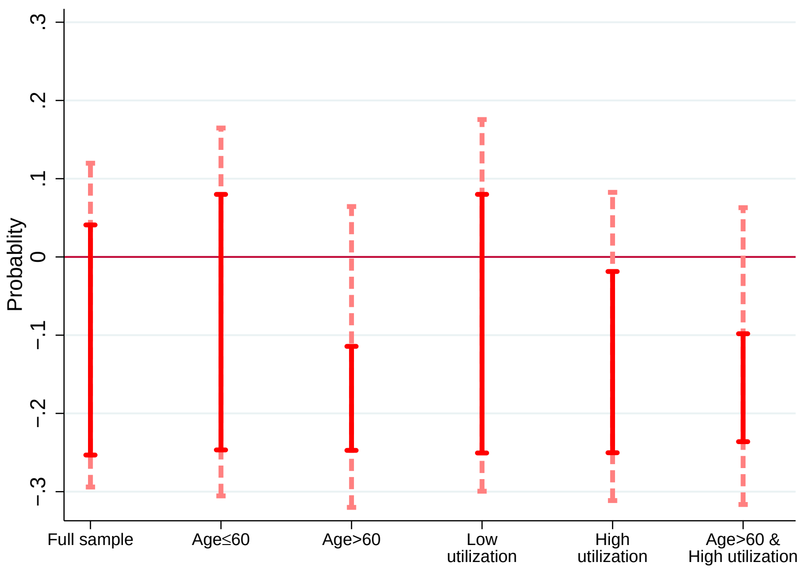
Note: The figure plots the coefficients and standard errors from the event study model akin to the model in Equation 3. The dependent variable is the number of a physician's own patient per hour.

Figure 4: Bounds on the effect of workload on the utilization of diagnostic inputs, IV

(a) All



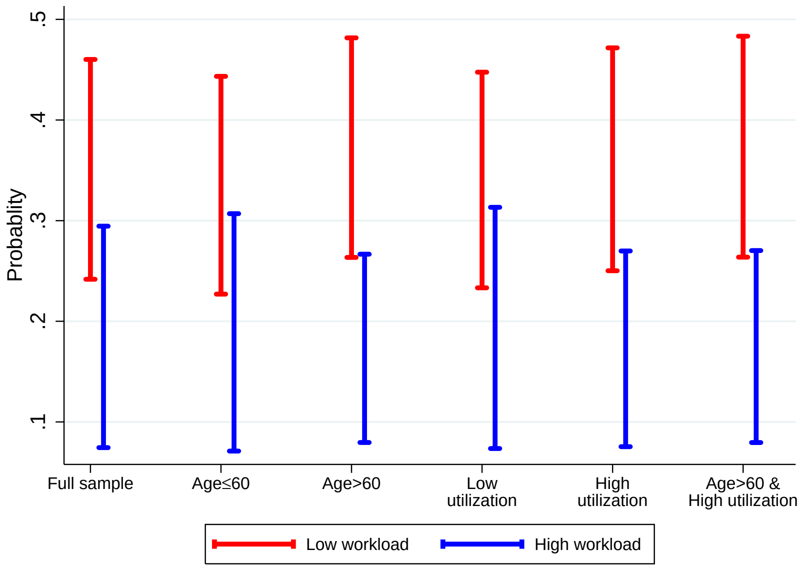
(b) ATE



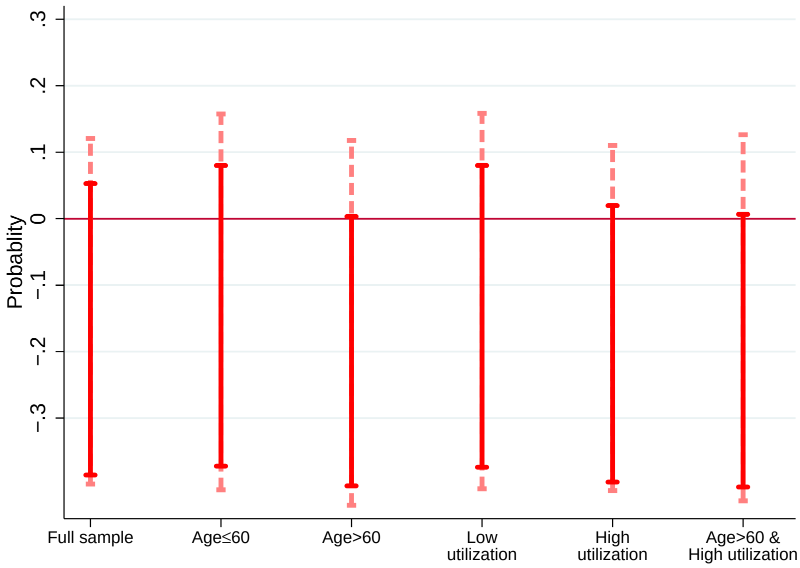
Note: Panels (a) and (b) of this figure report the estimates from the bounds analysis under the IV assumption.

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

(a) Low and high workload



(b) ATE



Note: Panels (a) and (b) of this figure report the estimates from the bounds analysis under the MIV assumption.

Table 1: Summary statistics of visits data

Patient characteristics	
Mean age	47.60
Share women	0.58
Share born in Israel	0.61
Share smokers	0.30
Share obese	0.26
Share hypertension	0.34
Share hyperlipidemia	0.45
Share ischemic heart disease	0.15
Office visits characteristics	
Visit length	11.56
Share referral to specialist	0.14
Share referral to imaging	0.08
Share referral to lab tests	0.20
Share referrals to ER	0.01
Share Painkiller	0.05
Share antibiotics	0.10
Number of patients	121,622
Number of physicians	98
Observations	825,660

Notes: The table includes face-to-face visits in the clinics used in this study in the period 2011-2014.

Table 2: 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)
B. IV 2		
Seeing absent physician's patients	-0.63**	-0.62**
	(0.05)	(0.05)
Year-month, day & physician FE	Yes	Yes
Patient age, gender & condition controls	No	Yes
Observations	825,658	823,349

Notes: All columns report estimates of effect of absence on workload, as per Equation (4). The Year-month fixed effects consists of a dummy variable for each of the calendar months in our data. Standard errors clustered by physician-day are reported in parentheses. One or two asterisks indicate significance at 5% or 1%, respectively.

Table 3: 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 =0.35)						
Mean visit length	0.48**	0.45**	1.62**	1.62**	1.85**	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)						
Mean visit length	0.16**	0.16**	0.76*	0.77*	0.91*	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)
Year-month, day & physician FE	Yes	Yes	Yes	Yes	Yes	Yes
Patient age, gender & condition controls	No	Yes	No	Yes	No	Yes
Observations	825,658	823,349	825,658	823,349	825,658	823,349

Notes: Panels (A), (B), (C) and (D) of this table report estimates of effect of workload on the probability of utilization of any of the diagnostic inputs, referral to a specialist, referral to a lab test and referral to imaging, respectively , as per Equation (5). The Year-month fixed effects consists of a dummy variable for each of the calendar months in our data. 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, by patient group

	Age>60			Age≤60			High utilization			Low utilization		
	OLS	IV1	IV2	OLS	IV1	IV2	OLS	IV1	IV2	OLS	IV1	IV2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
A. Dependent Variable: all diagnostic inputs												
	(mean=0.39)			(mean =0.33)			(mean=0.34)			(mean=0.34)		
Mean visit length	0.44**	2.04**	1.78*	0.45**	1.39**	1.81**	0.44**	1.86**	2.13**	0.44**	1.45**	1.57**
	(0.04)	(0.63)	(0.70)	(0.03)	(0.41)	(0.52)	(0.03)	(0.52)	(0.62)	(0.03)	(0.47)	(0.57)
B. Dependent Variable: referral to a specialist												
	(mean=0.17)			(mean=0.13)			(mean=0.17)			(mean=0.12)		
Mean visit length	0.35**	1.82**	1.94**	0.27**	0.81**	0.95**	0.34**	1.47**	1.72**	0.27**	0.88**	1.01**
	(0.03)	(0.49)	(0.53)	(0.02)	(0.26)	(0.35)	(0.03)	(0.40)	(0.48)	(0.02)	(0.29)	(0.37)
C. Dependent Variable: referral to a lab test												
	(mean=0.20)			(mean=0.19)			(mean=0.19)			(mean=0.20)		
Mean visit length	0.12**	1.28*	0.76	0.18**	0.46	0.92*	0.12**	1.06*	1.08*	0.18**	0.52	0.73
	(0.03)	(0.54)	(0.57)	(0.03)	(0.35)	(0.44)	(0.03)	(0.43)	(0.50)	(0.03)	(0.40)	(0.48)
D. Dependent Variable: referral to imaging												
	(mean=0.09)			(mean=0.07)			(mean=0.09)			(mean=0.07)		
Mean visit length	0.22**	0.24	0.36	0.19**	0.05	-0.01	0.22**	0.34	0.55	0.18**	-0.08	-0.23
	(0.02)	(0.38)	(0.39)	(0.02)	(0.19)	(0.26)	(0.02)	(0.29)	(0.35)	(0.02)	(0.22)	(0.29)
Year-month, day & physician FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Patient age, gender & condition controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	332,911	332,911	332,911	490,438	490,438	490,438	407,465	407,465	407,465	415,884	415,884	415,884

Notes: Panels (A), (B), (C) and (D) of this table report estimates of effect of workload on the probability of utilization of any of the diagnostic inputs, referral to a specialist, referral to a lab test and referral to imaging, respectively , as per Equation (5), by subsamples. Columns (1)-(6) report the results by age and Columns (7)-(12) report the results by patient condition. The Year-month fixed effects consists of a dummy variable for each of the calendar months in our data. 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)						
Mean visit length	0.04*	0.01	-0.48	-0.52	-0.80*	-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. 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	825,658	823,349	825,658	823,349	825,658	823,349

Notes: Panels (A), (B), (C) and (D) of this table report estimates of effect of workload on the probability of referral to any of the treatments, referral to the emergency room, prescription of pain killers and prescription of antibiotics, respectively, as per Equation (5). The Year-month fixed effects consists of a dummy variable for each of the calendar months in our data. 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 treatment decision, by patient group

	Age>60			Age≤60			High utilization			Low utilization		
	OLS	IV1	IV2	OLS	IV1	IV2	OLS	IV1	IV2	OLS	IV1	IV2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
A. Dependent Variable: all treatments												
	(mean=0.16)			(mean=0.15)			(mean=0.16)			(mean=0.15)		
Mean visit length	-0.02	-0.89	-1.25*	0.03	-0.30	-0.57	-0.04	-0.41	-0.76	0.05*	-0.58	-0.98*
	(0.03)	(0.48)	(0.53)	(0.02)	(0.31)	(0.39)	(0.03)	(0.40)	(0.47)	(0.02)	(0.34)	(0.42)
B. Dependent Variable: referral to the emergency room												
	(mean=0.01)			(mean=0.01)			(mean=0.02)			(mean=0.01)		
Mean visit length	0.05**	-0.15	-0.15	0.03**	0.07	-0.03	0.06**	-0.06	-0.13	0.02**	0.04	-0.02
	(0.01)	(0.16)	(0.17)	(0.01)	(0.08)	(0.11)	(0.01)	(0.13)	(0.16)	(0.01)	(0.08)	(0.11)
C. Dependent Variable: prescription of pain killers												
	(mean=0.07)			(mean=0.03)			(mean=0.07)			(mean=0.03)		
Mean visit length	0.02	-0.65	-0.72*	0.03*	0.06	0.00	0.00	-0.17	-0.29	0.04**	-0.20	-0.33
	(0.02)	(0.33)	(0.36)	(0.01)	(0.14)	(0.19)	(0.02)	(0.26)	(0.31)	(0.01)	(0.15)	(0.20)
D. prescription of antibiotics												
	(mean=0.08)			(mean=0.11)			(mean=0.08)			(mean=0.11)		
Mean visit length	-0.08**	-0.13	-0.43	-0.02	-0.42	-0.61	-0.10**	-0.22	-0.41	-0.00	-0.40	-0.68
	(0.02)	(0.35)	(0.38)	(0.02)	(0.28)	(0.34)	(0.02)	(0.31)	(0.35)	(0.02)	(0.31)	(0.37)
Year-month, day & physician FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Patient age, gender & condition controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	332,911	332,911	332,911	490,438	490,438	490,438	407,465	407,465	407,465	415,884	415,884	415,884

Notes: Panels (A), (B), (C) and (D) of this table report estimates of effect of workload on the probability of referral to any of the treatments, referral to the emergency room, prescription of pain killers and prescription of antibiotics, respectively, as per Equation (5), by subsamples. Columns (1)-(6) report the results by age and Columns (7)-(12) report the results by patient condition. The Year-month fixed effects consists of a dummy variable for each of the calendar months in our data. Standard errors clustered by physician-day are reported in parentheses. One or two asterisks indicate significance at 5% or 1%, respectively.

Table 7: The effect of workload on subsequent encounters

	OLS		IV 1		IV 2	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Dependent Variable: subsequent visit within 15 days (mean =0.38)						
Mean visit length	0.11**	0.09**	-0.44	-0.49	-0.44	-0.44
	(0.03)	(0.02)	(0.36)	(0.35)	(0.44)	(0.44)
B. Dependent Variable: subsequent visit within 30 days (mean =0.56)						
Mean visit length	0.14**	0.11**	-0.50	-0.56	-0.59	-0.58
	(0.03)	(0.03)	(0.37)	(0.36)	(0.45)	(0.45)
C. Dependent Variable: subsequent visit within 60 days (mean =0.73)						
Mean visit length	0.11**	0.08**	-0.59	-0.65*	-0.39	-0.39
	(0.02)	(0.02)	(0.33)	(0.32)	(0.39)	(0.38)
D. Dependent Variable: subsequent visit within 90 days (mean =0.81)						
Mean visit length	0.09**	0.07**	-0.50	-0.56*	-0.53	-0.54
	(0.02)	(0.02)	(0.29)	(0.29)	(0.34)	(0.33)
Year-month, day & physician FE	Yes	Yes	Yes	Yes	Yes	Yes
Patient age, gender & condition controls	No	Yes	No	Yes	No	Yes
Observations	825,658	823,349	825,658	823,349	825,658	823,349

Notes: Panels (A), (B), (C) and (D) of this table report estimates of effect of workload on the likelihood of a subsequent visit. The Year-month fixed effects consists of a dummy variable for each of the calendar months in our data. Standard errors clustered by physician-day are reported in parentheses. One or two asterisks indicate significance at 5% or 1%, respectively.

Table 8: The effect of workload on subsequent encounters, by patient group

	Age>60			Age≤60			High utilization			Low utilization		
	OLS	IV1	IV2	OLS	IV1	IV2	OLS	IV1	IV2	OLS	IV1	IV2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
A. Dependent Variable: subsequent visit within 15 days												
	(mean=0.30)			(mean =0.22)			(mean=0.31)			(mean=0.20)		
Mean visit length	0.07	-1.13	-1.36	-0.01	-0.15	-0.08	0.02	-1.14*	-1.21	0.02	-0.01	-0.06
	(0.04)	(0.65)	(0.71)	(0.03)	(0.37)	(0.48)	(0.03)	(0.53)	(0.64)	(0.03)	(0.41)	(0.49)
B. Dependent Variable: subsequent visit within 30 days												
	(mean=0.49)			(mean =0.34)			(mean=0.50)			(mean=0.31)		
Mean visit length	0.11**	-0.92	-0.42	-0.00	0.20	0.04	0.06	-0.78	-0.70	0.01	0.25	0.30
	(0.04)	(0.68)	(0.76)	(0.03)	(0.41)	(0.54)	(0.04)	(0.58)	(0.69)	(0.03)	(0.47)	(0.57)
C. Dependent Variable: subsequent visit within 60 days												
	(mean=0.68)			(mean =0.49)			(mean=0.68)			(mean=0.45)		
Mean visit length	0.07	-0.75	-0.72	-0.00	-0.40	-0.00	0.02	-0.34	-0.23	0.00	-0.69	-0.32
	(0.04)	(0.65)	(0.67)	(0.03)	(0.45)	(0.54)	(0.03)	(0.53)	(0.59)	(0.04)	(0.50)	(0.59)
D. Dependent Variable: subsequent visit within 90 days												
	(mean=0.77)			(mean =0.59)			(mean=0.78)			(mean=0.55)		
Mean visit length	0.05	-0.29	-0.19	0.00	-0.31	-0.42	-0.01	-0.15	-0.23	0.03	-0.45	-0.42
	(0.03)	(0.56)	(0.60)	(0.03)	(0.47)	(0.54)	(0.03)	(0.47)	(0.52)	(0.04)	(0.52)	(0.58)
Year-month, day & physician FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Patient age, gender & condition controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	332,911	332,911	332,911	490,438	490,438	490,438	407,465	407,465	407,465	415,884	415,884	415,884

Notes: Panels (A), (B), (C) and (D) of this table report estimates of effect of workload on the likelihood of a subsequent visit, by subsamples. Columns (1)-(6) report the results by age and Columns (7)-(12) report the results by patient condition. The Year-month fixed effects consists of a dummy variable for each of the calendar months in our data. Standard errors clustered by physician-day are reported in parentheses. One or two asterisks indicate significance at 5% or 1%, respectively.

Table 9: The effect of workload on non face-to-face encounters with patients

	OLS	IV 1	IV 2
	(1)	(2)	(3)
A. Dependent Variable:			
Response to online patient queries per hour (mean = 1.76)			
Mean visit length	-0.06***	0.07***	0.08***
	(0.00)	(0.03)	(0.03)
B. Dependent Variable:			
Phone calls with patients per hour (mean = 0.31)			
Mean visit length	-0.01**	0.03**	0.03**
	(0.00)	(0.01)	(0.01)
Observations	43,487	43,487	43,487

Notes: Panels (A) and (B) of this table report estimates of effect of workload on the number of online patients queries and phone calls with patients per hour, respectively, as per Equation (5). The Year-month fixed effects consists of a dummy variable for each of the calendar months in our data. One or two asterisks indicate significance at 5% or 1%, respectively.

Table 10: The effect of workload on utilization of diagnostic inputs, bounds

	Low workload		High workload		ATE (Low to High)		ATE CI	
	(Lower)	(Upper)	(Lower)	(Upper)	(Lower)	(Upper)	(Lower)	(Upper)
A. IV								
Full sample	0.254	0.458	0.205	0.295	-0.253	0.041	-0.294	0.120
Age $\leq$ 60	0.227	0.442	0.196	0.307	-0.247	0.080	-0.305	0.165
Age>60	0.381	0.480	0.233	0.267	-0.247	-0.114	-0.320	0.064
Low utilization	0.233	0.446	0.196	0.313	-0.251	0.080	-0.299	0.176
High utilization	0.288	0.470	0.219	0.270	-0.250	-0.019	-0.311	0.083
Age>60 & high utilization	0.368	0.482	0.246	0.270	-0.236	-0.098	-0.316	0.063
B. MIV								
Full sample	0.242	0.460	0.074	0.295	-0.386	0.053	-0.399	0.120
Age $\leq$ 60	0.227	0.443	0.071	0.307	-0.372	0.080	-0.408	0.158
Age>60	0.263	0.482	0.080	0.267	-0.402	0.003	-0.431	0.117
Low utilization	0.233	0.448	0.074	0.313	-0.374	0.080	-0.406	0.158
High utilization	0.250	0.472	0.075	0.270	-0.396	0.020	-0.409	0.110
Age>60 & high utilization	0.264	0.483	0.079	0.270	-0.404	0.007	-0.425	0.126

Notes: Panels (A), (B) and (C) of this table report the estimates from the bounds analysis. Sample sizes are the same as those reported for the corresponding subsamples in Tables 3-6.

A Appendix A

A.1 The reduced form regressions

Here we report the reduced form regression. Namely, we estimate the following equation:

$$y_{jsti} = \alpha + \nu_j + \nu_t + \beta_1 \cdot sa_{jst} + x_{jsti} \cdot \beta_2 + \epsilon_{jsti} \tag{A1}$$

Table A.1 reports the results for the diagnostic outcomes. As the table shows, consistent with the analysis in Section 4.3.1, the estimates in Panels (A) - (C) are negative and significant and the results in Panel (D) - referral to imaging are negative yet statistically insignificant.

Table A.2 reports the results for treatment outcomes. All estimates are statistically insignificant, in line with the estimates in Section 4.3.2, that do not show that a strong relationship between workload and treatment choice exists.

Table A.3 reports the results for the likelihood of subsequent visits. All estimates are statistically insignificant. This is quite consistent with the finding in Section 4.4 which indicate a small and mostly insignificant effect on subsequent visits.

Table A.4 reports the non face-to-face encounter results. The estimates are all negative and significant, in line with the results in section 4.5.

Table A.1: The effect of workload on utilization of diagnostic inputs, reduced form

	IV 1		IV 2	
	(1)	(2)	(3)	(4)
A. Dependent Variable: all diagnostic inputs				
	-7.80**	-7.93**	-1.16**	-1.16**
	(1.70)	(1.71)	(0.27)	(0.27)
B. Dependent Variable: referral to a specialist				
	-5.42**	-5.48**	-0.84**	-0.84**
	(1.20)	(1.19)	(0.20)	(0.20)
C. Dependent Variable: referral to a lab test				
	-3.65*	-3.65*	-0.57**	-0.56*
	(1.46)	(1.46)	(0.22)	(0.22)
D. Dependent Variable: referral to imaging				
	-0.51	-0.56	-0.10	-0.10
	(0.87)	(0.87)	(0.14)	(0.14)
Year-month, day & physician FE	Yes	Yes	Yes	Yes
Patient age, gender & condition controls	No	Yes	No	Yes
Observations	825,660	823,351	825,660	823,351

Notes: Panels (A), (B), (C) and (D) of this table report estimates of effect of absence on the probability of referral to any of the diagnostic inputs, referral to a specialist, referral to a lab test and referral to imaging, respectively , as per Equation (A1). The Year-month fixed effects consists of a dummy variable for each of the calendar months in our data. Standard errors clustered by physician-day are reported in parentheses. One or two asterisks indicate significance at 5% or 1%, respectively.

Table A.2: The effect of workload on treatment choice, reduced form

	IV 1		IV 2	
	(1)	(2)	(3)	(4)
A. Dependent Variable: all treatments				
	2.32	2.35	0.50*	0.52**
	(1.28)	(1.29)	(0.20)	(0.20)
B. Dependent Variable: referral to the emergency room				
	-0.02	-0.03	0.04	0.04
	(0.36)	(0.36)	(0.06)	(0.06)
C. Dependent Variable: prescription of pain killers				
	0.80	0.93	0.18	0.19
	(0.70)	(0.70)	(0.11)	(0.11)
D. prescription of antibiotics				
	1.56	1.47	0.32	0.33*
	(1.10)	(1.10)	(0.16)	(0.16)
Year-month, day & physician FE	Yes	Yes	Yes	Yes
Patient age, gender & condition controls	No	Yes	No	Yes
Observations	825,660	823,351	825,660	823,351

Notes: Panels (A), (B), (C) and (D) of this table report estimates of effect of absence on the probability of referral to any of the treatments, referral to the emergency room, prescription of pain killers and prescription of antibiotics, respectively, as per Equation (A1). The Year-month fixed effects consists of a dummy variable for each of the calendar months in our data. Standard errors clustered by physician-day are reported in parentheses. One or two asterisks indicate significance at 5% or 1%, respectively.

Table A.3: The effect of workload on the likelihood of subsequent encounters, reduced form

	IV 1		IV 2	
	(1)	(2)	(3)	(4)
A. Dependent Variable: subsequent visit within 15 days				
	2.13	2.19	0.37	0.34
	(1.69)	(1.68)	(0.27)	(0.27)
B. Dependent Variable: subsequent visit within 30 days				
	0.55	0.74	0.10	0.08
	(1.86)	(1.85)	(0.30)	(0.29)
C. Dependent Variable: subsequent visit within 60 days				
	2.03	2.38	0.16	0.15
	(1.92)	(1.89)	(0.28)	(0.27)
D. Dependent Variable: subsequent visit within 90 days				
	0.97	1.40	0.19	0.19
	(1.89)	(1.85)	(0.27)	(0.26)
Year-month, day & physician FE	Yes	Yes	Yes	Yes
Patient age, gender & condition controls	No	Yes	No	Yes
Observations	825,660	823,351	825,660	823,351

Notes: Panels (A), (B), (C) and (D) of this table report estimates of effect of absence on the likelihood of a subsequent visit. The Year-month fixed effects consists of a dummy variable for each of the calendar months in our data. Standard errors clustered by physician-day are reported in parentheses. One or two asterisks indicate significance at 5% or 1%, respectively.

Table A.4: The effect of workload on non face-to-face encounters, reduced form

	IV 1	IV 2
	(1)	(2)
A. Dependent Variable:		
Response to online patients queries per hour (mean =1.76)		
	-0.340**	-0.061**
	(0.119)	(0.021)
B. Dependent Variable:		
Phone calls to patients per hour (mean =0.31)		
	-0.149**	-0.024**
	(0.039)	(0.007)
Year-month, day & physician FE	Yes	Yes
Observations	43,489	43,489

Panels (A) and (B) of this table report estimates of effect of absence on the number of online patients queries and phone calls with patients per hour, respectively, as per Equation (A1). The Year-month fixed effects consists of a dummy variable for each of the calendar months in our data. One or two asterisks indicate significance at 5% or 1%, respectively.

B Appendix B

To create the "utilization score", we analyze the data at the patient year level. for each patient and each year we count the number of visits that patient made to a physician. This is the $\# - of - Visits_{it}$, the left hand side variable. We regress this measure of utilization against the set of patient personal characteristics using a model of the form:

$$\# - of - Visits_{it} = \alpha + X_{it} \cdot \beta + \epsilon_{it} \tag{A2}$$

Based on the results of the regression we create a predicted $\# - of - Visits_{it}$ for each patient. We use this variable as the "utilization score" for each patient in the data. As we noted above, this score reflects the number of visits per time period that a patient with these characteristics would have, on average, in a given time period.

Physician workload and treatment choice: the case of primary care

16 November 2017, 11:15-12:30, bld. 72, room 465

Ity Shurtz (HUJI)

Abstract: We examine how primary care physicians' treatment choices respond to physician workload, using detailed administrative data from eleven clinics of a large Israeli HMO. We use absences of colleagues at the clinic as a source of an exogenous increase in the physician's workload. Using a standard homogeneous-effects linear model, we find that physician time and utilization of diagnostic inputs are complements: during face-to-face visits, a one minute decrease in average (daily) visit length causes a 9 percent decrease in referrals to specialists, and a 3.8 percent decrease in referrals to lab tests. We find much smaller effects on the choice of treatment prescribed during the visit: our results imply no significant impact of workload on referrals to the emergency room, or on the prescription of painkillers, though there is some evidence that higher workload causes an increased prescription of antibiotics. Finally, when physicians experience higher workload they decrease the amount of non face-to-face encounters with patients. Our results are robust to relaxing the linearity and homogeneous-effects assumptions: following Manski and Pepper (2000), we compute nonparametric bounds on the Average Treatment Effects, resulting in qualitatively similar findings. Relaxing the exogeneity assumption of the instrument following a Monotone Instrumental Variable approach also results in similar conclusions. Our analysis provides important lessons to insurers and policy makers alike, as they reveal the channels via which practitioners respond to increased pressure brought about by limited capacity (the \primary care crunch"). In particular, we confirm that increased workload impairs primary care clinicians' ability to deliver preventive care, one of the key aspects of managed care health systems.