Working Hard or Hardly Working: Health Worker Effort and Health Outcomes

Edward N. Okeke*

Abstract

Effort by health workers in developing countries has been the subject of much recent attention, but the implications, if any, for health outcomes have remained largely a matter of conjecture. This paper provides new empirical evidence about the returns to health worker effort. In this experiment, outpatient providers in Nigeria were assigned to an intervention in which their consultations were monitored by a peer. To elicit behavior under the counterfactual, the same group of clinicians were covertly observed. Exploiting the sharp increase in effort induced by peer monitoring, I document large, positive, and statistically significant improvements in health. Using these estimates, I benchmark the cost of low effort to be around $350 million annually (or approximately 0.1% of Nigeria's GDP). These findings shed new light on potential gains from incentivizing health worker effort.

Key words: health workers; effort; health outcomes

JEL Codes: I11, I15

*RAND Corporation, 1200 South Hayes, Arlington, VA 22202. Email: eokeke@rand.org. Tel: (703) 413-1100.
This study was supported by a grant from the Eunice Kennedy Shriver National Institute for Child Health and Human Development (R01HD083444). I am grateful to all the field personnel – the field managers, supervisors, and data collectors, who worked very hard to implement this study. I am also grateful to all of the research participants who were generous with their time. Excellent research assistance was provided by Adeyemi Okunogbe and Juliana Chen-Peraza.

1 Introduction

Effort by health providers, primarily in developing countries, has been the subject of much recent attention (Das et al., 2016; Mohanan et al., 2015; Huillery and Seban, 2017; Das et al., 2012; Brock et al., 2018). Much of this extant literature finds that health workers often do much less than they are capable of.[1] Low effort has been documented along both intensive and extensive margins. Absenteeism, which reflects effort on the extensive margin, is pervasive – Chaudhury et al. (2006) found absenteeism rates for health workers ranging from 25 to 40 percent in six countries – and on the intensive margin, when health workers come to work, their interactions with patients are often short and cursory. In one study in India, as a vivid example, one-third of patient consultations lasted less than one minute, with the provider asking just one question *"What is wrong with you?"*, and carrying out no physical examinations (Das and Hammer, 2014). Similar findings have been documented in other countries (Das et al., 2008).

Low effort is not unique to the health sector; it has been shown in other sectors such as education (Banerjee and Duflo, 2006),[2] but effort by health providers is uniquely important because it can have literal life-or-death implications. Sick patients often have no idea what ails them and are completely dependent on the health care provider; in most cases, they have no idea whether the provider is doing the right thing for them or not.[3] *Ceteris paribus*, a health provider exerting low effort is more likely to miss something important, make the wrong diagnosis and mismanage the patient.[4] This is likely to have both short and long run implications. In the short run, individuals may remain sick for longer, impacting productivity and earnings (Dillon et al., 2014; Dobkin et al., 2018); they may incur additional costs seeking care from alternative sources, and in the long run, accumulate fewer assets (Poterba et al., 2017), suffer permanent disability and in the extreme, live shorter lives (Nardi et al., 2017).

There is very little empirical research on the relationship between health worker effort and patient outcomes. Bjorkman and Svensson (2009) and Gertler and Vermeersch (2012) study the effect of two

[1] The terms 'health provider', 'health worker', and 'clinician', are used interchangeably throughout this paper.

[2] These settings are often characterized by bureaucratic inertia, poorly functioning (or absent) accountability mechanisms, misaligned incentives, and low expectations. Das et al. (2016) show that the same providers exert more effort, and do better, in their private practices than in their public sector jobs.

[3] Information asymmetry is a well-known feature of health care going back to Arrow (1963).

[4] In these settings wrong diagnoses are common, and a patient is lucky if they receive treatment that is beneficial (Das et al., 2012).

interventions – community monitoring and pay-for-performance in Uganda and Rwanda respectively – and find that both led to improvements in child health. In both cases, they present some evidence that health improvements were mediated, at least in part, by increased health provider effort,[5] though given the complexity of potential causal pathways, one cannot draw a direct line from effort to health.[6] Goldstein et al. (2013) take a different approach, examining the effect of health worker absenteeism, and find that absenteeism by a HIV nurse led to lower uptake of beneficial treatment by patients.[7] However, no previous study to the best of my knowledge, has drawn a direct causal link between health worker effort along the intensive margin and health outcomes.[8]

One of the challenges in answering this question is finding exogenous variation in health worker effort. Despite the increasing number of pay-for-performance experiments and evaluations (Gertler and Vermeersch, 2012; Gertler et al., 2014; Peabody et al., 2013; Huillery and Seban, 2017), they do not quite get at the question of interest because they often move several other levers in addition to incentivizing provider effort, making it a package (or bundled) intervention. A related issue is that these programs often reward specific inputs (as against effort more generally) and multi-tasking theory (see Holmstrom and Milgrom, 1991; Eggleston, 2005) suggests that health workers will re-optimize, shifting effort away from other inputs (some of which might be unobserved) that may also affect health, clouding potential inference.[9] They may also affect provider knowledge, informing them about best practices (or making these more salient) or may lead to organizational changes which affect practice (Celhay et al., 2018).

This paper reports on a novel experiment conducted in Nigeria to try to answer this question. 242 outpatient clinicians across 124 primary health care clinics participated in the experiment. To induce effort, clinicians were assigned to a treatment condition in which their patient interactions were observed by a peer (peer monitoring). In the control condition, the same clinicians were

[5]Bjorkman and Svensson (2009) present evidence of increased effort along both intensive (more thorough examinations) and extensive (lower absenteeism rates) margins.

[6]In the case of the pay-for-performance program evaluated by Gertler and Vermeersch (2012), for example, Basinga et al. (2011) report transfers from clinics/providers to patients to incentivize them to use services because utilization was a rewarded metric.

[7]See also work by Chicoine and Guzman (2017) who showed that uncertainty about when the health provider was present decreased uptake of services and increased duration of child illness.

[8]From a policy standpoint, it is significantly harder to police what a provider does when at work than it is to police coming to work (Banerjee et al., 2008). The recent surge in pay-for-performance interventions in developing countries is aimed at addressing this agency problem by linking compensation to performance (Miller and Babiarz, 2013).

[9]For related empirical evidence in a healthcare and non-healthcare setting respectively, see Dumont et al. (2008) and Hong et al. (2013).

covertly observed to measure behavior under 'normal' conditions (a detailed discussion of the experiment is postponed until later). This experiment therefore generates data on the behavior of each participating clinician under treatment (high effort) and control (status quo effort) conditions. Since each clinician was exposed to both treatment and control conditions – a crossover design – they each serve as their own control. To study the effect of effort on the outcomes of their patients, follow-up phone interviews were conducted with these patients approximately three weeks after the clinic visit.

To summarize the main study findings: first, I show that the treatment led to a large, sustained increase in effort by clinicians. Using the length of the provider-patient interaction as a measure of effort, I show that consultations that were overtly monitored were 21% longer on average (equivalent to the difference between being seen by a doctor vs. a community health worker). Second, I show that longer consultations are associated with more care – more history-taking questions, more physical examinations, and greater use of diagnostic tests – and clinicians are more likely to make a diagnosis and provide treatment. Third, exploiting the increase in effort as a result of the treatment, I find large, positive, and statistically significant improvements in patient health measured using a follow-up survey. The intent-to-treat estimates imply an 8-percentage-point improvement in overall health (measured as the probability that the patient indicated their current state of health to be excellent) and a similarly-sized improvement in functional health (measured as the probability of having difficulty in climbing stairs or running the length of a football field). In the IV specification, using peer monitoring as an instrument for effort, I find that a 10% increase in clinician effort (measured by consultation time) leads to an approximately 4% improvement in the probability of being in excellent health. I carry out various robustness checks including estimating non-parametric Lee bounds to account for differential dropout. Using these estimated effect sizes and making various assumptions, I estimate the cost of low effort in terms of lost wages to be equivalent to about 0.1% of Nigeria's GDP.

Even though self-reported (as against clinically measured) health has its limitations,[10] this study provides some of the first robust evidence linking what health providers do in practice to patient outcomes. These results provide striking evidence that current levels of effort (and medical

[10]There is robust evidence that patient self-reported health is strongly correlated with, and predictive of, clinically measured outcomes. For references see Benjamins et al. (2004) and Jylhä (2009). A more detailed discussion and additional references are provided in Section 3 of this paper.

care) are sub-optimal, and suggest that there are substantial gains to be had from getting health providers to exert more effort when they are at work. This situation is in stark contrast to that in developed countries like the US where the debate is about 'flat of the curve' medicine (referring to practicing along the part of the productivity curve where the marginal returns to additional health care expenditure are close to zero (Fuchs, 2004)).[11] The situation in developing countries can perhaps more properly be described as 'base of the curve' medicine. One of the implications of the findings in this study is that even without changing any ancillary inputs, poor health outcomes in developing countries can be improved by getting health providers to do more of what they already know to, and importantly can, do.[12]

This paper makes a contribution to (and draws together) two distinct literatures. First, it makes a contribution to the literature on health worker effort (see for example Das and Hammer, 2007; Leonard and Masatu, 2010; Das et al., 2008). This literature has largely focused on describing and understanding effort (i.e., effort is on the left-hand-side). In contrast, this paper examines the relationship between effort and health outcomes (i.e., effort is on the right-hand-side).[13] Second, this paper makes a contribution to the literature on the returns to medical care (see Doyle, 2005, 2011; Almond et al., 2010). This literature has primarily been based in developed countries and has focused on care provided in a hospital or inpatient setting.[14] In contrast, this paper examines returns to outpatient care.[15] Outpatient spending accounts for about 41% of US health expenditures, and is the most significant driver of spending growth (Bradford et al., 2011). A novel aspect of this study is that variation in care comes from variation in provider effort, an important, but not previously studied, margin.

The results in this paper have relevance for ongoing policy discussions about ways of improving quality of service delivery and health outcomes in poor countries (see Das and Hammer, 2014; Das

[11] By some estimates up to one-third of US healthcare expenditure, or approximately 700 billion dollars per year, is wasteful spending that can be reduced without any adverse effects on health outcomes (Garber and Skinner, 2008; Berwick and Hackbarth, 2012).

[12] This is relevant for ongoing discussions in global health circles about improving quality (Kruk et al., 2016).

[13] There is some evidence that health providers when induced to exert more effort ask more questions, and carry out more examinations (Leonard and Masatu, 2010), but this does not necessarily imply better outcomes. Das et al. (2012) show that while private providers in India were more likely to ask the right questions and do the right exams, they were not more likely to prescribe the right treatment.

[14] For examples of developing country studies see Okeke and Chari (2018) and Adhvaryu and Nyshadham (2015).

[15] Existing work has also focused on narrow subsets of patients or conditions to better address endogeneity; for example, low birthweight infants (Almond et al., 2010), automobile accidents (Doyle, 2005), or travelers experiencing a medical emergency (Doyle, 2011). By examining a broad range of primary care conditions, the results in this paper are arguably more general.

et al., 2018). These implications are explored further in the discussion and concluding section of this paper. The rest of the paper proceeds as follows: Section 2 describes the study context and provides additional details about the experiment; Section 3 describes the data, Section 4 discusses identification and lays out the empirical strategy; results are presented in Section 5 and discussed in Section 6. Concluding remarks are in Section 7.

2 The Experiment

2.1 Study sites

This study was conducted in 124 public sector outpatient clinics in Nigeria.[16] The sample is stratified across five states representing three of the six geopolitical zones in Nigeria. Two of the states are in the northeast, two are in the northwest, and one is in the south-south.[17] These clinics provide a range of outpatient and some inpatient services. They also provide maternal and child health services.

2.2 Study design

The key intervention in this study is a peer monitoring intervention designed to induce health providers to work harder. The intervention leverages insights from previous work showing that individuals try harder when their behavior is being observed (Banerjee et al., 2008; Duflo et al., 2012; Rex et al., 2010; Leonard and Masatu, 2006; Muralidharan and Sundararaman, 2010). Specifically, the treatment consisted of an intervention in which participating providers were observed by an independent clinician (the peer) on the research team, as they provided care to patients.[18] The monitoring was entirely passive. The observing clinician did not interact with either the patient or the provider during the consultation, they merely took up a position in the consultation room where they could observe the interaction.

[16] Nigeria has approximately 30,000 primary health care clinics, 78% of which are in the public sector. These are the first point of entry for most patients into the health care system.

[17] The specific sites are Kano and Jigawa in the northwest, Gombe and Bauchi in the northeast, and Akwa Ibom in the south-south. The study clinics were drawn from a sample taking part in a randomized trial. The primary intervention in this trial was the assignment of an additional primary health care provider. All of these providers are included in this study. Ethical approval for this study was provided by RAND's Human Subjects Protection Committee and by the Ethics Committee of Aminu Kano Teaching Hospital, Nigeria.

[18] Observers were medical officers drawn from local secondary and tertiary health facilities.

One might wonder why, since there were no rewards or penalties attached to behavior, this would cause providers to work harder. There are several potential explanations: first, peer observation might be reminiscent of an examination and induce providers to want to perform well (medical training often involves examinations where the trainee is observed by other clinicians); second, monitored providers might, consciously or subconsciously, desire peer approval and would attempt to gain it by following what they knew (or believed) to be best practices; third, it is possible that the providers, despite assurances that the data would not be shared with managers or supervisors, wanted to insure themselves against any potential negative consequences in case we reneged on our promise, by again following what they knew (or believed) to be best practices.

Conceptually, the intervention is straightforward to implement, the harder part was figuring out how to collect data on 'normal' provider behavior, i.e., when not being observed (the counterfactual). In other words, we had to figure out a way to observe provider behavior without letting them know that they were being observed. The strategy we chose was to observe how long they spent with each patient (a detailed discussion of why consultation time is a good proxy for effort is left until later). Crucially for the design, we could collect data on this covertly without tipping off the provider.[19] Our strategy was to position a research assistant in the hallway outside the consultation rooms where they could observe the doors to the consultation rooms. This individual recorded time of entry and exit into each room. In order to be able to link each individual to a consultation, this required a bit of careful choreography. As I describe later, patients were approached upon presentation at the clinic, and those that agreed to participate were issued with a consent form with an identifying number. Before entering the consultation room, they gave this form to the research assistant in the hallway allowing me to link individual patients to individual consultations.[20]

Two outpatient providers participated in the study in each clinic.[21] Each provider was exposed to both treatment and control conditions (a crossover design). In the first period, one of the two providers was randomly selected to be monitored while the other provider was assigned to the control

[19]It is possible that the mere presence of the team in the clinic would impact behavior, even when a provider was not being overtly monitored. This would imply that the estimated effect of monitoring in this paper is a lower bound.

[20]This was ostensibly the reason why the research assistant was stationed in the hallway. If there was no form, it indicated that the patient did not give consent, and in such cases, the clinical observer was informed so he could exit the room. However, in addition to collecting consent forms, this individual recorded entry and exit times into the consultation room.

[21]Where there were more than two outpatient providers in the clinic, the research team endeavored to select those that saw the most patients. All providers gave consent. Clinic visits took place between March and October 2017.

condition in which they practiced as usual (unobserved providers were unaware of being in a control group, they were told they would be observed the next day). In the second period, usually the next work day, both providers switched places: the first provider went back to practicing (ostensibly) unobserved, while the second provider was now overtly monitored.

3 Data

3.1 Survey Data

Outpatient Providers: Each participating provider completed an interview with a trained interviewer (interviews were typically conducted in the morning before the provider started attending to patients, or at the end of the work day to minimize disruptions). The interview collected information about provider demographic characteristics (age, sex, ethnicity, marital status), medical qualifications (type of qualification and year obtained), and work experience (overall experience and work experience in the specific clinic). The health provider also completed a module intended to assess clinical knowledge. It included a set of multiple-choice questions covering a range of primary care conditions from malaria and child diarrhea to acute severe asthma.

Patients: Patients arriving at the clinic were briefly interviewed by an interviewer before taking their turn to wait to be seen.[22] The interview collected information about illness symptoms and severity (measured along a scale ranging from 0 – no pain, to 10 – worst pain possible), and about patients' current state of health. There is little consensus in the literature about how best to measure health, particularly in the short run. Health is multi-dimensional and it is difficult to find one measure that adequately captures all relevant dimensions. Measures used in previous studies have included morbidity (e.g., number of days sick in a reference period) and mortality, nutrition-related measures such as weight and height (and measures deriving from these such as height-for-age Z-scores), functional limitations, and self-reported health (Schultz, 2005). In this context, the measure of health needed to not only be sensitive to short-run changes but also general enough to include the wide variety of cases one might encounter in a typical outpatient clinic. Many of

[22]Because these were sick patients, interviews were very brief. Less than 0.5% refused consent. Cases were triaged by clinic staff and assigned to a provider following normal clinic protocol. When there were few patients available, however, preference was given to the provider being monitored (which is why there are more treated than control patients). The target was to observe up to 10 consultations per provider.

the measures mentioned previously fail to satisfy either one or both of these criteria. For example, nutrition-related measures are not suitable for measuring short-run changes in health. Previous studies measuring acute health have used days of ill-health (or days disabled due to ill-health), and self-reported health. The latter is what we chose to use in this study.

The relationship between self-rated health and objective health is very well-documented. For an extensive set of references, see Benjamins et al. (2004) and Jylhä (2009). Self-rated health has been shown to correlate well with clinically measured health, and is a significant predictor of mortality (McGee et al., 1999; Barger et al., 2016; DeSalvo et al., 2006). Importantly, it satisfies both of the criteria, and is commonly used in health assessments (Stewart et al., 1992; Jylhä, 2009). In this study we asked patients to rate their overall state of health along a categorical scale from excellent to poor. The key outcome used in the analysis is the probability that a patient rates their health as excellent.

I augment this measure of overall health with a measure of functional health. Functional health was assessed by asking patients how much difficulty they would have walking up a flight of stairs or running the length of a football field ('none', 'some', or 'a lot').[23] I define an indicator for patients that reported difficulty. The age and gender of each patient was also recorded, along with information about mode of transportation to the clinic and travel time, and a contact phone number if available. Several weeks after the clinic visit, patients were re-contacted for a follow-up phone interview (research assistants making the calls were blinded to treatment status). This interview once again measured patients' current state of health, and also collected information about treatment prescribed and adherence to the treatment, e.g., whether the patient obtained medicines prescribed. The interviewer also asked about satisfaction with the care they received.

3.2 Observational Data

Consultation time: I have data on the start and end times of each patient consultation (measured as in-out times from the consultation room). This is the primary measure of clinician effort in this study. Consultation times are a generally accepted proxy for how much was done by a clinician

[23]The answers to both questions are internally consistent as one would hope. For example, 97% of patients that said they would have a lot of trouble walking up a flight of stairs also said they would have at least some trouble running the length of a football field. Analogously, 99% of respondents who said they would have no trouble running the length of a football field also reported that they would have no trouble walking up a flight of stairs.

during the visit. In the CPT (Current Procedural Terminology) codes used to bill for physician services in the US, for example, time is explicitly used to determine the level of 'intensity' of the visit – the average consultation time for a Level 1 (the lowest intensity) new patient office visit (CPT code 99201) is 10 minutes, a Level 2 visit is 20 minutes, and a Level 5 visit (the highest intensity) is 60 minutes (Hill, 2008). Time is also an explicit component of the Relative value units (RVUs) that measure the amount of work done by a physician.[24] In developing countries, the length of the outpatient consultation is also frequently used as a key marker of quantity/quality of care provided (see Das et al., 2008; Das and Hammer, 2014; Irving et al., 2017). This is not surprising as there is an obvious correlation between time spent by a clinician during an office visit and how much he/she does. Consider that when a patient first presents, neither the patient nor the clinician knows what is wrong with the patient. It is the job of the clinician to make sense of the clues given by the patient and attempt to deduce what the underlying problem is. During the consultation, the clinician takes the patient's history, asking questions about the presence or absence of symptoms, and also often performs a physical examination. This information is cognitively processed by the clinician, leading to a tentative diagnosis (or diagnoses). The clinician may then request for laboratory tests to confirm this diagnosis, and prescribe medication. Clearly, the more that a provides does, the longer the consultation will take.

Consultation content: For monitored consultations, I have detailed information about the interaction (as observed by the monitoring clinician) including the specific questions asked (for patients presenting with fever, cough or diarrhea), any physical examinations performed, any diagnostic tests ordered, whether a diagnosis was made, and treatment provided. The data on the interaction also includes measures of the quality of provider communication; for example, whether the provider explained the diagnosis to the patient in common language, and whether they provided any health education relating to the patient's condition.

[24] RVUs are the basis of physician payment by Medicare (Coberly, 2015).

4 Empirical Strategy

4.1 Identification

The peer monitoring intervention generates the necessary variation in health provider effort. The assumption required for causal inference, as in any experimental setup, is that there is no correlation between the treatment and participant characteristics. Since each provider is exposed to both treatment and control conditions, I can include provider fixed effects to control for provider characteristics. Given the specific study design, the necessary identification assumption is that there is no sorting by patients across providers *and* time periods.[25] Sorting on either dimension alone does not threaten causal inference. To see why, imagine a scenario with sorting across providers. This is easily imagined: patients often have preferences for one provider over another and these preferences could be correlated with outcomes of interest; alternatively, clinic triage staff might assign patients to providers based on characteristics not observed by the research team, for example, based on information from prior visits.[26] This kind of sorting would, however, not vary across time periods when the provider switched treatment groups. The same reasoning applies to sorting across time periods. One can easily imagine scenarios in which the composition of patients would vary across time periods such that $E(X|T=1) \neq E(X|T=2)$, where X denotes patient characteristics and T denotes time periods. This might be the case, for example, if different clinics are held on different days of the week, e.g., antenatal clinics on Mondays, post-operative clinics on Tuesdays, etc. In each time period, however, since patients are seen by providers in both the treatment and control conditions, one would expect patient characteristics to be balanced across both conditions.[27]

Later on, I present empirical evidence that patient characteristics are balanced across treatment/control conditions.

[25] Providers did not choose their patients so we can rule this out. Patient assignment was made by clinic staff external to the consulting room. Once a provider was in the consulting room, they saw whichever patients that walked in the door.

[26] Clinic staff could also route more severe cases to certain providers, though this would not be an issue in this context because we observe, and can control directly for, severity.

[27] A strategy relying on temporal sorting might be problematic if patients that arrived later were systematically different from patients that arrived earlier. Cases that come in later may, for example, be more emergent cases; alternatively, assuming patients arrive in the morning, more emergent cases may be given priority and seen first. I find strong evidence of a correlation between consultation order and patient severity (results available on request).

4.2 Monitoring and effort

To estimate the effect of peer monitoring on effort, I run the following linear regression:

$$Ln(Time)_{ijks} = \alpha + \beta Treat_{ijks} + X_i'\delta + \varphi + \varepsilon_{ijks} \quad (1)$$

where $Ln(Time)$ is the natural log of consultation length for patient i seen by provider j in clinic k in state s. This is computed as the difference between entry and exit times from the consultation room. $Treat$ is a binary indicator for whether the consultation was monitored by the peer clinician. β, therefore, measures the increase in effort induced by monitoring. X_i' denotes a vector of control variables including patient age, sex, whether they arrived at the clinic in their own car or motorcycle (used to proxy for socioeconomic status), illness severity, whether the patient presented with fever (a common presentation), and whether the consultation was pregnancy-related.[28] To capture case complexity, I define a variable equal to 1 for patients presenting with generalized pain or weakness (without fever). This is a non-specific, more complex presentation that requires careful history taking and examination to figure out what is wrong with the patient. φ denotes either clinic or provider fixed effects.

The base specification is a regression of consultation duration on the treatment indicator, including only state dummies (recall that the sample was stratified by state). I gradually make it more restrictive by replacing the state dummies with clinic fixed effects (this compares treated and control patients within the same clinic, holding constant clinic factors such as prices, practice patterns, availability of drugs/diagnostic equipment, and area-level factors such as access to other facilities and providers), and then provider fixed effects. This is the most demanding specification, which compares patients seen by the same provider. It holds constant all provider characteristics including qualifications, motivation, and ability. This is the preferred specification because it takes full advantage of the crossover design, allowing each provider to serve as their own control. Lastly, I add-in controls for patient characteristics. All standard errors are adjusted for clustering at the clinic level.

[28] These generally take longer because an abdominal examination is often required.

4.3 Provider effort and health

This is the key relationship of interest. If increased effort by health providers leads to better patient outcomes, this would suggest that status quo levels of effort are too low. Observing that consultations are short or that providers do not carry out all the steps recommended by guidelines, as many studies have shown, does not necessarily mean that the level of care is too low. In a disease-endemic environment, diagnostic heuristics that treat this as the presumptive diagnosis may be efficient. Specifically, in countries like Nigeria where malaria is endemic there is a high probability that a patient presenting with fever in fact has malaria. In that case, once a provider ascertains that fever is the presenting complaint, treating that patient for malaria will more often than not be the correct choice. Asking many more history-taking questions and carrying out more examinations and tests may add little value and not lead to better outcomes.[29] This model would generate similar stylized facts – short consultations on average (because many patients will present with fever), but this would not necessarily be inefficient.

Providers could also be rationing effort across patients, choosing cases they spend more or less time on.[30] It is plausible that providers make quick judgements, based on presenting symptoms and other easily observed patient characteristics such as age and how sick the patient looks, about whether a case is 'simple' or 'complex', and then allocate effort accordingly.[31] Less time, for example, might be spent on cough and nasal congestion in an otherwise healthy-looking baby (likely a self-limiting viral disease), whereas a cachectic adult presenting with unexplained weight loss would be allocated more time. If a high proportion of cases are 'simple' cases, then this may generate a similar stylized finding of short consultation times. This may or may not be efficient depending on how accurate these judgements are.[32] If there are errors of judgement such that providers routinely misclassify serious cases as simple, then incentivizing or inducing them to spend more time might lead to overall improvements in patient health.

To examine the relationship between effort and health, I estimate the following linear intent-

[29] An oft-repeated saying in clinical training is that "common things occur commonly".

[30] These models are not necessarily mutually exclusive.

[31] This behavior might arise under capacity constraints (lots of patients are waiting to be attended to) or time constraints (e.g., because the provider wants to be done quickly so that they can leave). It may also arise because effort is an exhaustible, costly resource that needs to be 'spent' wisely (akin to a football player taking less critical plays off).

[32] One may think that better (or more experienced) clinicians will be right more often than they are wrong.

to-treat (ITT) model using an ANCOVA specification that controls for health status at the time of the clinic visit (later on I examine robustness to a non-linear specification). This dominates the differences-in-differences estimator in terms of efficiency when autocorrelation in the outcome is low (McKenzie, 2012).

$$Health_{ijks}^{T=1} = \alpha + \gamma Treat_{ijks} + Health_{ijks}^{T=0} + X_i'\delta + \varphi + \mu_{ijks} \tag{2}$$

I introduce a superscript T to differentiate between health at the time of the clinic visit (T=0) and health at the time of follow-up (T=1). I also estimate the following IV specification:

$$Health_{ijks} = \alpha + \theta Effort_{ijks} + X_i'\delta + \varphi + \upsilon_{ijks} \tag{3}$$

where $Effort$ is instrumented by the treatment indicator $Treat_{ijks}$. Given that I have a single endogenous variable and instrument, the IV estimate can also be computed from β and γ in Equations (1) and (2) as γ/β. The interpretation of θ is as the local average treatment effect for marginal patients (those cases where the outpatient provider was induced by the presence of the monitoring clinician to exert greater effort). The sample for Equation (2) and (3) consist of patients successfully interviewed at follow-up. Standard errors are adjusted for clustering at the clinic level.

In supplementary analysis, I examine whether there are heterogeneous treatment effects by specific patient characteristics – illness severity and case complexity. If health providers are rationing effort, for example, spending less time on simple cases, and more time on severe, more complex cases, then it is possible that we might see heterogeneous effects of monitoring on effort, with smaller effects for more severe cases since providers are already doing what they believe they should be doing. The predicted effects on health outcomes are unclear: there could be larger treatment effects for more severe cases because there are greater marginal returns to effort, or the opposite could be true if providers routinely misclassify severe cases as simple.

5 Results

5.1 Descriptives and validity checks

Provider sample: Summary statistics for the clinic and provider sample are presented in Table 1. Participating clinics are fairly small: they are staffed by just under six health providers, have about 15 beds, and attend to approximately 22 patients daily. 72% of clinics provide inpatient services, 58% offered laboratory services, and only 6% offered surgical services. Of the 242 participating providers, 58 are doctors, 16 have a nursing or midwifery certificate (or both), 28 are Community Health Officers, 128 are Community Health Extension Workers, and the remaining 14 providers have a variety of other qualifications.[33] The average provider is 34 years old, has eight years of experience working with their current qualification, and have been working in the study clinic for two years. 62% of the providers are male.

Patient sample: We enrolled 925 patients that provided a contact phone number in the study: 620 of these consultations were monitored.[34] Descriptives are presented in Table 2. Patients are predominantly women (64% of patients are female), and the average age is 23 years. As one might expect, patients were in poor health at the time of presentation: only 0.3% of patients described their health as 'excellent' at baseline; the mean illness severity score was nearly 6 out of 10, and 35% of patients reported a severity score of at least 7. Figure 1 shows the distribution of patient complaints sorted by frequency. Not surprisingly, given the endemic nature of malaria in this setting, the commonest complaint, by a wide margin, was fever. Headache, cough, weakness/tiredness, abdominal pain, vomiting, and diarrhea, were also frequent complaints. The ten most frequent complaints are reported in Table 2. Table 2 also reports t statistics from balance tests (the null is that group means are equal). We see that the sample is clearly balanced providing empirical validation for the study design.

[33] Community Health Officers (CHO) and Health Extension Workers (CHEW) are trained and licensed primary health care professionals in Nigeria. Their training, licensure and practice is regulated by the Community Health Practitioners' Registration Board of Nigeria, established by Decree 61 of 1992 (Ordinioha and Onyenaporo, 2010). Junior CHEWs receive two years of training and are awarded a Certificate in Community Health on completion. After a few years of work experience, they can go through a three-year training program to become CHEWs. On completion, they are awarded a Diploma in Community Health. CHEWs can also go on to become Community Health Officers. The CHO training program takes two years and successful trainees are awarded a Higher Diploma in Community Health.

[34] For purposes of the larger trial we also collected data on patients without a phone. We leverage these additional observations later.

We were able to reach and successfully complete interviews with 599 of these patients (396 of these consultations were monitored) – a 65% follow-up rate.[35] The remaining patients could not be reached for various reasons: the phone number was invalid or out-of-service, or no one ever answered (even after repeated attempts). On average, patients were interviewed 22.5 days after the initial visit. This varied from a low of 5 days to a high of 138 days (standard deviation of 23 days).[36] As expected the number of days between visit and follow-up interview is uncorrelated with treatment (mean of 21.9 days in the control group and 22.8 days in the treatment group; p =0.64).

Table 3 presents descriptives for this sample; I also present balance tests because even though the sample is balanced at baseline, differential patient dropout could potentially lead to imbalance. Overall, the sample remains balanced, though among patients successfully interviewed, those in the control group were slightly more likely to be in better health at baseline. To examine how large a difference this is, I convert it to a standardized effect size. Following guidelines in Cohen (1988), I conclude that this is a small difference (0.24 standard deviations). In the results that follow, I carefully examine whether the results are sensitive to including/excluding baseline health as controls. They are not. As I will show later, conditioning on patient characteristics affects the point estimates hardly at all. As a robustness check, I also estimate non-parametric Lee bounds on the treatment effect (Lee et al., 2009). The idea is to compute the largest and smallest possible effect sizes consistent with the observed data.

5.2 Monitoring and effort

The average patient consultation in the control group lasted 9 minutes (median of 8 minutes and standard deviation of 6 minutes). This is on the higher end of estimates for developing countries reported in Das et al. (2008), which ranged from a low of 4 minutes in India to 8.3 minutes in Paraguay.[37] Given how consultation length is measured in this study – in-out times from the consultation room – the length of the actual patient-provider interaction is likely to be shorter, allowing for the time it takes the patient to walk from the door to the chair on entry, and back to the door (on exit). Figure 2 shows the actual (Panel A) and log-transformed (Panel B) distribution

[35]This is substantially higher than for typical phone-based surveys (O'Toole et al., 2008; Keeter et al., 2017).

[36]This variation was due to logistics of scheduling (each week there were multiple follow-up calls, all of which obviously could not be made at the same time), availability of the research assistants making the calls, and variation in when a patient was reached for interview.

[37]For a more recent review of consultation times around the world, see Irving et al. (2017).

of consultation times.

Figure 3 provides a graphical examination of the effect of monitoring on consultation time. I plot kernel densities of consultation times separately for overtly vs. covertly monitored consultations. There is a noticeable shift in the distribution to the right for (overtly) monitored consultations, particularly for shorter consultations, indicating an increase in consultation time. A Kolmogorov-Smirnov test easily rejects equality of the two distributions (p-value=0.000). The unadjusted mean difference between overtly and covertly monitored consultations is about 1.9 minutes. Relative to the control mean this is a large increase – approximately 21%. Keep in mind that this includes the time to walk in and out of the room, which is technically not part of the consultation so the true effect on consultation length is even larger. In Figure 4, I plot the CDF of the treatment effect to examine the distribution. We see that while the average is just under 2 minutes, 20% of clinicians increased their consultation times by more than 50% (indicated by the dashed line), when they were being overtly monitored.

Table 4, which presents the results from Equation 1, confirms these graphical results. Column 1 is the base specification that includes only strata dummies. Column 2 is the clinic fixed effects specification, Column 3 is the provider fixed effects specification, and Column 4 adds-in patient-level controls. We see that across columns, the coefficients are nearly identical. Including patient-level controls makes no difference, as one would expect if characteristics are uncorrelated with treatment. From these results, I conclude that the peer monitoring treatment caused providers to increase consultation times by about 21 percent. What does this mean? The obvious conclusion is that providers exerted more effort in response to being monitored. One must, however, rule out other possibilities. It is possible, for example, that the increase in time simply reflect providers being more careful and proceeding more slowly because they were playing to the audience (of one), or perhaps the presence of the peer clinician made them more nervous so that things took longer. In this case a longer consultation would not indicate that providers did more, it would simply mean that they took longer to do the same things.[38]

First, evidence from prior work indicates that providers likely did more, i.e., they took substantive new actions, in response to being monitored. Leonard and Masatu (2010), for example, showed that clinicians asked more questions and carried out more recommended examinations when being

[38]Taking more time and being more careful may still be beneficial but this is not the primary posited mechanism.

observed.[39] Rowe et al. (2012) compared health worker adherence to Integrated Management of Childhood Illness (IMCI) guidelines under "conspicuous" observation to normal behavior assessed by fake or simulated patients. This is a particularly noteworthy study because the authors also collected data on consultation duration. They found that observed consultations were about 31% longer and observed clinicians exhibited greater adherence to IMCI guidelines (a median difference of 16 percentage points).[40] Second, data from our follow-up interviews confirm that patients in fact received more care when the clinician was being overtly monitored. In the follow-up survey we asked patients about the care they received, and specifically whether they were prescribed any medicines by the health provider.[41] Whether a patient receives treatment is obviously going to be related to their outcomes. We see from Table 5 that patients in the treatment group were about 8 percentage points more likely to report being prescribed treatment.[42] If we know that clinicians increased effort along this dimension, since medicines are not prescribed in a vacuum, it indicates that they increased effort along other dimensions.

Third, if clinicians responded to monitoring by doing more (rather than just moving more slowly), we would expect this to vary with ability. Higher-ability clinicians know more and therefore can be induced to do more.[43] To test this proposition, I construct a measure of clinician ability using principal component analysis to derive an index from the following variables: (i) the clinician's score on the multiple-choice assessment, (ii) percent of recommended history-taking questions asked,[44] (iii) percent of the clinician's patients that received a physical examination, (iv) percent of cases where a diagnostic test was ordered, and (v) percent of the clinician's patients that received a diagnosis.[45] The resulting index is divided into quartiles. I interact dummies for each quartile

[39] Interestingly, the magnitude of this increase was around 20%.

[40] For additional health care examples, see Rowe et al. (2006) and Campino et al. (2008). Outside of health care, see Muralidharan and Sundararaman (2010) who showed that teachers exerted more effort when being observed in the classroom.

[41] This was something we believed they would be able to recall with accuracy even months later.

[42] This could be biased upwards if patients whose consultations were monitored were more likely to recall receiving treatment. Since we observe, for monitored consultations, whether the patient was in fact prescribed any medicines, we can use the observational data instead of patient report. The resulting estimate is not very different – about 7 percentage points.

[43] If on the other hand, clinicians were simply unnerved by the presence of another clinician, one might expect lower-ability clinicians to be more affected, which would lead to a result in the opposite direction – lower-ability clinicians would take more time in response to the treatment.

[44] I first average within condition (for fever, cough, and diarrhea), and then average over all conditions for each provider.

[45] The last four are constructed using the observational data. As a check that this index captures ability, I relate it to medical school performance. In the survey providers were asked whether they: (a) attained a distinction or high pass on any of their courses during medical training, or (b) repeated a year at any point during their training (usually

with the the treatment indicator and re-estimate Equation 1. To make this point visually, I plot the coefficients from this interacted regression in Figure 5. For clarity the dependent variable is consultation time in minutes.[46] We can clearly see that the effect of monitoring increases with ability, as one would expect if clinicians did more. There is no effect of monitoring on clinicians in the bottom two quartiles (they are already doing what they know and cannot be induced to do more – they are constrained by their ability). However as ability increases, clinicians can be induced to do more because they know more. We see that the increase in consultation time for the highest-ability clinicians is nearly double the average, consistent with an increase in effort (higher-ability clinicians are practicing well inside their frontier and can be induced to do more).

Finally, I show that consultation time varies in the expected way with clinician effort. I regress various measures of clinical effort on consultation time.[47] The model includes provider fixed effects and controls for patients characteristics. The results are in Table 6. We see that longer consultations are associated with asking more recommended questions, carrying out more examinations, greater use of diagnostic tests, and a greater likelihood that the provider talked to the patient about their diagnosis and treatment. Asking more questions or performing more examinations is not a policy objective in and of itself, except to the extent that it increases the likelihood that the clinician will correctly deduce what is wrong with the patient, allowing them to prescribe treatment. In keeping with this, we see that longer consultations are associated with a greater likelihood that the provider makes a diagnosis, and an increase in the number of medicines prescribed consistent with patient reports from the follow-up survey (dependent variable means are reported at the bottom of the table). Based on these results, a 50% increase in consultation time would translate into a 0.8 percentage point increase in the proportion of recommended history-taking questions asked for patients with fever, a 3.5 percentage point increase in the probability of carrying out a physical examination, a 4.7 percentage point increase in the probability that a diagnostic test is ordered, and a 2.2 percentage point increase in the probability that the clinician provides health education related to the diagnosis.

due to poor grades). The former indicates 'high achievers' while the latter indicates potentially 'low achievers'. Given that these were face-to-face interviews, I expect some level of misreporting. In particular, I expect that the former will be biased upwards while the latter will be biased down. The results are, nevertheless, instructive. In Appendix Figure 1 I plot summary means for each variable within ability quartiles and find the expected positive (negative) gradients.

[46] The corresponding regression results are in Table 8.

[47] The sample includes all observed consultations.

5.3 Provider effort and health

Having provided evidence that monitoring induced health providers to exert greater effort, I turn to the effect on patient outcomes. Descriptively, patients were in better health at follow-up relative to baseline: 15.5% of patients, for example, reported being in excellent health at follow-up compared to 0.3% at the time of the initial visit. Patients also reported improvements in functional health: at follow-up 22% of patients reported difficulty walking up a flight of stairs or running the length of a football field compared to 63% at the time of the clinic visit. In Figure 6 I plot the distribution of self-reported health at follow-up, separately for the treatment and control groups. We can see a clear increase in the proportion of patients in the treatment group that rated their health as excellent at follow-up. The unadjusted difference is 4.1 percentage points (a 27% improvement relative to the mean). Figure 7 looks at changes between baseline and follow-up. We see that relative to the control group, patients in the treated group experienced a larger improvement in health between baseline and follow-up. The unadjusted relative difference is 4.5 percentage points.

Table 7 presents regression results. In Panel A, the dependent variable is a dummy indicator denoting excellent health on the day of interview based on the patient's or caregiver's rating. In Panel B, the dependent variable is a measure of functional health; an indicator for difficulty walking up a flight of stairs or running the length of a football field. All the specifications include clinician fixed effects. Column 1, the base specification, does not include any controls; Column 2 controls for baseline health but no additional controls; and Column 3 includes additional patient controls. Column 4 reports results from an IV specification where I instrument for clinician effort, measured by the amount of time spent by the clinician with the patient, with the treatment indicator - whether a peer was present during the consultation.

The results indicate that patients whose consultations were monitored were in better health at follow-up. Patients in the treatment group were 8 percentage points more likely to rate their health as excellent at follow-up (compared to a mean of 15%), and 7.3 percentage points less likely to report difficulty in walking up a flight of stairs or running the length of a football field (compared to a mean of 22%). As Column 2 shows, the results are not sensitive to controlling for baseline health; neither are they sensitive to including additional controls for patient characteristics (Column 3). In Table A.1, I show that the results hold up in a non-linear logit specification. I present results from

both a simple logit model with strata dummies and a logit model with provider random effects (the coefficients reported are average marginal effects).[48] We see that the coefficients are similar in both specifications and are in line with the linear estimates.

In Table A.2, I use as the dependent variable, a summary measure of health that aggregates information from all the health measures. Each indicator is normalized to have a mean of zero and a standard deviation of one in the control group, and I create a summary health index that is the mean of the normalized values. I present results using both the continuous index, and an indicator for a score above the median. In line with the previous results, patients in the treatment group score higher on the index and the probability that the patient's health score is above the median increases by 9.4 percentage points.

Next I examine whether there is a dose-response relationship between (induced) effort and health. The idea is to examine whether greater effort, induced by the treatment, results in larger improvements in health. If it does, it adds further credibility to the results as it indicates that the underlying mechanism is indeed effort. In Figure 5 I showed that the effect of monitoring on effort varied by clinician ability. Low-ability clinicians did not appear to respond to overt monitoring likely because they are constrained by their (lack of) knowledge, in contrast to high-ability clinicians. One way therefore to check that the health results are real is to examine whether patients seen by higher-ability clinicians (who experienced greater 'intensity' of treatment) experience better outcomes. Figure 8 provides strong supporting evidence. The distribution of health improvements in Figure 8 aligns with that of effort shown in Figure 5. The regression results are in Table 8.

In Figure A.2, I present some additional evidence. In the follow-up survey, we asked patients whether, since their visit to the clinic, they had sought care from a different provider or facility on account of the same illness.[49] Since patients are only likely to seek alternative care if they do not get better, we may expect that fewer patients in the treatment group will report seeking alternative care. Though only a small fraction of patients reported seeking other care overall (5.3%), as Figure A.2 shows, patients whose consultations were monitored were less likely to have sought care from a different provider or facility. Again, this result is consistent with treated patients being in better health.

[48] A fixed effects logit specification result in a loss of too many observations.

[49] The exact question was: "Since your visit to [clinic] on [date], have you consulted another health provider or visited another health facility on account of the same illness?"

IV: The IV coefficients indicate that a 10% increase in provider effort (measured by consultation time) leads to approximately a 4 percentage-point improvement in the probability of being in excellent health, and a 3.9 percentage point decrease in the probability of difficulty in walking up a flight of stairs or running the length of a football field. The latter result in the IV specification does not quite reach statistical significance.

Additional Robustness checks

In Table A.3, I examine whether the results might be driven by differential dropout. Though follow-up interview completion rates are similar – 67% in the control group vs. 64% in the treatment group (the p-value from a test of equality is 0.42), the pattern of dropout could still differ. To assess robustness I estimate non-parametric Lee bounds (Lee et al., 2009). To estimate bounds, I trim the outcome distribution in the treatment arm by the difference in retention rates as a proportion of the retention rate in the treatment arm. This requires trimming the upper or lower 4% of the outcome distribution. The resulting bounds are shown in Table A.3. Column 1 shows the estimated Lee bounds without any covariates while Column 2 includes the baseline health dummies as covariates (essentially splitting the sample into cells corresponding to each baseline health category). In both cases, the bounds are tight and exclude zero. The unadjusted (covariate-adjusted) lower and upper bounds imply a 3.6 and 7.8 (4.9 and 9.1)-percentage-point improvement respectively.[50]

In Table A.4, I attempt to rule out other confounding explanations. One possibility is that simply by being present, the observing clinician had an effect on patient outcomes. For example, it could be the case that the presence of a clinician from out-of-town (ostensibly there to monitor quality of services) made patients more confident that they were receiving good care and therefore more likely to adhere to the treatment prescribed. This might result in an improvement in health, but not as a result of providers exerting more effort (this would only be coincidental). To assess this I turn to the follow-up survey, where patients were asked whether they procured medicines prescribed and took them.[51] I find no evidence that treated patients were more likely to procure, or take medicines prescribed (conditional on medication being prescribed).

[50]This was estimated using the *leebounds* stata module (Tauchmann, 2014).

[51]The exact questions were: *Did you obtain these medicines?* ("Yes, all of it", "Yes, some of it", and "No"), and *Did you take these medicines?* ("Yes, completed the dose", "Yes, still taking medication", "Started but discontinued", and "Other").

Placebo effects are also possible: perhaps the presence of the observer made patients feel better about the care they received and made them more likely to feel better overall. These are harder to rule out. A weak test is to check whether treated patients reported being significantly happier with the care they received. If the result is negative, this would suggest that placebo effects are unlikely to be a driver. In Column 3 of Table A.4 I examine whether treated patients were more likely to report being satisfied with their care. Since nearly 97% of patients reported being either satisfied or very satisfied with their care, I define satisfaction as being very satisfied. There is no evidence that treated patients were more likely to report being very satisfied with their care. The fact that the treatment also led to improvements in functional health suggests that these were real health improvements.

5.4 Heterogeneous Effects

The results of the treatment heterogeneity analysis are in Table 9. I do not find any statistically significant evidence of heterogeneous effects by either patient illness severity or case complexity. Providers appear to increase effort by similar amounts regardless of illness severity. I also do not find any evidence of heterogeneity in terms of health outcomes. The coefficients in both cases are fairly small in magnitude and insignificant. There is suggestive evidence of some heterogeneity by case complexity but the results are not precisely estimated. These results run somewhat counter to the prediction of a rationing model.

6 Discussion

This study provides some of the first direct evidence of a causal link between health provider effort and patient outcomes. In this experiment, outpatient clinicians in Nigeria were exposed to an intervention designed to induce them to exert more effort: monitoring by a peer. The rationale was that providers would be more likely to follow what they believed to be proper practice when another clinician was in the room observing them (even if there were no rewards or penalties attached to their performance). To measure provider behavior under the counterfactual, the same providers were covertly observed. I show that when providers are overtly monitored, they respond by increasing the amount of time they spend with their patients. I marshal additional evidence to show that

consultation times are longer because providers are doing more for their patients. I find that this increase in effort results in health improvements. The ITT results indicate an 8-percentage-point improvement in the probability that a patient is in excellent health at follow-up. This is roughly equivalent to the difference in outcomes produced by clinicians in the second highest ability quartile compared to clinicians in the bottom quartile of ability.[52]

Are these health effects plausible given that the average increase in consultation time was just under two minutes? I have presented careful evidence supporting the credibility of these findings but it is important to put the increase in consultation time into context. First, while a two-minute increase in consultation time is not large in an absolute sense, it is large relative to average consultation times in this (and similar) settings. To put this into proper perspective, the estimated treatment effect is roughly the difference in time between a medical officer and a community health extension worker. Second, in the presence of diminishing marginal returns, the gains to increased time (and effort) are likely to be significantly larger at the base of the curve. A two-minute increase in time spent by a provider with a patient may not make a huge difference when clinicians are already spending 20 minutes with patients,[53] but it will when the median consultation lasts only 8 minutes. The extra two minutes may be the difference between a clinician examining a patient or asking an additional question that leads to a diagnosis and treatment. In a striking illustration of this, Rowe et al. (2012) found that majority of health workers during a sick child consultation failed to ask about diarrhea (unless it was spontaneously offered as a complaint); *"and that failure led to a cascade of errors, ultimately resulting in incorrect treatment for virtually all diarrhea cases."* Third, it is worth noting that while two minutes (a 20% increase) was the average, 1 in 5 clinicians increased effort by more than 50%.

The results in this paper suggest that there are significant welfare costs to low effort. On the one hand, the cost to providers of exerting additional effort appears to be small. Clinics see about 22 patients daily. Given an average consultation time of 8.5 minutes (in the control condition), this means that the average provider spends about 3.1 hours per day consulting with patients, suggesting that they have significant slack even if one factors in other activities such as administrative duties.

[52] Appendix Table 5 reports results from a regression of patient health on clinician ability. The model controls for other clinician characteristics (age, sex, qualifications, and years of experience), clinic characteristics (whether inpatient services are offered, whether the clinic has a lab and pharmacy, and level of cleanliness of the clinic as observed by the data collector), and patient characteristics. The model includes strata dummies.

[53] The average consultation length in the US (Irving et al., 2017).

I estimate that monitoring increased average consultation times by about 20 percent, which would be less than 40 additional minutes per day. The cost of low effort for patients are, however, large and significant. There is a large economic literature examining the relationship between health and productivity going all the way back to Schapiro (1919).[54] Good health is valuable because it increases productivity. Healthy individuals can work harder and longer, and can also think more clearly.[55] To quantify the cost of low effort in terms of wages, I turn to estimates from this literature.

Schultz and Tansel (1997) estimated 10-12% lower wages for each disabled day (days inactive due to illness) in Cote d'Ivoire and Ghana.[56] More recently, Dillon et al. (2014) estimated that a workplace malaria testing and treatment program in Nigeria increased daily wages by between 6 and 9% (Treatment-on-Treated effects for malaria-positive workers were smaller – about 2-3%). In Latin America, Cortez (2000) and Murrugarra and Valdivia (2000) estimated that an additional day ill during the four weeks prior to the survey was associated with a decline of between 1 and 4 percent in hourly earnings in Peru while Ribero and Núñez (2000) estimated that being sick in the last month decreased male/female earnings by 41/19 percent in rural areas in Colombia. Given the available range of estimates, assuming a 6% reduction in wages for each additional day of ill health does not seem unreasonable. Based on data from the 2015 Nigeria Living Standards Measurement Survey (NLSMS), 14% of Nigerians reported ill health in the last month and were unable to perform their usual activities for about 3.5 days, on average. Focusing on the working age population, 15-64 year olds (approximately 53% of Nigeria's 186 million people), and assuming that they each earn the statutory minimum wage (roughly $60 per month), this would translate to about two billion dollars in lost wages per year due to ill health.[57] Assuming that about half of sick individuals seek care from a health provider (author's calculation based on the 2015 NLSMS) and that providers exerting effort similar to that induced by the monitoring intervention would cut sick days by about a third (this seems reasonable given the estimates reported in this paper), this would imply that about $350 million in wages is lost annually because of sub-optimal levels of provider effort. This represents nearly 0.1% of Nigeria's GDP ($405.1 billion in 2016 according to World Bank estimates). To put

[54] See reviews of this literature in Bleakley (2010). There is also a macro literature looking at aggregate effects of health on development (Weil, 2007; Ashraf et al., 2009).

[55] Health also affects the marginal utility of consumption (see Finkelstein et al., 2013).

[56] They also estimate a 3% reduction in hours for each disabled day.

[57] This is calculated as lost wages annually due to ill-health (3.5 x $60 x .06 x 12) multiplied by the total number of working-age adults who are sick (186,000,000 x .53 x .14).

this into perspective, Nigeria currently spends about 0.3% of GDP on all social safety net programs combined (World Bank, 2018). This is an admittedly crude back-of-the envelope calculation but it provides an eye-opening benchmark of the potential cost of low effort.

What are the policy implications of these results? An obvious one is that getting clinicians to supply more effort is likely to significantly improve patient welfare. One must however temper this by noting that this appears to depend on underlying ability. In a setting where health workers with low levels of human capital are ubiquitous, patient welfare gains might be limited. It is worth speculating a bit though on the broader question of why health workers supply sub-optimal levels of effort, as this matters for thinking about policy implications. One possibility is that this reflects shirking behavior by health workers. Effort is costly (and hard to monitor) and in the absence of strong accountability mechanisms, health workers will supply lower-than-optimal levels of effort. In this scenario, health workers are earning rents on the job, and better monitoring is needed to eliminate these rents. A different possibility is that it represents an informal contract between employer and employee. Both have a tacit understanding that in exchange for accepting low wages and poor working conditions, maximum effort is not expected.[58] In that sense, low effort is a kind of non-wage benefit similar to other non-wage benefits such as vacation time. In this scenario increased monitoring would lead to greater output of effort in the short run, but in the long run would reduce the attractiveness of the job making it harder to attract candidates. A better alternative would be to improve salaries and working conditions (though monitoring could still be useful).

A completely different interpretation, alluded to earlier, is that this is not low effort *per se*, but rather a misallocation problem whereby health workers are spending too little time on cases where they should be spending more time (our results however do not support this explanation). When being observed by a peer they do more, not because they know that they should have been doing more in the first place, but because they are playing to their audience. Since they do more for all patients, as the results indicate, cases where too little effort was being expended initially, inevitably benefit. In this case, increased monitoring does not really address the underlying problem. While patients that were receiving too little care will benefit from the increased monitoring, others patients now receive too much care, which is also inefficient. In that sense, monitoring is too blunt of an

[58]This might be one reason why the same workers exert more effort in the private sector than to the public sector (Das et al., 2016). Because pay and conditions are better, it is implicitly understood that more effort is expected.

instrument. The right intervention would be better training of health workers; both to recognize specific cases that require greater attention, but to, in general, always be thorough and minimize the use of heuristics. This is an important area for future work.

In closing, I recognize that this study is not without limitations. One limitation, which has implications for generalizability, is that the study sample only included patients that provided a contact phone number. Phone ownership might, for example, be correlated with socioeconomic status, and more educated, wealthier patients might be more likely to carry out prescribed tests and procure prescribed medication. To the extent that phone ownership is correlated with other observable characteristics such as whether the patient arrived in their own car or motorcycle, including the latter as a control in the regressions helps in part to mitigate this concern.[59] Nevertheless, it will be interesting to see how well these results transfer to other settings. A second limitation is that I do not have data on clinically measured health outcomes. Even though, as alluded to earlier in this paper, there is very strong evidence linking patient self-reported health to clinically measured outcomes, it would be interesting to see if one would found similar effects using clinical measures of health.[60]

7 Conclusion

Over the last decade, a growing number of studies have found results suggestive of low levels of effort by health providers in developing countries (Das and Hammer, 2014). One of the hallmark findings is a divergence between what health providers know, and what they do in practice (referred to as the "know-do" gap). Sometimes lost in the discussion is that gaps also exist between knowledge and practice in highly developed countries (Cabana et al., 1999; Rethans et al., 1991). In practice, it is common for clinicians to use approximations, and to skip steps considered to be unnecessary for a given patient. As such, a gap between theory and clinical practice does not inexorably lead to the conclusion that providers are not exerting enough effort. The results in this paper, however,

[59] An alternative approach is to re-weight the sample. Since patients with a phone are a subset of all patients and I have data on the characteristics of all patients that visited the clinic while the research team was present, one can model the probability of being included in the study and then re-weight the sample using the inverse of the estimated probability (Busso et al., 2014). Coefficients from an inverse probability weighted regression for overall health are slightly smaller (0.073) but remain significant at the 5% level. The effect on functional health is similar but less precisely estimated and does not quite reach significance at the 10% level (p=0.12) I am therefore cautiously optimistic that these results are not idiosyncratic.

[60] This is the subject of ongoing work.

provide strong evidence of sub-optimal levels of effort. When clinicians exert more effort, patients' health improves, suggesting that initial levels of effort (and care) were too low. The estimates imply substantial returns to effort: the elasticity implied by the lower-end estimates is around 2. This is similar to the estimates of the returns to medical care spending in Almond et al. (2010).[61]

These results provide justification for programs to incentivize health providers to supply more effort. In the government sector, there may be limited scope for market forces to induce providers to supply optimal levels of effort, and even in the private health sector where market forces have a larger role, information asymmetries may limit the ability of markets to correct. This suggests that policy intervention will likely continue to prove necessary. There are various tools available to policy makers. Monitoring, as this paper demonstrates, is one such tool.[62] The kind of monitoring used in this study is resource-intensive and unlikely to be used widely, but 'lighter-touch' monitoring may also be effective. For example regular case reviews and clinical audits have been used to good effect in some settings (Benjamin, 2008), so have bottom-up strategies (Björkman Nyqvist et al., 2017). Brock et al. (2013) show that even a simple encouragement intervention can yield results. Another tool that has been used in developed countries, performance measurement (with or without reporting), may also have a role in developing countries, though research shows that this does not come without its own set of challenges (Smith et al., 2010). In the longer run, creating effective malpractice laws, and enforcing those on the books, may help to reinforce professional standards. In the short to medium term, programs offering providers incentives to improve effort and quality, and utilizing forms of contracting that include performance metrics may offer a viable pathway to improving outcomes.

References

Adhvaryu, A. and Nyshadham, A. (2015). Returns to treatment in the formal health care sector: Evidence from Tanzania. *American Economic Journal: Economic Policy*, 7(3):29–57.

[61]The authors estimate that one-year mortality for low birthweight infants falls by one percentage point as birth weight crosses 1500 grams from above (relative to mean infant mortality of 5.5%). Hospital costs increase by approximately $4,000 as birth weight crosses 1500 grams from above (relative to mean hospital costs of $40,000). The elasticity implied by their estimates is about 1.8.

[62]Monitoring, for example, has been used in the education sector (Duflo et al., 2012). Ultimately, however, monitoring may need to be accompanied by sanctions or penalties if it is to remain effective.

Almond, D., Doyle, J. J., Kowalski, A. E., and Williams, H. (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *Quarterly Journal of Economics*, 125(2):591–634.

Arrow, K. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*, 53(3):941–73.

Ashraf, Q. H., Lester, A., and Weil, D. N. (2009). When does improving health raise GDP? *Yale Economic Review*, 5(2):24.

Banerjee, A. and Duflo, E. (2006). Addressing absence. *Journal of Economic Perspectives*, 20(1):117.

Banerjee, A. V., Glennerster, R., and Duflo, E. (2008). Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System. *Journal of the European Economic Association*, 6(2-3):487–500.

Barger, S. D., Cribbet, M. R., and Muldoon, M. F. (2016). Participant-reported health status predicts cardiovascular and all-cause mortality independent of established and nontraditional biomarkers: evidence from a representative US sample. *Journal of the American Heart Association*, 5(9):e003741.

Basinga, P., Gertler, P. J., Binagwaho, A., Soucat, A. L., Sturdy, J., and Vermeersch, C. M. (2011). Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. *The Lancet*, 377(9775):1421–1428.

Benjamin, A. (2008). Audit: how to do it in practice. *BMJ*, 336(7655):1241–1245.

Benjamins, M. R., Hummer, R. A., Eberstein, I. W., and Nam, C. B. (2004). Self-reported health and adult mortality risk: An analysis of cause-specific mortality. *Social Science & Medicine*, 59(6):1297–1306.

Berwick, D. M. and Hackbarth, A. D. (2012). Eliminating waste in US health care. *JAMA*, 307(14):1513–1516.

Bjorkman, M. and Svensson, J. (2009). Power to the People: Evidence from a Randomized Field Experiment of a Community-Based Monitoring Project in Uganda. *Quarterly Journal of Economics*, 124(2):735–769.

Björkman Nyqvist, M., de Walque, D., and Svensson, J. (2017). Experimental evidence on the long-run impact of community-based monitoring. *American Economic Journal: Applied Economics*, 9(1):33–69.

Bleakley, H. (2010). Health, human capital, and development. *Annual Review of Economics*, 2(1):283–310.

Bradford, J. W., Knott, D. G., Levine, E. H., and Zemmel, R. W. (2011). Accounting for the cost of U.S. health care. pre-reform trends and the impact of the recession. Technical report, McKinsey Center for U.S. Health System Reform.

Brock, J. M., Lange, A., and Leonard, K. L. (2018). Giving and promising gifts: experimental evidence on reciprocity from the field. *Journal of Health Economics*, 58:188–201.

Brock, M. J., Lange, A., Leonard, K. L., et al. (2013). Generosity norms and intrinsic motivation in health care provision: evidence from the laboratory and field. European Bank for Reconstruction and Development Working Paper No. 147.

Busso, M., DiNardo, J., and McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics*, 96(5):885–897.

Cabana, M. D., Rand, C. S., Powe, N. R., Wu, A. W., Wilson, M. H., Abboud, P.-A. C., and Rubin, H. R. (1999). Why don't physicians follow clinical practice guidelines?: A framework for improvement. *JAMA: The Journal of the American Medical Association*, 282(15):1458–1465.

Campino, A., Lopez-Herrera, M. C., Lopez-de Heredia, I., and Valls-i Soler, A. (2008). Medication errors in a neonatal intensive care unit. influence of observation on the error rate. *Acta Paediatrica*, 97(11):1591–1594.

Celhay, P., Gertler, P., Giovagnoli, P., and Vermeersch, C. (2018). Long run effects of temporary incentives on medical care productivity. *American Economic Journal - Applied Economics*, Forthcoming.

Chaudhury, N., Hammer, J. S., Kremer, M., Muralidharan, K., and Rogers, H. F. (2006). Missing in action: Teacher and health worker absence in developing countries. *Journal of Economic Perspectives*, 20(1):91–116.

Chicoine, L. E. and Guzman, J. C. (2017). Increasing rural health clinic utilization with sms updates: Evidence from a randomized evaluation in Uganda. *World Development*, 99:419–430.

Coberly, S. (2015). Relative value units (RVUs). Technical report, National Health Policy Forum Brief. Available at `https://www.nhpf.org/library/the-basics/Basics_RVUs_01-12-15.pdf`.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Cortez, R. (2000). Health and productivity in Peru: Estimates by gender and region. In Savedoff, W. D. and Schultz, T. P., editors, *Wealth from Health: Linking Social Investments to Earnings in Latin America*, chapter 6, pages 189–218. Inter-American Development Bank, Washington DC.

Das, J. and Hammer, J. (2007). Money for nothing: the dire straits of medical practice in Delhi, India. *Journal of Development Economics*, 83(1):1–36.

Das, J. and Hammer, J. (2014). Quality of primary care in low-income countries: Facts and economics. *Annual Review of Economics*, 6(1):525–553.

Das, J., Hammer, J., and Leonard, K. (2008). The quality of medical advice in low-income countries. *Journal of Economic Perspectives*, 22(2):93–114.

Das, J., Holla, A., Das, V., Mohanan, M., Tabak, D., and Chan, B. (2012). In urban and rural India, a standardized patient study showed low levels of provider training and huge quality gaps. *Health Affairs*, 31(12):2774–2784.

Das, J., Holla, A., Mohpal, A., and Muralidharan, K. (2016). Quality and accountability in health care delivery: audit-study evidence from primary care in India. *American Economic Review*, 106(12):3765–3799.

Das, J., Woskie, L., Rajbhandari, R., Abbasi, K., and Jha, A. (2018). Rethinking assumptions about delivery of healthcare: implications for universal health coverage. *BMJ*, 361.

DeSalvo, K. B., Bloser, N., Reynolds, K., He, J., and Muntner, P. (2006). Mortality prediction with a single general self-rated health question. *Journal of general internal medicine*, 21(3):267–275.

Dillon, A., Friedman, J., and Serneels, P. M. (2014). Health information, treatment, and worker productivity: Experimental evidence from malaria testing and treatment among Nigerian sugarcane cutters. *IZA Discussion Paper No. 8074*.

Dobkin, C., Finkelstein, A., Kluender, R., and Notowidigdo, M. J. (2018). The economic consequences of hospital admissions. *American Economic Review*, 108(2):308–52.

Doyle, J. J. (2005). Health insurance, treatment and outcomes: Using auto accidents as health shocks. *The Review of Economics and Statistics*, 87(2):pp. 256–270.

Doyle, J. J. (2011). Returns to local-area health care spending: Using health shocks to patients far from home. *American Economic Journal: Applied Economics*, 3(3):221–243.

Duflo, E., Hanna, R., and Rya, S. P. (2012). Incentives work: Getting teachers to come to school. *The American Economic Review*, 102(4):1241–1278.

Dumont, E., Fortin, B., Jacquemet, N., and Shearer, B. (2008). Physicians' multitasking and incentives: Empirical evidence from a natural experiment. *Journal of Health Economics*, 27(6):1436–1450.

Eggleston, K. (2005). Multitasking and mixed systems for provider payment. *Journal of Health Economics*, 24(1):211–223.

Finkelstein, A., Luttmer, E. F., and Notowidigdo, M. J. (2013). What good is wealth without health? the effect of health on the marginal utility of consumption. *Journal of the European Economic Association*, 11(suppl_1):221–258.

Fuchs, V. R. (2004). Perspective: more variation in use of care, more flat-of-the-curve medicine. *Health Affairs*, page VAR104.

Garber, A. M. and Skinner, J. (2008). Is American health care uniquely inefficient? *The Journal of Economic Perspectives*, 22(4):27–50.

Gertler, P., Giovagnoli, P., and Martinez, S. (2014). Rewarding provider performance to enable a healthy start to life: evidence from Argentina's Plan Nacer. World Bank Policy Research Working Paper 6884.

Gertler, P. J. and Vermeersch, C. (2012). Using performance incentives to improve health outcomes. World Bank Policy Research Working Paper No. 6100.

Goldstein, M., Graff Zivin, J., Habyarimana, J., Pop-Eleches, C., and Thirumurthy, H. (2013). The effect of health worker absence and health clinic protocol on health outcomes: the case of mother-to-child transmission of HIV in Kenya. *American Economic Journal: Applied Economics*, 5:58–85.

Hill, E. (2008). Time is on your side: Coding on the basis of time. *Family practice management*, 15(9):17.

Holmstrom, B. and Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 7:24–52.

Hong, F., Hossain, T., List, J. A., and Tanaka, M. (2013). Testing the theory of multitasking: Evidence from a natural field experiment in chinese factories. Working Paper 19660, National Bureau of Economic Research.

Huillery, E. and Seban, J. (2017). Money for nothing? the effect of financial incentives on efforts and performances in the health sector. *Sciences Po Economics Discussion Papers*.

Irving, G., Neves, A. L., Dambha-Miller, H., Oishi, A., Tagashira, H., Verho, A., and Holden, J. (2017). International variations in primary care physician consultation time: a systematic review of 67 countries. *BMJ Open*, 7(10).

Johnson, E. M. and Rehavi, M. M. (2016). Physicians treating physicians: Information and incentives in childbirth. *American Economic Journal: Economic Policy*, 8(1):115–41.

Jylhä, M. (2009). What is self-rated health and why does it predict mortality? towards a unified conceptual model. *Social Science & Medicine*, 69(3):307–316.

Keeter, S., Hatley, N., Kennedy, C., and Lau, A. (2017). What low response rates mean for telephone surveys. Technical report, Pew Research Center.

Kruk, M. E., Larson, E., and Twum-Danso, N. A. Y. (2016). Time for a quality revolution in global health. *The Lancet Global Health*, 4(9):e594 – e596.

Lee, A. C., Lawn, J. E., Cousens, S., Kumar, V., Osrin, D., Bhutta, Z. A., Wall, S. N., Nandakumar, A. K., Syed, U., and Darmstadt, G. L. (2009). Linking families and facilities for care at birth: What works to avert intrapartum-related deaths? *International Journal of Gynecology & Obstetrics*, 107, Supplement(0):S65 – S88.

Leonard, K. L. and Masatu, M. C. (2006). Outpatient process quality evaluation and the hawthorne effect. *Social Science and Medicine*, 63(9):2330–40.

Leonard, K. L. and Masatu, M. C. (2010). Using the hawthorne effect to examine the gap between a doctors best possible practice and actual practice. *Journal of Development Economics*, 93(2):226–243.

McGee, D. L., Liao, Y., Cao, G., and Cooper, R. S. (1999). Self-reported health status and mortality in a multiethnic us cohort. *American journal of epidemiology*, 149(1):41–46.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more t in experiments. *Journal of development Economics*, 99(2):210–221.

Miller, G. and Babiarz, K. S. (2013). Pay-for-performance incentives in low-and middle-income country health programs. Technical report, NBER Working Paper No. 18932.

Mohanan, M., Vera-Hernández, M., Das, V., Giardili, S., Goldhaber-Fiebert, J. D., Rabin, T. L., Raj, S. S., Schwartz, J. I., and Seth, A. (2015). The know-do gap in quality of health care for childhood diarrhea and pneumonia in rural india. *JAMA pediatrics*, 169(4):349–357.

Muralidharan, K. and Sundararaman, V. (2010). The impact of diagnostic feedback to teachers on student learning: Experimental evidence from india. *Economic Journal*, 120(546):F187–F203. cited By 21.

Murrugarra, E. and Valdivia, M. (2000). The returns to health for Peruvian urban adults by gender, age, and across the wage distribution. In Savedoff, W. D. and Schultz, T. P., editors, *Wealth from Health: Linking Social Investments to Earnings in Latin America*, chapter 5, pages 151–188. Inter-American Development Bank, Washington DC.

Nardi, M. D., Pashchenko, S., and Porapakkarm, P. (2017). The lifetime costs of bad health. Working Paper 23963, National Bureau of Economic Research.

Okeke, E. N. and Chari, A. (2018). Health care at birth and infant mortality: Evidence from nighttime deliveries in Nigeria. *Social Science and Medicine*, 196(Supplement C):86 – 95.

Ordinioha, B. and Onyenaporo, C. (2010). Experience with the use of community health extension workers in primary care, in a private rural health care institution in South-South Nigeria. *Annals of African Medicine*, 9(4):240–245.

O'Toole, J., Sinclair, M., and Leder, K. (2008). Maximising response rates in household telephone surveys. *BMC medical research methodology*, 8(1):71.

Peabody, J. W., Shimkhada, R., Quimbo, S., Solon, O., Javier, X., and McCulloch, C. (2013). The impact of performance incentives on child health outcomes: results from a cluster randomized controlled trial in the philippines. *Health Policy and Planning*, pages 1–7.

Poterba, J. M., Venti, S. F., and Wise, D. A. (2017). The asset cost of poor health. *The Journal of the Economics of Ageing*, 9:172–184.

Rethans, J.-J., Sturmans, F., Drop, R., Van Der Vleuten, C., and Hobus, P. (1991). Does competence of general practitioners predict their performance? comparison between examination setting and actual practice. *British Medical Journal*, 303(6814):1377–1380.

Rex, D. K., Hewett, D. G., Raghavendra, M., and Chalasani, N. (2010). The impact of video-recording on the quality of colonoscopy performance: a pilot study. *The American journal of gastroenterology*, 105(11):2312.

Ribero, R. and Núñez, J. (2000). Adult morbidity, height, and earnings in Colombia. In Savedoff, W. D. and Schultz, T. P., editors, *Wealth from Health: Linking Social Investments to Earnings in Latin America*, chapter 4, pages 111–150. Inter-American Development Bank, Washington DC.

Rowe, A. K., Onikpo, F., Lama, M., and Deming, M. S. (2012). Evaluating health worker performance in benin using the simulated client method with real children. *Implementation Science*, 7(1):95.

Rowe, S. Y., Olewe, M. A., Kleinbaum, D. G., McGowan, Jr, J. E., McFarland, D. A., Rochat, R., and Deming, M. S. (2006). The influence of observation and setting on community health workers' practices. *International Journal for Quality in Health Care*, 18(4):299–305.

Schapiro, L. (1919). The physical and economic benefits of treatment for hookworm disease. *Journal of the American Medical Association*, 73(20):1507–1509.

Schultz, T. P. (2005). Productive benefits of health: Evidence from low-income countries. *Health and Economic Growth: Findings and Policy Implications. MIT Press, Cambridge MA*, pages 257–286.

Schultz, T. P. and Tansel, A. (1997). Wage and labor supply effects of illness in Cote d'Ivoire and Ghana: Instrumental variable estimates for days disabled. *Journal of development economics*, 53(2):251–286.

Smith, P., Mossialos, E., Papanicolas, I., and Leatherman, S. (2010). *Performance measurement for health system improvement: Experiences, challenges and prospects*. Cambridge University Press, United Kingdom.

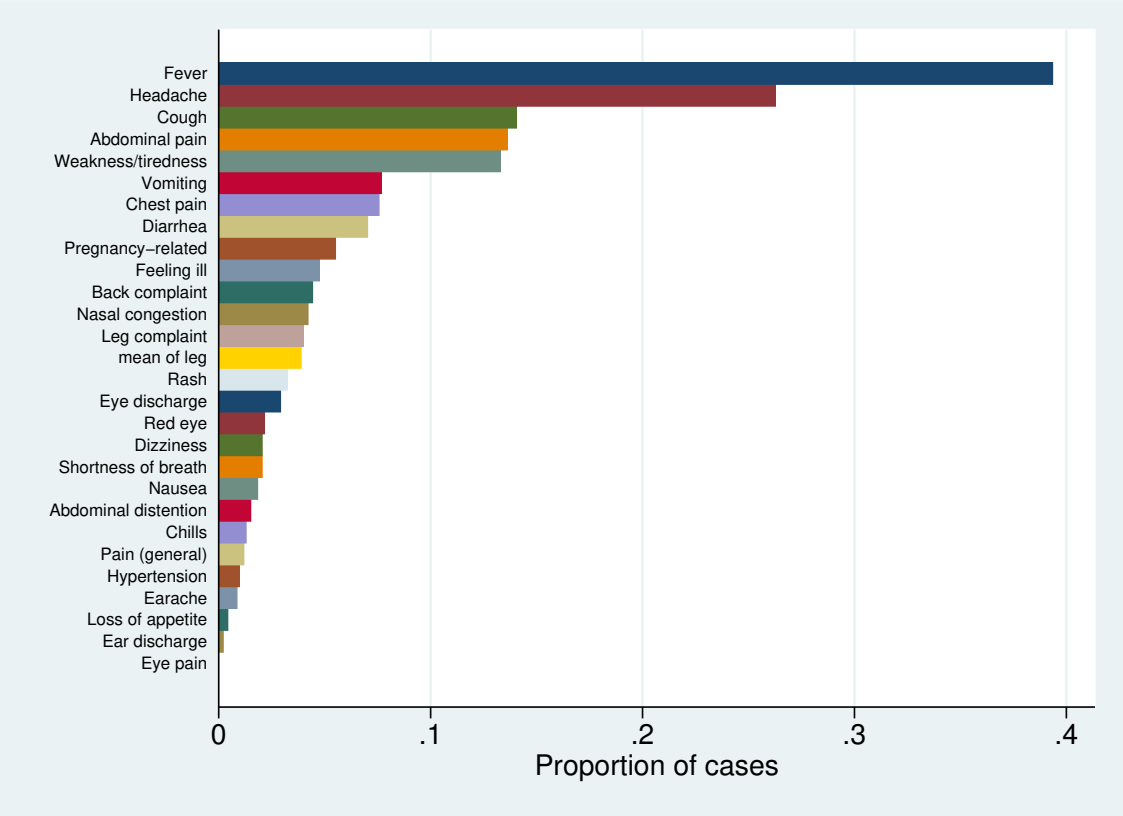
Stewart, A. L., Ware, J. E., and Ware Jr, J. E. (1992). *Measuring functioning and well-being: the medical outcomes study approach*. duke university Press.

Tauchmann, H. (2014). Lee (2009) treatment-effect bounds for nonrandom sample selection. *Stata Journal*, 14(4):884–894.

Weil, D. N. (2007). Accounting for the effect of health on economic growth. *The Quarterly Journal of Economics*, 122(3):1265–1306.

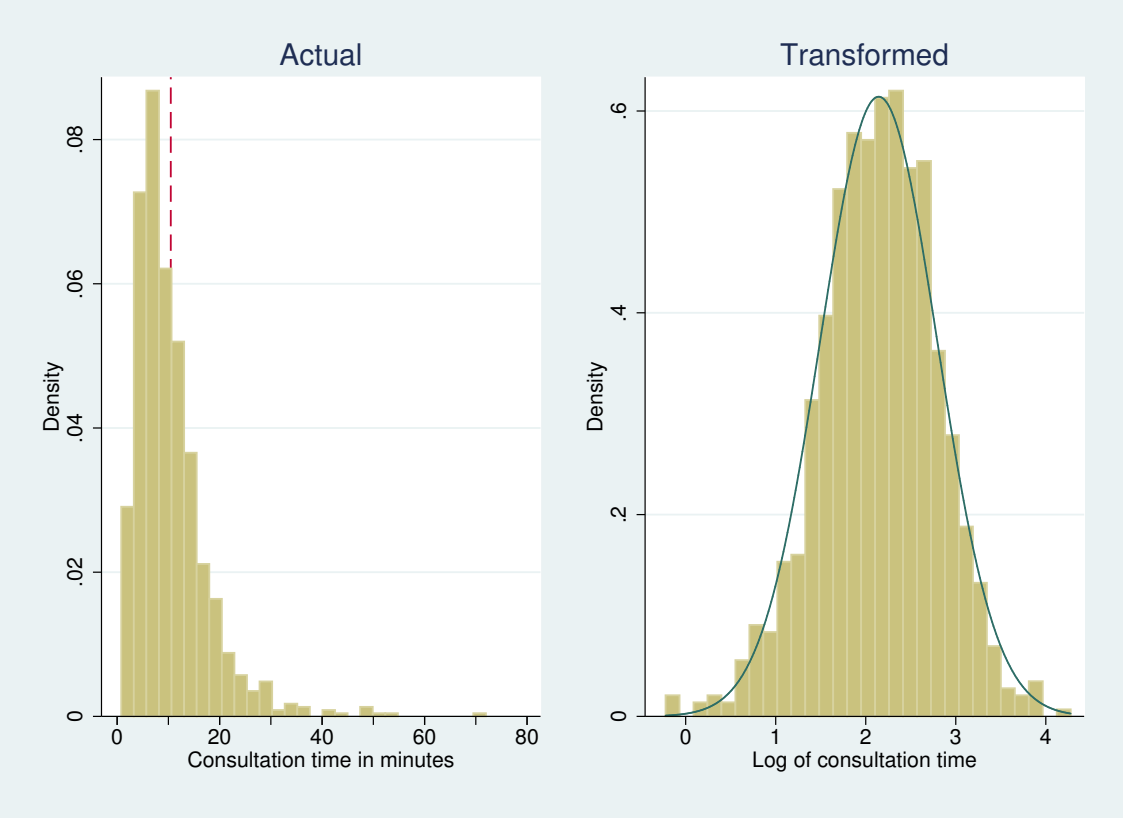
World Bank (2018). *The State of Social Safety Nets 2018*. World Bank, Washington DC.

Figure 1: Common Presenting Complaints



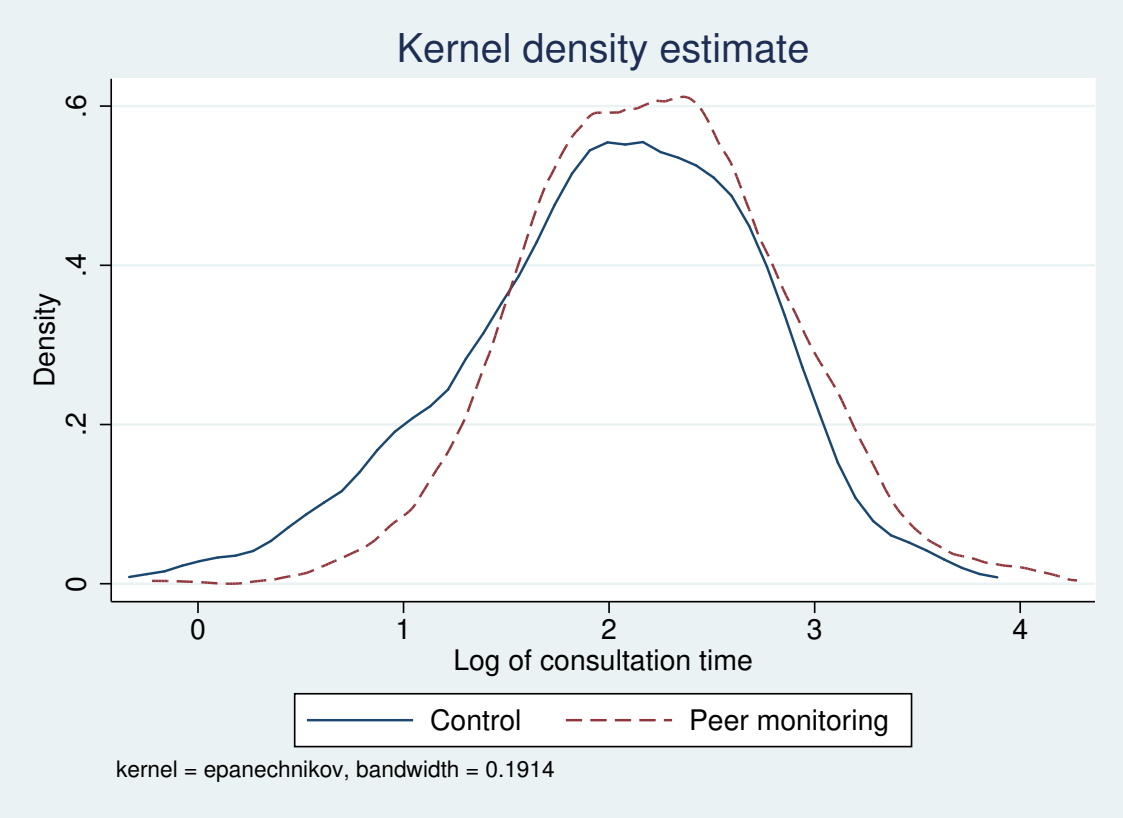
Note: Figure shows the distribution of presenting complaints based on the proportion of cases where a symptom was mentioned

Figure 2: Distribution of Consultation Time



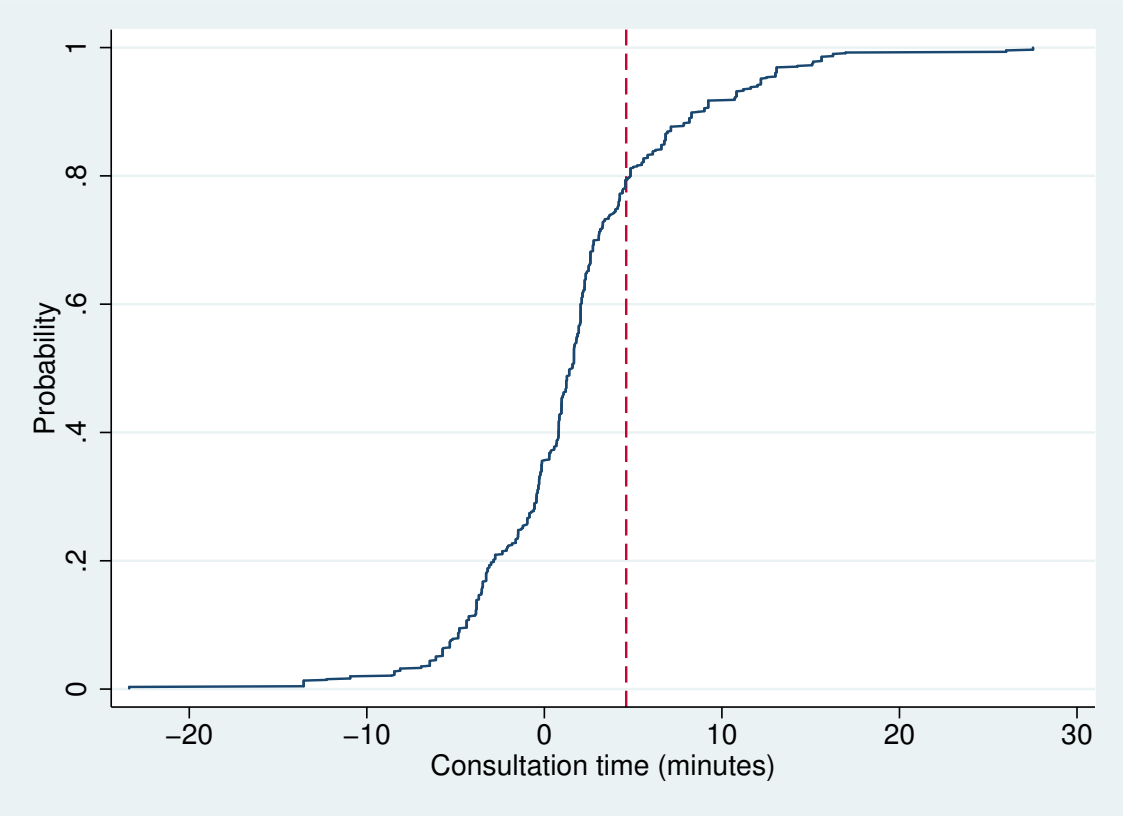
Note: Dotted line in Panel A denotes the mean consultation time

Figure 3: Distribution of Consultation Time by Treatment Status



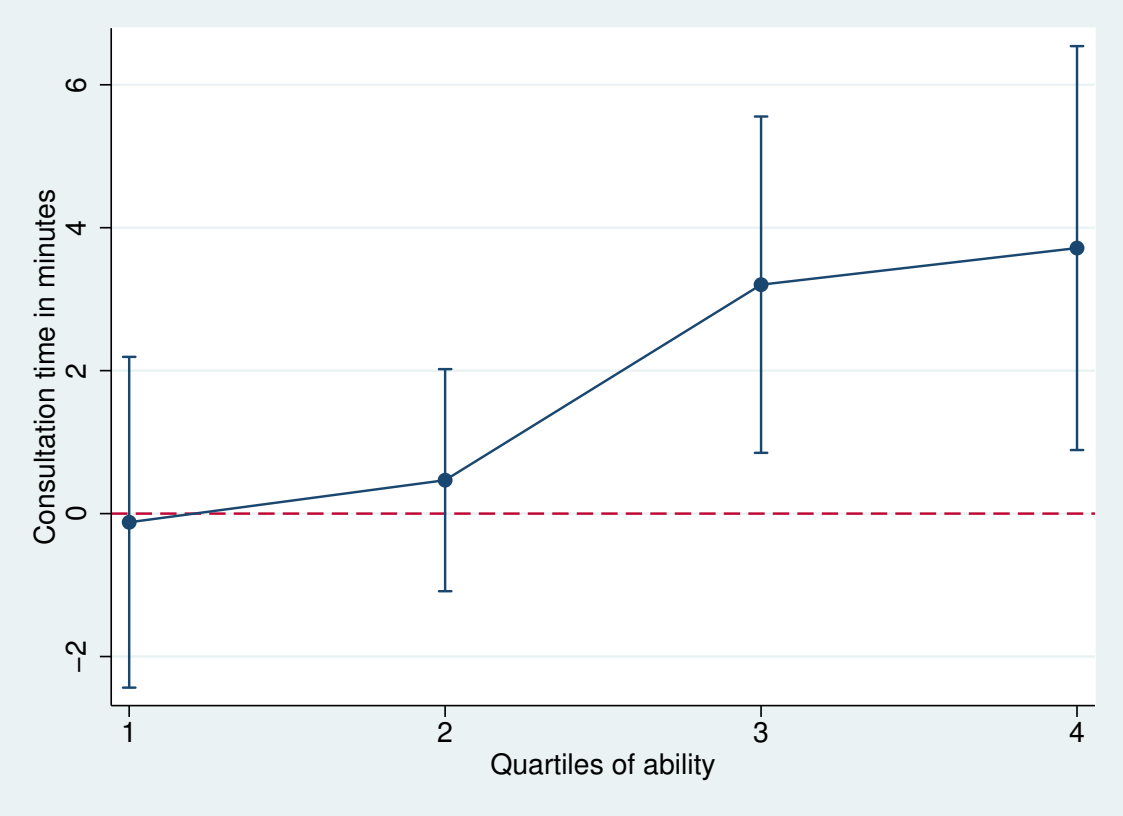
The figure shows the pdf of transformed consultation times for overtly vs. covertly monitored consultations.

Figure 4: CDF of Treatment Effect



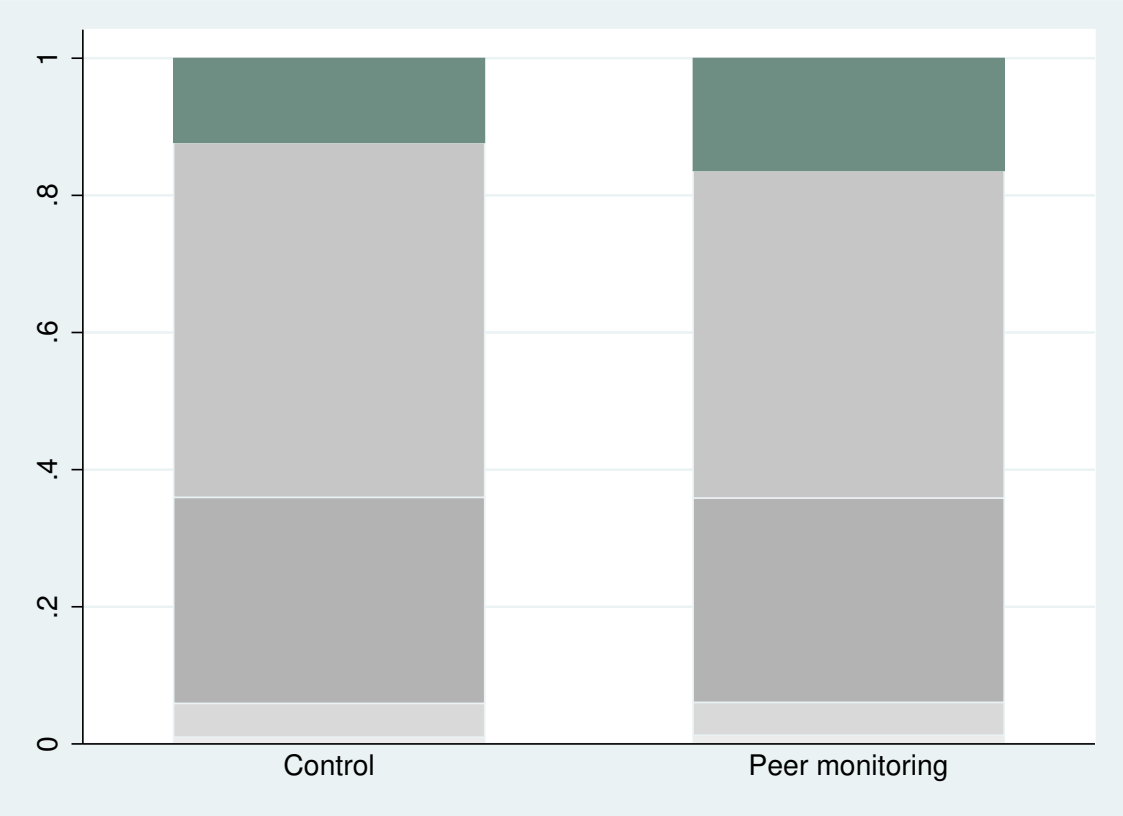
The figure shows the CDF of the difference in consultation time between the treatment and control groups. The dotted line indicates a 50% increase in consultation length.

Figure 5: Monitoring and Clinician Ability



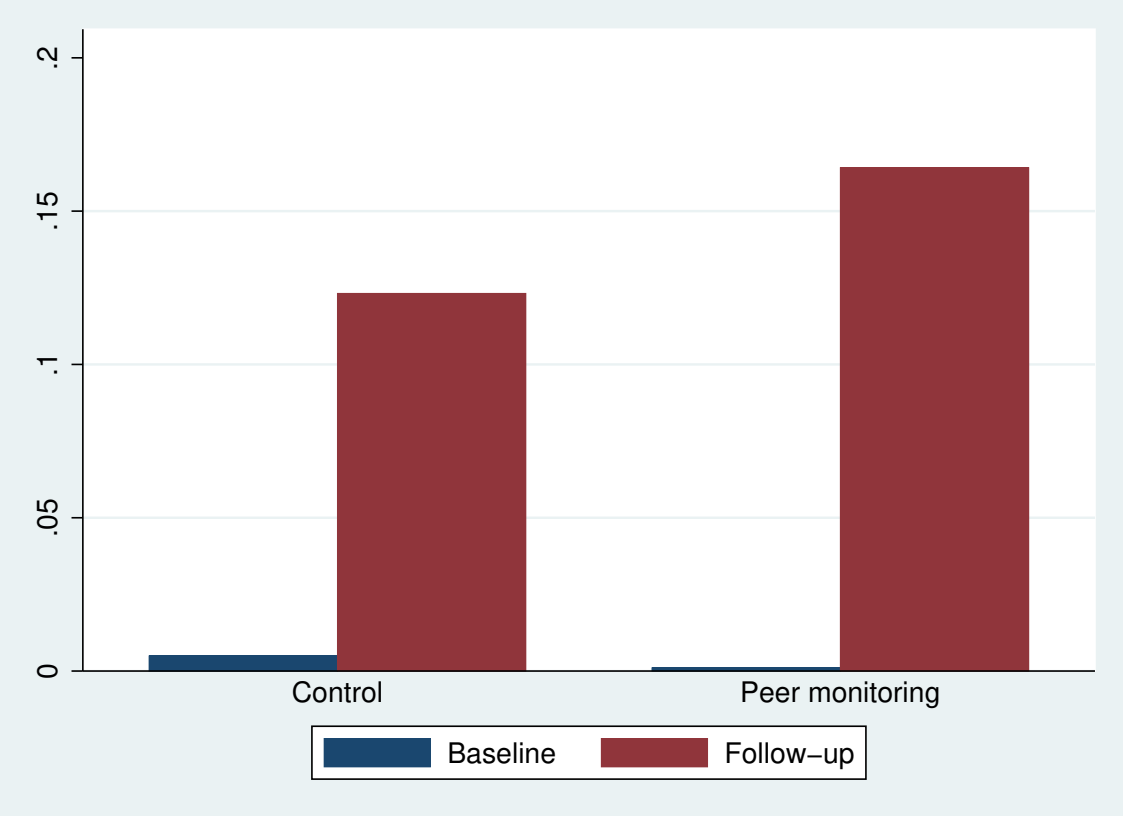
The figure examines how the effect of monitoring varies with clinician ability. It shows coefficients and 95% confidence intervals from a regression of consultation time on the treatment indicator interacted with dummies for each ability quartile. The ability index is derived by applying principal component analysis to the following variables: (i) the clinician's score on a multiple-choice assessment, (ii) percent of recommended history-taking questions asked, (iii) percent of the clinician's patients that received a physical examination, (iv) percent of cases where a diagnostic test was ordered, and (v) percent of the clinician's patients that received a diagnosis.

Figure 6: Patient Health at Follow-up



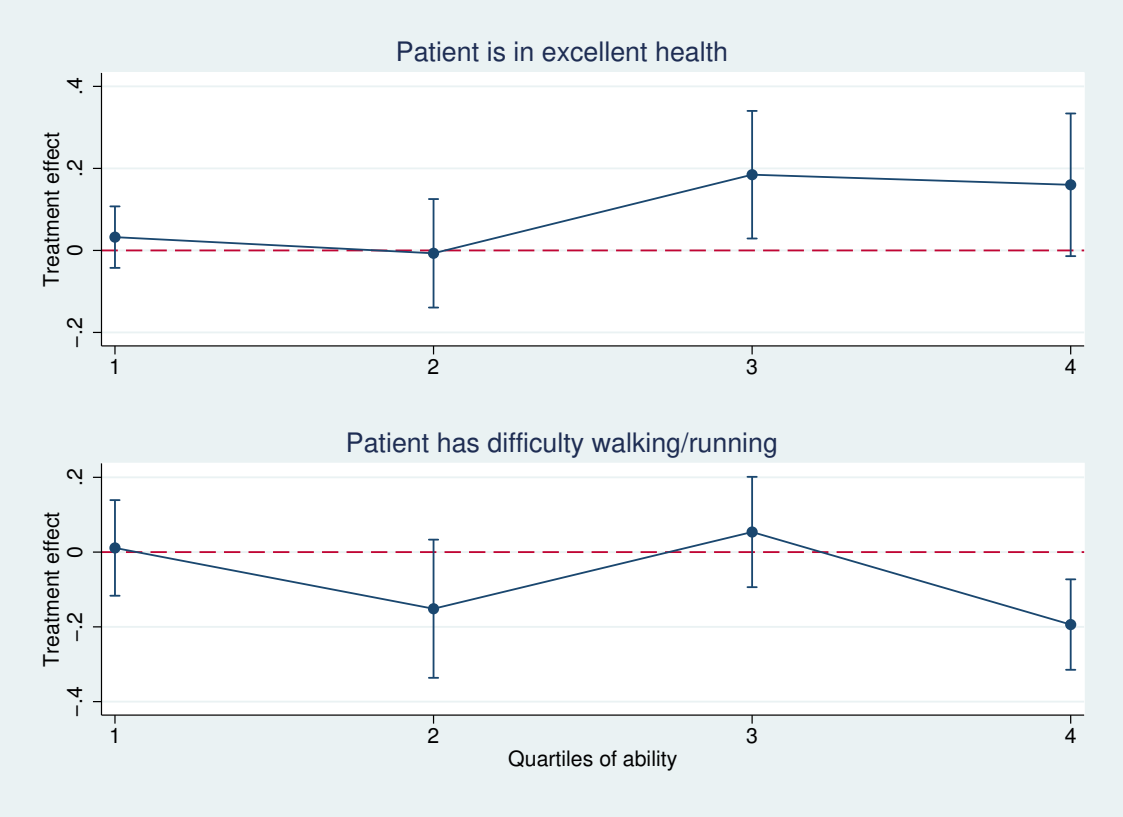
The figure shows the proportion of patients in each health category. From top to bottom: Excellent, Very good, Good, Fair, and Poor.

Figure 7: Fraction of patients in excellent health at baseline and follow-up by treatment status



The figure shows the proportion of patients reporting excellent health at baseline/follow-up by treatment status

Figure 8: Clinician Ability and Patient Health



The figure examines how the effect of the treatment on patient health varies with clinician ability. It shows coefficients and 95% confidence intervals from a regression of patient health at follow-up on the treatment indicator interacted with dummies for each ability quartile. In Panel A the dependent variable is an indicator for excellent health at follow-up and in Panel B the dependent variable is an indicator for difficulty in climbing up a flight of stairs or running the length of a football field. The ability index is derived by applying principal component analysis to the following variables: (i) the clinician's score on a multiple-choice assessment, (ii) percent of recommended history-taking questions asked, (iii) percent of the clinician's patients that received a physical examination, (iv) percent of cases where a diagnostic test was ordered, and (v) percent of the clinician's patients that received a diagnosis.

Table 1: Characteristics of Health Clinics and Providers

VARIABLES	Mean	SD
Clinic Characteristics (N=124)		
Number of health providers	5.839	2.724
Monthly patient volume	647.8	760.4
Offers inpatient services	0.716	0.453
Number of beds	15.452	9.116
Catchment area population	17366	17973
Offers PMTCT services	0.653	0.478
Offers C-section	0.073	0.260
Has electricity	0.274	0.448
Has backup generator	0.524	0.501
Has running water	0.460	0.500
Has Flush Toilet	0.355	0.480
Has ambulance	0.282	0.452
Has a laboratory	0.589	0.494
Has an operating theater	0.056	0.232
Provider Characteristics (N=242)		
Sex: Male	0.620	0.486
Age at last birthday	34.3	8.4
Married	0.620	0.486
Doctor	0.240	0.428
Community Health Officer	0.116	0.321
Nurse/midwife	0.066	0.249
Community Health Extension Worker (CHEW)	0.430	0.496
Junior CHEW	0.091	0.288
Other qualification	0.058	0.234
Years working with qualification	7.95	7.50
Years working in clinic	2.58	4.17
Percent of clinic outpatient consultations	26.01	17.58

Table shows characteristics of primary health clinics and providers included in the study. PMTCT stands for Prevention of Mother to Child Transmission of HIV.

Table 2: Baseline Patient Sample

VARIABLES	Total (N=925)	Control (N=305)	Treatment (N=620)	C-T
Male	0.356	0.341	0.363	-0.022
	(0.479)	(0.475)	(0.481)	[0.734]
Age at last birthday	23.26	21.60	24.08	-2.48
	(18.29)	(17.13)	(18.80)	[1.969]
Transportation: walked to the clinic	0.590	0.620	0.576	0.044
	(0.492)	(0.486)	(0.495)	[-1.317]
Transportation: own car or motorcycle	0.330	0.308	0.340	-0.032
	(0.470)	(0.463)	(0.474)	[1.000]
Travel time to clinic (minutes)	21.00	20.01	21.49	-1.48
	(20.69)	(16.16)	(22.59)	[1.089]
Illness Severity Score (0-10)	5.718	5.603	5.774	-0.171
	(1.953)	(2.066)	(1.895)	[1.211]
Severity Score ≥ 7	0.345	0.351	0.342	0.009
	(0.476)	(0.478)	(0.475)	[-0.284]
Overall health: Health is Excellent	0.0281	0.0393	0.0226	0.0167
	(0.165)	(0.195)	(0.149)	[-1.456]
Functional health: Has difficulty walking/running	0.658	0.656	0.660	-0.004
	(0.475)	(0.476)	(0.474)	[0.123]
Complex presentation	0.148	0.164	0.140	0.024
	(0.355)	(0.371)	(0.348)	[-0.998]
Presenting complaint: Fever	0.394	0.393	0.394	-0.001
	(0.489)	(0.489)	(0.489)	[0.003]
Presenting complaint: Headache	0.263	0.289	0.250	0.039
	(0.440)	(0.454)	(0.433)	[-1.125]
Presenting complaint: Cough	0.141	0.121	0.150	-0.029
	(0.348)	(0.327)	(0.357)	[1.105]
Presenting complaint: Abdominal pain	0.136	0.138	0.135	0.003
	(0.343)	(0.345)	(0.343)	[-0.110]
Presenting complaint: Weakness/Fatigue	0.133	0.144	0.127	0.017
	(0.340)	(0.352)	(0.334)	[-0.727]
Presenting complaint: Vomiting	0.0768	0.0721	0.0790	-0.0069
	(0.266)	(0.259)	(0.270)	[0.335]
Presenting complaint: Chest pain	0.0757	0.0656	0.0806	-0.015
	(0.265)	(0.248)	(0.273)	[0.954]
Presenting complaint: Diarrhea	0.0703	0.0557	0.0774	-0.0217
	(0.256)	(0.230)	(0.267)	[1.215]
Presenting complaint: Pregnancy-related	0.0551	0.0361	0.0645	-0.0284
	(0.228)	(0.187)	(0.246)	[1.810]
Presenting complaint: Feeling ill	0.0476	0.0426	0.0500	-0.0074
	(0.213)	(0.202)	(0.218)	[0.512]

The sample consists of 925 patients enrolled in the study. To measure overall health patients were asked to rate their using a categorical scale ranging from poor to excellent. The dependent variable is an indicator denoting excellent health. The variable shown in the table includes patients in very good health because only 3 out of 925 patients reported excellent health at baseline. Functional health was assessed by asking patients how much difficulty they would have walking up a flight of stairs or running the length of a football field (none, some, or a lot). The dependent variable is an indicator denoting at least some difficulty. A complex presentation is defined as a patient presenting with generalized pain or weakness. The table also includes the ten most frequently reported complaints. Standard deviations in parentheses. Column 4 reports tests of difference. t statistics are shown in brackets

Table 3: Follow-up Patient Sample

VARIABLES	Total (N=599)	Control (N=203)	Treatment (N=396)	C-T
Male	0.362	0.355	0.366	-0.011
	(0.481)	(0.480)	(0.482)	[0.295]
Age at last birthday	23.88	22.69	24.49	-1.8
	(18.82)	(18.28)	(19.08)	[1.150]
Transportation: walked to the clinic	0.614	0.635	0.604	0.031
	(0.487)	(0.482)	(0.490)	[-0.768]
Transportation: own car or motorcycle	0.309	0.310	0.308	0.002
	(0.462)	(0.464)	(0.462)	[-0.060]
Travel time to clinic (minutes)	20.85	19.31	21.63	-2.32
	(22.91)	(16.47)	(25.57)	[1.156]
Illness Severity Score (0-10)	5.603	5.512	5.649	-0.137
	(1.910)	(1.991)	(1.868)	[0.772]
Severity Score ≥ 7	0.319	0.330	0.313	0.017
	(0.466)	(0.471)	(0.464)	[-0.436]
Overall health: Health is Excellent	0.0267	0.0493	0.0152	0.0341
	(0.161)	(0.217)	(0.122)	[-2.078]
Functional health: Has difficulty walking/running	0.646	0.596	0.672	-0.076
	(0.479)	(0.492)	(0.470)	[1.780]
Complex presentation	0.140	0.163	0.129	0.034
	(0.348)	(0.370)	(0.335)	[-1.145]
Presenting complaint: Fever	0.382	0.365	0.391	-0.026
	(0.486)	(0.482)	(0.489)	[0.572]
Presenting complaint: Headache	0.250	0.291	0.230	0.061
	(0.434)	(0.455)	(0.421)	[-1.379]
Presenting complaint: Cough	0.129	0.123	0.131	-0.008
	(0.335)	(0.329)	(0.338)	[0.271]
Presenting complaint: Abdominal pain	0.120	0.133	0.114	0.019
	(0.325)	(0.340)	(0.318)	[-0.695]
Presenting complaint: Weakness/Fatigue	0.125	0.138	0.119	0.019
	(0.331)	(0.346)	(0.324)	[-0.648]
Presenting complaint: Vomiting	0.0785	0.0640	0.0859	-0.0219
	(0.269)	(0.245)	(0.281)	[0.825]
Presenting complaint: Chest pain	0.0768	0.0690	0.0808	-0.0118
	(0.266)	(0.254)	(0.273)	[0.539]
Presenting complaint: Diarrhea	0.0618	0.0394	0.0732	-0.0338
	(0.241)	(0.195)	(0.261)	[1.572]
Presenting complaint: Pregnancy-related	0.0484	0.0296	0.0581	-0.0285
	(0.215)	(0.170)	(0.234)	[1.388]
Presenting complaint: Feeling ill	0.0484	0.0197	0.0631	-0.0434
	(0.215)	(0.139)	(0.244)	[2.686]

The sample consists of 599 patients successfully interviewed at follow-up. To measure overall health patients were asked to rate their current state of health using a categorical scale ranging from poor to excellent. The dependent variable is an indicator denoting excellent health. The variable shown in the table includes patients in very good health because only 2 out of 599 patients reported excellent health at baseline. Functional health was assessed by asking patients how much difficulty they would have walking up a flight of stairs or running the length of a football field (none, some, or a lot). The dependent variable is an indicator denoting at least some difficulty. A complex presentation is defined as a patient presenting with generalized pain or weakness. The table also includes the ten most frequently reported complaints. Standard deviations in parentheses. Column 4 reports tests of difference. t statistics are shown in brackets

Table 4: Monitoring and Effort

VARIABLES	(1)	(2)	(3)	(4)
	Log of Consultation Time			
Monitoring	0.206***	0.216***	0.221***	0.210***
	(0.060)	(0.059)	(0.057)	(0.056)
Observations	925	916	891	891
R-squared	0.072	0.355	0.494	0.512
Fixed Effects	State only	Clinic	Provider	Provider
Additional Controls	No	No	No	Yes

The dependent variable is log-transformed consultation time. Monitoring denotes a consultation where the peer clinician was present. Additional controls in Column 4 include dummies for patient age and sex, an indicator for mode of transportation to the clinic (own car or motorcycle=1), an indicator for severe illness (severity score ≥ 7), an indicator for a fever presentation, an indicator for a pregnancy-related presentation, an indicator for a complex presentation (patients presenting with generalized pain or weakness), and indicators for baseline health.Standard errors in parentheses are clustered at clinic level *** p<0.01, ** p<0.05, * p<0.1

Table 5: Do clinicians do more when being monitored?

VARIABLES	(1)	(2)	(3)	(4)
	Was patient prescribed any medicines?			
Monitoring	0.075**	0.069**	0.094***	0.081***
	(0.029)	(0.030)	(0.033)	(0.030)
Observations	599	590	547	547
R-squared	0.198	0.384	0.484	0.519
Fixed Effects	State only	Clinic	Provider	Provider
Additional Controls	No	No	No	Yes

The dependent variable is an indicator for whether the patient was prescribed any medicines by the provider during their visit. The control group mean is 0.84. The sample consists of patients interviewed at follow-up. Monitoring denotes a consultation where the peer clinician was present. Additional controls in Column 4 include dummies for patient age and sex, an indicator for mode of transportation to the clinic (own car or motorcycle=1), an indicator for severe illness (severity score ≥ 7), an indicator for a fever presentation, an indicator for a pregnancy-related presentation, an indicator for a complex presentation (patients presenting with generalized pain or weakness), and indicators for baseline health. Standard errors in parentheses are clustered at clinic level *** p<0.01, ** p<0.05, * p<0.1

Table 6: Consultation Time and Clinician Effort

VARIABLES	(1) History-taking (Fever)	(2) Any Physical Exam	(3) Number of Physical Exams	(4) Any Lab Tests Ordered	(5) Number of Lab Tests	(6) Made a diagnosis	(7) Medicines Prescribed	(8) Explains treatment	(9) Provides health education
Log of consultation time	0.016*	0.068***	0.556***	0.091***	0.213***	0.037**	0.326***	0.016	0.045**
	(0.008)	(0.018)	(0.155)	(0.024)	(0.050)	(0.017)	(0.082)	(0.017)	(0.020)
Observations	899	1,905	1,905	1,905	1,905	1,905	1,905	1,905	1,905
R-squared	0.730	0.458	0.690	0.514	0.542	0.545	0.390	0.534	0.531
Mean of dependent variable	0.253	0.835	3.306	0.519	0.871	0.679	3.476	0.311	0.315

Table shows coefficients from regressions of various indicators of clinician effort on consultation length. The dependent variable in Column 1 is the percent of recommended history taking questions asked for patients presenting with fever; in Column 2 it is the probability that the clinician carried out a physical exam; in Column 3, it is the number of physical examinations; in Column 4 it is the probability that a lab test was ordered; in Column 5 it is the number of lab tests; in Column 6 it is the probability that the clinician made a diagnosis; in Column 7 it is the number of medicines prescribed, in Column 8 it is the probability that the clinician explained the treatment being provided, and in Column 9 it is the probability that the clinician provided health education related to the diagnosis. The sample consists of all patient consultations that were observed. All models include provider fixed effects and the following controls: dummies for patient age and sex, an indicator for mode of transportation to the clinic (own car or motorcycle=1), an indicator for severe illness (severity score ≥ 7), an indicator for a fever presentation, an indicator for a pregnancy-related presentation, an indicator for a complex presentation (patients presenting with generalized pain or weakness), and indicators for baseline health. Standard errors in parentheses are clustered at clinic level *** p<0.01, ** p<0.05, * p<0.1

Table 7: Clinician Effort and Patient Health

	Overall Health				Functional Health			
VARIABLES	ITT			IV	ITT			IV
Monitoring	0.077**	0.077**	0.080**		-0.061	-0.075	-0.073*	
	(0.032)	(0.034)	(0.033)		(0.047)	(0.047)	(0.042)	
Effort				0.425*				-0.388
				(0.214)				(0.290)
Observations	547	547	547	547	547	547	547	547
R-squared	0.454	0.479	0.500	0.218	0.343	0.355	0.425	0.254
Controls for baseline health	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Additional Controls	No	No	Yes	Yes	No	No	Yes	Yes

Monitoring denotes a consultation where the peer clinician was present. To measure overall health patients were asked to rate their current state of health using a categorical scale ranging from poor to excellent. The dependent variable is an indicator denoting excellent health (the mean is 0.15). Functional health was assessed by asking patients how much difficulty they would have walking up a flight of stairs or running the length of a football field (none, some, or a lot). The dependent variable is an indicator denoting at least some difficulty (the mean is 0.22). A negative coefficient therefore indicates an improvement in functional health. Additional controls include dummies for patient age and sex, an indicator for mode of transportation to the clinic (own car or motorcycle=1), an indicator for severe illness (severity score ≥ 7), an indicator for a fever presentation, an indicator for a pregnancy-related presentation, and an indicator for a complex presentation (patients presenting with generalized pain or weakness). Columns 1-3 in each panel are the ITT specifications and Column 4 is the IV specification where clinician effort (measured by length of the consultation) is instrumented with the treatment indicator. The Wald F-Statistic is 8.53. The sample consists of patients that were successfully contacted and interviewed. Standard errors in parentheses are clustered at clinic level *** p<0.01, ** p<0.05, * p<0.1

Table 8: Is there a dose-response relationship?

VARIABLES	(1) Log of Consultation Time	(2) Overall Health	(3) Functional Health
Monitoring	0.068	0.032	0.011
	(0.120)	(0.038)	(0.065)
Monitoring x Ability Quartile 2	0.043	-0.039	-0.163
	(0.160)	(0.078)	(0.112)
Monitoring x Ability Quartile 3	0.316*	0.152*	0.043
	(0.173)	(0.083)	(0.101)
Monitoring x Ability Quartile 4	0.204	0.128	-0.205**
	(0.152)	(0.095)	(0.084)
Observations	836	516	516
R-squared	0.517	0.521	0.424
Fixed Effects	Provider	Provider	Provider
Additional Controls	Yes	Yes	Yes

The results shown are from a regression of the dependent variable on the treatment interacted with clinician ability (dummies for each quartile). The ability index is derived by applying principal component analysis to the following variables: (i) the clinician's score on a multiple-choice assessment, (ii) percent of recommended history-taking questions asked, (iii) percent of the clinician's patients that received a physical examination, (iv) percent of cases where a diagnostic test was ordered, and (v) percent of the clinician's patients that received a diagnosis. Monitoring denotes a consultation where the peer clinician was present. To measure overall health patients were asked to rate their current state of health using a categorical scale ranging from poor to excellent. The dependent variable is an indicator denoting excellent health (the mean is 0.15). Functional health was assessed by asking patients how much difficulty they would have walking up a flight of stairs or running the length of a football field (none, some, or a lot). The dependent variable is an indicator denoting at least some difficulty (the mean is 0.22). A negative coefficient therefore indicates an improvement in functional health. Additional controls include dummies for patient age and sex, an indicator for mode of transportation to the clinic (own car or motorcycle=1), an indicator for severe illness (severity score ≥ 7), an indicator for a fever presentation, an indicator for a pregnancy-related presentation, an indicator for a complex presentation (patients presenting with generalized pain or weakness), and indicators for baseline health. The sample consists of patients that were successfully contacted and interviewed. Standard errors in parentheses are clustered at clinic level *** p<0.01, ** p<0.05, * p<0.1

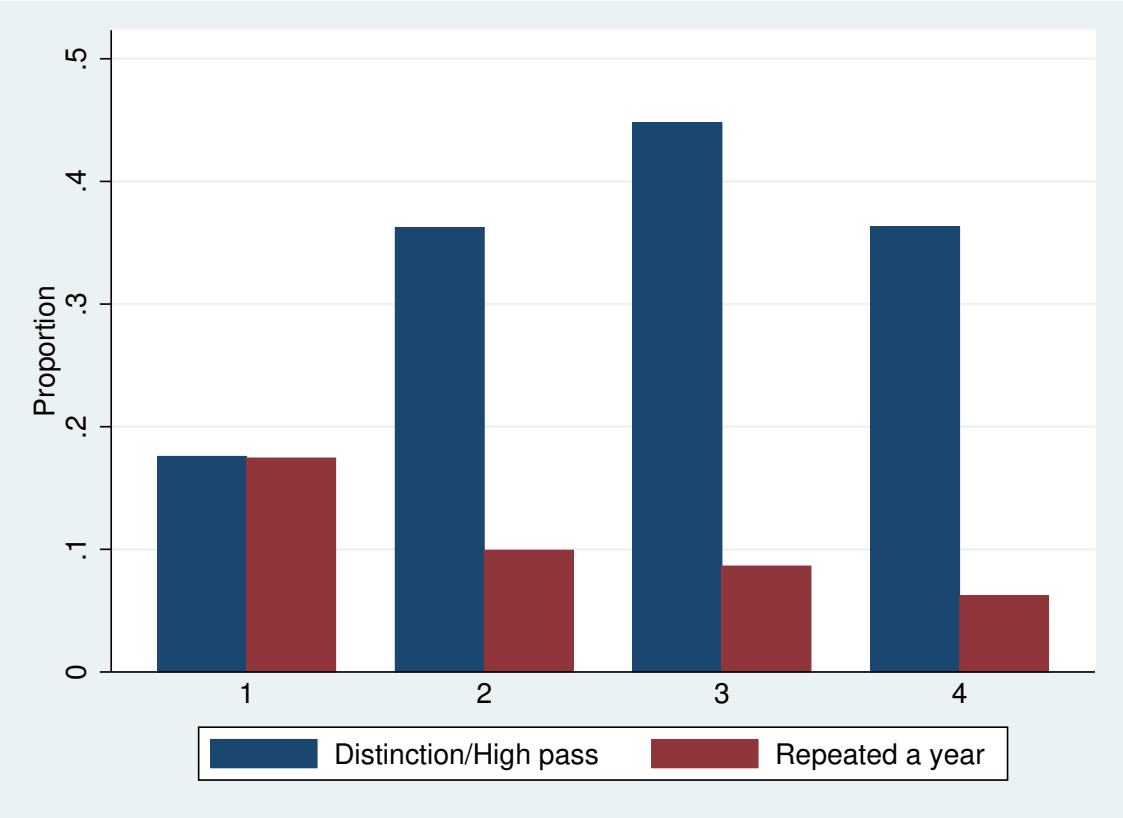
Table 9: Treatment Effect Heterogeneity by Illness Severity and Case Complexity

	(1)	(2)	(3)	(4)
VARIABLES	Log of Consultation Time	Overall Health	Log of Consultation Time	Overall Health
Monitoring	0.200***	0.081**	0.176***	0.074**
	(0.062)	(0.038)	(0.059)	(0.036)
Severity score ≥ 7	-0.056	0.041		
	(0.107)	(0.067)		
Monitoring x Severity score ≥ 7	0.031	-0.003		
	(0.092)	(0.066)		
Complex presentation			-0.138	-0.033
			(0.117)	(0.076)
Monitoring x Complex presentation			0.208	0.042
			(0.135)	(0.084)
Observations	891	547	891	547
R-squared	0.512	0.500	0.514	0.500
Fixed Effects	Provider	Provider	Provider	Provider
Additional Controls	Yes	Yes	Yes	Yes

Column header indicates the dependent variable. Monitoring denotes a consultation where the peer clinician was present. Illness severity was assessed on a scale ranging from 0 to 10. I define a case as severe if the severity score is greater than or equal to seven. A complex presentation is defined as a patient presenting with generalized pain or weakness. To measure overall health patients were asked to rate their current state of health using a categorical scale ranging from poor to excellent. The dependent variable is an indicator denoting excellent health (the mean is 0.15). Functional health was assessed by asking patients how much difficulty they would have walking up a flight of stairs or running the length of a football field (none, some, or a lot). The dependent variable is an indicator denoting at least some difficulty (the mean is 0.22). A negative coefficient therefore indicates an improvement in functional health. Additional controls include dummies for patient age and sex, an indicator for mode of transportation to the clinic (own car or motorcycle=1), an indicator for a fever presentation, an indicator for a pregnancy-related presentation, and indicators for baseline health. Standard errors in parentheses are clustered at clinic level *** p<0.01, ** p<0.05, * p<0.1

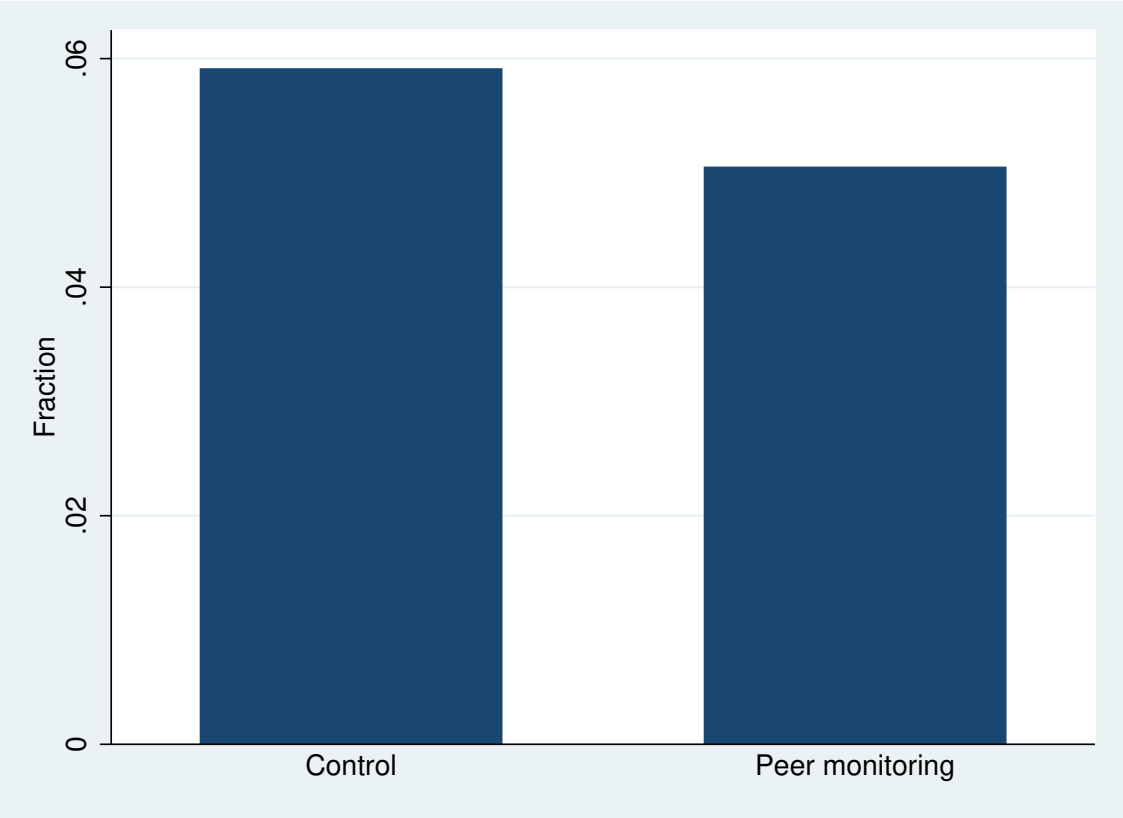
Appendix

Figure A.1: Association between Clinician Ability Index and Academic Performance



Note: The ability index was derived using principal component analysis where the model inputs are (i) the clinician's score on a clinical knowledge module, (ii) percent of recommended history-taking questions asked, (iii) percent of the clinician's patients that received a physical examination, (iv) percent of cases where a diagnostic test was ordered, and (v) percent of the clinician's patients that received a diagnosis. The index is divided into quartiles (shown on the X-axis). *Distinction* denotes a distinction or high pass on a course during training while *Repeated a year* denotes repeating a full year of training because of poor grades.

Figure A.2: Fraction of patients seeking care from a different provider or facility for the same illness



Note: The figure shows the proportion of patients in each group that answered in the affirmative to the following question: "Since your visit to [clinic] on [date], have you consulted another health provider or visited another health facility on account of the same illness?"

Table A.1. Monitoring and Health (Non-linear model)

	(1)	(2)	(3)	(4)
VARIABLES	**Overall health**		**Functional health**	
Monitoring	0.060*	0.060*	-0.032	-0.034
	(0.031)	(0.031)	(0.034)	(0.035)
Observations	599	599	599	599
Model	Logit with strata dummies	Logit with provider random effects	Logit with strata dummies	Logit with provider random effects

Monitoring denotes a consultation where the peer clinician was present. Coefficients shown are average marginal effects from logit models. Column headers indicate the dependent variable. To measure overall health patients were asked to rate their current state of health using a categorical scale ranging from poor to excellent. The dependent variable is an indicator denoting excellent health (the mean is 0.15). Functional health was assessed by asking patients how much difficulty they would have walking up a flight of stairs or running the length of a football field (none, some, or a lot). The dependent variable is an indicator denoting at least some difficulty (the mean is 0.22). A negative coefficient therefore indicates an improvement in functional health. Models include the following controls: dummies for patient age and sex, an indicator for mode of transportation to the clinic (own car or motorcycle=1), an indicator for severe illness (severity score ≥ 7), an indicator for a fever presentation, an indicator for a pregnancy-related presentation, an indicator for a complex presentation (patients presenting with generalized pain or weakness), and indicators for baseline health. Standard errors in parentheses are clustered at clinic level *** p<0.01, ** p<0.05, * p<0.1

Table A.2: Health Index

VARIABLES	(1) Index Score	(2) Index Score > Median
Monitoring	0.118	0.094***
	(0.093)	(0.033)
Observations	547	547
R-squared	0.369	0.483

To create the health index, each health measure is normalized to have a mean of zero and a standard deviation of one in the control group. The index is the mean of the normalized values. In Column 1 the dependent variable is the continuous index, and Column 2 it is an indicator for a score above the median. All models include provider fixed effects and the following controls: dummies for patient age and sex, an indicator for mode of transportation to the clinic (own car or motorcycle=1), an indicator for severe illness (severity score ≥ 7), an indicator for a fever presentation, an indicator for a pregnancy-related presentation, an indicator for a complex presentation (patients presenting with generalized pain or weakness), and indicators for baseline health. Monitoring denotes a consultation where the peer clinician was present. The sample consists of patients that were successfully contacted and interviewed. Standard errors in parentheses are clustered at clinic level *** p<0.01, ** p<0.05, * p<0.1

Table A.3. Lee Bounds

VARIABLES	Overall Health	
Monitoring	[0.036 0.078]	[0.049 0.091]
Includes covariates	No	Yes
Observations	925	925

Table shows lower and upper non-parametric Lee bounds. Monitoring denotes a consultation where the peer clinician was present. The dependent variable is an indicator denoting excellent health (the mean is 0.15). Column 2 includes baseline health dummies as covariates.

Table A.4. Treatment Adherence

VARIABLES	(1) Patient obtained medicines	(2) Patient took medicines	(3) Patient was very satisfied with care
Monitoring	-0.017	-0.063	0.008
	(0.047)	(0.047)	(0.041)
Observations	474	455	547
R-squared	0.435	0.487	0.441
Additional Controls	Yes	Yes	Yes

Monitoring denotes a consultation where the peer clinician was present. Column headers indicate the dependent variable: in Column 1 it is the probability that the patient obtained medicines prescribed by the provider; in Column 2 it is the probability that they took the medicines; and in Column 3 it is the probability that the patient was very satisfied with the quality of care provided. Satisfaction was assessed using a Likert scale. All models include provider fixed effects and the following controls: dummies for patient age and sex, an indicator for mode of transportation to the clinic (own car or motorcycle=1), an indicator for severe illness (severity score ≥ 7), an indicator for a fever presentation, an indicator for a pregnancy-related presentation, an indicator for a complex presentation (patients presenting with generalized pain or weakness), and indicators for baseline health. Standard errors in parentheses are clustered at clinic level *** p<0.01, ** p<0.05, * p<0.1

Table A.5. Clinician Ability and Patient Outomes

VARIABLES	Overall Health
Clinician ability (Quartile 2)	0.027
	(0.045)
Clinician ability (Quartile 3)	0.089**
	(0.042)
Clinician ability (Quartile 4)	0.136**
	(0.062)
Clinician Characteristics	
Female clinician	-0.065
	(0.043)
Clinician age	0.008**
	(0.003)
Clinician years of experience	-0.005
	(0.003)
Community Health Officer	-0.032
	(0.077)
Nurse/midwife	0.114
	(0.087)
Community Health Extension Worker (CHEW)	0.073
	(0.064)
Junior CHEW	0.059
	(0.110)
Other qualification	-0.003
	(0.077)
Clinic Characteristics	
Offers inpatient services	0.125**
	(0.049)
Has a laboratory	-0.069
	(0.059)
Has a pharmacy	0.030
	(0.101)
Clinic is somewhat dirty[1]	0.182*
	(0.097)
Clinic is clean	0.124*
	(0.068)
Clinic is very clean	0.380**
Observations	534
R-squared	0.237

The results shown are from a regression of patient health on clinician ability. The omitted group is clinicians in the bottom quartile. The ability index is derived by applying principal component analysis to the following variables: (i) the clinician's score on a multiple-choice assessment, (ii) percent of recommended history-taking questions asked, (iii) percent of the clinician's patients that received a physical examination, (iv) percent of cases where a diagnostic test was ordered, and (v) percent of the clinician's patients that received a diagnosis. Monitoring denotes a consultation observed by the peer clinician. To measure overall health patients were asked to rate their current state of health using a categorical scale ranging from poor to excellent. The dependent variable is an indicator denoting excellent health (the mean is 0.15). The model controls for the following patient characteristics: patient age and sex, an indicator for mode of transportation to the clinic (own car or motorcycle=1), an indicator for severe illness (severity score ≥ 7), an indicator for a fever presentation, an indicator for a pregnancy-related presentation, an indicator for a complex presentation (patients presenting with generalized pain or weakness), and indicators for baseline health. The sample consists of patients that were successfully contacted and interviewed. Standard errors in parentheses are clustered at clinic level *** p<0.01, ** p<0.05, * p<0.1

[1] Omitted group is clinic is very dirty.