The Market for Healthcare in Low Income Countries

Abhijit Banerjee, Abhijit Chowdhury, Jishnu Das, Jeffrey Hammer,
Reshmaan Hussam, Aakash Mohpal

July 16, 2023

Extended Abstract

Patient trust is an important driver of the demand for healthcare. But it may also impact supply: doctors who realize that patients may not trust them may adjust their behavior in response. We assemble a large dataset that assesses clinical performance using standardized-patients (akin to audit studies in economics) in low-income countries to investigate this possibility; most of these data are on healthcare providers who practice in the private sector on a fee-for-service basis. We establish that patients receive low quality of care, with a generous definition suggesting that fewer than 50% of cases are correctly managed, and between 70% and 90% of expenditures are medically unnecessary. Strikingly, and in contrast to literature suggesting that the main problem with fee-for-service provision is over-treatment, the majority of these unnecessary expenditures are incurred because patients are incorrectly rather than over-treated.

We then rule-out two plausible explanations for low quality of care: low levels of medical knowledge and low market incentives to invest effort. In our data, there are many healthcare providers who know how to correctly treat the patient and could substantially increase their revenue by doing so given the price-quality gradients we estimate, but still treat the patient incorrectly.

A model of the patient-provider relationship in which patients have incomplete information about the quality of providers generates predictions consistent with our findings. The theory additionally suggests that issuing a credible signal of quality should raise average quality of care among providers, even if their underlying ability remains unchanged. We assess this prediction through an evaluation of a highly-publicized training program with informal healthcare providers in West Bengal, India. The program has no impact on knowledge, yet substantially raises quality of care, leading to an increase in the likelihood of correct treatment, a 19% decline in unnecessary expenditures for patients and a 9% increase in revenues for providers. We conclude that low trust undermines clinical performance in an economically and medically significant manner.

1 Introduction

A patient walks into the doctor's office complaining of a headache. After examining her, the doctor recommends an expensive MRI test, telling the patient that she may have a "serious" problem. For the patient, this can pose a conundrum. She has heard stories of doctors taking side-payments from MRI providers and fears that the recommendation may be motivated by profit rather than her wellbeing. She is paying out-of-pocket for the consultation. How should she think of her doctor, and what should she do?

This problem of trust or 'credence' lies at the heart of most clinical interactions, and research from multiple settings shows that bundling diagnostics and treatment services increases the use of medically unnecessary procedures (Chen, Gertler, and Yang (2016); Yi et al. (2015)). Doctors react to (profit) incentives, even when it is not in the best interest of the patient.[1] Simultaneously, a lack of trust in the health system can depress demand, a phenomenon that has been documented globally (Lowes and Montero (2019); Martinez-Bravo and Stegmann (2018); Das and Das (2003)), and recently by Alsan and Wanamaker (2018) in the United States, who show a precipitous decline in the demand for curative care among African Americans in the aftermath of the infamous Tuskegee experiment.

What we know less about is how a patient's level of trust in a provider affects the behavior of the provider themselves. In a low-credence environment where doctors recognize that patients are suspicious of their motives, (s)he may withhold recommendations for a test or treatment even when it is medically necessary in order to build the confidence of his/her patients. As a result, the fear of overtreatment on the part of a patient may lead to under- or even incorrect treatment by a provider in the market for healthcare.[2]

This insight frames our investigation of the market for primary care in low-income countries, with a focus on India. We first establish a set of stylized facts regarding primary care in these settings using unique data from 5000+ standardized patient interactions (akin to in-person audits) as well as direct tests of medical knowledge among healthcare providers collected from five sites in India, a sixth in China, and seventh in Kenya. The studies from India include 4000+ interactions in the unregulated private sector where entry barriers are minimal, doctors operate on a fee-for-service basis, insurance for primary care is nonexistent, and prices and to some extent, quality, are market determined.[3]

[1] Policies often restrict the ability of doctors to provide ancillary services for precisely this reason, and the Sunshine Act in the U.S., which provides detailed information on every doctors' links to the pharma industry, are designed to further weaken this link.

[2] There is a literature on defensive medicine in the U.S., where doctors overprescribe tests because of the fear of malpractice suits especially in a context where patients do not bear the full price of treatment (McGuire and Pauly (1991); Dulleck and Kerschbamer (2006)). This is less of an issue in most low income healthcare settings–such as those in our sample—where patients pay out-of-pocket for their healthcare.

[3] Consistent with low entry barriers, Das et al. (2020) show that providers without formal medical training comprise 80% of the rural market and provide 70% of primary care. An average village in the country has three such "informal" providers. The fee-for-service private sector provides care for more than 70% of primary care episodes and more than 70% of primary healthcare expenditures is financed out-of-pocket by consumers. This is consistent with Grépin (2016) who find that across multiple low- and middle-income countries, the private sector provides 50% of primary care, a fraction that has remained remarkably steady since 1970. Data on provider qualifications typically show a

From this expansive sample, the first stylized fact we document is that the fraction of standardized patient (SP) interactions in which the provider correctly managed a case - across all samples - ranges from 0-2%. This very low fraction reflects, in part, the indiscriminate use of unnecessary medicines. But, more generous definitions of "correct case management" that do not penalize the prescription of unnecessary medicines still increase this fraction only to 10-50%, leaving a majority of SPs receiving the wrong treatment. This incorrect treatment in turn implies that 70-90% of all primary healthcare expenditures in our sample can be classified as medically avoidable.[4] Strikingly, the vast majority of avoidable expenditures is due to under- and incorrect treatment, in contrast to the concern that fee-for-service healthcare systems incentivize overtreatment.

We then consider three plausible explanations for this ubiquity of low quality care: low medical knowledge, overwhelmingly high caseloads, and inadequate price incentives.

To assess provider knowledge, we use 'medical vignettes', in which enumerators present the provider with a medical case that is identical to what is later presented through the SP and ask the provider for his or her preferred treatment. In this case, providers know they are being tested and we expect their recommendations to reflect their knowledge (Das and Hammer, 2005).

Combining medical vignettes with SP data reveals a substantial and persistent 'know-do' gap. Providers tell us, for instance, that they would recommend a chest x-ray and a sputum test to patients with three weeks of persistent coughing and a pattern of diurnal fever. This is the correct course of action for a patient suspected of having tuberculosis, which the case history should lead them towards. When a patient with the identical case appears at their clinic, however, they recommend neither. This know-do gap would be less concerning were its magnitude to decline with knowledge, a pattern that would suggest healthcare systems can "upskill" their way out of the gap. We find the opposite pattern: as knowledge increases, so does the know-do gap.

One explanation for this persistent know-do gap is high patient volumes. An optimizing provider who is time constrained will equalize the marginal benefit of time spent with every patient, thus potentially reducing the amount of effort they invest per patient and practicing below their knowledge frontier. However, we find that clinics in our sample are operating at substantial excess capacity. In the two samples where we observe providers for a full day, we find that providers spend four to eight hours in their clinics but less than one hour actively seeing patients. This is not only a feature of our sample. The World Bank's Service Delivery Indicators similarly exhibit massive excess capacity in primary care across eight Sub-Saharan African countries: the average primary health care center in Nigeria sees a single patient per day.

We then consider whether the know-do gap is due to inadequate returns to effort in the market.

dominance of non-physician clinicians in most rural areas while data on provider locations suggests that multiple providers practice close to each other in highly competitive market clusters.

[4]We assume that (a) if a patient receives correct and unnecessary treatments, only the cost portion of the unnecessary part is medically avoidable, but (b) if patient received incorrect treatment, the entire expenditure was medically avoidable. This could be an upper-bound if doctors use how patients react to a treatment as input into further diagnostics, and a lower bound given the financial costs of the health consequences of incorrect or medically avoidable treatment. Nevertheless, the approach will be useful to understand variation across healthcare providers and to benchmark the results from our experiment.

Using time spent and the completion of necessary checklist items as measures of effort, both of which are strongly associated with correct treatment in our data, we show that there is a positive price-effort gradient in the market. The gradient suggests that providers would have to value their time at $94 per day for the cost of effort to be higher than the revenue benefits. This compares to an estimated wage of $7.3 per day for male urban workers with secondary schooling, the sample most akin to our own, making it unlikely that the low net return to effort is the driving force behind the low quality we observe. Data on clinical interactions with real patients, which allows us to include provider fixed-effects and therefore compare different patient interactions for the same provider, exhibit similar patterns. In fact, the provider fixed-effect and the OLS specifications yield nearly identical effort-price coefficients, a first indication that patient sorting is not an obvious reason for this positive effort-price gradient.

One implication of this last observation is that providers with the same level of knowledge may exhibit variation in their clinical practice, which is associated with variation in effort. We observe this in our data. Relative to provider-patient interactions in the top 5th percentile of effort exerted, those in the bottom 5th percentile report 50% lower revenue, and this relationship is robust to the inclusion of provider fixed-effects.[5] It is also robust to the inclusion of patient characteristics, indicating that the variation we observe is not driven by (observed) differences in patient characteristics that are correlated with their demand for quality.

This set of patterns is consistent with ethnographic evidence that underscores the bilateral nature of healthcare provision. Das (2015) and Saria (2020b) argue that, in atmospheres of low trust, the clinical interaction is not an exercise of hierarchical power in which the doctor treats and the patient follows, but rather a negotiation. The patient is skeptical of (costly) treatment; the doctor takes into account the anticipated reactions of the patient in determining his/her own clinical recommendations. As Saria (2020a) describes,

> *Corruption in the health sector in Patna was perceived by patients as the prescribing of diagnostic tests and medicines that resulted in "cuts" (commissions) for the doctors.... A gynecologist remarked to me: "If I tell the patient straightaway that they have to go get tests that cost 10,000 rupees, he will run away and not come back. So I give them some medicines and then tell them to return after 3 days. Then I have to slowly ask them to get tests done."* (p.24)[6]

This anthropological evidence motivates a simple model of healthcare provision that is neither

[5] In this back-of-the-envelope calculation, effort is measured by the number of questions asked and exams completed out of a recommended list with standardized patients. Providers in the 5th percentile complete 6.25% of the recommended checklist and make approximately $4 per day, while those in the 95th percentile complete 50% and make approximately $8.6 per day.

[6] In another illustrative example (emphasis added), "The poor patients who came to the many clinics in the densely medical landscape of Patna with cough and fever were never tested for TB but were given medicines that would alleviate the symptoms associated with smoking and breathing the heavily polluted air. *This was not because doctors did not understand the severity of the epidemic, which they knew about it only too well. They hesitated in initiating investigation because immediate investigation and treatment would, as we have seen, raise suspicions in the minds of patients* and make them vulnerable to defaulting on their medication or worse." (Saria (2020a))

driven purely by supply (provider knowledge, incentives, or capacity constraints) nor just by demand (patient trust in the healthcare system), but is bilateral and mediated by the patient's belief in the quality of the provider. In the model, patients choose between multiple healthcare providers and can contract on the effort that providers exert, but lack information on the provider's ability. Providers can choose to (truthfully) reveal their 'type', but neither side can make commitments about future contracts. We show that the equilibrium of this signaling game cannot sustain full separation. In the resulting pooling equilibrium, providers with higher levels of knowledge will be contracted to exert *low* effort, producing a know-do gap and under-treatment by high types despite there being a positive reward for extra effort. We further demonstrate that any other robust equilibrium must involve at least partial pooling, and, as a consequence, there is always a know-do gap and under-treatment by high type providers despite a positive earnings gradient on effort. This pooling equilibrium can be broken or partly broken if patients' beliefs can be altered through a credible signal of quality. As patients beliefs about providers become more optimistic, they are willing to contract higher effort, resulting in more accurate treatments without any change in actual knowledge or underlying ability.

With this in mind, in the final part of the paper, we turn to a secondary analysis of a prominent and highly-publicized training program for informal healthcare providers in West Bengal, India. Our first analysis of the program (Das et al., 2016a), which was offered in a randomized fashion, established that the intervention significantly raised both effort invested and quality of care as measured by checklist items completed, time spent, and rates of correct case management. We now combine data from direct tests of medical knowledge with data on standardized patients and clinical interactions to explore the source of this improvement and document a series of patterns consistent with our theory of bilateral relationships between the patient and provider.

We first find that, despite its purported intent, the training program had little to no impact on provider knowledge. What, then, generates the large improvements in effort and quality of care? We consider two possibilities: the training program may have shifted providers onto a new production function such that the same amount of knowledge and effort produces higher quality of care (through, for example, more efficient patient management). Alternatively, the program may have shifted quality *as perceived* by patients, thereby enabling high-type providers to move *along* their existing effort-care production function, as patients are now willing to pay higher fees for higher effort exerted by such providers. The experimental nature of the training program, with the the control group serving as a counterfactual, permits a direct test of the first possibility. We find that the relationship between effort and practice remains unchanged between control and treated providers. Rather, we find evidence consistent with the second story. Treated providers are more likely to practice what they know in theory. Consistent with the model, treated providers who possess greater knowledge ('high-type' providers) exhibit larger improvements in quality of care, and a mediation analysis suggests that these improvements are largely the result of increased effort. Though imprecisely estimated, we also document higher prices charged by such providers, consistent with a shift in patient beliefs and demand.

This paper makes three contributions towards a broader understanding of healthcare provision in low-income countries. First, within the literature on quality of care in low- and middle-income countries, we gather together data from multiple studies and document striking similarities across samples, countries and (illness) conditions. The combined dataset also allows us, for the first time in the literature, to estimate the relative costs of overtreatment versus incorrect treatment in the data. In a literature that has largely focused on the perils of fee-for-service in incentivizing overtreatment, this exercise underscores the role of incorrect or under-treatment as the primary source of negligent healthcare provision.[7]

Beyond the systematization of these findings, we make two novel contributions. The first is to demonstrate that (a lack of) price incentives alone cannot account for the know-do gap. To date, the know-do gap has been linked to insufficient incentives to exert effort in the public sector. Das and Hammer (2007) and Das et al. (2016b) show that the know-do gap is lower in the private fee-for-service market compared to the public sector. What has received less attention is that even in the private sector, the know-do gap remains substantial. For example, Das et al. (2016b) finds a gap of 43.4 percent for fully qualified doctors in the private sector between what they do and what they know. In this study, we show that the know-do gap persists despite a steeply sloped price-quality gradient on the market. Our calibrations pose a very specific puzzle on why, in a context of excess capacity, doctors fail to increase effort when it would seem highly profitable to do so.

Our second contribution is the progress we make towards resolving this puzzle. Our theoretical model of bilateral reputation differs from two canonical formulations in the literature. The provider-induced-demand literature of McGuire and Pauly (1991) posits a doctor's utility function that trades off revenues and a direct utility cost of over-treatment; as patients never reject a treatment, the utility cost is the only disciplining device on a doctor's propensity to over-treat. The possibility of rejection is explicitly permitted in the credence good formulation of Dulleck and Kerschbamer (2006). In their model, doctors recommend a treatment after paying a diagnostic cost, and patients can choose to accept or reject the treatment. Over-treatment is disciplined through the possibility of rejection, and in a variant of the model, through liability in the case of under-treatment. Doctors' costs are known, and therefore the announcement of prices fully eliminates any informational asymmetry between doctors and patients. Even if patients suffer from a credence goods problem, they are able to predict, based on prices and costs, what action the doctor will undertake. Therefore, joint surplus is maximized when the doctor "does the right thing," since it eliminates the social waste from over- or under-treatment.

Our model re-introduces asymmetric information through uncertainty over costs, resulting in a richer set of doctor-patient interactions, and critically, the possibility of incorrect treatment—even when the doctor knows what the correct treatment is and would seemingly make more money by pursuing such an action. To our knowledge, this is the first formulation in this literature that is able

[7] As a by-product, we show that these results on under-treatment hold after using repeat observations to account for the measurement error in vignettes-based measures of knowledge.

to generate incorrect treatment in fee-for-service markets, the most common context for primary healthcare provision in low income countries.

Finally, this study contributes to a long-standing and fundamental debate in the literature: are private healthcare markets efficient? Friedman and Friedman (1962) advocate vociferously for the deregulation of healthcare:

> *I am myself persuaded that licensure has reduced both the quantity and quality of medical practice...that it has forced the public to pay more for less satisfactory medical service, and that it has retarded technological development both in medicine itself and in the organization of medical practice. I conclude that licensure should be eliminated as a requirement for the practice of medicine.* (p.158)

while Arrow (1963) draws our attention to the nuances of market interactions where trust is required:

> *One consequence of such trust relations is that the physician cannot act, or at least appear to act, as if he is maximizing his income at every moment of time. As a signal to the buyer of his intentions to act as thoroughly in the buyer's behalf as possible, the physician avoids the obvious stigmata of profit-maximizing. Purely arms-length bargaining behavior would be incompatible, not logically, but surely psychologically, with the trust relations.* (p.965)

The evidence emerging from our broad swathe of provider and patient settings elevates Arrow's insight, suggesting a fundamental market failure in the market for healthcare in low income countries.

2 Data and Samples

2.1 Data

We draw from three types of data: standardized patients (SPs), clinical observations, and medical vignettes. Unannounced standardized patients, our primary means of assessing quality of care, are regarded as close to a "gold-standard" measure of clinical practice. The method has been increasingly employed to measure healthcare quality in primary care settings following the first validation study in a large population-based sample in India in 2015 (Das et al. 2015).[8] Standardized patients are individuals recruited from the local community and extensively trained to present identical prespecified conditions to various providers. For instance, an SP may be trained to portray angina, reporting to the doctor with 'crushing chest pain' when he woke up and accurately responding to questions and examinations that the doctor then performs. This method is a substantive improvement upon other proposed quality measures, both in the richness of the data collected and

[8]See for instance, Sylvia et al. (2015); Mohanan et al. (2015); Das et al. (2016b, 2020) and Goodman in Tanzania.

its ability to avoid typical biases that plague quality measurement. Most importantly, given the low rates of correct case management that we observe across our settings, SPs offer the only viable method for measuring clinical practice against a condition-specific benchmark of appropriate care.[9]

The seven conditions that we examine across our multiple samples - chest pain, respiratory distress, child diarrhea, and TB1-4 - allow us to assess the ability of the providers across a range of basic and necessary skills, from triaging a patient with a severe condition (chest pain), to managing a chronic condition (respiratory distress), to offering simple primary level care (child diarrhea), to advising further appropriate testing for diagnostic purposes (TB). For each condition, SPs present with a standard set of symptoms and rehearse answers to anticipated questions that maximize the likelihood that providers, if engaging in proper questioning and examinations, will diagnose correctly. The process for hiring and training the SPs is extensive and described in greater detail in the Appendix; there we also discuss the research that validates the use of SPs in terms of (a) low detection rates; (b) suspicion among doctors that that patient does not have the illness that they claim to be presenting with; (c) the ability of SPs to recall their interaction accurately.

There may be a concern that SPs are not a marker of regular practice precisely because they do not reflect the equilibrium sorting of patients and case types in the market. While rich in depth, the SP methodology also limits the breadth of case types (conditions) that providers can be evaluated on. To extend our results to the typical patient, we therefore also present data from direct clinical observations in which trained observers remain with the provider for a full day and record clinical practice.

Finally, to compare provider practice with provider ability, we utilize data from medical vignettes, in which enumerators return to providers several weeks after the SP and evaluate provider knowledge over the identical set of cases tested by the SPs (Leonard and Masatu, 2005; Das and Hammer, 2005). Enumerators express an opening statement of symptoms identical to that expressed by the SP, and providers are invited to proceed as they would with a real patient, with enumerators providing the same answers to questions and examinations as the SP. As in the SP evaluation, providers are evaluated against a checklist of history and exam questions drawn from an established government medical protocol (Jindal et al. (2005); Ingle and Malhotra (2007)).[10]

[9]The alternative measures of quality of care, namely exit interviews, case records, or clinical observations, preclude researchers from assessing the appropriateness of care since one cannot know what disease the patient truly has. With SPs, one can assess the care received, including misdiagnosis, overtreatment, and undertreatment, against prespecified benchmarks for the condition of interest. SPs also allow researchers to minimize measurement challenges related to patient sorting and case mix, which confound observed relationships in administrative and patient data. Further, since providers are unaware when they are interacting with an SP, biases from the Hawthorne effect and social desirability are likewise minimized. Finally, the ability to design various aspects of the condition presented—down to the way the SP dresses and behaves—allows researchers to tailor the mix of conditions presented to a given context and a given research question. See the Appendix for further details.

[10]One may expect providers to have been "primed" from the SP visit and therefore have since learned how to better deal with the case presented. Vignettes are typically conducted two to six months after the SP interactions, and as providers only suspect SPs as being actors instead of real patients in less than 2% of all observations, this learning is a possibility only if providers are continuously learning on every case that they thought they performed poorly on.

2.2 Samples

Data for this paper is drawn primarily from five SP studies in India, to which we supplement two additional samples from China and Kenya, all conducted between 2010 and 2016 (Appendix Table **??**). Of the five studies from India, two are from rural settings (in Madhya Pradesh and West Bengal) and three from urban centers (Delhi, Mumbai and Patna).[11]

The samples span a diverse set of providers. The Madhya Pradesh sample consists of 127 public and 225 private providers, the majority of whom do not have formal medical qualifications. The West Bengal study includes 267 private providers, none of whom have formal qualifications. In contrast, the samples from Delhi (106 providers), Mumbai (831 providers) and Patna (591 providers) have a higher proportion of qualified private providers and include some of the best practitioners in each of these cities. The China sample covers 253 public providers in rural and small towns, and the Kenya sample covers 14 public and 28 private providers in urban Nairobi. Across the samples, providers tend to be overwhelmingly male with an average age in their mid-40's.

Measures of both knowledge (via vignettes) and effort (via standardized patients) are available for Madhya Pradesh, West Bengal, Delhi, and China; these sites therefore inform our stylized facts around the know-do gap. For a subset of providers from Madhya Pradesh and the full sample in West Bengal, we have two vignette measurements several months apart, which we employ to estimate the measurement error in our know-do gap estimates (as described in greater detail in Section III). In the remaining three sites of Kenya, Mumbai, and Patna, only SP data is available. The Appendix presents extensive detail on sampling frames and provider features for each study site.

3 Stylized Facts

We combine these seven sites of data to establish a set of stylized facts around the quality of healthcare in low income countries. We begin by defining our measure of quality of care. Our beleaguered patient walks into the doctor's office and complains of "crushing chest pain" and "anxiety" (as the standardized patient does in the case of angina). A high-quality provider would ask several questions, conduct some examinations, and strongly suspect either stable or unstable angina. He would immediately prescribe aspirin and refer the patient to the hospital, perhaps with a specific recommendation to complete an ECG test. He would also explain to the patient that he or she may have suffered a heart attack and an immediate visit to the hospital is critical to his or her health. The patient does not require any further medication at the primary care level, and therefore the triad of (1) aspirin, (2) referral to higher order care and (3) referral for an ECG, combined with (4) no other medications, constitutes correct case management. This happens so infrequently that we adopt a less stringent definition of correct case management. Let A be the

[11]The sources for each sample are as follows: for Madhya Pradesh, Das et al. (2016b); for West Bengal, Das et al. (2016a); for Delhi, Das et al. (2015); for Mumbai and Patna, Kwan et al. (2018); for China, Sylvia et al. (2017); and for Kenya, Daniels et al. (2017).

vector of care that is required and B be the vector of care that is not required. We define:

- **Correct Case Management:** The SP receives at least one component of A and no component of B. For the purposes of apportioning expenditures, we regard all expenditures incurred in such interactions as "medically necessary."[12]
- **Overtreatment:** The SP receives at least one component of A and at least one component of B. Note that neither in theory, nor in practice, are seemingly senseless combinations (such as aspirin and a knee x-ray) ruled out. Our data allow for all such combinations to be assessed. We separately count the expenditures on items in B and regard these as "medically unnecessary due to overtreatment."
- **Incorrect Case Management:** The SP receives no component of A and at least one component of B. In this case, we classify all expenditures as "medically unnecessary."

We highlight three caveats to our approach. First, provider recommendations may be following an optimal dynamic treatment path ("Try this and come back if you do not improve"), such that a visit we classify as "medically unnecessary" may in fact inform better case management in a subsequent visit. Should this be the case, we may overestimate "medically unnecessary" expenditures. Although we cannot eliminate the possibility that providers learn from prior visits, our hedonic pricing regressions will demonstrate that more qualified doctors manage more cases correctly and charge higher prices within single visits. This suggests that both medical training and patient demand (as proxied by price) work against the use of treatments as a primary source of diagnostic learning.

Second, our cost estimates consider only the price of doctor visits and medicine; they abstract away from the likely health costs incurred by incorrect or delayed treatment, which may be substantial and result in a significant underestimate of the true cost of healthcare in our settings.

Third, an estimate of the magnitude of medically unnecessary costs is arguably uninformative without a clear policy counterfactual (Feldstein, 1967). However, existing estimates of overtreatment in the medical literature are computed in similar ways, and our estimates are therefore at least directly comparable to those that are frequently employed to justify policy,[13] adding knowledge to what we know about medical expenditures in low and middle-income countries. More directly, the final part of our paper provides a policy-relevant counterfactual: our data from a randomized field experiment shows that a provider-level intervention can have an economically meaningful impact on medically avoidable expenditures defined in this manner (see Section V).

[12]We would have liked to incorporate the specific diagnosis that doctors provide into this definition, but in most SP interactions in India, doctors choose not to voice a diagnosis. Consequently, 61% of diagnoses are missing in our India subsamples. This also raises a concern around what to do with patients who are referred away without any further treatment, most often without a spoken diagnosis. This happens in 10% of cases, and we do not include "only referral" in correct treatment, choosing instead to present the robustness of our results to such alternate definitions in the Appendix.

[13]There are also attempts to classify costs due to diagnostic errors and incorrect treatments, but these are much harder to glean from administrative data (they are typically based on malpractice claims or second-guessing based on treatment paths) and therefore combine the lack of a policy counterfactual with uncertain data (CITE).

3.1 Incorrect treatments and medically avoidable expenditures are ubiquitous

Our first stylized fact, presented in Table 1, is that the fraction of SPs who receive all components of the correct treatment vector A and nothing else is less than 5%; the two exceptions are the studies in China and Kenya where the fraction rises to 23.7% and 21.1% (Column 2). A larger fraction of SPs receive at least one component of correct treatment, but even this remains below 35% in India, with China rising to 36% and Kenya to 52%. Most stark in these data is the high fraction of patients who receive entirely incorrect treatments; in Kenya this is 47.6%, and in all other samples it exceeds 63%.

Given the high incidence of incorrect treatments, our computations of medically avoidable expenditures range from 69.6% (Madhya Pradesh) to 84.8% (China), of which the vast majority is due to incorrect treatment (Table 2).[14] This contrasts sharply with the fraction of medically avoidable expenditures attributed to overtreatment, which falls below 15% for all samples with the exception of Kenya (reflecting, in part, the low cost of drugs in India). As we have noted before, such low costs from overtreatment are striking, given that the bulk of providers in our sample are private providers in fee-for-service practices, a context in which models of supplier-induced demand would predict that the bulk of improper case management and avoidable costs arise from overtreatment.

Robustness tables with alternate definitions are shown in Appendix Table **??**. Appendix Table **??** show that the patterns are similar across public and private providers in our sample, but the fraction of medically avoidable expenditures is lower and the fraction attributed to incorrect treatment is markedly lower for qualified private providers. Typically, 35% to 60% of medically avoidable expenditures are due to incorrect treatment among qualified providers compared to 72% to 95% among unqualified private providers.

3.2 Poor knowledge cannot explain poor treatment

One potential explanation for the high prevalence of incorrect treatments is a lack of underlying medical knowledge. To assess this, we employ data from medical vignettes to directly examine providers' knowledge on the same cases as those assessed through SPs. We focus here on the samples in Madhya Pradesh, Birbhum, China, and Kenya, all of which include matching data from SPs and vignettes.

We first assess the unconditional know-do gap; that is, the gap in the likelihood of prescribing a correct treatment in the vignettes relative to the SPs. Figure 1 depicts the relative probabilities of correct treatment under the two scenarios, disaggregated by sample and provider level of training. Two patterns emerge. First, we observe a significant know-do gap across all samples: correct case management is 20 to 80 percentage points higher in the medical vignettes than with the SPs. Second, the amount that providers "know" always increases with training, but the know-do gap

[14] These numbers weight the incorrect treatment by the cost of the recommended treatment and therefore differ from the proportions of cases that are incorrectly treated.

also tends to be higher for trained providers (40 to 80 percentage points) relative to the untrained (20 to 30 percentage points).

While a comparison of raw means establishes the existence of a know-do gap, the measure most relevant to policy is rather the *conditional* mean: what fraction of an increase in knowledge translates into an increase in practice? A know-do gap is less concerning if one can upskill providers out of it: by endowing providers with greater medical knowledge, they may progressively close the gap with practice. Our observation that the know-do gap tends to be *higher* among providers with a medical degree relative to untrained providers suggests that this is not the case. We next derive this conditional estimate using our direct measure of knowledge from the vignettes. We estimate the following regression:

$$Do_{ic} = \alpha + \theta Know_{ic} + \gamma_c + \epsilon_i \tag{1}$$

in which Do_{ic} represents a binary variable that equals one if provider i issues a 'correct' treatment to the standardized patient for case-type c and zero otherwise, $Know_{ic}$ is the analogous binary variable for the vignette, and γ_c is case-type fixed effects. Standard errors are clustered at the level of the healthcare facility. Our coefficient of interest, θ, represents the know-do gap, or the extent to which a one unit increase in knowledge translates into practice.

Results are presented in the first row of Panel A in Table 3: in Madhya Pradesh, for example, providers who "know" how to correctly manage a case will only do so in practice 14% of the time.[15] In other words, we estimate a conditional know-do gap in Madhya Pradesh of 86%. The know-do gap in Birbhum is approximately 97%; in Delhi, 86%,; and in China, 81%.

Our measure of knowledge using the vignettes, however, is vulnerable to measurement error: as is evident in Appendix Table **??**, 6.7% of treatments which are incorrectly prescribed in vignettes in Madhya Pradesh, for example, are *correctly* prescribed in the corresponding SPs. Across our samples, this number ranges from four to eight percent. The presence and interpretation of measurement error in estimation of the know-do gap has not been examined in the literature to date. The exercise is complicated by the fact that our outcome variable of correct case management is a binary variable, resulting in OLS and IV estimates presenting lower and upper-bounds, respectively, of our coefficient of interest.

In order to address this measurement error, we exploit the repeated vignette measures taken in the Madhya Pradesh and Birbhum samples, in which enumerators returned several months after the completion of the first vignettes for a re-test with the same providers. The correlations between the first and second vignettes in MP and Birbhum are 0.26 to 0.31, respectively, suggesting that measurement error may be quite severe. We first consider three simple adjustments to decrease the error. The remaining rows in Panel A of Table 3 presents results for these three alternative specifications, in which our measure of 'know' is derived from (1) the average of the two vignette reports, (2) the maximum of the two reports, or (3) a measure of 'know for sure' (in which the provider prescribes the correct treatment in both tests, following Bollinger (1996)). Relative to the

[15]These estimates are lower than the 0.2 reported previously in Das and Hammer (2007) and could reflect a larger fraction of unqualified providers in the estimating samples.

base specification estimates of 0.1 (MP) and 0.026 (Birbhum), these simple adjustments increase the coefficient of interest (and thereby reduce the estimated 'know-do' gap) to 0.13 to 0.16 (MP) and 0.12 to 0.14 (Birbhum).

Panel B then employs two stage least squares (2SLS) and generalized method of moments (GMM) techniques, which instrument for provider knowledge using either the second knowledge test or provider's adherence to the checklist. Unlike the OLS, for which θ is biased downward (producing an upper bound on the know-do gap), or the 2SLS, for which θ is biased upward (thereby producing a lower bound on the gap), GMM produces consistent estimates (CITE; see further details in Appendix D). We run the following regression:

$$Do_{ic} = \alpha + \theta \widehat{Know}_{ic} + \gamma_c + \epsilon_i \tag{2}$$

In which $\widehat{Know}_{ic}$ is estimated in the first stage regression:

$$\widehat{Know}_{ic} = \alpha + \theta Z_{ic} + \gamma_c + \epsilon_i \tag{3}$$

In which Z_{ic} is either the a binary variable which equals one if the provider prescribed a correct treatment in the second knowledge test and zero otherwise, or the fraction of checklist history and exam questions asked during the first vignette. We consider the latter instrument because the measure can be obtained with a single administration of the vignettes rather than requiring a vignette re-test. Panel C, which reports the coefficient on knowledge in the first stage of each IV regression, shows that both instruments are strongly correlated with correct case management in the first vignette.

Panel B reports the results. Our GMM estimates suggest that if providers 'know' how to correctly manage a case, they will do so in practice 17 to 22 percent of the time in MP (Panel B, Columns 2 and 3) and 23 percent of the time in Birbhum (Panel B, Columns 4 and 5).

Reassuringly, the results in Birbhum are consistent across the two instruments, and while they vary more in MP, they are well within the standard errors of one another. This opens up the possibility of instrumenting using adherence to checklist for samples in which we lack a second test of the vignettes. In Columns 5 and 6, we employ this technique for the China and Delhi samples. The OLS and IV linear bounds generate estimates between 0.14 and 0.36 in Delhi and 0.19 and 0.42 in China. In no case do we exceed 0.42, suggesting that across multiple conditions and samples, no more than 42% of knowledge translates into practice. While substantially larger than the initial OLS estimate of 14%, the estimate implies that less than half of the knowledge that providers demonstrate in vignettes manifests in clinical practice. The distance between what a provider knows and what he does therefore *grows* with knowledge, such that healthcare systems cannot simply "upskill" their way out of the gap.

3.3 High caseloads cannot explain poor treatment

Why do providers who know how to provide higher quality care still choose not to do so? Perhaps providers are constrained by high patient caseloads. Total revenue depends on both the price charged per patient and the number of patients a provider sees, and providers may choose to lower their time and quality with some patients in order to increase their volumes.

Our data again suggests otherwise. In two samples, MP and Birbhum, enumerators sit with providers in their clinics for the duration of their practice over one day; we employ these clinical observations to compute the total number of patients that providers see in a day. Figure 2 shows the distribution of total time spent with patients in these two samples. The mean total amount of time that providers see patients during a work day is 15 minutes in MP and just over 30 minutes in Birbhum, despite sitting in their clinics (otherwise unoccupied) for four to six hours. Providers in these primary care settings appear to have considerable excess capacity.

This basic fact is consistent with multiple other studies across multiple low- and middle-income countries (Maestad, Torsvik, and Aakvik, 2010): effort and treatment quality do not appear to be constrained by high patient caseloads.

3.4 Insufficient incentives cannot explain poor treatment

A third leading explanation for poor quality of care is that providers face inadequate incentives to effortfully transform knowledge into practice. As in the case of knowledge, poor incentive structures do have some impact on quality of care: Das et al. (2016a) send SPs to the same providers in their public and private clinics and demonstrate that correct case management rates increase from 37.3% in the provider's public clinic to 56.6% in the same provider's private clinic, arguably due to the shift from salaried to fee-for-service payment regimes.

The fact remains, however, that upon entering the private sector, these providers exhibit a still substantial know-do gap of 40%. If poor incentives are to explain this remaining gap, it must be that the return in the market to providing correct treatment falls below the provider's marginal cost of effort. To investigate this channel, we utilize direct measures of prices and clinical performance collected through SP interactions in each of our samples.

3.4.1 Evidence from standardized patients

Our starting point is that effort and knowledge are complementary in the production of clinical quality, but that correctly managing a case requires costly effort. Then, whether a provider chooses to exert effort depends on his or her compensation for doing so. The following set of analyses seeks to estimate the returns to effort. We proxy for effort using two measures: time spent with a patient and adherence to a condition-specific checklist. While the former is a pure measure of effort, the second captures the possibility that proper case management requires cognitive attention and is therefore more costly.

Figure 3 plots the non-parametric relationship between these two proxies for effort, with time

spent per patient standardized within samples. The relationship is upward sloping, concave, and similar across samples. At consultation lengths below the mean, a one standard deviation increase in time spent is associated with a one standard deviation increase in adherence to the checklist; above the mean, that relationship is halved. Figure 4 then shows that in all samples except for Kenya, greater adherence to the checklist is associated with a greater likelihood of correct case management. Here, the relationships vary across countries, but an increase of one standard deviation in adherence to the checklist increases correct case management by 10-25%.[16] Correct management and effort are correlated, perhaps because correct management requires a provider to incur some cost in terms of time and effort.

Figure 5 then looks at the potential monetary benefits of effort investment. We examine the non-parametric relationship between provider fees (amount paid by the SP to the provider) and checklist adherence. The relationship between fees and checklist adherence is positive, robust across samples (except for China, where we have data only from public clinics), and linear across the range of fees. At the mean, a one standard deviation increase in checklist adherence is associated with an increase in fees of 0.3 to 0.5 standard deviations. The patterns are substantively the same when we disaggregate our sample by knowledge in Appendix Figures **??** through **??**.

The superimposed histogram plots of adherence to checklist demonstrate that, across both our MP and Birbhum samples, there exists substantial variation in whether doctors who presumably "know" how to correctly manage a case in fact do so in practice. As the upward sloping lines indicate, those who correctly manage the case *in practice* indeed charge higher fees.

The price-effort gradient is replicated when we use time spent with patient, rather than checklist items completed, to proxy for effort (Appendix Figure **??**). A one standard deviation increase in time spent with the SP is associated with an increase in fees of 0.5 standard deviations; the pattern remains approximately linear up to one standard deviation of time spent, but then disperses (likely due to conversation that is uncorrelated with the quality of care).

To estimate the magnitude of returns to various measures of quality and effort in the market, we run the following regression:

$$Price_{ic} = \alpha + \theta Quality_{ic} + \phi Know_{ic} + \gamma_c + \delta_s + \epsilon_i \tag{4}$$

in which $Price_{ic}$ is the price charged to an SP on an interaction with provider i for case c, $Quality_{ic}$ is a measure of case management quality or effort invested during the interaction (correct case management checklist adherence, time spent per patient, number of medicines given, and whether the patient is referred elsewhere), and $Know_{ic}$ is a binary variable equal to one if the provider prescribed the correct treatment for case c in the vignette. We include both case (γ_c) and SP (δ_s) fixed-effects, precluding the possibility that observed variation arises from differences

[16]The Kenya exception is driven by the high incidence of doctors immediately recommending patients with three weeks cough for sputum testing without asking any questions. This is correct case management and points to an obvious potential problem—there may be "no questions asked" protocols in countries (such as "refer immediately for ECG if a patient reports with chest pain") that will complicate our understanding of this relationship. How to understand and analyze these differences in a cross-country analysis remains an open question.

in patient characteristics. We also control for whether the patient was referred, as referrals must be discounted, since the degenerate equilibrium is to always refer if there is no discount in the private market. Standard errors are clustered at the facility level. Finally, since we are using the equivalent of audit data, concerns arising from omitted case and patient characteristics are substantially reduced (but not eliminated if patients sort and doctors react to their average patient population).

Results are reported in Table 4 and underscore the patterns of the non-parametric plots. We present bivariate correlations in odd-numbered columns and estimates from the OLS regression above in the even-numbered columns.[17] Across all samples, the amount paid by the SP is positively correlated with each of our quality measures; China (where the sample is smaller and includes only the public sector) is somewhat different, and smaller sample sizes in Kenya reduce the statistical precision of the results, but the directions remain consistent. Referrals are always discounted, and more medicines are associated with higher payments. All quality measures are positively correlated with payments across all samples in the OLS specification, with the exception of several measures in Birbhum, where directions are consistent but precision is low, the adherence to checklist measure in China, which is negatively signed and marginally significant, and the correct treatment measure in Kenya, which is negatively signed and insignificant.[18]

3.4.2 Evidence from clinical observations

While the patterns we observe remain robust to disaggregation by knowledge as measured through vignette performance, one may be concerned that unmeasured components of provider-specific knowledge may be correlated with quality and price, or that the price-quality relationship is a result of unobserved variation in patient demand. We therefore utilize clinical observations, as available in our Birbhum and MP samples, from which we can obtain multiple real patients' data per provider. Patterns in these data reveal whether the audit relationships hold with real patients; they also allow us to include provider-level fixed-effects, which we cannot include in our SP specifications. A drawback of these data is that, unlike SPs for whom we specify the conditions they present with, we lack the precision of condition-specific metrics in clinical observations since the true conditions with which patients present the provider are unknown. We therefore cannot assess correct case management, but instead use two proxy measures for quality: time spent with patient and a count of the total number of questions asked and examinations completed. Note that, in these samples, 50% of the variation along both margins is *within* providers.

Table 5 presents the regression results. As with the SP data, we present both bivariate correlations (odd columns) and multivariate OLS regressions (even columns). We control for a wide variety of patient characteristics that may be correlated with quality of care, including indicators of patient health (the Activities of Daily Living score) and patient wealth as assessed through an asset

[17] In the multiple regression framework, given the high correlations between each of our quality measures, the precise quality measure that loads will depend on the variance-covariance matrix across these measures in each of our sample.

[18] The latter is again due to the high use of sputum tests (where the provider takes less money) for tuberculosis.

index, both of which limit the extent to which variation can be explained by patient-level price discrimination. As in the SP data, we find that prices are positively correlated with consultation length and questions asked using real patients. Surprisingly, there is little difference in the relationship with and without provider fixed-effects; the price-quality association we find is consistent between the SPs and real patients, with providers charging more when they provide higher quality.

3.4.3 Calibrating the cost of effort

These two sets of exercises - one using standardized patients and the other using real patients - now allows us to calibrate the relative gains to higher effort among private providers (Appendix Table **??**), and therefore understand whether poor quality of care can be explained by insufficient incentives to invest effort. In the MP and Birbhum samples, we find that an additional ten minutes spent with a patient is associated with approximately an $0.80 USD gain in revenue, after taking into account both price and quantity correlations. This implies an implicit wage rate of $4.80 USD per hour in 2011/2014, relative to a minimum wage of $3 USD *per day* in India's workfare scheme, and $7.3 daily for urban secondary school workers, which is the sample closest in educational qualifications to the providers in these studies. The calibrations suggest that providers might increase their daily revenue by 30-50% simply by spending an additional 30 minutes with the patients they already see, in a context where they spent 30 minutes seeing patients in their clinic and the bulk of the day sitting idle within the clinic. Greater effort and the resulting improved quality of care thus appear to have substantial returns in the market.

4 Conceptual Framework

Why, then, do we see such variation in quality of care even conditional on provider and patient type. And why do we remain in an equilibrium of low quality of care for the vast majority of observed patient-provider interactions, despite high levels of knowledge? We argue that our story requires incorporating how patients' beliefs may directly shape the level of care offered by providers.

4.1 Model basics

We propose a dynamic model of the patient-provider interaction in which a patient's expectation of provider type informs the quality of care a provider offers. Consider an individual who currently has symptoms of some illness that she is unable to diagnose. Suppose, for such a person, that there is a probability p of having a mild condition and a probability $1-p$ of having a serious condition. If the mild (serious) condition is treated incorrectly by a doctor, it has a cost $C_m(C_s)$. Assume $C_m < C_s$. Assume that all utilities are measured in units of money.

Whether or not the outcome is treated correctly or not depends on the doctor (and on luck). Doctors come in two types: high (H) and low (L). The prior probability that a randomly chosen doctor is a type H is π. Given the patient's symptoms, a type H (L) doctor is able to correctly diagnose the patient's symptoms with probability $p_{iH}(p_{iL})$, with $p_{iH} > p_{iL}$, if the disease is of type

i where i can either be a mild disease (m) or a serious one (s). In other words, there are multiple mild conditions and multiple serious ones, and each requires a different treatment and therefore the right diagnosis.

We assume that doctors are able to treat diseases appropriately if they can correctly diagnose them, though this assumption is easily relaxed. However, whether or not the doctor treats the disease correctly depends on how hard he tries. In particular, a doctor can either spend τ_s amount of effort-cost to diagnose the illness, in which case he can detect and treat *both* mild and serious conditions with the probabilities given above, or he can spend $\tau_m < \tau_s$ amount of effort-cost, in which case he can only detect and treat mild conditions with the probabilities given above. The effort cost is observable and can be contracted on.

We operate in a thick market of doctors,[19] and the doctor's outside option is assumed to be zero. In the first period, he discounts his second-period utility by a factor δ, which can also capture the possibility that the patient may not be sick in the second period.

The individual lives two periods and has symptoms of some illness in both (without loss of generality). In each period, she must decide which doctor to see. Her outside option is to go to the nearest town, which gives her a utility $\underline{V}$. In the first period, she discounts her second-period utility by a factor δ, which can also capture the possibility that she may not be sick in the second period. Alternatively, she can pick a doctor in the village. In the first period, she knows nothing about any doctor and chooses at random among those in her village. In the second period, she can draw upon the experience of her first period interaction, either returning to the doctor she first saw or choosing someone else.

4.2 Contracting

Assume that this is an environment with asymmetric information, in which the doctors know their own types but the patients do not know the doctors' types. The contract between the patient and the doctor is proposed by the doctor at the beginning of each period, but there is no commitment on either side beyond the time of the visit. In other words, there is no possibility of paying the doctor after the disease is cured, the doctor can always change the contract at the beginning of the second period, and the patient can always walk away and take her outside option.

In the second period, the doctor begins with the knowledge that the patient has placed a probability of π' on his being a high type *before he proposes the second period contract.* A doctor of type $T = H, L$ facing such a patient can propose a contract or probability distribution over contracts that have three elements: his announced type A, which lies in $\{H, L\}$, his proposed fee f, which lies in some interval $[\underline{f}, \bar{f}]$, and the effort he proposes to invest τ, which lies in $\{\tau_s, \tau_m\}$. His choice of contract or distribution over contracts may in principle depend on both T and π'.

In the first period, the doctor begins with the knowledge that the patient has placed a probability of π on his being a high type *before he proposes the first period contract.* A doctor of type $T = H, L$

[19] This closely approximates the context of the bulk of our empirical work: the average village in India has 4.4 healthcare providers a patient may choose from (Das et al. (2020)).

proposes a contract or a probability distribution over contracts again composed of the three elements of his announced type, the proposed fee, and the proposed effort level. We do not impose the restriction that the announced type has to be the same in the two periods. Note that the doctor always has the choice of offering a contract that the patient will never take if he does not wish to serve the patient.

Given this structure, the choice of contracts in equilibrium in each period can either involve complete pooling, partial separation, or full separation. The uncertainty is then resolved and the patient is offered a single contract, updates her beliefs based on the offer, and chooses to either accept the contract or take her outside option.

4.3 Properties of the model

This defines a Provider - Patient game. This is a signalling game with many sequential equilibria. For our purposes however, it is not necessary to pin down a single equilibrium; rather, it will be enough to rule out fully separating equilibria.[20] We describe here the intuition behind the results, with a formal explication of the model in the Appendix.

Our process for ruling out the existence of a fully separating equilibrium rests on a few key observations. First, in the second period, both provider types have identical preferences. This implies that, at this stage in the game, we will only observe pooling equilibria. Moreover, given that the contract is proposed by the provider, we argue that there is no reason why the provider would propose anything other than the revenue maximizing pooling contract.[21] This is convenient since it pins down the second period outcome: the standard refinements of sequential equilibrium are only defined for one period games, and this assumption essentially transforms our two-period game into a single-period game.

Second, observe that in any equilibrium where there is full separation in the first period, the outcome of the treatment in the first period (cure or no cure) does not affect the patient's beliefs. This is because, by virtue of the fully separating equilibrium, the provider's type has already been revealed, and there is no additional information that the treatment outcome can offer to the patient about her provider's type.

The next step is a direct consequence of the previous observation. In a fully separating equilibrium, if an L type provider claims that he is an H type, his payoff will be exactly the same as that of an H type. This is because the only difference between the types is the difference in their probability of delivering a cure, and the resulting fact that the patient updates negatively about those providers who fail to cure them. However, because of the previous observation, this is not true once there is full separation.

The final step is to observe that, since an L type provider receives a net utility of zero in any

[20] We will, however, also discuss the robustness of complete pooling, and in that context refer to the well-known Cho-Kreps "Intuitive Criterion" as a restriction on possible beliefs. The formal discussion of this model is in the Appendix.

[21] The only reason they will not propose the revenue maximizing contract is if the patient assumes that the provider is a low type. However, this makes little sense given that the two types both gain equally by moving to this contract.

fully separating sequential equilibrium, he may switch and claim to be an H type and receive the same payoff as the H type (from the previous observation). Therefore, H types must also receive zero in equilibrium (otherwise L would deviate). We then introduce a certain Condition *, which rules out this possibility given that the fee that the H charges is bounded below by zero.

This brings us to the main result of this section: a fully separating equilibrium at the beginning of the first period can be ruled out if a certain condition that we call Condition * holds. Therefore, the only possible equilibria in this game are ones that involve at least some degree of pooling in both periods.

4.4 Implications

This result delivers a number of predictions that we can bring to our data.

Prediction 1 First, equilibria with some degree of pooling imply a know-do gap. The knowledge and action space we consider is a particular one: providers may know the right steps to take in order to deliver a diagnosis or treatment, but they may vary in their ability to do them well enough to be useful. For example, all providers may be capable of tying the cuff of a blood pressure gauge around a patient's arm, but only a few may be able to obtain and interpret a meaningful reading from it: it is along this margin that H and L types are differentiated. In an efficient equilibrium, L types will be aware of their abilities they choose not to expend effort performing the blood pressure test at all, while H types will invest such effort. However, in a pooling equilibrium, *neither* type will take the blood pressure reading in the first period, since H types cannot differentiate themselves from L types and be rewarded for their effort, and both types will take it in the second. If we take the knowledge of the right steps to be a measure of "know", then we will observe a know-do gap in period one of the pooling equilibrium: H types will not enact what they know. With partial pooling, this remains true whenever the practitioner actually treats the patient, which is when there is pooling.

Prediction 2 Second, in any of the equilibria, when a provider spends more time with a patient (a proxy for investing more effort), he gets paid more and is more likely to get the diagnosis and treatment correct.

Prediction 3 Third, consider an intervention that raises π, positively shifting patients' beliefs about the abilities of a provider. An increase in π will shift the equilibrium of the Provider - Patient game. If $\pi > \pi*$, then there is a pooling equilibrium where the practitioner will choose τ_s in the first period instead of τ_m, and charge more. Diagnosis, or quality of care, will likewise improve. In the case of partially pooling equilibria, the increase in π may again shift the equilibrium to one where the the pooling outcome in the first period involves τ_s rather than τ_m.[22]

[22]Note that Condition * does not involve π and a separating equilibrium will still not exist.

Prediction 4 Finally, note that the entire improvement in provider's quality of care in this setting arises from an increase in the patient's perceived π: the change in patient beliefs about the provider's type permits the provider to suggest a higher level of effort to the patient. The provider's actual ability (type) has not changed. We should therefore observe a movement *along* the pre-existing relation between a provider's effort and his quality of care, rather than shift *of* this curve, should a provider be exposed to a positive change in π.

Predictions 1 and 2 are borne out in the data already presented, namely Tables 5 - 8. We now turn to a field experiment in West Bengal to assess Predictions 3 and 4.

5 Field Experiment in Birbhum, West Bengal

An experimental training program for informal providers in the Indian state of West Bengal (Das et al. (2016a)) presents a unique opportunity to assess the theoretical framework we proposed above. Our intervention of interest is a highly publicized nine-month medical training program administered to rural healthcare providers by a prominent public health organization. The program was offered to a randomly chosen 152 providers from a pool of 304 providers in West Bengal. Those who did not receive the program served as the control group and qualified for the program two years later.

Held between January and October of 2013, the program provided generalized instruction on a broad range of topics, from anatomy and physiology to first-aid and public health. Sessions were held two days per week for approximately two hours per session, totaling 150 hours of interaction. Providers maintained their clinics for the duration of training but were often forced to close their practice during the hours or days that the classes took place due to the long travel time to the district capital where classes were taught.[23]

Our initial evaluation of the intervention indicated a highly effective program (Das et al. (2016a)). The program increased the proportion of cases that were correctly managed for SPs presenting with three diverse conditions: angina, asthma, and diarrhea in a child. However, as we demonstrate below, the program generated no meaningful improvements in provider knowledge.

Our theory proposes a channel for the quality of care improvements we observe in this experiment in the absence of a shift in underlying provider knowledge (type). If the intervention raised patients' *beliefs* about the quality of attending providers, then those who participated in the program could be enabled to exert greater effort, with a commensurate reduction in their know-do gap and an increase in their quality of treatment and fees earned. These predictions should hold *without* an increase in the underlying ability of the provider.

Why might a program that randomly selected providers to be trained alter patient beliefs about their quality? We document self-selection into attending the training program (conditional on being randomized into the opportunity): providers with higher knowledge at baseline exhibit higher rates

[23]Our qualitative work elucidated that many patients were aware that their provider was in a training program, and many admiringly mentioned the reputation of the public health organization administering it.

of attendance.[24] Patients were likely to have picked this up as well: the training schedule was demanding and required regular closure of providers' offices, and those who chose to regularly attend were likely to be those who could gain from an opportunity to signal their type. Patients would therefore be correct in assuming that the training signaled high quality and updating their beliefs accordingly.

Our aim in this final section is to examine the source of the large improvements in quality of care that Das et al. (2016a) find from this training program and assess the patterns we document against the predictions of our model. We first ask: to what extent can the improvements in quality of care can be attributed to changes in the underlying knowledge, or ability, of providers? We show that there is virtually no increase in provider knowledge as a result of the training program: the program rather enables high type providers to translate more of their *existing* knowledge into practice, consistent with the theory. The presence of a viable counterfactual, via the control group of providers, further allows us to rule out that the improvements we observe are due to changes in the provider's underlying production function (treated providers do not become more 'efficient' at providing quality care), and the use of the same standardized patients across providers precludes the possibility of patient sorting, wherein patients with certain conditions may shift to or from treated providers. What remains, then, is a shift in patient perception of provider type. Our theory predicts that an upward shift in patient beliefs should result in a narrowing of the know-do gap among high type providers, as defined by those who exhibit high levels of knowledge at baseline. Indeed, we find that the impact of the program on quality of care is almost entirely driven by those who possessed high levels of knowledge at baseline, with their high-type control counterparts exhibiting no parallel changes in quality of care.

Taken together, this set of empirical patterns is strongly consistent with the theory. First, there exists a sizeable know-do gap among providers in Birbhum, as estimated from the control sample. Second, the program raises treated providers' effort invested, quality of care, and fees collected (the latter estimated imprecisely). Third, among those offered the program, those with greater knowledge at baseline ("high types") are more likely to avail themselves of it, suggesting they have more to gain. Do they gain in learning or signalling? We find, third, that the program has no impact on knowledge or productivity, suggesting that providers are learning little in substance. Rather, it appears to serve a signalling purpose: we find, fourth, that improvements in quality of care are explained by a greater ability to translate existing knowledge into practice: in essence, treated providers are enabled to reduce their know-do gap.

We demonstrate that the quantitative effects of such a shift are sufficiently large to induce meaningful welfare gains, and this proposed channel therefore merits continued investigation in studies beyond our own. While Das et al. (2016a) was designed to examine provider quality, future work may focus more attention on the patient side. Our analysis employs revealed preference outcomes like fees and provider caseload to estimate changes in patient demand; while these move

[24] A one percent higher level of knowledge at baseline is associated with a 0.75 higher rate of attendance to the training program; results available upon request.

in the direction predicted by the theory, the experiment in Das et al. (2016a) was not powered to pick up the small effect sizes predicted on these margins. Future work may expand sample size and pair price and quantity measures with patient health outcomes, to estimate welfare effects, and patient beliefs, though such self-reports are admittedly vulnerable to demand effects.

5.1 Data

The preceding part of this paper employs data from the control group of this experiment ("Birbhum data"); we now introduce treatment group data and describe the data collection process of the full experiment in detail.

A baseline survey, which collected data on provider demographics, clinical practice (clinic history, patient caseload, fees charged), and clinical knowledge was conducted two months prior to the commencement of the training program. The endline survey was administered between three and six months after the completion of training. Data collected at endline mirrored that of baseline, supplemented by two measures of the quality of care: standardized patients and clinical observations. Importantly, SPs were first sent to providers, and vignettes were conducted approximately two months *after* in order to minimize the likelihood of cross-contamination. The study was conducted as a "triple-blind" evaluation: trainers in the program were blinded to what cases the SPs would present; SPs were blinded to the group assignment of the providers they visited; and SPs were neither anticipated by providers nor ex-post recognized as actors. Each SP presented the provider with one of three conditions representative of the broad range common to the providers' typical patient base (respiratory distress, chest pain, and child diarrhea; see Das et al. (2016a) for details). We complement the SP data with day-long clinical observations. While clinical observations provide a picture of clinical practice for a broader sample of patients, we note that they are vulnerable to Hawthorne effects, as treated doctors were aware they were being evaluated and had been repeatedly told in their training program to not charge excess fees.

5.2 Results

5.2.1 Impact of training program on clinical practice

We turn first to the overall impact of the intervention on clinical practice of the treatment group, which we replicate from Das et al. (2016a). We estimate the following regression.

$$Y_{ic} = \alpha + \beta * Treatment_i + \delta_c + \gamma_b + \theta X_i + \epsilon_i \tag{5}$$

where Y_i represents various measures of clinical practice for provider i as elicited from the SPs; $Treatment_i$ is a binary variable which equals one if the provider was offered the training program and zero otherwise, X_i is a vector of provider-level characteristics (baseline vignette performance, age, gender), δ_c represents case-type c fixed effects, and γ_b represents geographic (block-level b) fixed effects. Standard errors are clustered at the level of the provider facility.

ITT results are presented in Columns 1-4 of Table 6. Columns 5-8 of Table 6 present the analog IV (LATE) estimates, in which *Treatment* serves as an instrument for attendance to training (which averaged at 56% across all sessions and providers).[25]

Effort invested, as measured by checklist adherence (time spent) with SPs at endline, was higher among treated providers by 15 to 25% (7 to 12%), who exhibited 14 to 23% higher rates of correct case management. This improvement in quality appears to have translated into greater demand as evident in equilibrium prices: fees charged per interaction rise on the order of 10 to 16% among treated providers, although estimates are imprecise.[26] [27]

5.2.2 Impact on provider knowledge

We now explore whether the impact of the intervention that we observe among high-ability providers may be a consequence of a change in the knowledge-base of these providers. Indeed, improving medical knowledge was a primary motivation of the public health organization in administering the training program.

Notably, such improvements in knowledge would not preclude our posited mechanism, as knowledge and effort are complements in our data; however, we would then need to rely on a parametric calibration of the provider's quality of care production function in order to disentangle the knowledge channel from an effort channel alone.[28]

To test for the impact of the training program on knowledge, we estimate regression (5) with provider knowledge as our outcome of interest. We condition on the value of the outcome variable in the baseline vignette. We further explore heterogeneity by type: perhaps the program was such that only those who were sufficiently knowledgeable at baseline could benefit from the information imparted in the classroom. Any improvements in the quality of care among high types, as predicted by our theory, could in this case be due to a differential acquisition of knowledge. To do so, we run the following regression:

[25] No individual from the control group ever attended training, permitting the interpretation of the LATE estimate as the treatment on the treated (TOT).

[26] Imprecision in fees data is unsurprising. As Figure 7 depicts, checklist adherence and fees charged exhibit a slope of approximately 1.5 along the mean of the distribution; treated providers asked 4.1% more checklist items, implying a fee increase of 0.062. In magnitude, this matches closely with the treatment effect on fees we observe; however, the experiment would require substantially greater power to pick up these effects precisely.

[27] Appendix Table **??** presents the parallel estimates using clinical observation data. Results remain largely consistent. The improvement in quality is further reflected in the change in patient caseload observed within treated providers' clinics: although imprecise, treated providers appear to experience a 6% increase in patient load per day (Column 7). Clinical observations data also report a decrease, though imprecise, in fees charged. Although we report this measure, we interpret fees data from clinical observations warily because experimenter demand effects are likely to be substantial. Avoiding high consultation fees was a topic of repeated instruction in the training program, and providers on several occasions noted to the enumerators who were recording their behavior during clinical observations some version of: "You see? I did not charge the patients much, just as you had asked of us."

[28] An increase in knowledge that in turn improves quality would also imply a *reduction* in fees charged, since a knowledge improvement is a productivity shift that would result in an outward shift in the local supply of quality healthcare, thereby leading to fees decreasing in equilibrium. SP data instead exhibits an increase in fees, although results are imprecise.

$$Y_i = \alpha + \beta Ability * Treatment_i + \delta Ability_{ic} + \gamma Treatment_i + \delta_c + \gamma_b + \theta X_i + \epsilon_i \quad (6)$$

where Y_i is provider i's knowledge score, as proxied by checklist adherence or correct case management in the endline vignettes, $Ability_i$ is the provider's initial ability as proxied by baseline vignette performance, $Treatment_i$ is a binary variable equalling one if the provider was offered the training program and zero otherwise, X_i is a vector of provider-level characteristics (baseline vignette performance, age, gender), δ_c represents case-type c fixed effects, and γ_b represents geographic (block-level b) fixed effects. Standard errors are clustered at the level of the provider facility. Our coefficient of interest is β, or the differential impact on knowledge that the intervention has among high-ability providers relative to their lower ability counterparts.

ITT results are presented in Panel A of Table 7. Panel B of Table 7 presents the analog IV (LATE) estimates, in which attendance to the training program is again instrumented by treatment status.

Columns 1 and 3 present average effects. The training program raised the likelihood of asking a checklist item by 6% (ITT) to 10% (LATE), marginally significant at the ten percent level. This failed to translate into improved knowledge of correct treatment: effects on the likelihood of prescribing any correct treatment within the vignettes are noisy and close to zero. Columns 2 and 4 present the corresponding results on heterogeneity by ability. High ability providers are no more likely to adhere to the checklist or exhibit correct case management in the endline vignettes than their lower ability counterparts; if anything, the treatment appears to reduce knowledge gains, though the coefficient is imprecise.

The training appeared to have little to no meaningful impact on the knowledge level of exposed providers, and no differential impact on high ability providers, suggesting that the improvements we observe in quality of care were due minimally, if at all, to a shift in provider ability, insofar as can be measured through knowledge acquisition in the vignettes.

5.2.3 Impact on the effort-care production function

While the training program did not increase knowledge *per se*, it may have made providers more efficient in their practice conditional on their knowledge; for example, they may have become more deft at interpreting cues of their patients. This would be represented by a shift to a new and steeper production function of effort (input) to quality of care (output): each unit of effort invested would transform into higher correct case management. The experiment presents an opportunity to disentangle this mechanism from that of patient trust: with the production function of the control group serving as a counterfactual, we can observe whether the intervention generated shifts in the provider's underlying ability to produce high quality care or simply enabled him to move along an unchanged curve.

Figure 6 plots the non-parametric relationship between the number of checklist items completed in standardized patient interactions and the likelihood of any correct treatment, separately for

treated and control providers. The relationship between checklist items and the quality of care remains essentially unchanged between control and treatment: the intervention does not appear to shift treated providers to a steeper effort-quality production function. Figure 7, which plots the non-parametric relationship between the number of checklist items completed and the average fees charged per visit, reflects the same pattern: treatment providers exhibit a nearly identical production function to that of the control.

Finally, we can perform a mediation analysis to assess the extent to which the improvements in correct case management are mediated through increases in effort, or a movement along the curve. The results, presented in Appendix Table **??**, confirm that correct case management, time spent, and fees charged are all mediated by checklist adherence in SPs. The estimates suggest that 46% of the increase in correct case management, 90% of the increase in time spent, and 59% of the increase in fees charged can be traced to an increase in checklist adherence among treated providers, though the latter two estimates are imprecise.

5.2.4 Impact on the know-do gap

We now assess our theoretical prediction directly. If the intervention did indeed increase patient perceptions of the quality of treated providers, the treatment should differentially enable higher ability providers, as measured by knowledge at baseline, to increase their effort with commensurate increases in quality of treatment and fees. We therefore estimate regression (6) with provider effort at endline, as proxied by checklist adherence, and provider quality of care at endline, as proxied by correct case management with standardized patients, as our outcomes of interest.

Columns 1-6 of Table 8 demonstrate that individuals who are higher ability at baseline are enabled by the treatment to improve their effort and quality of care, whereas those who are untreated do not exhibit commensurate increases in quality. For example, providers who exhibit knowledge of correct case management at baseline are 18.9 percentage points more likely to correctly treat their SP at endline *if* they are treated; their high-ability counterparts in the control group see a noisy 1.9 percentage point increase in correct treatment with SPs.

This relationship is even more pronounced when we define provider ability using *endline* knowledge, or checklist adherence in endline vignettes.[29] Columns 7-9 of Table 8 demonstrate that among untreated providers, the correlation between vignettes and SP checklists at endline is weak and small in magnitude, while among treated providers, this correlation increases substantially (to 56 percentage points) and is significantly different from that of their control counterparts. The treatment therefore substantially reduces the know-do gap, enabling those who know more in theory to do more in practice.[30]

Moreover, while the impact of the training program on fees charged is imprecise, both the direction and the magnitude are consistent with a channel of increased patient demand and provider

[29] This measure of knowledge is of course endogenous to the intervention and therefore not our preferred specification; we investigate in the next section whether the intervention impacted endline knowledge.

[30] Appendix Table **??** shows similar, albeit weaker results using time spent with the SP as an alternate measure of effort.

effort as predicted by the theory. Namely, in Birbhum, a 100% increase in checklist completion is correlated with a $1.5 increase in fees. Treated providers asked 4.1% more checklist items, implying a fee increase of 0.062. In magnitude, this matches closely with the treatment effect on fees we observe; however, the experiment would require substantially greater power to pick up these effects precisely.

Taken in sum, the evidence from the Birbhum context is strongly consistent with the theory: a public and extensive training program led to a significant increase in treated providers' effort exerted, with a commensurate increase in quality of care and patient demand, but no corresponding change in underlying provider ability. High-ability providers, who may have formerly been stuck in a low-effort equilibrium, shift differentially more along each of these margins. The training program appears to have served primarily as a means of signalling provider quality, enabling underlying high-type providers to exercise in practice more of what they already possessed in knowledge.

5.2.5 Counterfactual reduction in medically avoidable costs

As correct case management increased without a decline in unnecessary treatments (as proxied by the number of medicines prescribed, reported in Das et al. (2016a)), we can now provide a clear policy counterfactual on avoidable medical expenditures. As Table 9 reports, the fraction of spending by SPs on the "worst case" scenario of incorrect treatment and no referral decreased by eight percentage points due to the intervention, and spending on correct treatment increased by five percentage points (ITT estimates). These estimates reflect purely economic costs, not considering the potential health benefits of increasing correct treatment rates and avoiding incorrect treatment.

6 Conclusion

Using a large dataset of provider and patient interactions across a wide swathe of geographies, we document systematically low quality of healthcare with the vast majority of expenditures being medically unnecessary. Importantly, the bulk of these unnecessary expenditures arise from under- and incorrect treatment, rather than the overtreatment that is the common concern in fee-for-service healthcare systems. This equilibrium of low quality care is inconsistent with profit maximization in fully functioning markets: healthcare providers have the knowledge to correctly treat patients, they can increase their revenue by doing so, and they operate in an environment with considerable excess capacity.

We suggest an underlying market failure which hinges on the bilateral and asymmetric nature of the medical interaction; as Arrow (1963) describes, the physician's work is dictated not only by profit maximization through higher-quality care but also by the necessity of preserving 'trust relations' with his or her patient.

Our theory shows that, in a signaling game where full revelation may drive low-type providers below their reservation wage, equilibria are always either fully or partially pooling. High-quality doctors would like to propose more complex treatments, but cannot due to patient uncertainty

about their type. This forces high-types on a dynamic path where more complex treatments can be introduced only once uncertainty over type has (sufficiently) resolved.

We are able to verify this prediction in a unique experimental setting in West Bengal, in which we find a prominent training program significantly improves clinical practice with no underlying change in provider ability. Notably, our training experiment reduces unnecessary expenditures and increases provider revenue, establishing the existence of an underlying market failure. These changes are both medically and economically significant, suggesting that a lack of trust not only depresses demand on the extensive margin, but also meaningfully limits the clinical practice of doctors.

References

Alsan, Marcella and Marianne Wanamaker. 2018. "Tuskegee and the Health of Black Men." *The Quarterly Journal of Economics* 133 (1):407–455.

Arrow, Kenneth J. 1963. "Uncertainty and the Welfare Economics of Medical Care." *The American Economic Review* 53 (5):941–973.

Bollinger, Christopher R. 1996. "Bounding Mean Regressions When a Binary Regressor Is Mismeasured." *Journal of Econometrics* 73 (2):387–399.

Chen, Brian K., Paul J. Gertler, and Chun-Yuh Yang. 2016. "Physician Ownership of Complementary Medical Services." *Journal of Public Economics* 144:27–39.

Daniels, Benjamin, Amy Dolinger, Guadalupe Bedoya, Khama Rogo, Ana Goicoechea, Jorge Coarasa, Francis Wafula, Njeri Mwaura, Redemptar Kimeu, and Jishnu Das. 2017. "Use of Standardised Patients to Assess Quality of Healthcare in Nairobi, Kenya: A Pilot, Cross-Sectional Study with International Comparisons." *BMJ Global Health* 2 (2):e000333.

Das, Jishnu, Abhijit Chowdhury, Reshmaan Hussam, and Abhijit V. Banerjee. 2016a. "The Impact of Training Informal Health Care Providers in India: A Randomized Controlled Trial." *Science* 354 (6308).

Das, Jishnu, Benjamin Daniels, Monisha Ashok, Eun-Young Shim, and Karthik Muralidharan. 2020. "Two Indias: The Structure of Primary Health Care Markets in Rural Indian Villages with Implications for Policy." *Social Science & Medicine* :112799.

Das, Jishnu and Saumya Das. 2003. "Trust, Learning, and Vaccination: A Case Study of a North Indian Village." *Social Science & Medicine (1982)* 57 (1):97–112.

Das, Jishnu and Jeffrey Hammer. 2005. "Which Doctor? Combining Vignettes and Item Response to Measure Clinical Competence." *Journal of Development Economics* 78 (2):348–383.

———. 2007. "Money for Nothing: The Dire Straits of Medical Practice in Delhi, India." *Journal of Development Economics* 83 (1):1–36.

Das, Jishnu, Alaka Holla, Aakash Mohpal, and Karthik Muralidharan. 2016b. "Quality and Accountability in Health Care Delivery: Audit-Study Evidence from Primary Care in India." *American Economic Review* 106 (12):3765–3799.

Das, Jishnu, Ada Kwan, Ben Daniels, Srinath Satyanarayana, Ramnath Subbaraman, Sofi Bergkvist, Ranendra K. Das, Veena Das, and Madhukar Pai. 2015. "Use of Standardized Patients to Assess Quality of Tuberculosis Care: A Pilot, Cross-Sectional Study." *The Lancet. Infectious diseases* 15 (11):1305–1313.

Das, Veena. 2015. *Affliction*. Fordham University Press.

Dulleck, Uwe and Rudolf Kerschbamer. 2006. "On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods." *Journal of Economic Literature* 44 (1):5–42.

Feldstein, Martin. 1967. *Economic Analysis for Health Service Efficiency: Econometric Studies of the British National Health Service*. Amsterdam: North-Holland Publishing Company.

Friedman, Milton and Rose D Friedman. 1962. *Capitalism and Freedom*.

Grépin, Karen A. 2016. "Private Sector An Important But Not Dominant Provider Of Key Health Services In Low- And Middle-Income Countries." *Health Affairs (Project Hope)* 35 (7):1214–1221.

Ingle, G. K. and Chetna Malhotra. 2007. "Integrated Management of Neonatal and Childhood Illness: An Overview." *Indian Journal of Community Medicine* 32 (2):108.

Jindal, S. K., D. Gupta, A. N. Aggarwal, R. Agarwal, World Health Organization, and Government of India. 2005. "Guidelines for Management of Asthma at Primary and Secondary Levels of Health Care in India (2005)." *The Indian Journal of Chest Diseases & Allied Sciences* 47 (4):309–343.

Kwan, Ada, Benjamin Daniels, Vaibhav Saria, Srinath Satyanarayana, Ramnath Subbaraman, Andrew McDowell, Sofi Bergkvist, Ranendra K. Das, Veena Das, Jishnu Das, and Madhukar Pai. 2018. "Variations in the Quality of Tuberculosis Care in Urban India: A Cross-Sectional, Standardized Patient Study in Two Cities." *PLoS medicine* 15 (9):e1002653.

Leonard, Kenneth L. and Melkiory C. Masatu. 2005. "The Use of Direct Clinician Observation and Vignettes for Health Services Quality Evaluation in Developing Countries." *Social Science & Medicine* 61 (9):1944–1951.

Lowes, Sara and Eduardo Montero. 2019. "Traditional Medicine in Central Africa." *AEA Papers and Proceedings* 109:516–520.

Maestad, Ottar, Gaute Torsvik, and Arild Aakvik. 2010. "Overworked? On the Relationship between Workload and Health Worker Performance." *Journal of Health Economics* 29 (5):686–698.

Martinez-Bravo, Monica and Andreas Stegmann. 2018. "In Vaccines We Trust? The Effects of the CIA's Vaccine Ruse on Immunization in Pakistan." Tech. Rep. wp2017_1713, CEMFI.

McGuire, T. G. and M. V. Pauly. 1991. "Physician Response to Fee Changes with Multiple Payers." *Journal of Health Economics* 10 (4):385–410.

Mohanan, Manoj, Marcos Vera-Hernández, Veena Das, Soledad Giardili, Jeremy D. Goldhaber-Fiebert, Tracy L. Rabin, Sunil S. Raj, Jeremy I. Schwartz, and Aparna Seth. 2015. "The Know-Do Gap in Quality of Health Care for Childhood Diarrhea and Pneumonia in Rural India." *JAMA pediatrics* 169 (4):349–357.

Saria, Vaibhav. 2020a. "The Fallen Idol: Mistrust and Medicine in India."

———. 2020b. "New Machine, Old Cough: Technology and Tuberculosis in Patna." *Frontiers in Sociology* 5.

Sylvia, Sean, Yaojiang Shi, Hao Xue, Xin Tian, Huan Wang, Qingmei Liu, Alexis Medina, and Scott Rozelle. 2015. "Survey Using Incognito Standardized Patients Shows Poor Quality Care in China's Rural Clinics." *Health Policy and Planning* 30 (3):322–333.

Sylvia, Sean, Hao Xue, Chengchao Zhou, Yaojiang Shi, Hongmei Yi, Huan Zhou, Scott Rozelle, Madhukar Pai, and Jishnu Das. 2017. "Tuberculosis Detection and the Challenges of Integrated Care in Rural China: A Cross-Sectional Standardized Patient Study." *PLOS Medicine* 14 (10):e1002405.

Yi, Hongmei, Grant Miller, Linxiu Zhang, Shaoping Li, and Scott Rozelle. 2015. "Intended And Unintended Consequences Of China's Zero Markup Drug Policy." *Health Affairs* 34 (8):1391–1398.

Tables

Table 1: Patterns of treatment

	(1) Any correct treatment	(2) Correct treatment	(3) Over-treatment	(4) Incorrect treatment	(5) Gave an antibiotic (excl. diarrhea)	(6) Gave a steroid (excl. asthma)	(7) Referred to another provider	(8) Number of cases
MP	0.302	0.048	0.255	0.698	0.350	0.032	0.180	939
Birbhum	0.237	0.015	0.222	0.763	0.331	0.015	0.321	396
Delhi	0.108	0.008	0.100	0.892	0.540	0.092	0.104	250
Mumbai	0.292	0.033	0.258	0.708	0.566	0.198	0.086	1,583
Patna	0.310	0.051	0.259	0.690	0.679	0.096	0.057	1,019
China	0.361	0.237	0.124	0.639	0.512	0.000	0.191	299
Kenya	0.524	0.211	0.313	0.476	0.548	0.016	0.164	166

Notes: All figures are unweighted. In correct treatment definitions, referrals to a higher level of care alone are NOT considered the right treatment. The Birbhum data includes observations from the control group only.*** p<0.01, ** p<0.05, * p<0.1.

Table 2: Necessary and avoidable costs of treatment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Cost breakdown (US dollars)				**Fraction of total cost that is avoidable**			**Number of cases**
	Total	Consultation	Medicines	Avoidable	Total	Attributed to over-treatment	Attributed to incorrect treatment	
MP	0.646	0.294	0.354	0.414	0.696	0.041	0.655	939
Birbhum	0.689	0.183	0.506	0.562	0.861	0.129	0.731	396
Delhi	2.094	1.616	0.478	1.462	0.805	0.026	0.778	250
Mumbai	4.898	3.247	2.206	2.859	0.763	0.068	0.695	1,583
Patna	5.503	2.674	2.545	3.990	0.771	0.096	0.675	1,019
China	3.763	0.224	2.855	3.238	0.848	0.116	0.732	299
Kenya	4.330	1.553	3.846	3.335	0.746	0.228	0.518	166

Notes: We assume that all costs are necessary when a provider recommends the correct treatment and only the correct treatment. When a provider over-treats, we assume that the cost of consultation, and cost of indicated medicines are necessary, while the costs of unnecessary medicines are avoidable. Finally, when a provider recommends an incorrect treatment, we assume all costs (consultation and medicines) are unnecessary.** p<0.01, ** p<0.05, * p<0.1.

Table 3: Empirical Estimates of the Know-Do Gap

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Madhya Pradesh			**Birbhum**		**Delhi**	**China**
	Full sample	Two reports sample					
Panel A: Base specification							
OLS	0.138***	0.104***		0.026		0.140***	0.193***
	(0.031)	(0.037)		(0.058)		(0.050)	(0.059)
Average of two reports		0.158***		0.117*			
		(0.047)		(0.070)			
Max of two reports		0.128***		0.126*			
		(0.043)		(0.068)			
Bollinger specification		0.161***		0.132*			
		(0.048)		(0.072)			
Number of observations	939	525		393		69	267
Panel B: Structural Estimates							
OLS (upper bound)	0.386**	0.159***	0.232***	0.040	0.085		
	(0.163)	(0.061)	(0.086)	(0.052)	(0.150)		
IV Linear (lower bound)	0.360**	0.198*	0.197**	0.268	0.207*		
	(0.103)	(0.114)	(0.084)	(0.314)	(0.120)		
IV Linear	0.593**	0.253*	0.355**	0.318	0.293*	0.358	0.426**
	(0.177)	(0.147)	(0.153)	(0.376)	(0.171)	(0.246)	(0.188)
GMM	0.334***	0.168***	0.218***	0.235	0.226***		
	(0.086)	(0.049)	(0.060)	(0.240)	(0.090)		
a0	0.156	0.044	0.401	0.332	0.060		
a1	0.146	0.109	0.118	0.158	0.193		
Number of observations	939	525	525	393	393		
Instruments	Percentage checklist	Percentage checklist	Second report	Percentage checklist	Second report	Percentage checklist	Percentage checklist
Panel C: First Stage for IV							
Percentage checklist	0.553***	0.648***		0.863***		1.642***	1.549***
	(0.088)	(0.119)		(0.273)		(0.437)	(0.228)
Second report			0.263***		0.321***		
			(0.052)		(0.057)		
Number of observations	939	525	525	393	393	69	267
F-statistic	37.380	30.420	32.490	133.880	130.670	14.082	45.880
Panel D: Summary statistics							
Mean of vignettes	0.706	0.644		0.450		0.724	0.764
Mean of SPs	0.302	0.240		0.239		0.108	0.361

Notes: Robust standard errors clustered at the facility level in parenthesis. All regressions include a constant and case fixed effects. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Price regressions by study using SP data

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	
	Outcome: Price charged in SPs (USD)									
	Madhya Pradesh		Birbhum		Delhi		Mumbai		Patna	
	Binary	Multiple	Binary	Multiple	Binary	Multiple	Binary	Multiple	Binary	M
Time spent with SPs (mins)	0.044***	0.020***	0.158***	0.066**	0.143***	0.029	1.073***	0.183**	0.273***	-
	(0.007)	(0.007)	(0.021)	(0.028)	(0.034)	(0.039)	(0.090)	(0.084)	(0.079)	(
Checklist completed (percent)	0.755***	0.524***	1.511***	0.009	6.195***	3.333**	28.209***	16.068***	23.532***	16
	(0.087)	(0.097)	(0.245)	(0.274)	(1.084)	(1.323)	(1.257)	(1.362)	(1.358)	(
Any correct treatment	0.167***	0.055	0.342***	0.071	1.948***	1.623***	12.332***	7.806***	9.552***	7
	(0.044)	(0.047)	(0.089)	(0.082)	(0.409)	(0.362)	(0.490)	(0.524)	(0.531)	(
Number of medicines	0.129***	0.077***	0.268***	0.237***	0.265***	0.220***	-0.676***	-0.298***	0.388**	
	(0.012)	(0.014)	(0.025)	(0.028)	(0.060)	(0.069)	(0.107)	(0.093)	(0.188)	(
Referred/asked to see provider	-0.349***	-0.297***	-0.150**	0.000	-1.781***	-2.069***	-8.022***	-11.663***	-9.609***	-1
	(0.038)	(0.056)	(0.064)	(0.085)	(0.442)	(0.572)	(0.776)	(0.917)	(0.822)	(
Number of patients waiting	0.011	0.014**	0.032	0.015	-0.038	-0.036	0.192***	0.028	0.107**	
	(0.008)	(0.007)	(0.032)	(0.026)	(0.039)	(0.037)	(0.062)	(0.041)	(0.048)	(
R-squared		0.325		0.449		0.324		0.477		
Number of observations		684		396		250		1,575		
Mean of fees charged (USD)		0.543		0.634		2.060		8.161		
St dev. of fees charged		0.452		0.591		1.818		9.087		

Note: Robust standard errors clustered at the facility level are in parenthesis. Observations are at the SP-provider interaction level. Interpretation o regressions needs caution. Each coefficient represents a separate regression of prices on the row variable and case fixed effects. Multiple regressions inc All multiple regressions include a coefficient, not show for brevity. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 5: Fees Charged (USD) using real patients' data

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Madhya Pradesh		Birbhum		Madhya Pradesh		Birbhum	
	Binary	Multiple	Binary	Multiple	Binary	Multiple	Binary	Multiple
Time spent with patients (mins)	0.073***	0.048***	0.065***	0.037***	0.090***	0.071***	0.073***	0.038***
	(0.014)	(0.016)	(0.013)	(0.013)	(0.017)	(0.019)	(0.014)	(0.012)
Number of questions and exams	0.074***	0.035**	0.040***	0.020**	0.047***	0.013	0.055***	0.023**
	(0.015)	(0.015)	(0.009)	(0.009)	(0.012)	(0.014)	(0.010)	(0.009)
Correct treatment rate in vignettes	0.449**	0.165	0.306	0.272				
	(0.179)	(0.128)	(0.211)	(0.197)				
Number of medicines	0.167***	0.129***	0.249***	0.217***	0.133***	0.117***	0.255***	0.217***
	(0.026)	(0.027)	(0.028)	(0.027)	(0.019)	(0.027)	(0.031)	(0.028)
Referred patients	-0.444***	-0.218	-0.107	-0.115	-0.154	-0.064	-0.297*	-0.210
	(0.141)	(0.140)	(0.142)	(0.130)	(0.137)	(0.136)	(0.173)	(0.132)
Patients ADL score	0.041	0.024	0.039**	0.026*	0.018	0.007	0.030**	0.021*
	(0.031)	(0.025)	(0.019)	(0.016)	(0.020)	(0.018)	(0.015)	(0.013)
Patients Assets Level	0.028*	0.026*	0.020	0.025**	0.001	0.001	0.011	0.014
	(0.015)	(0.015)	(0.014)	(0.012)	(0.012)	(0.014)	(0.011)	(0.011)
R-squared		0.179		0.189		0.607		0.510
Number of observations		821		981		821		981
Provider fixed effects					X	X	X	X
Mean of fees charged (USD)		0.786		0.828		0.786		0.828
SD of fees charged		0.698		0.819		0.698		0.819

Note: Robust standard errors clustered at the facility level are in parenthesis. Observations are at the interaction level and correspond to patient visits. Interpretation of coefficients in binary regressions needs caution. Each coefficient represents a separate regression of prices on the row variable and provider fixed effects if applicable. All multiple regressions include a coefficient, not show for brevity. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 6: Effect of treatment on SP Outcomes

	(1) Checklist in SPs	(2) Any correct treat-ment	(3) Time spent (mins)	(4) Fees charged (USD)	(5) Checklist in SPs	(6) Any correct treat-ment	(7) Time spent (mins)	(8) Fees charged (USD)
	ITT Estimates				LATE Estimates			
Treatment group	0.040***	0.073*	0.239*	0.067				
	(0.011)	(0.038)	(0.140)	(0.078)				
Training attendance					0.067***	0.122*	0.401*	0.113
					(0.018)	(0.063)	(0.229)	(0.130)
R2	0.089	0.053	0.119	0.035	0.118	0.061	0.129	0.034
Number of observations	790	790	790	790	790	790	790	790
Mean of dependent variable: Control	0.273	0.520	3.252	0.689	0.273	0.520	3.252	0.689
Mean of dependent variable: Treatment	0.313	0.594	3.495	0.757	0.313	0.594	3.495	0.757

Note: Robust standard errors clustered at the facility level are in parenthesis. Observations are at the provider-SP interaction level and correspond to SP visits. All multiple regressions include a constant and case and block fixed effects, not show for brevity. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 7: Effect of treatment on knowledge

	(1)	(2)	(3)	(4)
	Checklist		Any correct treatment	
Panel A: ITT Estimates				
Treatment	0.012*	0.038**	0.002	0.089
	(0.007)	(0.018)	(0.033)	(0.076)
Ability		0.260***		0.421
		(0.073)		(0.259)
Ability * Treatment		-0.121		-0.430
		(0.104)		(0.389)
R2	0.011	0.091	0.023	0.037
Number of observations	784	778	784	778
Mean of DV: Control	0.202	0.202	0.774	0.774
Mean of DV: Treatment	0.214	0.216	0.775	0.774
Panel B: LATE Estimates				
Training attendance	0.020*	0.071**	0.003	0.161
	(0.012)	(0.031)	(0.055)	(0.140)
Ability		0.263***		0.424
		(0.072)		(0.260)
Ability * Treatment		-0.240		-0.771
		(0.173)		(0.688)
R2	0.052	0.133	0.023	0.036
Number of observations	784	778	784	778
Mean of DV: Control	0.202	0.202	0.774	0.774
Mean of DV: Treatment	0.214	0.216	0.775	0.774

Note: Ability is proxied by checklist adherence in baseline vignettes for columns 1-2 and any correct treatment in baseline vignettes for columns 3-4. Robust standard errors clustered at the facility level are in parenthesis. Observations are at the interaction level and correspond to SP visits. All regressions include a constant, case and block fixed effects; columns 2 and 4 also include controls for age and gender of provider, not show for brevity. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 8: Treatment effects by underlying ability, SP data

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Checklist in SPs			Any correct treatment			Checklist in SPs		
	Control	Treatment	Both	Control	Treatment	Both	Control	Treatment	Both
Knowledge	0.084	0.243**	0.074	0.019	0.189***	0.060	0.167	0.560***	0.173
	(0.119)	(0.111)	(0.117)	(0.058)	(0.057)	(0.050)	(0.145)	(0.135)	(0.140)
Treatment group			0.014			0.034			-0.045
			(0.029)			(0.035)			(0.042)
Knowledge * Treatment			0.159			0.093			0.396**
			(0.158)			(0.059)			(0.195)
R2	0.059	0.115	0.099	0.203	0.168	0.182	0.063	0.147	0.122
Number of observations	396	394	790	396	394	790	393	391	784

Note: Knowledge is proxied by checklist adherence in baseline vignettes for columns 1-3; any correct treatment in baseline vignettes for columns 4-6; and checklist adherence in endline vignettes for columns 7-9. Robust standard errors clustered at the facility level are in parenthesis. Observations are at the interaction level and correspond to SP visits. All multiple regressions include a constant, case and block fixed effects, not show for brevity. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 9: Impact of intervention on revenues and avoidable expenditures

	(1)	(2)	(3)	(4)
	Control		Treatment	
	Total (USD)	Percentage of total spending	Total (USD)	Percentage of total spending
Spending by SPs	$273.0		$298.1	
Spending by treatment type				
Incorrect treatment, no referral	$115.6	42.3%	$102.4	34.3%
Incorrect treatment, with referral	$50.2	18.4%	$64.2	21.5%
Correct treatment	$107.2	39.3%	$131.6	44.1%

Figures

Figure 1

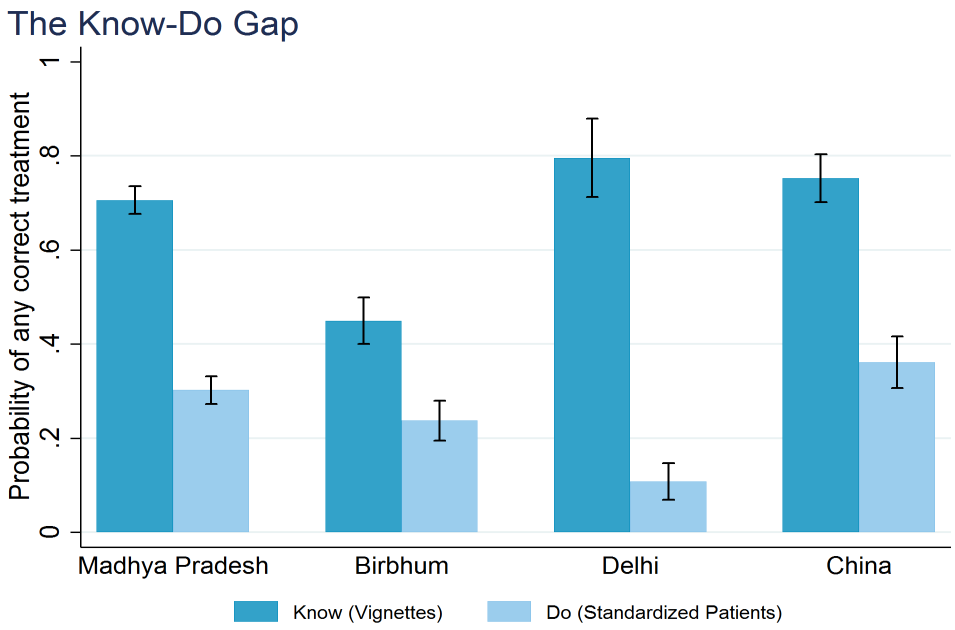


Figure 2

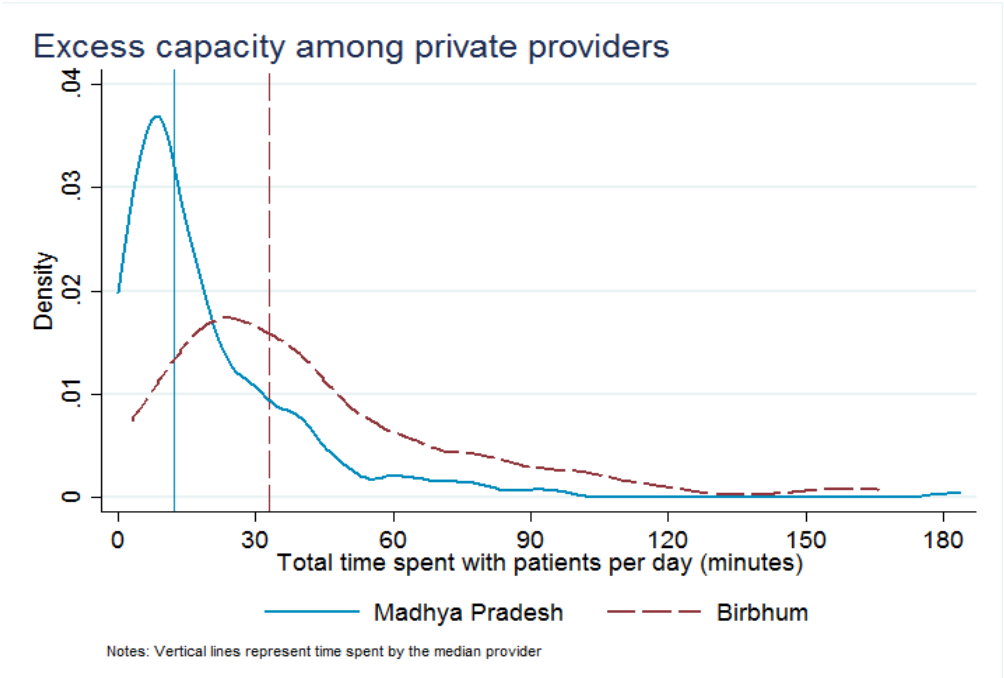


Figure 3

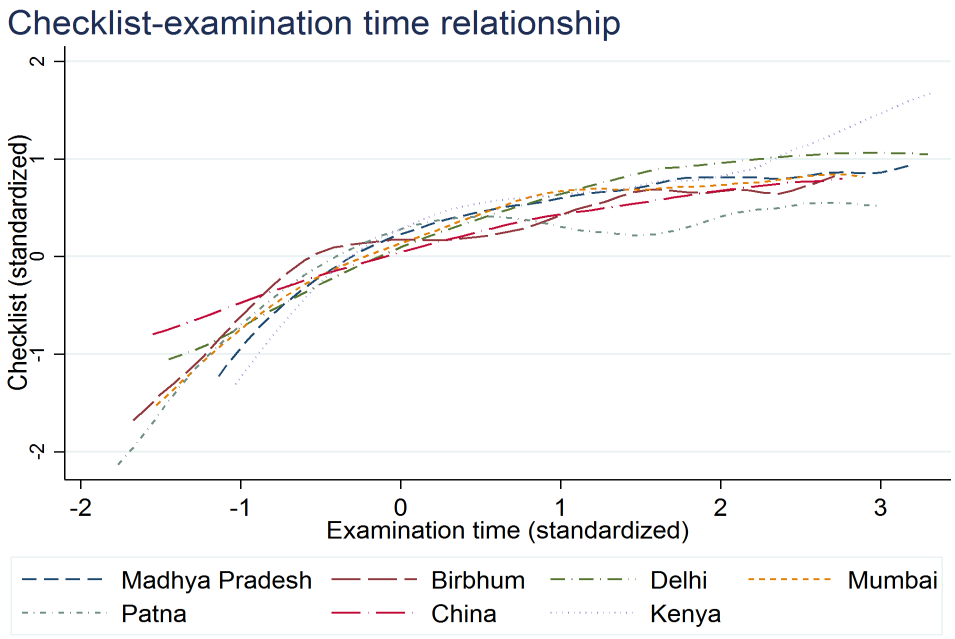


Figure 4

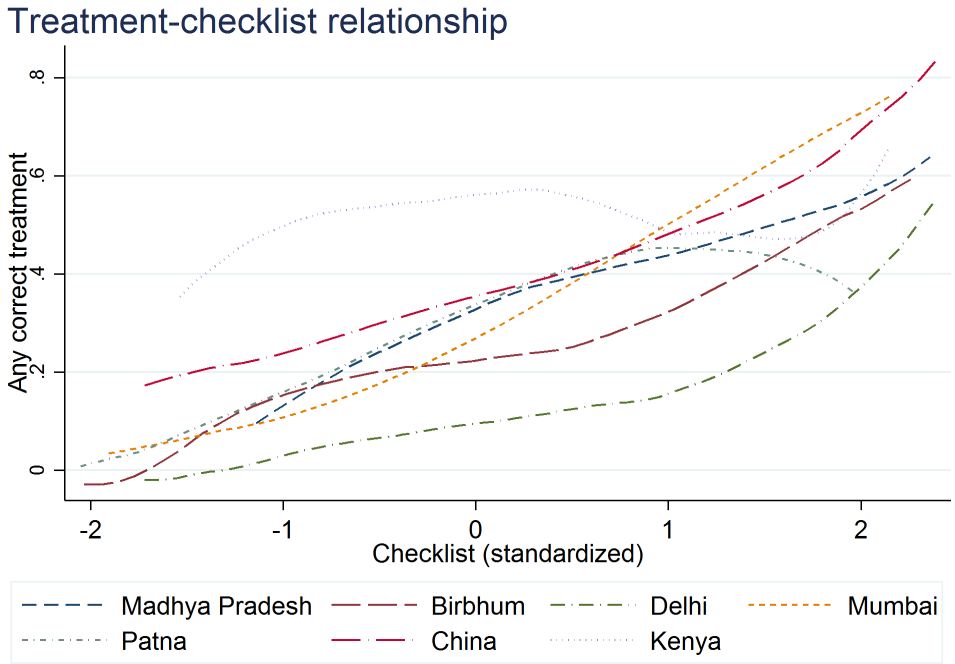


Figure 5

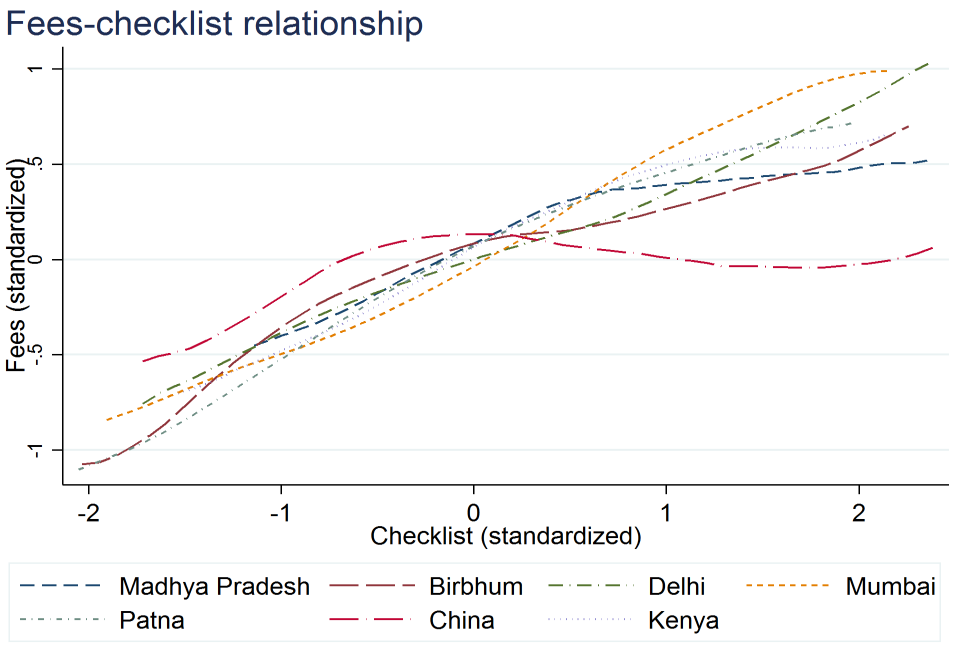


Figure 6

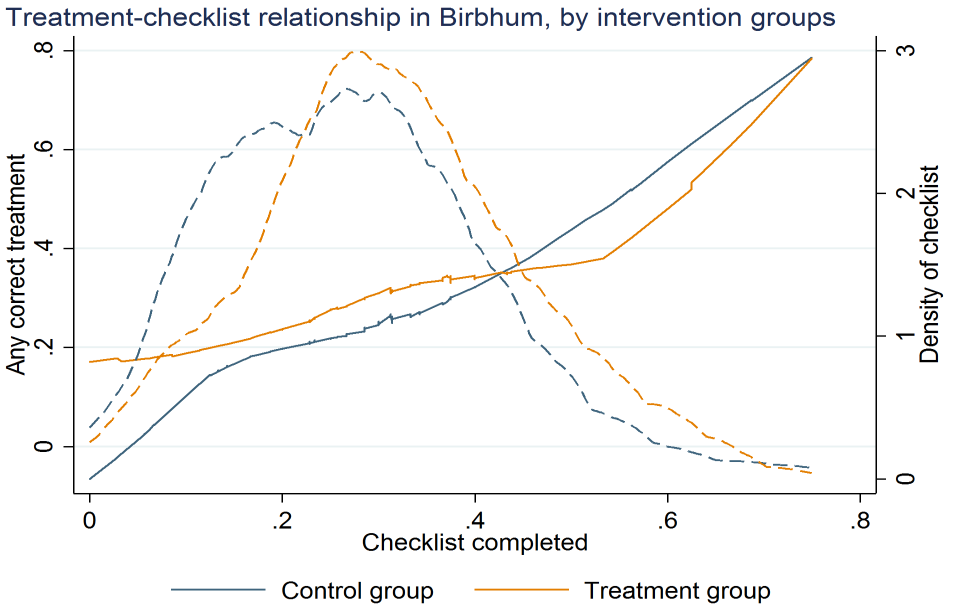
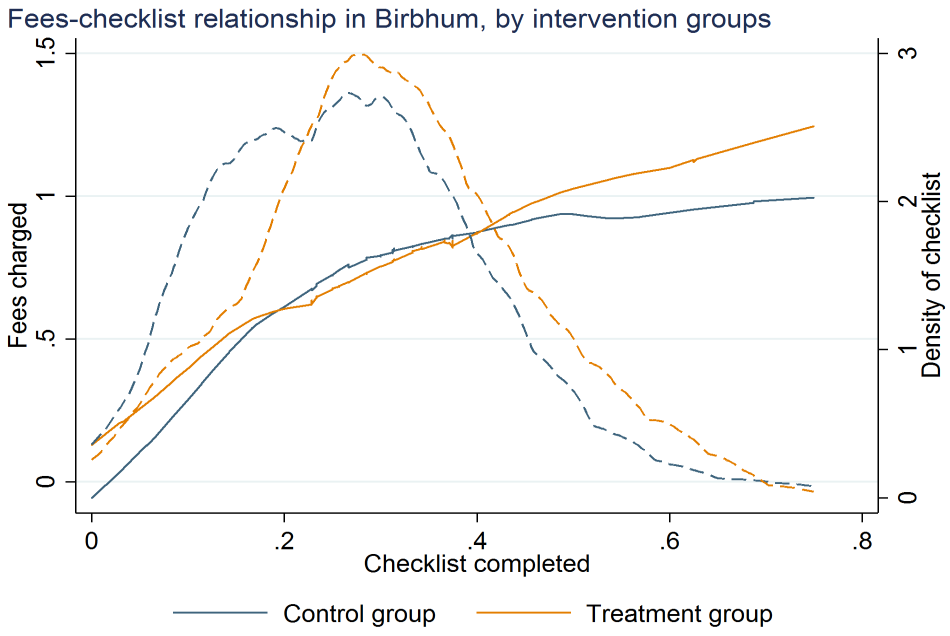


Figure 7


Fees-checklist relationship in Birbhum, by intervention groups
Fees charged
0
.5
1
1.5
Density of checklist
0
1
2
3
0
.2
.4
.6
.8
Checklist completed
Control group
Treatment group

7 Appendix A: Sample Description and Construction of Data for Each Site

We use data from seven sites collected through six different studies between the 2010-2016 period. Five of the seven sites are in India, of which two are rural – Madhya Pradesh and Birbhum, and three are urban – Delhi, Mumbai and Patna. We completed these with data from two other countries – China and Kenya. Below we describe each sample and the characteristics of the data.

7.0.1 Madhya Pradesh

The data for Madhya Pradesh was collected through the Medical Advice, Quality and Availability in Rural India (MAQARI) project funded by the Bill and Melinda Gates Foundation between 2010-2011 (Das et al., 2016). At per capita income of $1,100 in 2016, Madhya Pradesh is one of the poorest states of India (ranked 27 out of 33). Around the time of fieldwork (2010-2011), Madhya Pradesh was also among the worst performing state in terms of health outcomes. Infant Mortality was 62 per 1,000 live births, the highest rate among all states. The original study consists of two separate sample of providers – the first is a sample of public and private providers operating in a set of 100 representative villages in 5 districts and the second is a sample of public MBBS providers (along with their private clinics) in the same five districts.

To generate the first sample, the study first drew 100 villages across 5 districts of the state, which were stratified by geography and an index of health outcomes. In the first phase, the authors enumerated a household census in each village, which among other things, asked households to identify all providers (regardless of their practice locations) they sought primary care services from in the past 30 days. Subsequently, the study surveyed all providers serving the villages, regardless of whether providers were located inside and outside the village, to form "health markets" for the sampled villages. The average number of providers per market is 11 and the majority are private (64 percent). Forty-six percent of all providers and 77 percent of all private providers have no formal medical training, yet they account for 77 percent of all primary care visits. In the second phase of the study, standardized patients were administered to a subset of randomly sampled providers (stratified by practice sector and qualification), in 3 districts (60 health markets). Column 1 of Appendix Table **??** reports characteristics of providers in the sampling universe. Only twenty percent of 706 providers are public providers, 7 percent had MBBS degrees, and 17 percent had qualifications in alternative systems of medicine. The majority has no formal medical qualifications. The average age of providers is 43 years and 88 percent are males. Providers see 16 patients per day on average. The authors randomly sampled 247 unique provider-clinic combinations (5 providers had multiple practices) to receive unannounced SPs for three tracer conditions. SP interactions were completed in 224 provider-clinic combinations, representing a completion rate of 91 percent. Medical vignettes for the same three tracer conditions were administered in a third phase (3 months after the SP interactions were completed). Finally, in a fourth phase (six months after the third phase), providers revisited providers to complete a second round of vignettes. Crucially for our purpose, these data

allow us to estimate and correct for measurement error in the Know-Do estimates. Column 2 of Table A1 reports characteristics of providers included in this paper and Column 3 reports results from t-tests of comparison between the universe and the data. There are no major systematic differences between the universe and the sample. One key difference is that public providers are less likely to be sampled. This is because public providers often operate in clinics with multiple providers, and to minimize the risk of detection (i.e. many angina patients arriving in a short time frame), only one provider was sampled from each public clinic.

The study supplemented this data with a separate sample of qualified providers who operate in the public sector in all 5 districts included in the study. The authors obtained a list of all Primary Health Centers (PHCs) and Community Health Centers (CHCs) and the providers mapped to those facilities. Excluding PHCs/CHCs that were mapped as part of the representative sample, 200 additional facilities were encountered. Of these, 40 did not have an MBBS provider posted, and there were 216 providers mapped to the remaining 160 facilities. The team then undertook fieldwork to find out if these providers operated private practices and, if yes, to locate their private practices. Private practices were located for 132 of the 216 providers (61.1 percent). For the SP study, the authors sampled one provider from each PHC/CHC with a preference for a provider with a private practice in cases where there were multiple providers in the same clinic. In cases where providers were posted to multiple public facilities, and where there were no additional MBBS providers to sample from, the authors randomly sampled providers from one of the facilities. This sampling strategy yielded 139 providers of which, 91 providers (65.5 percent) operated private practices. The average of providers in this sampling frame is 44.5 which is comparable to the age of providers in the representative sample (Column 4 of Table A1). However, providers in this sample have shorter tenure practicing in their current location (7.9 years versus 13.6 years in the representative sample).

All sampled providers were assigned to receive asthma and diarrhea cases in both public and private facilities. Providers with private facilities were randomly assigned to receive angina either one of their public and private facilities. Providers without a private practice were assigned angina cases in their public facilities by default. SPs completed at least one interaction with 116 of the 139 providers (consistent with high rates of absenteeism in the public sector). Of the 48 providers without private practices, SPs completed interactions with 32 (66.7 percent). Of the 91 providers with private practices, SPs completed interactions with 84 (92.3 percent). The total number of unique provider-practice combinations with at least one completed SP interaction is 187. Vignettes were administered to all sampled providers in a subsequent phase. Of course, because vignettes measure knowledge (which is assumed same across sectors), these were administered only once. Of the 187 providers with completed SP interactions, vignettes were completed for 139 (74 percent). Column 5 of Table A1 reports characteristics of providers included in this paper and Column 5 shows that are no major differences between the universe and the sample.

In this paper we combine the two samples to generate a paired vignettes-SP data set of 939 case presentations to 352 providers. In this final sample, thirty percent of providers are from the

public sector, 45 percent have an MBBS degree, 11 percent have some other qualification, and 44 percent have no formal training (Table A1). The average provider is 44 years old, has been in her current location for 11 years, and sees 18 patients per day on average. Appendix Table A1 also shows that the pooled sample preserves the average provider characteristics of the original sampling universe. The combined sample allows us to maximize statistical power and generate evidence on the entire gamut of health care provider types – those with no formal training, some training and fully-qualified.

7.0.2 Birbhum

The Birbhum data was collected as part of the randomized evaluation of a 9-month long multitopic training program for informal providers in Birbhum district in the state of West Bengal. The training was designed and executed by the Liver Foundation, a public health organization based in West Bengal (Das et al., 2016). The program invited 360 providers to participate in the training, of whom 304 responded positively. These were randomized equally into treatment and control groups (stratified by blocks), and members of the control group were told that they would be eligible for the program in the following year.

Providers in Birbhum are very similar to informal providers encountered in other states. The average provider is 40 years old, has 13 years of experience and is male (95 percent) (Appendix Table **??**). About 62 percent of providers have completed high school and 75 percent have no formal training while the remaining 25 percent have some formal training. Appendix Table **??** also shows internal validity of the initial randomization as there are no differences in provider characteristics between treatment and control groups (Columns 2 and 3).

Baseline data were collected prior randomization and included information of provider backgrounds and practice characteristics. Vignettes were also administered at baseline which provides the baseline estimate of providers' knowledge. The training consisted of 72 sessions and 150 teaching hours over a 9-month period. It included a wide variety of topics, with an emphasis on basic medical conditions, triage, and avoidance of harmful practices. Full details of the training program are available from Das et al. (2016). Endline data collection took place within 3 to 6 months within the end of the training program. Standardized patients made unannounced visits to all providers who participated in the evaluation. Endline follow-up rate was 85.5 percent for the control group and 88.2 percent for the treatment group, and there is no evidence of differential attrition (Columns 4-6 of Appendix Table **??**). Vignettes were administered again at endline to estimate training impact on knowledge. The two rounds of vignettes data for the control group also allows for the estimation and correction of measurement error in the Know-Do estimates.

7.0.3 China

Medical care in rural China is provided primarily by the public sector through a three-tiered health care system comprising of village clinicians (VCs), township health center clinicians (THCs), and county hospitals (CHs). Although patients are free to choose to among any of the three tiers,

VCs and THCs are often the first point of contact with the health system for the 600 million rural residents of China and often the only source of medical care (Barbiarz et al., 2010). Before 2011, guidelines stipulated that suspected TB patients be referred to separate county -level facilities under the CCDC that diagnosed and treated patients under the Directly Observed Therapy (DOTs) strategy (State Council of China, 2011). Since then these functions has been transferred to county-level hospitals within the regular health system.

Starting from 2002, China has significantly expanded health insurance coverage in rural areas through the New Rural Co-Operative Medical Scheme (in conjunction with the Urban Employee Basic Medical Insurance and Urban Resident Basic Medical Insurance in urban areas). As a result, insurance coverage nationwide increased from less than 10 percent in 2002 to over 95 percent in 2017 (Liu, Vortherms and Hong, 2017). Nevertheless, while providers in VCs and THCs receive a nominal base salary and a capitation payment to provide treatment for a set of basic health conditions from the public sector, they mostly rely on revenues from drug sales and user fees (see for example, Sylvia et al., 2010). Consequently, while health insurance expansion increased coverage, patient costs and the incidence of inappropriate treatment remain high, including the high use of antibiotics (Currie et al., 2011). This has promoted a series of health reforms starting in 2009 to address issues related to inefficiency an fragmentation of the health system (China Joint Study Partnership, 2016).

The data used in this paper was collected as part of a larger cross-sectional study on quality of care in rural China (see Sylvia et al., 2017 for more details). The study sampled providers from rural areas in three provinces – Sichuan, Saanxi, and Anhui. In each province, the authors selected one prefecture (the administrative division below the province level but above county, from a total of 47 prefectures). The sample was selected as triplets of VCs, THCs and CHs so that the final sample is representative of rural health systems. The researchers first sampled 21 of the 24 counties in the three prefectures and included the primary CHs in the study. Within each county, the study then sampled 10 THCs randomly. A total of 311 THCs were available to sample from, and 209 THCs were included in the study (one county only had 9 THCs). Then for each THCs, the study randomly sampled an associated VC, yielding 209 VCs. Twenty-percent of THCs had no associated VCs and these were replaced by a randomly chosen backup VC. The final sample thus comprises of 439 facilities – 21 CHs, 209 THCs and VCs each – which is representative of set of facilities serving 12.23 million people.

Data were collected in three waves. In the first wave implemented in June 2015, an initial facility and provider survey were completed among village and township providers (but not county hospitals). SP made unannounced visits in the second wave in August 2015. SP visits were made to all 209 THCs and 21 CHs but only to randomly sampled 49 VCs. This is done to minimize the risk of detection as the the larger study include multiple cases of SPs and sending many SPs to the same provider at the village level could have aroused suspicion. The targeted number of health clinics across the three levels for the implementation of TB SPs is 279. The final wave of data collection was conducted in September 2015 in which providers in village clinics and township

health centers (but not county hospitals) were asked if they detected any SPs and vignettes were administered.

Since clinical vignettes were not implemented in the county hospitals, the sample used in this paper only uses data from the VCs and THCs. Of the 258 possible SP interactions, successful visits were made to 253 (98 percent), and of the 253 paired SP-vignettes cases possible, vignettes were completed with 243 providers (96 percent). The reason for non-completion in all cases was the absence of providers at the time of the visit.

Appendix Table **??** reports provider and clinic characteristics. Overall education levels among village providers was low: only 9 percent had completed upper secondary or higher education. Eighty-five percent had completed the "Rural Physician Certificate" training, the minimum qualification required to practice as a village clinician. Another 13 percent reported completing the "Assistant Physician Certificate" training. Education levels were higher among township providers. Sixty percent had completed upper secondary of higher general education, and 61 percent had a "Practicing Physician Certificate." Twenty-three percent reported completing the "Assistant Practicing Physician Certificate," and 9 percent had the "Rural Physician Certificate." In both types of clinics, over 85 percent of providers were male, and the average age was 52 years in VCs and 44 years in THCs. Average monthly salaries for village clinicians was $318 and for township health center clinicians was $503.

7.0.4 Delhi

The sample from Delhi was drawn from a convenience sample of 106 private healthcare providers practicing in outpatient settings in low- and middle-income neighborhoods of Delhi (Das et al., 2015). A total of 250 SP cases were assigned to 100 consenting providers who were informed that over the subsequent six months, they may visit someone who is not a real patient. The study was designed as a pilot to validate the use of SP methodology for the assessment of quality of care for TB, and thus only administered four variants of TB cases (more on this later). The two naïve TB cases (TB1 and TB2) were assigned to 75 providers each randomly (50 providers were to receive both TB1 and TB2 cases). Then 50 providers were randomly TB3 cases and finally, those who were not assigned TB3 received TB4 cases. Fieldwork was conducted in April 2014, and all providers received two or three SPs (the study ensured that no providers received more than 3 SPs).

In a subsequent phase, a brief provider survey and vignettes were administered which was completed by 93 out of 100 providers. This paper uses data for these 93 providers for whom there exists paired SP-vignettes interactions. Twenty-eight percent had a four-year medical degree (MBBS) and 43 percent had degrees in alternative systems of medicine. The remainder were informal providers with minimum or no qualifications. An overwhelming majority of these providers were male (97 percent) and the average age was 46.8 years (Appendix Table **??**). Although providers were not asked to self-report the number of patients they see, the average number of patients waiting at the clinics during the time of SP visits was 1.27 which suggests low caseload or patients in this setting.

7.0.5 Mumbai and Patna

Data for Mumbai and Patna comes from the Private Provider Interface Agency (PPIA) pilot programs funded by the Bill and Melinda Gates Foundation (Kwan et al., 2016). Mumbai, with a population of 12 million, is the capital of the state of Maharashtra and has an annual per capita income of $2,845. Patna is the capital of one of the least developed states – Bihar, and its 2 million population has a per capita income of $470. In both cities, like much of urban and rural India, the private sector is the dominant source of care despite the existence of nominally priced public system.

In both cities, to map providers and facilities, field officers conducted a street-by-street mapping exercise to construct a comprehensive universe of all provider and facilities in the private sector. All providers known to see adult outpatients with respiratory problems were eligible for the study (in practice, this includes all providers but orthopedists, gynecologists, ophthalmologists, and pediatricians). This exercise yielded 7,115 providers in Mumbai (3,591 or 50 percent did not have MBBS degrees) 3,179 providers in Patna (2,716 or 85 percent did not have an MBBS degree) (see Kwan et al., 2016). Eligible providers were then sampled randomly after stratification by geographical areas within each city. In Mumbai, the researchers sampled 831 providers (331 facilities with MBBS providers and 500 non-MBBS providers) and in Patna, 591 providers (471 MBBS providers and 120 non-MBBS providers). In both settings, all providers were assigned the TB1 case. In addition to TB1, in Mumbai, providers were also assigned to receive TB2, TB3 or TB4 (1:1:2 ratio). In facilities with MBBS providers, facilities were assigned 1 to 3 walk-in TB1 cases. In addition, facilities were also randomly assigned TB2, TB3 and TB4 cases. In Patna, all providers were assigned two cases – TB1 was assigned to all providers, and one of TB2, TB3, and TB4 was randomly assigned to each provider (1:1:1 ratio).

SP interactions were completed with high success rates: 727 of 831 providers in Mumbai (87 percent) and 473 of 591 providers in Patna (80 percent). Appendix Table **??** present further details on provider characteristics. The fraction of fully qualified providers (with MBBS degrees) in Mumbai is 31 percent which is lower than that in Patna (75 percent). Most providers are male (83 percent in Mumbai and 96 percent in Patna) and are in the age category of 30-50 years (71 percent). The average caseload of providers is higher than that observed in Delhi. At the time of the SP visit, an average of 2.07 and 2.58 patients were waiting to be seen in Mumbai and Patna, respectively.

7.0.6 Kenya

The Kenya data consists of both private and public providers from neighborhoods of Nairobi and comes from Daniels et al. (2017). Forty-six health facilities in low-, middle- and high-income neighborhoods were approached in a convenience sample, of which 46 agreed to participate in the SP study. Of these 46 facilities, 42 facilities were sampled to receive up to 4 SP cases each (one each of asthma, diarrhea, angina and TB). Fourteen of these facilities were public facilities, 28 were private, 5 were operated by faith-based organizations and 4 were operated by social franchise

operations. According to the government data, 61.5 percent of all 858 facilities operating in the neighborhoods of Nairobi are public facilities, therefore this study under-sampled public facilities (33 percent). From the maximum of 168 interactions possible, 166 were completed (98.5 percent).

Public facilities tend to be larger than the private ones (there were 4 staff members in public facilities versus 2.6 in private facilities). The average daily number of patients is also much higher in public than private, 113 vs. 22. In Kenya, facilities are also designated levels depending on their size and technical capacity. While Level 2 and Level 3 facilities are smaller facilities offering basic primary care and preventative services, Level 4 and 5 facilities offer integrated care and inpatient services as well. The data collected in this study excludes Level 4 and Level 5 facilities.

7.0.7 Selection and Design of Cases

For the Madhya Pradesh, Birhbum and Kenya studies, cases were selected because of their relevance to the local context, and also because the respective regulatory bodies (National Rural Health Mission in India and XX in Kenya) have established protocols for the triage, management and treatment of those cases in the public primary care sector, suggesting clear guidelines for patients presenting with those symptoms. In India, the incidence of cardiovascular and respiratory diseases has been increasing, and diarrheal diseases kill more than 200,000 children per year (Black et al., 2010, Patel et al., 2010). Similarly in Kenya, diarrheal diseases is the third leading cause of deaths in children under 5 years, followed by lower respiratory tract infections (IHME, 2010).

In the three urban settings of India, the studies focused on only Tuberculosis, as the goal of the studies were to obtain representative measures of quality of Tuberculosis care at the city level. India has the highest incidence of TB compared to all other countries: it accounts for over 25 percent of an estimated 10.4 million new TB cases worldwide each year, and nearly a third of the 1.7 million annual TB deaths (WHO, 2017). Similarly, the study in China built upon previous studies using other tracer conditions to assess the quality of care available for TB in rural areas (Sylvia et al., 2010). Although China has substantially prevalence of TB from 170 per 100,000 people in 1990 to 59 per 100,000 people in 2010, China in only behind in India in terms of disease-burden. Moreover, prevalence rate in rural areas – the focus of the study – is three times the national average (Li et al., 2013).

Details of the six cases used across the seven sites are presented in Table A4 and Table A5. Table A4 provides a short description of the cases and the opening statements made and information provided by the SPs to the providers. Responses to all cases were standardized such that appropriate history taking and examinations on the part of the provider would unambiguously lead to the unique diagnosis associated with the case. The list of history questions and examinations were developed in consultation with an international and local panel of qualified medical providers. These were supplemented with extensive pilots to enumerate questions providers may ask patients which may be specific to local contexts. Responses to all such questions were standardized and included in SP training modules and exit interviews. The list of associated history questions and examinations, as well as recommended correct treatments are available in Table A5.

8 Appendix B: Coding of Treatment and Robustness to Alternative Definitions

8.0.1 Coding of Treatment

We partition treatment outcomes in SPs and Vignettes cases into mutually exclusive bins of correct case management, over-treatment and incorrect case management by examining outcomes of each interaction to confirm if providers recommended any or all components of the correct treatment vectors described in Appendix A and Table A4 above. Such an exercise requires researchers to make some judgment calls which could lead to varying results. In this section we describe the decisions made, and present results using alternative definition of correct treatments.

The first complication arises from the treatment of referrals to other providers or higher levels of care. In all three cases used in India (both Madhya Pradesh and Birbhum), while the Indian Government's National Rural Health Mission (NRHM) has stipulated guidelines for the triage, management, and treatment protocols at primary care level public clinics, providers may or may not refer these cases to higher levels of care. The conundrum is therefore if we should consider referrals alone (without accompanying correct medicine or treatment) as correct case management. Take the case of diarrhea for example, where ORS and a prescription for Zinc is the WHO recommended course of action. Nevertheless, if the provider refers without any medication, it is impossible to discern if the provider (erroneously) believed that the condition in the child may have been serious or they did so because they were simply unable to diagnose (or were too lazy to put in effort). One option would have been to use the diagnoses given to ask if providers referred only when they correctly diagnosed the case. But in these settings, a diagnosis is provided in less than one-third of the observations, rendering this option infeasible (while we do have diagnosis for vignettes, because providers were required to give one). Comparing referral rates in vignettes and SPs provides help shed some light on which types of providers decide whether to refer or not. If providers were indeed referring patients because they thought these were serious cases, we would have expected to higher rates of referrals among more competent providers. In Madhya Pradesh, while overall referral rates in vignettes and SPs are 37 percent and 18 percent respectively, there are no differences in referral rates conditional on provider competence. Providers below and above median use of checklist refer patients with similar frequencies: 34.6 percent versus 38.7 percent in vignettes and 17.6 percent versus 18.4 percent in SPs. The patterns are remarkably similar across cases.

The second complication arises because many rural providers in our sample often dispense medicines from their own clinic and some of these medicines may be unlabeled or unidentifiable (we collectively called these unlabeled). This naturally poses a problem both in the partitioning of treatment outcome and costs (we discuss costs in the next section). Providers gave at least one unlabeled medication in 24.8 percent of the 939 cases in Madhya Pradesh and 9.6 percent of the 790 cases in Birbhum. Nevertheless, most providers who gave unlabeled medication often did so in conjunction with other labeled medications. In only 5.2 percent of interactions in Madhya Pradesh and 1.5 percent of interactions in Birbhum, providers gave only unlabeled medications.

Providers gave at least one unlabeled medications (all unlabeled medications) in 67.6 percent (11.6 percent), 54.9 percent (20.5 percent) and 8.3 percent (0.0 percent) of interactions in Delhi, Mumbai and Patna, respectively. Finally, the figure for Kenya are 9.0 percent for at least one unlabeled medication and 0.6 percent for all medications.

While this could potentially bias the results, the data provides some evidence that unlabeled medications are more likely to constitute incorrect treatments rather than correct treatments. First, in Delhi, Mumbai and Patna, only TB cases were administered. The first-line recommended treatment for TB consist of a 4-drug therapy for an extended period so unlabeled medicines are unlikely to represent correct treatments. Second, in Madhya Pradesh, where we have variation in provider qualifications, we observe that MBBS providers are 25 percentage points less likely and providers with alternative qualifications are 2.7 percentage points less likely to give unlabeled medications than providers without any qualifications. Finally, in Madhya Pradesh, Birbhum and Delhi where we have vignettes data available, we also find that the likelihood of dispensing unlabeled medications decreases with provider competence as measured by vignettes. These patterns provide strong evidence in support of treating unlabeled medicines as not correct treatments.

Our preferred definition of correct case management therefore (a) does not consider referrals only as correct treatments, and (b) considers all unlabeled medications as unnecessary. Nevertheless, below, we also present results with alternative definitions of correct treatment. Details of correct treatments for each case are available in Table A4. In asthma cases, a recommendation for any of the following – bronchodilators, theophylline, inhaled or oral corticosteroids, leukotriene inhibitors, cromones, and inhaled anticholinergics – constitute correct treatment. For unstable angina, any of an aspirin, clopidogrel, anti-platelet agents, recommendation for an ECG or referral to higher level of care is considered correct treatment. Correct treatment for TB varies depending on the variant of the condition presented, but includes sputum testing, chest x-ray, DST, CXR, referral to public DOTS center or appropriate provider or specialist.

8.0.2 Robustness to alternate measures including referrals

Given the complications in coding treatment due to referrals and unlabeled medications, we use two alternative definitions of correct treatment. The first alternative definition treats all referrals as correct case management. This however, does not make a difference to how we separate correct case management from over-treatment. For example, if a provider only referred and did not do anything else, this would reclassify the case from incorrect to correct. Instead if the provider had referred and given an unnecessary medication, the interaction will be categorized as overtreatment. The second alternative definition treats all cases where providers gave only unlabeled medications as correct case management. Here we categorize all such instances as correct treatments only. This in practice, increases the proportion of correct case management while leaving the proportion of overtreatment the same. Results are available in Table A5. As expected, the coding of referrals alone as correct case management increases the proportion of cases with any correct treatment across all samples. In Madhya Pradesh, Mumbai, Patna, China and Kenya, the increases are

modest (ranging from 1 percent to 42 percent). In Birbhum and Delhi, the rate of any correct treatment doubles. Still, the fraction of cases with any correct case management reaches only 52 percent in Birbhum (from 24 percent) and 21 percent in Delhi (from 11 percent). The proportion of cases with correct treatment only increases 1.5 percent to 11 percent in Birbhum and from 0.8 percent to 9 percent in Delhi. Most cases get reclassified as over-treatment types, i.e. providers gave some unnecessary medications and referred. The fraction of overtreatment cases reaches 41 percent (from 22 percent) in Birbhum and 13 percent (from 10 percent in Delhi).

Reclassifying interactions where providers gave only unlabeled medications increases the fraction of correct treatment in all settings (except China where we do not have this data available). The biggest gains are observed in Delhi and Mumbai (11 percentage points and 18 percentage points respectively), where large proportions of providers gave only unlabeled medications. While these adjustments on our categorizations of case outcomes likely represent upper bounds on the true estimates, the rate of any correct treatment remains far below from unity.

9 Appendix C: Coding of Costs and Robustness to Alternative Definitions

9.0.1 How we do this

We apportion costs of each interaction into necessary and avoidable components following the partitioning of treatments into mutually exclusive categories of correct case management, over-treatment and undertreatment. For interactions with correct case management and undertreatment as final outcomes, this follows naturally. We assume that when providers correctly managed cases (i.e. gave only the necessarily treatments and no additional medicines), both consultation and medicines costs are necessary. Conversely, we assume both consultation costs and medicines costs to be unnecessary and avoidable when providers incorrectly treat patients. Finally, when providers overtreat patients, we assume that consultation costs are necessary, costs arising from medically indicated pharmaceuticals are necessary, and the costs arising from medicines deemed unnecessary are avoidable. We compute necessary and avoidable components of costs in such cases by estimating prices of all medicine recommended in the interactions.

There could be several concerns associated with our approach. First, our approach does not allow us to make distinctions between quantities and/or qualities of correct treatment. For instance, in an asthma case, if a provider chooses to recommend both oral and inhaled corticosteroids when one would have sufficed, our approach considers costs of both as necessary. We do this because there it is impossible if providers gave the "right amount" of treatment. We attempted to do this by asking a panel of three qualified doctors to independently rate treatment outcomes in Birbhum, but the doctors failed to arrive at a consensus.

Second, our approach does not make distinctions between palliative treatments or treatments that could provide symptomatic relief and medically harmful treatments. For instance, a patient with TB may feel temporary relief from the ingestion of acetaminophens (which reduces fever) but

not antibiotics (which could be harmful in the long-term). While patients may derive utility from palliative treatments, from a medical point of view, they are still deemed unnecessary. Relatedly, it may be problematic to categorize consultation costs in cases with undertreatment as the outcome as unnecessary. If diagnosis and treatment could be decoupled into two-stage transactions, diagnosis costs would become truly unavoidable in all instances. However, in the settings we study, diagnosis and treatments are often bundled together – providers dispense medicines from their own or from a pharmacy attached to the clinics. Therefore, the costs that we report and discuss in the paper need to be viewed from a medical perspective, rather than a utilitarian perspective.

Given our definitions of necessary and avoidable costs and the study contexts, we need to partition total costs in consultation costs and costs of medicines whenever we observe overtreatment. Two types of complications arise in doing so. The first is because many providers in our settings dispense medicines from their own rather than prescribing them for purchase from a pharmacy. When they do so, we do not observe prices of individual medicines. In the SP interactions across all study sites, providers dispensed medicines in 42 percent of the interactions in Madhya Pradesh, 81 percent of interactions in Birbhum, XX percent in Delhi, 58 percent in Mumbai, 21 percent in Patna and 61 percent in Kenya (we do not have this data available for China). Pricing individual medicines becomes difficult when providers dispense medicines because they often charge below market prices. Anecdotal evidence suggests that this happens for two reasons – providers often receive free samples for medical representatives which they sell to patients, and providers purchase medicines from wholesalers and distributors at prices below suggested retail prices. The second complication arises due to unlabeled and unidentifiable medicines (as discussed in Appendix B). Here we simply do not have a way to price such medicines. We use market rates of individual medicines to impute unit prices of medicines recommended by providers. As an example, across the Madhya Pradesh samples, providers recommended 3,104 unique medicines. Through surveys with chemists and desk research, we were able to obtain prices for 82.5 percent of these medicines. We impute prices of individual medicines by taking the average costs of medicines belonging to the same drug class (analgesics, anti-allergy, anti-ulcer, antibiotics, ayurvedic, cardiac, homeopathic, household remedies, psychiatric/neural, steroids, other type). For unlabeled medicines, we assume prices to be zero. Note that this potentially understates the true fraction of avoidable costs. Since unlabeled medicines are categorized as unnecessary treatments, assigning zero prices inflates consultation costs. In turn, consultation costs get categorized as necessary costs in cases with overtreatment. Because market rates for medicines are different from costs to the provider, using market rates to parse out medicine costs from consultation costs also pushes consultation costs below zero in some interactions (1.88 percent of interactions in Madhya Pradesh and XX percent of interactions in Birbhum). In these instances, we assume consultation costs to be zero and normalize costs of individual medicines by total fees charged by the provider. The implicit assumption is that providers derive income from markups they charge for medicines. This potentially overstates avoidable costs since consultation costs are now built into medicine costs. In our final data set, consultation costs are estimated to be zero in 33 percent of interactions in Madhya Pradesh, 13

percent in Birbhum, 10 percent in Delhi, 7 percent in Mumbai, 29 percent in Patna and 31 percent in Kenya.

9.0.2 Bounds on Avoidable Costs

Given the issues discussed with alternative definitions of correct treatment (those that include referrals, those that include cases where providers only give unlabeled medications) and the problem of pricing unlabeled medications, we present two alternative estimates of the fraction of avoidable costs. In the first measure, we code referrals as correct treatment. In the second measure, we code cases where providers gave only unlabeled medicines as correct treatment. Thus, both measures lower the fraction of cases with incorrect treatments and the fraction of costs that is avoidable.

Results are available in Table A9. Coding referrals as correct treatments lowers the fraction of avoidable costs modestly. In Madhya Pradesh, the fraction of avoidable costs decreases by 6 percentage points (from 70 percent to 64 percent). In Birbhum, the change is 6 percentage points (from 86 percent to 80 percent). The figures are 1.2 percentage points for Delhi (from 80.5 percent to 79.2 percent), 3.7 percentage points for Mumbai (from 76.3 percent to 72.7 percent) and 0.8 percentage points for Patna (from 77.1 percent to 76.2 percent).

Similarly, coding cases where providers gave only unlabeled and unidentifiable medicines as correct treatment also lowers avoidable costs modestly (Table A9, Panel B). In Madhya Pradesh and Birbhum, the fraction of avoidable costs lowers by 1.5 and 1.7 percentage points respectively. Larger reductions are seen in Delhi (11 percentage points) and Mumbai (19 percentage points) where a large fraction of providers give unlabeled medicines.

9.0.3 Treatment of Lab Costs

In the TB studies undertaken in Mumbai and Patna, SPs also collected prices of laboratory tests recommended by providers (although they did not get tested). SPs were advised to do laboratory tests in 38 percent of interactions in Mumbai and 46 percent of interactions in Patna. The average costs of laboratory exams (conditional on receiving a recommendation) are USD 9.38 and USD 9.83 respectively. Because SPs did not follow-up with providers with results of laboratory tests they recommended, we do not know if and how providers would have changed their recommendations for treatment upon receiving new information. Note however that, (a) SPs were designed to present lab results in two of the four variants of case presentations (TB2 naive+ and TB3 AFB+ve), (b) laboratory tests were not necessary for providers to diagnose the TB cases and recommend appropriate actions. In our main estimation of necessary and avoidable costs, we exclude laboratory costs. Inclusion of laboratory costs increases the fraction of avoidable costs from 76.3 percent to 78.4 percent in Mumbai and from 77.1 percent to 83.3 percent in Patna.

9.0.4 Adding Salary Costs of Public Sector Providers

In our estimation of necessary and avoidable costs, so far, we have only considered the "out-of-pocket" costs. This makes sense for private providers who operate on a fee-for-service basis, but public providers in our sample (in Madhya Pradesh, China and Kenya) receive salaries instead of consultation fees. While we do not have salary data for providers in China and Kenya, we can estimate necessary and avoidable costs incorporating salary costs in Madhya Pradesh. These estimates would then represent "cost to the taxpayer" rather than "out-of-pocket" costs. Since per patient costs in the public sector in Madhya Pradesh was 5 times higher than in the private sector (Das et al., 2016), the fraction of avoidable costs is higher when salaries are considered. The average cost of an interaction increases by USD 1.672 (from USD 0.646 to USD 2.318), and the fraction of avoidable costs increases by XX percent.

10 Appendix D: Measurement Error Correction in Know-Do Gap

In this section, we describe the measurement error in the Know-Do estimation and discusses feasibility of available methods for correction.

Assume the true model is given by: $y^* = \alpha + \beta \cdot x^* + \epsilon$, where y^* is the SP outcome and x^* is the vignette outcome. Since both are binary deterministic variables, that is $y^*,\ x^* \in \{0, 1\}$, providers either "know for sure" or they "do not know for sure." This is like education attainment reports in the labor literature, e.g. a respondent either completed high school or they did not.

Now instead, of observing y^* and x^*, we observe error ridden measures y and x, such that:

$$y = y^* + \nu$$

$$x = x^* + \mu$$

We assume no error in the LHS (because the provider sees a real patient) and error in RHS (because the provider sees a hypothetical patient) and focus on LHS. As such, the measurement error is non-classical by design, that is, $\sigma_{x^*,\ \mu} \neq 0$. Observe that $x^* = 1 \Rightarrow \mu = x - x^* \leq 0$ and $x^* = 0 \Rightarrow \mu = x - x^* \geq 0$, and thus $\sigma_{x^*,\ \mu} < 0$. The measurement error is non-classical and mean-reverting.

In an OLS estimation, instead of estimating $y^* = \alpha + \beta \cdot x^* + \epsilon$, we estimate:

$$y = \alpha + \ \beta \cdot (x - \ \mu) + \epsilon = \alpha + \beta \cdot x + (\epsilon - \beta \cdot \mu)$$

This result is that:

$$\widehat{\beta}_{OLS} = \ \frac{\sigma_{y,x}}{\sigma_x^2} = \ \frac{\sigma_{\alpha+\beta\cdot x^*+\ \epsilon+\mu}}{\sigma^2_{x^*+\mu}} = \ \frac{\beta \cdot \left(\sigma^2_{x^*} + \ \sigma_{x^*,\mu}\right) + \ \sigma_{x^*,\epsilon} + \sigma_\epsilon}{\left(\sigma^2_{x^*} + \ \sigma_{x^*,\mu}\right) + \left(\sigma^2_\mu + \ \sigma_{x^*,\mu}\right)}$$

By construction, $\sigma_{x^*,\epsilon} = 0$ and assuming the measurement error in x is uncorrelated with the

error term in the main equation (i.e. $\sigma_{\mu,\epsilon} = 0$), we have:

$$\widehat{\beta}_{OLS} = \frac{\beta \cdot \left(\sigma^2_{x^*} + \sigma_{x^*,\mu}\right)}{\left(\sigma^2_{x^*} + \sigma_{x^*,\mu}\right) + \left(\sigma^2_{\mu} + \sigma_{x^*,\mu}\right)}$$

Since, $\sigma^2_{\mu} + \sigma_{x^*,\mu} \geq 0$, we have $\widehat{\beta}_{OLS} < \beta$. See Brown, Bound and Matheowitz (1996) for further discussion.

One could estimate the size of the measurement error and correct for it if the true fraction of x^* were known. Let $\pi = Pr(x^* = 1)$ be the true proportion of providers who know. Define $\pi_{01} = \Pr(x = 0 \mid x^* = 1)$ be the fraction of false negatives and $\pi_{10} = \Pr(x = 1 \mid x^* = 0)$ be the fraction of false positives. Bound, Bound and Matheowitz (1996) show that:

$$\beta_{yx} = \left[1 - \frac{\pi_{01} \cdot \pi}{\pi_{01} \cdot \pi + (1 - \pi_{10}) \cdot (1 - \pi)} - \frac{\pi_{10} \cdot (1 - \pi)}{\pi_{10} \cdot (1 - \pi) + (1 - \pi_{01}) \cdot \pi}\right]$$

If we had estimates of π, π_{01} and π_{10}, we could correct for measurement error directly. However, in practice we often do not know these quantities. Later we will discuss how these can be estimated in the GMM framework.

10.1 Improving OLS bounds with two reports

The OLS is clearly biased downwards and represents a lower bound. Bollinger (1996) shows that the OLS lower bound can be improved when two error ridden reports of x^* such that: $x_1 = x^* + \mu_1$ and $x_2 = x^* + \mu_2$, are available. One can instead run the regression:

$$y = \alpha + \beta_{01} I\left(x_1 = 0,\ x_2 = 1\right) + \beta_{10} I\left(x_1 = 1,\ x_2 = 0\right) + \beta_{11} I\left(x_1 = 1,\ x_2 = 1\right) + \epsilon$$

Then we have $\beta_{yx} \leq \beta_{11} \leq \beta$. The intuition here is that we can compare the mean differences between observations where both reports were incorrect with observations where both reports were correct. Presumably, these contain fewer misclassified observations and the estimated coefficient is closer to the true β.

10.2 Instrumental Variable estimation with two reports

In the case where the true variables are binary, using one error-ridden measure to instrument for the other does not yield unbiased estimates. Without loss of generality, suppose we use x_2 to instrument for x_1. The IV estimator is given by:

$$\beta_{iv} = \frac{\sigma_{y,\ x_2}}{\sigma_{x_1,x_2}}$$

Which can be expanded and rewritten as:

$$\beta_{iv} = \frac{\sigma_{\beta \cdot x* + \epsilon,\ \ x^* + \mu_2}}{\sigma^2_{x^*} + \sigma_{x^*,\ \mu_2} + \sigma_{x^*,\ \mu_1} + + \sigma_{\mu_1,\ \mu_2}}$$

$$= \frac{\beta\left[\sigma^2_{x^*} + \sigma_{x^*,\ \mu_2}\right] + \sigma_{\epsilon,\mu_2}}{\left[\sigma^2_{x^*} + \sigma_{x^*,\ \mu_2}\right] + \sigma_{x^*,\mu_1} + \sigma_{\mu_1,\mu_2}}$$

For $\widehat{\beta}_{iv}$ to be unbiased, we need that $\sigma_{x^*,\mu_1} = \sigma_{\mu_1,\mu_2} = \sigma_{\epsilon,\mu_2} = 0$. Even if we assume that $\sigma_{\mu_1,\mu_2} = 0$ (measurement error in two reports is uncorrelated) and $\sigma_{\epsilon,\mu_2} = 0$ (measurement error in the second report is uncorrelated with the error term in the main equation), $\sigma_{x^*,\mu_1} < 0$ and $\widehat{\beta}_{iv}$ will be biased upwards.

The same argument extends to using the IRT score (z) as an instrument. Assume z is highly correlated with x^*, such that $\sigma_{x^*,\ z} > 0$, and uncorrelated with the error term (ϵ) such that $\sigma_{\epsilon,\ z} > 0$. We can show the usual IV coefficient to be:

$$\beta_{iv} = \frac{\sigma_{y,z}}{\sigma_{x,z}} = \frac{\sigma_{\beta\cdot x^* + \epsilon,\ z}}{\sigma_{x^* + \ \mu,\ z}} = \beta \cdot \left[\frac{\sigma_{x^*,\ z}}{\sigma_{x^*,\ z} + \sigma_{\mu,z}}\right]$$

Again, for $\widehat{\beta}_{iv}$to be identified, we need $\sigma_{\mu,z} = 0$. However, because $\sigma_{x^*,\mu} \leq 0$ and $\sigma_{x^*,z} > 0$, we have $\sigma_{\mu,z} \leq 0$. The term in the parenthesis is less than unity, and the IV estimate will still be biased upward.

10.3 GMM Estimation to Correct for Measurement Error

11 Appendix E: The Provider - Patient Game

11.1 Some useful notation and definitions

Let the contemporaneous (as against life-time) value (or net cost) of going to a doctor who is type H with probability π_t *at the patient's decision point* in period $t = 1, 2$ be represented by $V(\pi_t, \tau_s)$ if the high level of effort is chosen and $V(\pi_t, \tau_m)$ if the low level of effort is chosen. From above,

$$V(\pi_t, \tau_s) = -p[\pi_t(1 - p_{mH}) + (1 - \pi_t)(1 - p_{mL})]C_m - (1 - p)[\pi_t(1 - p_{sH}) + (1 - \pi_t)(1 - p_{sL})]C_s$$

and

$$V(\pi_t, \tau_m) = -p[\pi_t(1 - p_{mH}) + (1 - \pi_t)(1 - p_{mL})]C_m - (1 - p)C_s$$

We now impose a set of conditions on the $V(\bullet, \bullet)$ function. The first assumption is that $V(\pi, \tau_s) - \tau_s < \underline{V}$, which says that getting a mean-quality doctor to put in high effort is dominated by the social surplus of the outside option. The second assumption is that $V(\pi, \tau_m) - \tau_m > \underline{V}$, so a mean-quality doctor putting in low effort is better than the social surplus of the outside option in that period.

Next, define π^* to be the value of π for which the social surplus is the same regardless of whether the effort invested is high or low. That is to say

$$V(\pi^*, \tau_m) - \tau_m = V(\pi^*, \tau_s) - \tau_s.$$

$$\tau_s - \tau_m = (1-p)[1-\pi(1-p_{sH})-(1-\pi)(1-p_{sL})]C_s.$$

for $\pi = \pi^*$ Note that since $p_{sH} > p_{sL}$, the expression for $V(\pi,\tau_s)-\tau_s - V(\pi,\tau_m)-\tau_m$ is increasing in π. For a probability of encountering an H type doctor $\pi > \pi^*$, the high-effort option therefore yields a higher social surplus. Likewise, define $\underline{\pi}$ to be the value of π for which exerting low effort yields the same social surplus as the outside option:

$$V(\underline{\pi},\tau_m) - \tau_m = \underline{V}.$$

For $\pi < \underline{\pi}$, the patient's outside option yields a higher social surplus.

We now define the second period probabilities (according to the patient) of having encountered an H type doctor in the first period *if the patient was cured*, where $\pi_{cm}(\pi_1)$ = Prob[Doctor type is H — disease in period one was cured, low effort] and $\pi_{cs}(\pi_1)$ = Prob[Doctor type is H — disease in period one was cured, high effort]. From our definitions,

$$\pi_{cm}(\pi_1) = \frac{\pi_1 p_{mH}}{\pi_1 p_{mH} + (1-\pi_1)p_{mL}}$$

$$\pi_{cs}(\pi_1) = \frac{\pi_1[pp_{mH} + (1-p)p_{sH}]}{\pi_1[pp_{mH} + (1-p)p_{sH}] + (1-\pi_1)[pp_{mL} + (1-p)p_{sH}]}.$$

We make the additional simplifying assumption that $\frac{p_{mH}}{p_{sH}} = \frac{p_{mL}}{p_{sL}}$. Under this assumption, $\pi_{cs} = \pi_{cm}$, which means that the patient does not learn more (or less) about the doctor's type if the doctor chooses high effort (and the patient is cured).[31]

If the patient was not cured in the first period, we likewise define $\pi_{nm}(\pi_1)$ = Prob[Doctor type is H — disease in period one was not cured, low effort] and $\pi_{ns}(\pi_1)$ = Prob[Doctor type is H — disease in period one was not cured, high effort]. From our definitions,

$$\pi_{nm}(\pi_1) = \frac{\pi_1[p(1-p_{mH})+1-p]}{\pi_1[(1-p_{mH})+1-p] + (1-\pi_1)[(1-p_{mL})+1-p]}$$

$$\pi_{ns}(\pi_1) = \frac{\pi_1[p(1-p_{mH})+(1-p)(1-p_{sH})]}{\pi_1[p(1-p_{mH})+(1-p)(1-p_{sH})] + (1-\pi_1)[p(1-p_{mL})+(1-p)(1-p_{sL})]}$$

It is easy to check that our assumption that $\frac{p_{mH}}{p_{sH}} = \frac{p_{mL}}{p_{sL}}$ implies that $\pi_{nm}(\pi_1) > \pi_{ns}(\pi_1)$. Since we also have $\pi_{cm}(\pi_1) = \pi_{cs}(\pi_1)$, this tells us that putting in a higher level of effort for the sake of increasing the informativeness of the outcome (cure, no cure) is never worthwhile. Reducing effort does reduce the informativeness of the no cure outcome–but we will make assumptions that imply that if there is no cure the patient does not return to that practitioner. Hence it also does not pay to strategically reduce effort. This helps simplify our analysis below.

Finally, assume that $0 < \underline{\pi} < \pi < \pi^* < 1$. In addition to the assumptions already made, says that a known L type is worse than taking the outside option, and a known H type dominates

[31] This rules out the possibility that it may be useful to choose (say) high effort because at high effort there is greater separation between the types. It is however possible to easily relax this assumption without changing the spirit of our results

the outside option Further assume that it is socially efficient to ask for higher effort in the second period from a doctor who started with the mean type but cured you in the first period ($\pi_{cs}(\pi) = \pi_{cm}(\pi) > \pi^*$); but going to a doctor who started with the mean type but failed to cure in the first period disease is worse than taking the outside option ($\pi_{ns}(\pi) < \underline{\pi}$ and $\pi_{nm}(\pi) < \underline{\pi}$).

11.2 Characterization of Equilibria

11.2.1 Period 2

Consider the second period outcome in any sequential equilibrium starting with a belief π', $\underline{\pi} > \pi'$. Then the only way a patient could be persuaded to see the practitioner is if the fee (f) is such that

$$V(\pi', \tau_m) - f \geq \underline{V}.$$

. But this implies that

$$f - \tau_m \leq V(\pi', \tau_m) - \underline{V} - \tau_m < 0$$

since $\underline{\pi} > \pi'$. But then the practitioner is better off not treating the patient and getting a payoff of zero. On the other hand as long as

$$V(\pi', \tau_m) - f > \underline{V}.$$

the patient will always see the practitioner.

As long as $\underline{\pi} \leq \pi'$ at the start of the second period, there can be second period outcome in a sequential equilibrium where the patient gets treated. Note however there is a continuum of possible Nash Equilibria outcomes with different levels of fees. For example, it is a Nash Equilibrium for both types to announce *any* pair (f, τ) such that

$$V(\pi', \tau) - f \geq \underline{V}.$$

and $f - \tau \geq 0$. This equilibrium is sustained by the belief that anyone who deviates is an L type, which is what, for example, prevents a profitable deviation where both types announce the same higher f that is still consistent with

$$V(\pi', \tau) - f \geq \underline{V}.$$

. However given that in the second period the two types have identical incentives we feel that the very pessimistic belief that support the proposed equilibrium are hard to justify.

For the same reason (the fact that the two types have identical incentives) the two types must earn the same period 2 payoffs in any separating or semi-separating second period outcome. Otherwise the type that was earning less would simply switch to the other option. Suppose there are two distinct contracts that get offered in equilibrium (there may be more) represented by (f_1, τ_1) and (f_2, τ_2). Let the beliefs of the patient associated with these two offered contracts be π_1 and π_2

and wlog let $\pi_1 \geq \pi_2$. From the fact that the two contracts must produce the same payoff for the practitioner, we know that if $f_1 \neq f_2$, $\tau_1 \neq \tau_2$.

Given second period belief π' about the type of the practitioner, define the revenue maximizing contract (from the point of view of the practitioner) $C(\pi) = (f(\pi), \tau(\pi))$ to be the contract that sets $f = max_\tau V(\pi', \tau) - \tau$ and $\tau = argmax_\tau V(\pi', \tau) - \tau$. We will now argue that it is not possible that $(f_1, \tau_1) = C(\pi_1$ and $(f_2, \tau_2) = C(\pi_2)$ unless $\pi_1 = \pi_2$.To see this note that from above it must be the case that $V(\pi_1, \tau_1) - \tau_1 = V(\pi_2, \tau_2) - \tau_2$. But since $V(\pi, \tau)$ is strictly increasing in π for fixed τ, and $V(\pi_1, \tau_1) - \tau_1 \geq V(\pi_1, \tau_2) - \tau_2 > V(\pi_2, \tau_1) - \tau_1$, we have a contradiction unless $\pi_1 = \pi_2$.

In fact this argument tells us that if there are at least two contracts proposed in equilibrium the one associated with the higher value of π cannot have a revenue maximizing contract.

Of course if $\pi_1 = \pi_2$, then for a generic π_1, $C(\pi_1) = C(\pi_2)$ and therefore this case does not arise. The one exception is when $\pi_1 = \pi_2 = \pi^*$ so that the two levels of effort are both optimal. We will treat this as a case of revenue-maximizing pooling, since there is no additional separation in period 2.

Now consider any equilibrium where at least one proposed contract is not revenue maximizing and suppose a practitioner offering that particular contract switches to the corresponding revenue maximizing contract. This would clearly make the practitioner better off, Clearly to prevent this, the patient needs to place a higher probability of it coming from an L type than the original contract. But both types have exactly the same incentive to deviate, so such a belief seems hard to justify.[32] Therefore we will rule out such beliefs and make the assumption that in the second period the only possible pooling equilibrium outcome is that where both types propose revenue maximizing pooling contracts. Moreover we assume that this equilibrium selection rule applies *in and out of equilibrium*–if there is a deviation in the first period and therefore we reach the second period with the patient having a belief π' then the second period equilibrium will have payoffs given by the contract $C(\pi')$. It is worth also stating that our main result does not depend on this exact assumption. Indeed, it should become apparent that all it needs is: (a) that the second period contract is a pooling contract (which is natural, given that the two types have identical incentives); (b) that the second period contract only depends on the patient's second period beliefs (and, for example, not on the outcome of the treatment in period 1, conditional on those second period beliefs); and (c) that if the patient's second period belief becomes more favorable to the practitioner being an H type ($\pi' goesup$) then the payoff to the practitioner associated with the second period equilibrium goes up.[33] From now on we take this as given and try to characterize the period 1 outcome.

11.2.2 Period 1: Separating equilibria

Suppose there is sequential equilibrium of the above game where in the first period there is a non-zero probability that the L type proposes a contract that the H will never offer and vice versa. We

[32]It fails, for example, the criterion of Divinity, as proposed by Banks and Sobel (1987).

[33]This extends to the case where the patient randomizes in period 2

describe this as an equilibrium with the possibility of full separation.

In such an equilibrium type L (irrespective of which message she actually sends) gets zero in both periods since as soon as she is identified to be a type L no patient visits her. As a result this is also the maximum utility he could get by choosing the contract that in equilibrium is chosen *only* by a H type. However a patient who receives that contract in period 1 must assign probability one to the provider being type H at the beginning of the second period *irrespective of the whether the disease in period 1 gets cured or not.* This is because conditional on knowing that the contract was offered by a type H the outcome of the first period offers no additional information. It follows that the two types must have identical payoffs in both periods in this equilibrium since the only wedge between them comes from their differential ability to cure the disease in period 1, but, as just argued, that information from that outcome is ignored in this particular case.

Given $\pi_{cm}(\pi_1) = \pi_{cs}(\pi_1) = \pi_{nm}(\pi_1) = \pi_{ns}(\pi_1) = 1$, for $\pi_1 = 1$, there is an unique optimal second period choice for the type H doctor in any sequential equilibrium , which is to announce $A = H$, choose

$$f_2 = V(1, \tau_s) - \underline{V}$$

and

$$\tau = \tau_s.$$

Given that two types have identical payoffs in this potential equilibrium, the H type must also get zero over the two periods. This is only possible if the separating contract for period 1 proposed by the type H, (H, f_1, τ), is such that

$$f_1 - \tau + \delta[V(1, \tau_s) - \underline{V} - \tau_s = 0.$$

This is not possible if $-\tau_s$ which is the lower bound of $f_1 - \tau$ is less than $\delta[V(1, \tau_s) - \underline{V} - \tau_s$. Therefore we have

Result 1: There is an sequential equilibrium with the possibility of full separation where some of the type Hs choose a first period contract (H, f_1, τ) which L types do not choose, and some of the type Ls choose a first period contract (L, f_1', τ') that no H type chooses if and only if Condition * holds:

$$-\tau + \delta[V(1, \tau_s) - \underline{V} - \tau_s \leq 0.$$

However even when the above condition holds, there is a problem with this equilibrium.

11.2.3 Period 1: Partial Pooling Equilibria A

One class equilibria that may exist is what we call Partial Pooling A. In this class of equilibria the type L chooses to separate in period 1 with some probability strictly less than 1, but pools with the H type with the remaining probability and the H type always pools. In this equilibrium the type L gets a payoff of zero over the two periods, but the type H gets a strictly positive payoff.

To see how such such an equilibrium can exist, as long as the fraction Ls is positive, $\pi_{cm}(\pi_1) >$

$\pi_{nm}(\pi_1)$ and $= \pi_{cs}(\pi_1) > \pi_{ns}(\pi_1)$ and as result, patients will be willing to pay more for providers who have effected a cure in the previous period. As a result, the fact that the L types have a lower probability of a cure now implies that the L types have a lower expected payoff from mimicking the H type than the H type itself. That wedge means that we can have an equilibrium where the L types get exactly zero in both periods, but the H types do not.

However it can be shown that as long Condition * holds:

$$-\tau + \delta[V(1, \tau_s) - \underline{V} - \tau_s > 0.$$

in any sequential equilibria of this class the fraction of L types that pool with H types (and therefore treat patients in equilibrium) must be bounded away from zero. In other words, there is no partial pooling equilibrium type A which is arbitrarily close to a fully separating equilibrium.

To see why, consider the case where the fraction of Ls choosing the same outcome as the H is α which is close to zero. In this case, $\pi_{cm}(\pi_1) \approx \pi_{nm}(\pi_1)$ and $\pi_{cs}(\pi_1) \approx \pi_{ns}(\pi_1)$. Therefore, in this case, the amount a patient is willing to pay someone who cured them in the previous period will be very similar to what they will accept to pay someone who failed to cure them. Hence the expected two period payoff of a H type in such a equilibrium has to be very close to the expected two period payoff of a L type who adopts the strategy of an H type. Therefore if L type earns zero in this equilibrium the two-period expected payoff to the H type must be close to the zero when α is close to zero, and converges to zero when α converges to zero.

This is where Condition * comes in. As in the argument for Result 1, the H type's two-period payoff is still bounded below by a number that converges to

$$-\tau + \delta[V(1, \tau_s) - \underline{V} - \tau_s$$

when α goes to zero. This contradicts the condition, derived above, that the two-period expected payoff to the H type must be close to the zero when α is close to zero.

Result 2: There is an $\alpha^* > 0$ such that in any Partial Pooling Equilibrium A the fraction of L types who pool with a type H is not less than α^*.

11.2.4 Period 1: Partial Pooling Equilibrium B

There can also be partial pooling equilibria which we call Partial Pooling B, where the H fully separate with a positive probability in period 1, but the L always pools with other H types who are indifferent between separating and pooling.

In such an equilibrium L types must make more than zero in expectation. If not we have the same problem as in the case of Partial Pooling A–that the type H payoff cannot be low enough to discourage switches away from the pooling outcome as long as

$$-\tau + \delta[V(1, \tau_s) - \underline{V} - \tau_s > 0.$$

But the pooling payoff can only be positive if the patient is willing to accept a contract proposed by the provider in period 1 which requires that the beliefs associated with the pooling offer, $\pi \geq \underline{\pi}$. This means that the fraction of H types choosing this offer is bounded away from zero. Result 3: There is an $\alpha^{**} > 0$ such that in any Partial Pooling Equilibrium A the fraction of H types who pool with a type L is not less than α^{**}.

11.2.5 Period 1: Pooling Equilibrium

Consider a Pooling Equilibrium where both types offer the contract f_1, τ_m in period 1, and if the patient is cured, both offer the contract f_2, τ_s, where $f_1 = V(\pi, \tau_m) - \underline{V}$ and $f_2 = V(\pi_{cm}(\pi), \tau_s) - \underline{V}$. If the patient is not cured, no contract gets offered, since none that is acceptable to the practitioner will be taken up.

The standard concern with pooling equilibria is whether one type will benefit by deviating from the equilibrium play and sending a message that is different. So assume that the practitioner deviates and offers a different contract (f_d, τ_d). Let the patient's belief at his/her first period decision point after being offered this contract be denoted by π_{1d}. The patient updates π_{cm} and π_{nm} based on π_{1d}. Denote the updated values by π_{cmd} and π_{nmd}, and note that both of these increase with π_d.

The key observation is that there is a range of values of π_d for which type H benefits more than type L from deviating, but there is also a range in which the reverse is true. This is key to arguing that the pooling equilibrium is robust to the standard "Intuitive Criterion" for beliefs (CITE Kreps and Cho (1987)).

The practitioner's net loss from deviating for $i = H, L$ is

$$\begin{aligned}(f_1 - \tau_1) - (f_d - \tau_d) + \delta\{p_{mi} V(\pi_{cm}(\pi), \tau_s) - \underline{V} - \tau_s\} \\ -\delta p_{mi}\{V(\pi_{cmd}, \tau(\pi_{cmd})) - \underline{V} - \tau(\pi_{cmd})\} \\ -\delta(1 - p_{mi}) max\{0, V(\pi_{nmd}, \tau(\pi_{nmd})) - \underline{V} - \tau(\pi_{nmd})\}\end{aligned}$$

Step 1: The first step is to show that if there is a deviation that type H benefits for some set of values of π_d, there are also values of π_d for which type L benefits. Consider the case where $\pi_d \approx 1$. Then both $\pi_{cmd} \approx 1$ and $\pi_{nmd} \approx 1$. Hence the expression from the loss from deviating can be rewritten as

$$\begin{aligned}(f_1 - \tau_1) - (f_d - \tau_d) + \delta p_{mi}\{V(\pi_{cm}(\pi), \tau_s) - \underline{V} - \tau_s\} \\ -\delta\{V(1, \tau_s) - \underline{V} - \tau_s\}\end{aligned}$$

The net loss is clearly increasing in p_{mi}. When $\pi_d \approx 1$ if H types actually gain by deviating (net loss is negative), L types will gain even more.

Step 2: The previous step establishes that type L cannot be eliminated at this stage and therefore $\pi_d \approx 0$ remains an admissible belief. Given that, it is not a dominant strategy for type H to propose this alternative contract. Therefore under the "Intuitive Criterion" the potential cannot break the pooling equilibrium.