NBER WORKING PAPER SERIES

RANDOMIZED REGULATION: THE IMPACT OF MINIMUM QUALITY STANDARDS ON HEALTH MARKETS

Guadalupe Bedoya
Jishnu Das
Amy Dolinger

Working Paper 31203
http://www.nber.org/papers/w31203

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2022

Extended members of the Kenya Patient Safety Impact Evaluation (KePSIE) project team include Jorge Coarasa, Ana Goicoechea, Njeri Mwara, Khama Rogo, and Frank Wafula from The World Bank; John Kabanya, Charles Kandie, Mary Wangai, and representatives of the Kenya Medical Boards and Councils from the Ministry of Health; and the County Executive Members and Directors for Health of Kakamega, Kilifi, and Meru Counties. We are grateful to Sherlene Chatterji, Benjamin Daniels, Rebecca De Guttry, Thomas Escande, Seungmin Lee, Maria Camila Ayala Guerrero, Garima Sharma, Chex Yu, and Tatiana Zarate for excellent research assistance throughout the project. Rodgers Kegode, Purity Kimuru, Andrew Muriithi, Pheliciah Mwachofi, Salome Omondi, Pamela Kuya, and Leah Adero provided field support. We thank Jay Bhattacharya, Paolo Belli, Mickey Chopra, Daniel Chen, Ana Goicoechea, James Habyarimana, Arianna Legovini, Nikolas Mittag, Edit Velenyi, as well as seminar participants at UCLA, Georgetown University, George Mason University, Harvard School of Public Health, Indian Statistical Institute, Warwick University, University of Washington, University of Maryland, World Bank and the 2020 ASSA meeting for providing valuable comments. Funding was provided by the DIME Impact Evaluation to Development Impact (i2i) fund, the Trade and Competitiveness Impact program (Compel), the Strategic Impact Evaluation Fund (SIEF), the Knowledge for Change Program (KCP), the Korea World Bank Group Partnership Facility (KWPF), the Development Economics Research Support Budget (RSB) and the Development Research Group (DECRG) at The World Bank. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the view of the World Bank, its executive directors, or the countries they represent. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper. This study was pre-registered in the AEA RCT registry with RCT ID AEARCTR-0001001. Ethical clearance was granted by the Ethics and Scientific Review Board at the African Medical and Research Foundation (Approval no. AMREF-ESRC P94/2013), the Kenyan Ministry of Health and authorities at participating facilities. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Randomized Regulation: The Impact of Minimum Quality Standards on Health Markets
Guadalupe Bedoya, Jishnu Das, and Amy Dolinger
NBER Working Paper No. 31203
May 2023
JEL No. H75,I11,I18,O17

ABSTRACT

We report results from the first randomization of a regulatory reform in the health sector. The reform established minimum quality standards for patient safety, an issue that has become increasingly salient following the Ebola and COVID-19 epidemics. In our experiment, all 1348 health facilities in three Kenyan counties were classifiéd into 273 markets, and the markets were then randomly allocated to treatment and control groups. Government inspectors visited health facilities and, depending on the results of their inspection, recommended closure or a timeline for improvements. The intervention increased compliance with patient safety measures in both public and private facilities (more so in the latter) and reallocated patients from private to public facilities without increasing out-of-pocket payments or decreasing facility use. In treated markets, improvements were equally marked throughout the quality distribution, consistent with a simple model of vertical differentiation in oligopolies. Our paper thus establishes the use of experimental techniques to study regulatory reforms and, in doing so, shows that minimum standards can improve quality across the board without adversely affecting utilization.

Guadalupe Bedoya
Development Impact Evaluation, DIME
The World Bank
1818 H Street N.W.
Washington, DC 20433
gbedoya@worldbank.org

Amy Dolinger
The World Bank
1818 H Street, NW
Washington, DC 20433
adolinger@worldbank.org

Jishnu Das
McCourt School of Public Policy
Georgetown University
37th & O Street NW
Washington, DC 20057
and NBER
jishnu.das@georgetown.edu

A randomized controlled trials registry entry is available at:
https://www.socialscienceregistry.org/trials/1001
An appendix is available at:
http://www.nber.org/data-appendix/w31203

I Introduction

Despite frequent calls for increased regulation, the difficulty of randomizing regulations in the health sector has meant that there is currently no experimental evidence on its impacts.[1] The lack of well identified studies is particularly worrying because theoretical models and empirical research both yield ambiguous results. On the one hand, regulatory reforms like minimum standards can be extremely beneficial in low- and lower-middle-income countries (LLMICs) where the quality of care is low and variable and a non-negligible fraction of health facilities in the private sector may be illegal and/or unlicensed.[2] On the other, even well implemented reforms can reduce geographical access and lead to higher prices as facilities are forced to close if they do not meet minimum standards, changes that have been shown to disproportionately hurt the poor.[3]

In this paper, we bring regulatory reforms firmly within the ambit of experimental techniques and show that doing so yields novel and important insights into the functioning of health markets. A minimum quality standard, accompanied with inspections and sanctions, raises quality without any decline in utilization; the quality increase reflects improvements within facilities rather than entry or exit and; mechanisms privilege market power rather than lack of information as a source of the underlying inefficiency that the standards address. Taken together the results provide a powerful illustration of how government regulation and stewardship can significantly improve the quality of care in LLMICs.

The specifics are as follows. Between 2013 and 2015, as part of a World Bank team, we worked with the Ministry of Health in Kenya and its nine regulatory boards and councils to develop a new regulatory mechanism for both public and private providers. The reform established minimum quality standards (MQS) that changed the content, frequency and consequences of facility inspections. In terms of content, it established a standardized inspection protocol called the "joint" health inspection checklist or JHIC that was used to assess the facility's compliance with patient safety protocols. Further, it replaced an earlier system of infrequent and ad-hoc inspections with

[1] See, for instance, WHO (2006). Two systematic reviews on the impacts of healthcare regulation found only two studies that met the eligibility criteria (Flodgren et al., 2011, 2016). Both studies examined the impact of inspections with additional support rather than a broader regulation that combined inspections with sanctions (but nothing else) and were "uncertain" on the impact of inspections. Outside the scope of these reviews, recent observational studies examine the impact of regulations that restrict physician's economies of scope. Chen et al. (2016) show that restricting physician ownership of pharmacies in Taiwan reduced drug prescriptions, although loopholes in the policy attenuated this effect. Yi et al. (2015) show that a similar policy in China reduced drug sales, but increased inpatient days driven by changes in producer behavior.

[2] In India, 75% of primary care is delivered by providers without any formal medical training (Das et al., 2022). This fraction is similar to what is found in other low-income contexts with the difference that in Sub-Saharan Africa, many countries allow non-physician clinicians to practice and prescribe medicines, including antibiotics. Multiple audit studies in primary care show severe deficits in the diagnosis and management of basic conditions in LLMIC. See Das et al. (2012), Banerjee et al. (2020), Mohanan et al. (2015), Daniels et al. (2017), Kovacs et al. (2022), King et al. (2021), and Kwan et al. (2022) for evidence from India, Senegal, Tanzania and Kenya. For hospital care, Siam et al. (2019) document substantial variation in the quality of obstetric care within a single city, Nairobi, Kenya.

[3] For instance, Chipty & Witte (1997) and Hotz & Xiao (2011) show that childcare regulations in the United states disproportionately reduced access for the poor.

regular inspections. Finally, the scores generated through the JHIC triggered well-defined warnings and sanctions ranging from immediate closure for unlicensed or very low scoring facilities to less frequent inspections for those with higher scores.

With cabinet approval, we implemented this new regulation in an experimental manner in three counties across the country for 13 months from November 2016 to December 2017. These counties (Meru in the center, Kakamega in the lakes region and Kilifi in the East coast) were chosen in consultation with health executives from all 47 counties in Kenya to represent the variation across the country in terms of geography and market structure. Inspections were carried out by government inspectors and fealty to the experimental allocation and protocol was maintained through the period of the evaluation, albeit with delays. Facilities did not receive any financial or in-kind support as part of the inspections, although importantly, the reform was published in the national gazette and therefore publicly available from March 2016 onward. The regulation and checklist were delivered to *all* facilities prior to the first inspection.

We coupled the experimental allocation of the regulation with a market-level randomization, where we first allocated all 1348 health facilities (including unlicensed providers) in the three counties to 273 distinct health markets and then assigned markets to one control and two treatment groups. In Treatment Group 1 (T1) all facilities were inspected, with warnings and closures implemented as necessary. In Treatment Group 2 (T2) we additionally displayed the results from the inspection on a health facility report card that prominently assigned a letter grade (A to D) to the facility. This market-level allocation of experimental treatments allows us to estimate the causal effects of the regulation for multiple outcomes despite (as we document) substantial exit and entry during the evaluation period, some of which was due to the treatment itself. The outcome measures we focus on include patient safety as measured by the facilities score on the JHIC, patient volume, and prices, all measured independent of the inspections by our team between March and August 2018, or three to eight months after the inspections ended.

We first show that the regulation (treating T1 and T2 as a combined treatment) successfully increased our main measure of patient safety, the JHIC score, which measures compliance with the items on the inspection checklist. This score increased by 0.49 SD for the average facility or 0.33 SD for the average patient in treated markets, the difference reflecting the use of patient load as weights. At the facility level, improvements were larger for the private sector (0.58 SD), licensed versus unlicensed private facilities (0.61 SD vs. 0.52 SD) and for facilities that had been in the program longer (0.50-0.65 SD). Improvements of 0.31 SD in the public sector were also substantial and an important demonstration that bringing public facilities under a uniform regulation can yield positive results, even without any additional resources as part of the intervention. Finally, in contrast to a concern that facilities may have focused on those areas of the checklist that were easiest to improve but not critical for patient safety, an item-by-item enumeration shows that the largest improvements were in facility infrastructure, equipment, and supplies—all of which required

substantial investments and are arguably necessary to deliver a minimal level of patient safety.

We then show that the intervention meaningfully altered the market structure. In treated markets, private facilities that were unlicensed at baseline were 8.9 percentage points more likely to exit, and visits to public facilities at endline increased by 19%. Interestingly, even though facilities that were unlicensed at *baseline* lost patients, the intervention did not decrease the patient load in unlicensed facilities at *endline*, as closed facilities were replaced by new ones or facilities re-opened often without obtaining a license after being closed. The regulation also did not increase prices for the average patient or decrease the use of health facilities, even among the poor.

Despite the increased exits and the reallocation of patients, an accounting decomposition based on Chandra et al. (2016) combined with the market-level randomization shows that 87% of the improvement in the JHIC score was due to improvements within-facilities, with another 5% due to the exit of facilities with lower than the mean market quality. We thus conclude that this regulatory reform improved patient safety without deleterious impacts on the population, specifically the poor, with changes within facilities driving the bulk of the improvements.

Our final set of results explores potential mechanisms. Here, we are guided by a literature that studies how MQS can influence market outcomes through a direct regulatory channel, an information channel (Shapiro, 1986), and/or a market power channel arising from vertical differentiation in oligopolies (Ronnen, 1991). The predictions from these models differ: if facilities were interested only in meeting the regulatory requirements, they should have minimized the costs of their investments. Similarly, if information was the main channel, we should see larger improvements in the information treatment (T2) as well as a decline in the use of facilities among the poor. Uniquely among these theories, Ronnen (1991) is the only one who suggests that even well performing facilities may see quality improvements in response to the regulation as they increase investments in order to maintain market power.

In order to establish the plausibility of each of these channels, we establish that (a) facilities invested in improvements that were (far) more costly than what was required under the regulation and were not optimizing decisions to meet compliance thresholds; (b) there was no difference in treatment outcomes between the inspection only (T1) and the inspection + information (T2) arms and; (c) quantile treatment effects by market density show that impacts were highest at the *top-end* of the distribution of patient safety, where facilities were least affected by regulatory requirements, as well as in markets with greater competition from public facilities. Therefore, in addition to a direct regulatory channel, we conclude that the data are consistent with market power as a source of inefficiency; nevertheless, we caution that the experiment was not designed to test a specific mechanism and we consider several alternate explanations in our discussion.[4]

[4]Our results on the mechanism are speculative because most facilities could have been sanctioned under the regulation and, therefore, beliefs over how the regulation functions and what other facilities, in turn, believe will determine facility investments. While previous work uses rational expectations to model beliefs regarding inspections (Duflo et al., 2018), in the case of a new system like the one we evaluate, such an assumption is harder to sustain.

In terms of the theory of regulation, our results elevate the relative importance of a market-power based explanation, like in Ronnen (1991) with facility investments potentially responding to (derived) demand in markets with multiple facilities. Given that fundamental problems of healthcare are often tied to a poor informational environment, it is surprising that we are unable to find a clear role for information constraints. This could be because patient safety as measured through the JHIC is one of the few dimensions of quality that is broadly observable and not patient specific—using a new sterile needle is observable and always good for the patient, but whether the patient is given an antibiotic is both harder to ascertain and may be good or bad depending on the underlying condition.

Our results also offer an interesting response to the vexing challenge of how to implement minimum standards in LLMICs given that low entry costs allow many low-quality and unlicensed providers to enter the market. Regulators worry that in this context, closing down one low-quality facility may mean that it is just replaced by another. This is in fact what we see in the data as the number of outpatients do not decline significantly in unlicensed facilities in treated markets at endline. However, the regulators inability to fully control what happens at the bottom of the market may still be consistent with improvements in quality for the average patient. In our experiment, it is improvements in the public sector and at the higher end of the private sector that drive an increase in the JHIC score for the average patient.[5] These are also the facilities that arguably faced the lowest regulatory pressure to improve, showcasing that minimum quality standards may lead to a broader set of impacts across the range of the quality distribution.

Our contributions to the literature are then three-fold. First, we show that the study of regulatory changes—one of the most significant functions of the state—can be brought under the scanner of experimental methods. The unit of randomization will be an important consideration in these studies; in our case, intervening experimentally at the market level was critical as regulations altered the market structure, and these effects would have been harder to identify if the treatment unit was the facility. We are not aware of previous work on health markets in LLMICs that either experimentally evaluates a regulation or randomizes at the level of the market.[6]

Second, we show that regulation without additional resources can improve patient safety without decreasing utilization. This contrasts with more common and expensive models of mentoring

Perhaps facilities invested in costly infrastructure because they believed they would be closed down or because others were doing so–even if these beliefs are inconsistent with the actual pattern of government-enforced closures in the data and would thus violate rational expectations.

[5]That inspections alone can improve quality in the public sector without additional financing or support is consistent with Dizon-Ross et al. (2017) observation of the (good) governance of public subsidies in a similar context.

[6]An established tradition examines health markets and market dynamics in the literature on OECD countries using natural experiments. Recent contributions include Dafny et al. (2019) and Chandra et al. (2016). A lack of data has hampered similar investigations in LLMICs, although recent contributions by Bennett & Yin (2019), Banerjee et al. (2020), Siam et al. (2019), Jain (2022) and Jain & Dupas (2022) all point to the importance of market dynamics for facility investments and patient choice. In education, Andrabi et al. (2017) and Andrabi et al. (2020) introduced the idea of market-level randomizations.

and financial assistance in the health sector that surprisingly yield worse results. Two previous experimental evaluations of a program called SafeCare sought to improve patient safety using mentoring and supervision. One of these evaluations, among primary public facilities in Nigeria, used similar measures to ours but found no impacts one year after the intervention (Dunsch et al., 2022). The other targeted private formal facilities in Tanzania, reporting a 4.4 pp or 8.5% increase over control facilities (King et al., 2021). That increase compares to an 8.8 pp or 23% improvement for a comparable group of licensed private and non-profit facilities in our study.[7] What is striking is that the cost per facility in their case was more than $8,000, which we will show is 26 to 28 times the cost of our intervention (King et al., 2021).

Third, the study allays concerns that even if MQS regulation improves quality, it does so by hurting the poor as the cost of care increases, either in terms of distance or price (Leland, 1979; Shapiro, 1986; Klein & Leffler, 1981). Our finding that quality increased across the board without increases in prices for the average patient or declines in utilization is consistent with theoretical predictions from the literature on vertically differentiated oligopolies, mediated in our case by the presence of the public sector. It is also consistent with recent evidence, also from Kenya, that healthcare providers do not face a perfectly elastic demand curve and therefore enjoy some market power in their pricing decisions.[8]

While we thus make progress in understanding the impacts of regulation, our assumption is that improvements in the JHIC score will improve downstream health outcomes, such as a decline in mortality or nosocomial infections. We do not have independent data to verify this claim, as the coverage of administrative data on health outcomes (such as mortality) is limited and not linked to health facilities or geographical areas at a sufficiently granular level in Kenya (WHO, 2021; Arudo et al., 2003).[9] One alternative we pursue to understand the benefits of the program uses demand-based measures of welfare instead. Specifically, we show that quality, as measured by the checklist, is positively correlated with price and the gains in consumer surplus from the intervention appear to be at least 10 times its cost.

The remainder of the paper is as follows. Section II discusses the setting and context. Section III presents the intervention and data collection. Section IV presents the results, Section V presents a discussion of possible mechanisms, and Section VI concludes.

[7] Another intervention to improve quality in Kenya's private sector, Contreras-Loya et al. (2021) also finds relatively smaller effects of a large and costly intervention designed to improve business management and care delivery on healthcare quality, although it increases facility investments.

[8] Contreras-Loya et al. (2021) show that a management consulting intervention improved structural quality but decreased clinical quality, a result they attribute to providers marking down quality in the face of an inelastic demand curve.

[9] The World Health Organization estimates that 45% of deaths in Kenya were unregistered in 2021 (WHO, 2021).

II Setting and Context

Primary healthcare in Kenya is delivered through tax-funded public (61%) and fee-charging private (39%) facilities.[10] Public facilities are managed independently by each of 47 counties following a process of devolution of responsibilities in 2010. Patients can choose what facility to visit. Prices in public facilities are administratively determined and substantially lower than prices in the private sector, which are set independently by each facility. Finally, facilities are divided by levels with Levels 2 and 3 providing primary care while Levels 4 and 5 also offer inpatient and advanced care (Figure S1, Supplemental Material shows examples of facilities at different levels of care).

Most health facilities operate in settings with some competition. In our study counties, 79% of all health facilities are in markets with 4 or more facilities (we define "market" more precisely in Section III) and 15% in markets with 2 to 3 facilities. The remaining 7% are "singleton" facilities, which tend to be publicly-owned and located in rural areas. A public sector option is available in 88% of markets catering to 98% of all patients, implying that even if all private sector facilities were closed, patients could still access healthcare. Mirroring the market structure, 70% of patients seek care in markets with 4 or more providers and 11% from singleton facilities. This distribution of markets in the study counties is similar to the rest of the country, although with more private facilities and greater competition (see Table S1 in Section 1 of the Supplemental Material).[11]

Patient safety is regulated by the national government through nine "Boards and Councils," each responsible for a different facet of healthcare delivery (for instance, the Kenya Medical Practitioners and Dentists Council licenses most health facilities, while the Kenya Medical Laboratory Technicians and Technologists Board addresses lab safety). Prior to the intervention, facilities were visited by inspection teams on an ad hoc basis based on the quota for the inspection period or by individual boards and councils, usually following a complaint or a serious adverse event. Four percent of facilities were inspected annually and the likelihood of two inspections in one year was zero.[12] Section 2 of the Supplemental Material provides details of the old inspection system.

Concerns around patient safety were raised after a national survey in 2012 reported that 2% of health facilities were compliant with minimum patient safety standards and systems. A subsequent study that used clinical observations thankfully suggested a more nuanced situation with variation across specific tasks. For instance, compliance was 87% with safe injections and blood draw practices but 2% for hand-hygiene. Even then, outpatients faced an average of 5.1 violations of infection, prevention and control (IPC) safety practices out of 7.5 observed indications where a safety action should have been taken (Bedoya et al., 2017). Despite these deficits, the quality of care in Kenyan

[10] Faith-Based and Non-Government Organizations account for 11% of facilities and 9% of patients in our baseline survey. These operate similarly to private facilities, except location decisions may be taken at a higher level.

[11] Differences between the data collected in the study counties and administrative data in other counties could also reflect under-counts of unlicensed providers in the latter (Table S1 in the Supplemental Material).

[12] Private communication with the Kenya Medical Practitioners and Dentists Council.

facilities is among the best in LLMICs, both in terms of the clinical knowledge of healthcare providers and the diagnosis and management of patients (Gatti et al., 2021; Daniels et al., 2017).

III Intervention, Experimental Design and Data

We now describe the intervention, experimental design and data collection.

III.1 Intervention

As part of a regulatory reform, in 2016 the government legislated a new framework, which included a Joint Health Inspection Checklist (JHIC) for facility inspections along with a scoring system and warnings and sanctions resulting from that score. Under the new inspection regime, both public and private facilities were to be inspected regularly—only private facilities were inspected before—and facilities could be closed if they failed to improve or lacked the appropriate licenses to operate. We discuss three facets of the reform—the JHIC instrument and implementation, the scoring and warning system and the implementation of the inspections with details presented in the Supplemental Material, Section 2.

The JHIC instrument: The JHIC focuses on input-driven measures of patient safety with 471 individual items across 14 sections.[13] The standards included in the JHIC represent widely validated minimum expectations for safe care by multiple international institutions including the World Health Organization, the US Centers for Disease Control, and the Joint Commission, which accredits hospitals in the United States. Meeting these standards is expected to reduce nosocomial infections in health facilities (WHO, 2011; Pittet et al., 2000). Scores are computed by equally weighting each section of the checklist, certain subsections, and components within subsections, and aggregating across sections to emerge at an aggregate percentage of the maximum score. This scoring system was a considered decision by the boards and councils after debating multiple options on the basis of pilot inspections and scoring systems developed by our team. The boards and councils felt that a system that was easy to understand was more important at this stage. What this means in practice is that items with very different compliance costs may receive the same weight in the JHIC. For instance, printing and posting a standard operating procedure receives the same weight as introducing a costly waste management system.

Sanction and Warning System: Following an inspection, facilities scoring less than 10% or those without a valid license to operate are categorized as "non-compliant" and recommended for

[13] See more in Supplemental Material Section 3. JHIC sections for all facilities include administrative and licensing information, health facility infrastructure, general management and recording of information, infection prevention and control, and medical consultation. Further sections are activated for facilities that provide additional services including labor ward, medical and pediatric wards, theater, pharmacy, laboratory, radiology, nutrition and dietetics and mortuary. A final section includes findings and recommendations. The complete checklist can be found in the 2016 Kenya Gazette Supplement No. 31 as part of Legal Notice No. 46 Public Health Act, Cap. 242.

immediate closure. Facilities scoring 11-40% are considered "minimally compliant" and receive a 3-month notice for improvement and re-inspection, while facilities with scores between 41-60% are classified as "partially compliant" and receive a 6-month notice for improvement and re-inspection. For these two categories, facilities are supposed to be closed if they do not improve to a higher category by the third inspection. Facilities that score above 60% do not face any risk of closure. Those classified as "substantially compliant" (61-75% of maximum score) are re-inspected every 12 months and facilities in the "fully compliant" category (above 75%) face inspections every 24 months (Table A1). These standards are very ambitious and in multiple pilots over 2 years, we documented that almost all facilities would fall in the "minimally compliant" category with very few scoring above 60%. The boards and councils nevertheless insisted on maintaining these high standards, which therefore departs quite strikingly from the focus in economic theory on marginal changes.

Implementation: The new regulation was implemented by full-time inspectors nominated and seconded by the Boards and Councils and County Governments for one year. Candidates went through a standardized training course developed as part of the intervention with classroom and field assessments and the top 12 candidates were selected. Our results should be viewed in the light of this stringent selection and training process, which is known to affect performance (Ashraf et al., 2020). There were very few instances of corruption and/or rude behavior and inspectors were able to frame the inspections as an exercise carried out together with the facility in the face of considerable challenges to improve healthcare for Kenyans.[14] Inspections were carried out on a tablet and the inspection protocol and scoring system was publicly available, allowing facilities to evaluate themselves as required, even prior to the inspection. A monitoring system, including real-time reports, was also put in place to facilitate planning and follow-up visits according to the regulation schedule.

III.2 Experimental Design, Timing and Data, Design Integrity

We discuss three components of the experimental design: The construction of markets, the allocation of treatment and control arms and the timing of inspections. Section 4 of the Supplemental Material documents IRB approvals of the trial and a discussion of the ethical issues.

Construction of Markets: We started with a census of 1,258 facilities that we could locate in the 3 counties between January and September 2015 and a census update conducted between October and November 2016 (see Section 1 of the Supplemental Material). We defined a market using a "z-center" clustering algorithm that assigned facilities to markets such that no facility was

[14] In the endline survey, 76% of facility in-charges commented on their experience with inspections and of these, only 2% of the comments were related to corruption. In addition, random inspection quality checks performed during the implementation showed minor discrepancies with inspectors results. Finally, a third-party qualitative assessment, separate from our team, similarly found few facility complaints with the inspection process (Tama et al., 2021).

more than 4km from the centroid of its assigned market, with the centroid computed recursively from the location of all facilities mapped to the market. The 4km radius was based on data from the baseline, which showed that 73% of patients lived within 4km of the health facility. This algorithm yielded a total of 273 markets of which 30% had one facility, 28% had 2-3 and 42% had 4 or more (Figure S2 in the Supplemental Material shows mapped examples of each type of market). This distribution also implies, as discussed previously, that 79% of facilities are located in markets with 4 or more providers, and 70% of care is sought in such markets.

Allocation of treatments: Having defined markets, we used a stratified cluster randomized experimental design to allocate markets to treatments. Clusters are healthcare markets and the cluster size is the number of health facilities per market. We stratify by market size and county for a total of 16 strata.[15] All 273 markets were randomly allocated to one of three arms:[16]

1. The Inspection Only or T1 Arm: 90 markets were assigned to high-intensity inspections with enforcement of warnings and sanctions for non-compliant facilities.

2. The Inspections plus Information or T2 Arm: 96 markets were assigned to the T2 arm, which combines the T1 arm with the public disclosure of inspection results.

3. Control Group: 87 markets were assigned to the "business-as-usual" low-probability inspections arm. Although inspections could have been carried out if there was a serious complaint, in practice, there were no joint inspections in the year of the intervention.

The scorecard system in T2 consisted of 4 letter grades ranging from A (fully compliant, or more than 75% or the maximum score) to D (minimally compliant, or 11-40%). See Panel A in Figure A1 of the Appendix. After each inspection, the inspector posted the scorecard in a prominent area, such as the patient waiting area, together with an explanatory poster (Panel A, Figure A2). In additional visits to all health facilities, quality officers distributed 65,000 flyers explaining the inspection results to community members, patients and other residents in the market areas (Panel B of Figure A2).

In cases where a facility was marked for closure (whether in T1 or T2) an additional red closure scorecard was posted at the facility or department during visits by the national team and county health officials (Panel B in Figure A1 of the Appendix). Closure events often led to extended discussions with the in-charge and people from the catchment area, where the government explained the reasons for the closure and why this was important for the population. The team also provided in-charges with information about the licensing process.

[15] We have 5 strata by market size for markets with 1, 2, 3, 4-10, and 11+ health facilities for the 3 counties, and an additional stratum for market size 34 or more (extreme values) in Meru for a total of 16 strata.

[16] Section 5 of the Supplemental Material includes tables presenting baseline and endline surveys (Table S3), the census of health facilities (Table S4), and details by treatment arm and county at randomization and endline (Table S5).

Data Collection Timeline and Sample: Figure 1 shows the timeline for data collection. Between January and September 2015, we located 1,104 facilities in the three counties and completed the baseline in 1,027 for a response rate of 93%. Following a delay of 15 months between the completion of the baseline and the start of the intervention we updated the census between October and November 2016, increasing the number of facilities to 1,258. For this update we collected basic characteristics such as ownership, level and location, but did not complete a full baseline survey. These are the facilities we used for the randomization.

The intervention then took place between November 2016 and December 2017 and the endline was completed between March and August 2018. The average time elapsed between the last inspection or closure visit, and the endline for all facilities was 7 months, although this varied from 4 to 18 months, a variation that we exploit when we examine the impact of program duration on impact.

During the endline survey we counted 1,322 facilities and completed the endline in 1,285 facilities for a 97% response rate.[17] Of these, 173 were new facilities which we allocated to existing markets using a nearest-neighbor algorithm and 90 were facilities that had been missed previously, with 4.5% market share at endline.[18]

For the treatment impacts, we always use the 1285 facilities surveyed at endline. When we examine impacts on facilities that were open at the baseline, we use the 1258 facilities we located at baseline or during the pre-randomization update plus the 90 missed facilities, for a total of 1348 facilities. When we estimate impacts on exit/entry, we use all facilities operational at randomization (1348) and/or endline (1319) regardless of whether they have a completed survey.

III.3 Data Sources and Description of Main Outcomes

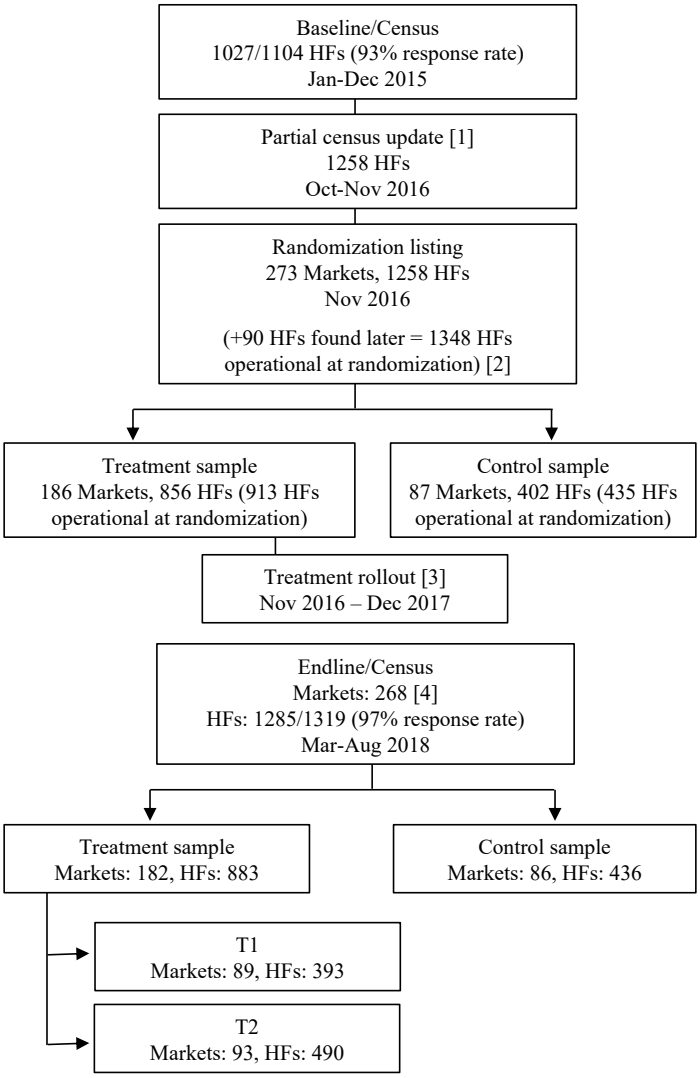
Our primary data sources are surveys of health facilities and their staff, exit surveys of patients, and direct clinical observations. At endline (baseline) we surveyed 1,285 (1,027) health facilities, 11,098 (8,577) patients, 2,098 (1,625) healthcare workers, and observed 19,178 (18,698) clinical interactions. We augment these survey data with additional administrative information on licensing status. Section 6 in the Supplemental Material lists the outcome variables and key covariates, along with details on how they were constructed.

In our study counties, 70% of facilities were private and 30% public, although higher patient

[17] At endline all facilities in 5 markets had closed, reducing the total number of markets to 268. We also exclude 3 of 1322 facilities that were more than 4km from our existing markets, which results on a total of 1319 facilities at endline.

[18] A difficulty with undertaking a census of this magnitude is that many of the facilities were small, one-roomed clinics and not included in administrative databases. In addition, 23 of the 90 facilities that we had "missed" were closed during the initial surveys, but during the endline survey, the facility in-charges gave us a facility opening date prior to the randomization. If we exclude these facilities, the market share of facilities that were missed is 2.7%. We assign these 90 facilities to a market using a closest-neighbor algorithm preserving the 4km clustering rule. Therefore, in total, there were 1348 facilities in the 273 markets at randomization.

Figure 1: Timeline of Study



Notes. [1] Due to the high turnover of facilities and delay in the implementation, we conducted a partial update of the census in markets of size 1, 2, and 3 between October and November 2016. We used this partial update of the census of 1,258 facilities located with available GPS coordinates for the randomization. [2] 90 facilities were missed or listed as temporarily or permanently closed during the randomization census. These facilities were added using a nearest-neighbor algorithm to the nearest market by endline. [3] Another partial update to the census was conducted at the end of July 2017 when the first round of inspections was completed in all counties. At this stage, only the new facilities were assigned to the markets as per randomization. [4] 268 of the randomized markets were still active at endline, or those with at least one health facility found in the market. Five markets were dropped because all the facilities permanently closed. HF = health facility.

volumes of 49 patients per day in the public facilities implied that they accounted for 71% of all outpatient visits at baseline (Table 1). We highlight that private providers saw an average of only 11 patients a day and 53% either did not have a valid operating license or were operating with an expired license before the intervention. Out-of-pocket expenditures per visit were USD 0.7 PPP in public compared to USD 8.4 PPP in private facilities and a wealth index of patients visiting private facilities was 1.36 units or 0.65 SD higher than for those visiting public facilities. Table 1 also shows that 97% of facilities at baseline were below the government threshold for full compliance, scoring 60% or less of the JHIC maximum score. JHIC scores did not differ by market size (Table S8 of the Supplemental Material), although public facilities scored 7.69 points or 0.67 SD higher than private facilities.

Table 1: Summary Statistics at Baseline

	All	Public	Private	N
	(1)	(2)	(3)	(4)
Panel A: Facility-level characteristics				
Facility is public/private	1.00	0.30	0.70	1348
Facility is:				
Level 2: Dispensaries and clinics	0.85	0.74	0.90	1348
Level 3: Health centers and maternity and nursing homes	0.11	0.19	0.07	1348
Level 4 or 5: Primary and secondary hospitals	0.04	0.07	0.02	1348
Facility is unlicensed (or has an expired license) (private)	NA	NA	0.53	944
Daily outpatients, mean [SD]	24.76 [39.03]	49.41 [52.29]	11.01 [17.95]	1025
Share of total outpatients	1.00	0.71	0.29	1025
Patients' OOP, mean [SD] USD PPP	5.47 [8.50]	0.70 [0.98]	8.39 [9.67]	958
JHIC score x 100 (% of max score) mean [SD]	36.24 [11.53]	41.18 [10.20]	33.49 [11.32]	1027
Facility is in JHIC category:				
Minimally compliant (11-40% of max score)	0.66	0.49	0.76	1027
Partially compliant (41-60% of max score)	0.31	0.47	0.21	1027
Substantially compliant (61-75% of max score)	0.03	0.03	0.02	1027
Fully compliant (>75% of max score)	0.00	0.01	0.00	1027
Panel B: Patient-level indicators				
Patients reporting zero OOP, proportion	0.49	0.65	0.23	8523 (958 HFs)
Patients reporting facility distance from home <=4km, proportion	0.73	0.72	0.75	8116 (966 HFs)
Patient's wealth index is, mean [SD] (-4 to 12)	0.87 [2.09]	0.34 [1.71]	1.70 [2.35]	8477 (960 HFs)
IPC indications in outpatient visit, mean [SD]	7.50 [5.61]	7.18 [5.46]	8.28 [5.90]	14108 (926 HFs)
Violations of IPC practices in outpatient visit, mean [SD]	5.11 [3.33]	4.85 [3.18]	5.72 [3.58]	14108 (926 HFs)
Panel C: Indication-level indicators from patient-HCW interactions				
Compliance with all IPC practices measured, mean [SD]	0.32 [0.47]	0.32 [0.47]	0.31 [0.46]	105876 (929 HFs)
Injection and blood draw safety practices	0.87 [0.33]	0.89 [0.32]	0.84 [0.36]	17541 (796 HFs)
Hand hygiene practices	0.02 [0.15]	0.02 [0.14]	0.04 [0.19]	41118 (879 HFs)

Notes. Standard deviations reported in brackets. The sum of proportions across categories may not add up to one due to rounding issues. Indicators at the patient level are unweighted. Infection prevention and control measures follow Bedoya et al. (2017). The variables and corresponding samples are described in detail in Supplemental Material Section 6. HF = health facility; JHIC = Joint Health Inspection Checklist; OOP = out-of-pocket payments; PPP = purchasing power parity, IPC = infection prevention and control.

One aspect of these markets that we had not anticipated was the significant churn in the private sector. Of the 301 private facilities in the control group operational at randomization, 57 (19%) had exited by August 2018 and 55 (15%) new facilities had entered. These closure rates far exceed the 8.2% reported by McKenzie & Paffhausen (2019) for small firms in LMICs. In our 2015 census itself, we were able to identify 202 (21%) facilities from the government master facility list in February 2015 that were no longer operational, and 379 (40%) facilities that were not part of the 938 facilities listed in the government records.

A second key feature of our data is the close link between the JHIC score, licensing status and market outcomes, which shows up in every aspect of facility performance. In the private sector at baseline, the JHIC score for unlicensed relative to licensed providers was 21% lower.[19] JHIC scores and licensing status were also strongly correlated with facility exits in the control group, with a 1 SD increase in the JHIC score (9.6 percentage points) associated with a 7.7 percentage point decline in exits (Table 2). However, facilities that exit the market by endline tend to be small and represent only 3% of all patients in the data (Table S9 of the Supplemental Material).

Column 1, Table 2 then shows that a 1 SD increase in the JHIC score (12.1 percentage points) was correlated with an increase of USD 1.9 (PPP) in out-of-pocket (OOP) payments per visit, a correlation that remains robust to the inclusion of machine-selected controls. As is well understood, this association between prices and the JHIC score does not identify the structural hedonic parameter in the presence of patient sorting. While we cannot address patient sorting fully, we can assess the sensitivity of our estimates to select features of the patient population in each facility, as shown in Column 2, Table 2. Here, in addition to the machine-selected controls from Column 1, we also include patient wealth, education, self-reported health status and distance traveled to the health facility, all characteristics that are likely correlated with the demand for higher quality care. Although these variables are positively associated with OOP payments, there is virtually no change in the price premium for higher quality as measured by the JHIC score, which retains its strong statistical significance.

Vertical differentiation requires a positive price-quality correlation in the private sector (which we find) but not necessarily a quantity-quality correlation as some facilities could be niche high-end facilities. Nevertheless, we do find a positive, but insignificant correlation between market share and the JHIC score in the private sector. We can also ask whether the positive valuation of quality extends to patients visiting public sector clinics. Since prices in the public sector are administratively determined and therefore uncorrelated with the JHIC score, a positive valuation of quality should show up in demand and we indeed find a strong quantity response with a 1 SD increase in the JHIC score associated with a 3.1 percentage points increase in outpatients among public facilities.

These results strongly suggest that (a) consumers placed a premium on safety as measured by

[19] Throughout the paper, unlicensed refers to facilities that do not have a license or have an expired license.

Table 2: Baseline Quality Association with OOP, Market Share, and Facility Exits
By Facility Ownership

	Private				Public	
	OOP (USD PPP) at Baseline	OOP (USD PPP) at Baseline	Market Share at Baseline (x100)	Exit by Endline (Control Facilities)	OOP (USD PPP) at Baseline	Market Share at Baseline (x100)
	(1)	(2)	(3)	(4)	(5)	(6)
JHIC Score at Baseline	0.155***	0.146***	0.036	-0.008**	-0.006	0.308**
	(0.049)	(0.046)	(0.044)	(0.004)	(0.007)	(0.149)
Unlicensed at Baseline	0.216	0.300	0.849	0.103**		
	(0.632)	(0.641)	(0.868)	(0.051)		
Patient Wealth Index		0.394**				
		(0.189)				
Patient Years of Education		0.061				
		(0.042)				
Patient Health Status (Bad or Very Bad)		2.680***				
		(0.667)				
Distance from Home (in Km)		0.683***				
		(0.152)				
Observations	3201	2938	648	189	5260	367
R^2	0.07	0.09	0.77	0.08	0.21	0.79
Dependent Variable Mean	8.16	8.13	9.66	0.15	0.70	53.05
Mean (SD) JHIC Score at Baseline	36.63 (12.14)	36.68 (12.22)	33.55 (11.28)	32.66 (9.56)	42.90 (10.31)	41.18 (10.20)
Total Controls Selected by PDF (out of 23)	6	8	12	2	3	6

Notes. Robust standard errors reported in parentheses and clustered at the market level. *** (**) (*) denotes significance at 1% (5%) (10%) level. Controls are selected by PDSLASSO out of a list of 23 variables. Market size, facility opening year, facility levels and strata FE at baseline are partialled out for all regressions (imposed as controls in the regression) so not included in the list of 23 variables. The indicator for unlicensed at baseline is partialled out for the private facilities regressions. In Column 2, patient wealth index, years of education, health status, and distance from home are partialled out. HF = health facility; JHIC = Joint Health Inspection Checklist; OOP = out-of-pocket payments; PPP = purchasing power parity.

the JHIC score and (b) that there was at least some (perhaps substantial) information about this score available to consumers. As we will see in Section V, this is consistent with a diminished role for information as the mechanism for the improvements we observe. These patterns also suggest that regulating facilities at the low-end may be very costly given their high rates of churn and low patient loads, an observation we return to in the conclusion.

III.4 Design Integrity

III.4.1 Balance, Attrition, and Accretion

There are no systematic differences across treatment and control groups in baseline main outcome variables and key covariates with the exception of out-of-pocket payments at the facility level and the test of joint significance yields an F-stat of 1.020 (p=0.425) (Table A2 of the Appendix). Response rates were 93% at baseline and 97% at endline and non-response is balanced between treatment and control at endline with an estimated null difference (p-value = 0.974). At baseline there is a small 4 percentage point higher response rate among facilities in treatment markets

(p-value < 0.001) as shown in Table S6 of the Supplemental Material.

III.4.2 Compliance with Treatment

Table S7 (Panel A) in the Supplemental Material shows that we reached 90% facilities in randomized markets in the T1 arm (95% of facilities still open at first inspection), 85% in the T2 arm (95% of facilities still open at first inspection) and 97% of facilities in the control group did not receive the intervention (3% contamination). A small number of facilities in the treatment arms did not received an inspection because they were found (or opened) at some point after the randomization. This is a plausible reflection of how an actual inspection process works in markets with considerable churn. Fidelity to the implementation protocol was maintained through the period of the evaluation with compliance of 94% or higher with the delivery of different intervention components (Figure A3) and in the T2 arm random quality checks showed that 89% of facilities left the scorecards displayed after the inspection (Bedoya et al., 2020).

Departures from the planned intervention were due to delays. It took 7.5 months to complete the first inspection in 90% of the facilities (versus a projected 4 months) due to delays in the starting date, absences (inspector absences implied that an average of 6 full-time inspectors conducted the inspections during 13 months of intervention), vehicle breakdowns and general strikes (Figure A4). These delays had two repercussions for our study. First, cabinet approval for the intervention allowed us to maintain a control group for one year. Therefore, the full cycle of three inspections could be completed only for 6% of treated facilities. Second, most facility closures reflected the lack of operating licenses rather than a lack of improvement and the time elapsed between the report for closure and its enforcement by a federal team averaged 70 days versus a stated 1-day protocol. Facility in-charges may have realized that enforcement capacity was weak, affecting their incentives and subsequent beliefs, an issue that we discuss further below.

IV Results

IV.1 Econometric Specifications

We estimate the impact of the program as the mean difference in the outcomes of interest between all facilities in treatment and control markets at endline, as in Equation 1:

$$Y_h = \alpha + \delta T_{m(h)} + \sum_{j=1}^{n-1} \theta_j V_{hj} + \omega X_h + \epsilon_h \tag{1}$$

Here, Y_h is the outcome of interest for health facility h in market m at endline and $T_{m(h)}$ is the treatment indicator at the market level that equals one when facility h is in a market m that receives the intervention. The parameter of interest, δ, is the impact of the regulation on facilities

in treated markets and it captures both the impact on existing facilities as well as changes in facility composition due to exit or entry.[20] X_h are facility or market-level covariates, and ϵ_h are unobserved characteristics. Since we stratified by county-market size groups, we follow Bruhn & McKenzie (2008) and include V_{hj}, which is a dummy variable equal to one if the facility is in one of the randomization strata j, where $n = 16$. Standard errors are clustered at the market level, unless otherwise stated. To account for multiple hypothesis testing, we also report sharpened two-stage q-values for the main outcomes of interest in braces, following Benjamini et al. (2006) and as described in Anderson (2008). Finally, we present both unweighted and weighted estimates at the facility level, where the weights are the patient load. The former relates to standard models in the IO literature, where quality and price are facility characteristics and demand is endogenous, while the latter show the impact on the average patient and is therefore what is important for the patient's welfare.

We further estimate the heterogeneity of impacts, using the following specification:

$$Y_h = \alpha + \delta_k T_{m(h)} + \gamma_k T_{m(h)} W_{hk} + \rho_k W_{hk} + \sum_{j=1}^{n-1} \theta_{jk} V_{hj} + \omega_k X_h + \epsilon_{hk} \tag{2}$$

Here, W_{hk} is a binary variable, indicating whether the observation belongs to one of the sub-groups over which we are running the heterogeneity analysis, for instance, whether a facility is private or unlicensed. All other notations are similar to Equation 1. We first report the impact of the treatment on facilities with *endline* characteristic k in treated markets. This is the relevant policy parameter of interest, and answers questions of the type: "What is the difference in the quality of unlicensed facilities in treatment versus control markets?" It is not the causal impact of the treatment on facilities with characteristic k, which at endline is endogenous to the treatment itself.

We therefore also report the causal impact of the treatment on facilities with characteristic k at baseline. In this case, the treatment effect is most precisely reported for the likelihood of exit and patient load; for the latter, we can correctly assign a value of zero when the facility is closed. For other characteristics, such as the JHIC score, we will have missing data for the 16% of all facilities in the census at randomization that exited by endline, and although we present these results in the appendix, they come with the caveat that they pertain only to surviving facilities. With this high rate of exit, any estimates based on bounds will be quite imprecise, underscoring the importance of the market-level randomization, which still allows us to back out the policy relevant impact of the treatment on regulated markets.

[20] The treatment estimators thus correspond to population intent-to-treat, but due to the high take-up and adherence to treatment status, as well as the high response rate at endline (97% of the census of facilities), they are unlikely to differ from treatment-on-the-treated effects.

IV.2 Impacts on Main Outcomes

Panel A of Table 3 presents the main reduced-form results from the regulatory reform, where we pool the T1 and the T2 arms into a single treatment allocation. We emphasize that there was *no change* in the JHIC score among control facilities between baseline and endline, either in the mean or at any point of the distribution (Figure S3 in the Supplemental Material). The treatment effect therefore accrues entirely from improvements in the treated facilities.

Health facilities in the treated markets improve their JHIC score by 5.2 percentage points (0.49 SD, q-value < 0.010) or 15% (Column 1). There is no significant change in daily outpatients or in the entry of new facilities (Columns 5 and 6). At the facility level, prices which are measured as out-of-pocket payments (OOP) per visit increased by USD 0.97 PPP or 24% (q-value $= 0.022$) (Column 2). However, when weighted by patient load in Column 4, these increases are negligible and never statistically significant. The impact on the weighted JHIC score is also smaller (Column 3), suggesting larger effects among smaller facilities.

Panel B, Table 3 then shows how private facilities at endline differed between treated and control markets. We highlight three important results. First, compared to facilities in control markets the JHIC score for private facilities in treated markets is 6.3 percentage points higher (0.58 SD, p-value < 0.010) and for public facilities 2.8 percentage points (0.31 SD, p-value < 0.010) higher (Column 1). Second, the intervention increases daily outpatients in public facilities by 7.8 patients of 19% (0.25 SD, p-value $= 0.021$), while it decreased daily outpatients in private facilities by 1.5 patients or 13% (0.06 SD, p-value $= 0.436$) (Column 5). Again, weighted impacts on prices are statistically insignificant for patients attending both public and private facilities (Column 4).

Finally, Panel C, Table 3 examines heterogeneity by licensing status at endline. JHIC scores were similarly higher for both licensed and unlicensed facilities in treated versus control markets (Column 1). Further, there is no significant difference in the patient load of licensed or unlicensed facilities in treated compared to control markets—if anything, the decline in patient load among private facilities seems to have come from licensed facilities at endline (Column 5). This could in part reflect the fact that unlicensed facilities were prompted to obtain a license and in fact, we see that in treated markets, the proportion of private facilities with a license increases by 7.7 percentage points (0.15 SD, p-value $= 0.061$), compared to 50% in control markets.

We present multiple checks in Figure A5 in the Appendix that confirm the robustness of these results to the inclusion of market baseline controls or keeping randomization strata alone.

IV.2.1 What did facilities invest in?

One concern is that, in the absence of data on health outcomes, improvements in the JHIC score could have been cosmetic with little likelihood of affecting downstream outcomes. As Section 3 of the Supplemental Material shows, several checklist items could be fulfilled simply by printing

Table 3: Treatment Effects on JHIC Score, OOP, Outpatients, and Entry:
Overall and Interacted with Indicators for Private and Unlicensed Health Facilities at Endline

	Unweighted		Weighted			
	JHIC Score (pp of max) (1)	OOP (USD PPP) (2)	JHIC Score (pp of max) (3)	OOP (USD PPP) (4)	Daily Outpatients (5)	New (6)
Panel A: *Overall Impact*						
Treatment	5.159***	0.973**	3.926***	0.138	1.484	0.006
	(0.836)	(0.419)	(1.319)	(0.553)	(1.741)	(0.022)
	{0.001}***	{0.022}**	{0.007}***	{0.474}	{0.247}	[0.785]
Observations	1285	1285	1285	1285	1285	1319
R^2	0.317	0.126	0.517	0.178	0.247	0.049
Control Mean	35.493	4.069	42.526	3.136	20.793	0.133
Impact: {%; SD}	{15%; 0.49}	{24%; 0.20}	{9%; 0.33}	{4%; 0.03}	{7%; 0.05}	{5%; 0.02}
Panel B: *Interaction with Private*						
Treatment	2.798***	-0.052	2.965*	0.364	7.803**	0.015
	(1.058)	(0.242)	(1.600)	(0.249)	(3.349)	(0.016)
	[0.009]	[0.829]	[0.065]	[0.144]	[0.021]	[0.349]
Private HF	-5.929***	4.373***	-0.038	5.485***	-28.353***	0.146***
	(1.011)	(0.377)	(2.364)	(1.012)	(2.989)	(0.030)
	[0.000]	[0.000]	[0.987]	[0.000]	[0.000]	[0.000]
Private HF x T	3.498***	1.509***	3.091	0.048	-9.303**	-0.013
	(1.176)	(0.569)	(2.505)	(1.072)	(4.117)	(0.036)
	[0.003]	[0.008]	[0.218]	[0.965]	[0.025]	[0.726]
Observations	1285	1285	1285	1285	1285	1319
R^2	0.337	0.219	0.524	0.350	0.409	0.078
Control Mean Public	39.760	0.643	42.236	0.808	41.060	0.022
Control Mean Private	33.463	5.698	43.033	7.211	11.151	0.184
Impact Public: {%; SD}	{7%; 0.31}	{-8%; -0.06}	{7%; 0.32}	{45%; 0.32}	{19%; 0.25}	{68%; 0.10}
Impact Private: {%; SD}	{19%; 0.58}	{26%; 0.28}	{14%; 0.39}	{6%; 0.07}	{-13%; -0.06}	{1%; 0.01}
Test T + Private x T = 0 (p-value)	0.000	0.007	0.003	0.712	0.436	0.944
Panel C: *Interaction with Unlicensed (Private and active at endline only)*						
Treatment	6.766***	1.094	7.712***	-0.135	-1.986	-0.036
	(1.222)	(0.762)	(1.956)	(1.418)	(2.786)	(0.038)
	[0.000]	[0.153]	[0.000]	[0.924]	[0.477]	[0.355]
Unlicensed HF at Endline	-3.815***	-1.295**	-2.148	-3.496***	-3.859	0.014
	(1.050)	(0.567)	(2.155)	(1.087)	(2.641)	(0.056)
	[0.000]	[0.023]	[0.320]	[0.002]	[0.146]	[0.805]
Unlicensed HF at Endline x T	-1.906	0.502	-4.303	1.515	1.515	0.095
	(1.427)	(0.831)	(2.904)	(1.517)	(3.221)	(0.064)
	[0.183]	[0.547]	[0.140]	[0.319]	[0.639]	[0.141]
Observations	872	872	872	872	872	905
R^2	0.372	0.090	0.602	0.077	0.302	0.056
Control Mean Licensed	36.703	6.393	45.718	8.083	15.821	0.161
Control Mean Unlicensed	30.086	4.974	35.991	4.924	6.283	0.207
Impact Licensed: {%; SD}	{18%; 0.61}	{17%; 0.19}	{17%; 0.48}	{-2%; -0.02}	{-13%; -0.06}	{-22%; -0.10}
Impact Unlicensed: {%; SD}	{16%; 0.52}	{32%; 0.35}	{9%; 0.29}	{28%; 0.35}	{-8%; -0.04}	{29%; 0.15}
Test T + Unlicensed x T = 0 (p-value)	0.000	0.008	0.062	0.041	0.650	0.247

Notes. Robust standard errors reported in parentheses and clustered at the market level. *** (**) (*) denotes significance at 1% (5%) (10%) level. "Naive" p-values are reported in brackets with stars next to the estimated coefficients. Sharpened q-values are reported in braces, following Benjamini et al. (2006), with stars next to the braces. Missing values for OOP in 5.8% of observations are imputed using means defined by level, ownership, treatment, license status at randomization, and daily outpatients. Regressions include randomization strata controls (by county and market size) and health facility level controls. HF = health facility; JHIC = Joint Health Inspection Checklist; OOP = out-of-pocket payments; PPP = purchasing power parity.

and pasting one-page operating instructions and even though checklists can improve medical care, they typically require a further process of integration into the care process Bosk et al. (2009). To assess precisely what items changed, we therefore estimated the impact of the intervention for seven different groups: Infrastructure, equipment, supplies (low-cost and medium-cost separately), management, medical records, and standard operating procedures (SOPs). While some of these items are simple to improve, others such as infrastructure, equipment and medium-cost supplies require substantial investments that are more likely to improve patient safety outcomes.[21]

Table 4, Panel A, shows that there were improvements in item compliance of 3.4 to 8.6 percentage points across these categories. Interestingly, the gains were the highest for infrastructure, equipment and medium-cost supplies (Columns 1, 2 and 4) and the lowest for improvements in SOPs (Column 7), which is the opposite of what we would have expected if the improvements were primarily cosmetic. The gains were higher among private sector facilities in treated markets for the categories of infrastructure, equipment and supplies; for medium cost supplies there was a 44% increase relative to a baseline of 28% compliance. In public facilities, the gains were again higher in the domains of infrastructure, equipment and supplies. These types of gains suggest that facilities did not focus just on the categories that were simple to improve but not critical for patient safety. Instead, the regulation led facilities—both public and private—to invest in areas that could have a genuine impact on patient well-being. Table A3 in the Appendix suggests that their investments could reflect demand as baseline quality-price correlations by functional category are statistically significant and higher for infrastructure and supplies, compared to SOPs, and remain robust to the inclusion of machine-selected controls.

IV.2.2 Heterogeneity by baseline characteristics

We now turn to the causal impact of the regulation on the likelihood of facility exits and on the number of outpatients, focusing on the facilities that were open at baseline. Prior to doing so, it is useful to understand the descriptive evidence on how the intervention could have directly affected facility exits through closures. Similar to what we presented in Table 2 on the correlation of facility exits and quality in the control markets, Table A4 now shows private facility exits (inactivity) in *treated* markets, again separated by licensing status and by quintiles of JHIC score. We also include an additional column showing the facilities that were closed by the government.

[21] Infrastructure items include items such as adequate ventilation, lighting, water, and physical structure requirements for emergency rooms and medicine storage. Equipment includes medical devices and equipment like neonatal incubators and delivery beds. Medium-cost supplies include specialized obstetrics and medical ward supplies (e.g., drip stands), as well as radiology supplies. Low-cost supplies include hygiene supplies (disinfectant or waste bins) and personal protective equipment as well as equipment like thermometers, stethoscopes, and sphygmomanometers used to measure blood pressure. Management includes items related to staff management, quality management, and information systems such as patient register systems, equipment service contracts, and quality assurance programs. Medical records include systems to record patients' medical history and records. Standard operating procedures include facility protocols across departments, such as waste management and cleaning charts for infection prevention and control (IPC), and for the handling, labeling and storage of samples in the laboratory.

Table 4: Treatment Effects on JHIC Item Compliance by Functional Categories:
Overall and Interacted with Indicators for Private and Unlicensed Health Facilities at Endline

	Infrastructure	Equipment	Supplies (Low cost)	Supplies (Medium cost)	Management	Medical Records	SOPs
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: *Overall Impact*							
Treatment	0.063***	0.072***	0.062***	0.086***	0.034***	0.049**	0.035***
	(0.012)	(0.010)	(0.010)	(0.019)	(0.008)	(0.024)	(0.007)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.042]	[0.000]
Observations	50927	16726	53711	2892	56321	6337	29617
R^2	0.045	0.047	0.017	0.078	0.042	0.096	0.033
Control Mean	0.409	0.278	0.383	0.364	0.289	0.467	0.078
Impact: {%; SD}	{15%; 0.13}	{26%; 0.16}	{16%; 0.13}	{24%; 0.18}	{12%; 0.08}	{10%; 0.10}	{45%; 0.13}
Panel B: *Interaction with Private*							
Treatment	0.035**	0.056***	0.036***	0.029	0.024*	0.012	0.030**
	(0.015)	(0.015)	(0.014)	(0.034)	(0.013)	(0.037)	(0.011)
	[0.020]	[0.000]	[0.008]	[0.392]	[0.054]	[0.749]	[0.011]
Private HF	-0.092***	-0.005	-0.023	-0.217***	-0.128***	0.006	-0.035***
	(0.014)	(0.013)	(0.015)	(0.037)	(0.013)	(0.042)	(0.011)
	[0.000]	[0.687]	[0.121]	[0.000]	[0.000]	[0.885]	[0.001]
Private HF x T	0.043***	0.025	0.040**	0.092**	0.014	0.064	0.008
	(0.017)	(0.017)	(0.017)	(0.040)	(0.016)	(0.046)	(0.013)
	[0.010]	[0.136]	[0.015]	[0.022]	[0.377]	[0.166]	[0.509]
Observations	50927	16726	53711	2892	56321	6337	29617
R^2	0.048	0.047	0.017	0.099	0.054	0.098	0.034
Control Mean Public	0.481	0.288	0.398	0.499	0.390	0.463	0.106
Control Mean Private	0.370	0.272	0.375	0.276	0.236	0.470	0.062
Impact Public: {%; SD}	{7%; 0.07}	{20%; 0.12}	{9%; 0.07}	{6%; 0.06}	{6%; 0.05}	{3%; 0.02}	{28%; 0.10}
Impact Private: {%; SD}	{21%; 0.16}	{30%; 0.18}	{20%; 0.16}	{44%; 0.27}	{16%; 0.09}	{16%; 0.15}	{61%; 0.16}
Test T + Private x T = 0 (p-value)	0.000	0.000	0.000	0.000	0.000	0.011	0.000
Panel C: *Interaction with Unlicensed (Private and active at endline only)*							
Treatment	0.079***	0.100***	0.072***	0.122***	0.048***	0.053	0.047***
	(0.016)	(0.015)	(0.015)	(0.025)	(0.014)	(0.042)	(0.011)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.001]	[0.208]	[0.000]
Unlicensed HF at Endline	-0.068***	-0.035**	-0.054***	-0.122***	-0.032***	-0.146***	-0.013
	(0.015)	(0.014)	(0.018)	(0.033)	(0.011)	(0.050)	(0.009)
	[0.000]	[0.012]	[0.003]	[0.000]	[0.003]	[0.004]	[0.164]
Unlicensed HF at Endline x T	-0.016	-0.057***	0.004	0.001	-0.030*	0.035	-0.028**
	(0.019)	(0.019)	(0.024)	(0.048)	(0.017)	(0.060)	(0.014)
	[0.400]	[0.003]	[0.883]	[0.976]	[0.084]	[0.557]	[0.039]
Observations	33125	10700	33929	1752	36640	3646	18352
R^2	0.052	0.057	0.023	0.098	0.043	0.132	0.034
Control Mean Licensed	0.425	0.310	0.402	0.340	0.273	0.560	0.076
Control Mean Unlicensed	0.304	0.224	0.332	0.138	0.188	0.318	0.038
Impact Licensed: {%; SD}	{19%; 0.16}	{32%; 0.22}	{18%; 0.15}	{36%; 0.26}	{18%; 0.11}	{9%; 0.11}	{61%; 0.18}
Impact Unlicensed: {%; SD}	{21%; 0.14}	{19%; 0.10}	{23%; 0.16}	{89%; 0.36}	{10%; 0.05}	{28%; 0.19}	{48%; 0.10}
Test T + Unlicensed x T = 0 (p-value)	0.000	0.001	0.000	0.002	0.034	0.036	0.004

Notes. Robust standard errors are reported in parentheses and clustered at the market level. P-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. Regressions include randomization strata controls (by county and market size) and health facility level controls. HF = health facility; JHIC = Joint Health Inspection Checklist.

We note first that 24% of all private facilities in treatment markets that were open at our baseline were closed by the government at some point in time. These facilities were mostly unlicensed (45% of unlicensed facilities were closed by the government compared to 7% of licensed facilities), and even though *all* unlicensed facilities were supposed to be closed, actual closure rates were much higher (61%) among facilities in the lowest quintile of JHIC scores compared to the top quintile (11%). Among licensed facilities, facilities in the bottom two quintiles experienced a 11% to 21% rate of closures, compared to a negligible 1% to 3% among facilities in the top quintiles. Finally, overall exit rates in treatment markets are smaller than the closure rate: this is because many facilities reopened after being closed by the government and most of them do so without obtaining the required licenses. Both because the patterns of exits in treatment markets are very similar to what we see in the control group and because closed facilities seem to re-open, the impact of the treatment on exit rates will be smaller than the rate of government closure—emphasizing the difference between the impact of regulation from its proximate effect, which is what regular monitoring data would provide.

In Table 5, we use the census of facilities at randomization to estimate a 3.4-percentage point increase (p-value = 0.238) in exits among treated private facilities. This impact is not statistically significant and it is zero for public facilities. It is only when we look at variation by licensing status that significant differences arise, with unlicensed facilities 8.9 percentage points (37%, p-value = 0.046), more likely to exit in treated compared to control markets. Coding all outpatients as zero for inactive facilities shows that facilities that were unlicensed at randomization also see a decline in their outpatient load of 3.1 patients (p-value $<$ 0.010) or 43% compared to an average of 7.1 in control, with no impact on the outpatient caseload for licensed facilities. We conclude that facilities unlicensed at randomization were most affected by the regulation in terms of closures and loss of business. Again, this is consistent with unlicensed facilities *at endline* maintaining their patient load, as facilities that were closed were replaced by new unlicensed facilities or simply reopened, often without obtaining their licenses.

Table A5 shows that overall results on JHIC score and OOP for facilities open at randomization remain the same as those reported for the whole sample at endline (Table 3), with impacts slightly higher for the former. These differences widen further for private facilities that show an increase of 21% (p-value $<$ 0.010) and even more so for licensed facilities that report an increase in the JHIC score of 8.8 percentage points (p-value $<$ 0.010), or 23%—the highest impact on patient safety reported across all groups. While we do not emphasize these results as they pertain only to surviving facilities, they presage two important discussions below. First, they suggest that improvements in treated markets mostly reflect gains in existing facilities (rather than exit or entry) and second, they show that even as licensed facilities experienced *lower* rates of government closures, they improved the most. This will guide our discussion when we turn to mechanisms below.

Table 5: Treatment Effects on Outpatients and Inactivity: Overall and Interacted with Indicators for Private and Unlicensed Health Facilities at Randomization

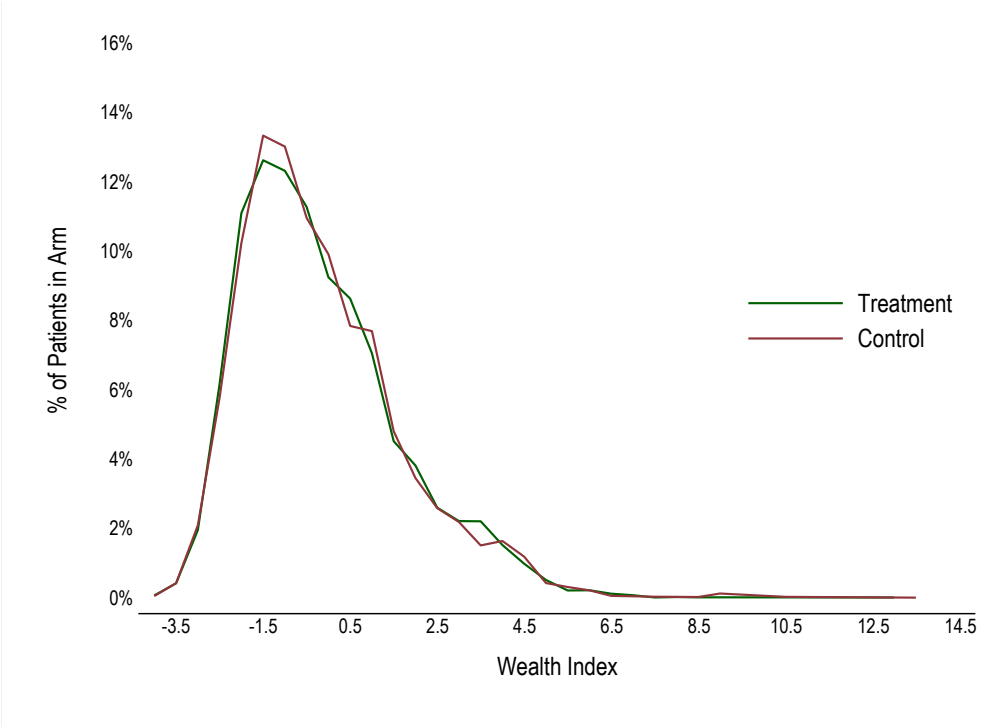
	Daily Outpatients	Inactive
	(1)	(2)
Panel A: *Overall Impact*		
Treatment	0.682	0.027
	(1.629)	(0.021)
	[0.676]	[0.199]
Observations	1322	1348
R^2	0.253	0.042
Control Mean	20.114	0.131
Impact: {%; SD}	{3%; 0.02}	{21%; 0.08}
Panel B: *Interaction with Private*		
Treatment	7.620**	0.003
	(3.449)	(0.009)
	[0.028]	[0.709]
Private HF at Randomization	-29.214***	0.170***
	(3.082)	(0.025)
	[0.000]	[0.000]
Private HF at Randomization x T	-9.321**	0.031
	(4.239)	(0.030)
	[0.029]	[0.299]
Observations	1322	1348
R^2	0.419	0.090
Control Mean Public	41.424	0.000
Control Mean Private	10.267	0.189
Impact Public: {%; SD}	{18%; 0.24}	{ .%; .}
Impact Private: {%; SD}	{-17%; -0.07}	{18%; 0.09}
Test T + Private x T = 0 (p-value)	0.365	0.238
Panel C: *Interaction with Unlicensed (Private and active at randomization only)*		
Treatment	0.004	-0.015
	(2.665)	(0.034)
	[0.999]	[0.657]
Unlicensed HF at Randomization	-0.628	0.092*
	(2.014)	(0.050)
	[0.756]	[0.067]
Unlicensed HF at Randomization x T	-3.089	0.104*
	(2.662)	(0.057)
	[0.247]	[0.071]
Observations	919	944
R^2	0.311	0.080
Control Mean Licensed	14.378	0.124
Control Mean Unlicensed	7.109	0.238
Impact Licensed: {%; SD}	{0%; 0.00}	{-12%; -0.05}
Impact Unlicensed: {%; SD}	{-43%; -0.23}	{37%; 0.21}
Test T + Unlicensed x T = 0 (p-value)	0.002	0.046

Notes. Robust standard errors reported in parentheses and clustered at the market level. P-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. Regressions include randomization strata controls (by county and market size) and health facility level controls. HF = health facility.

IV.2.3 Impacts on healthcare utilization among the poor

Did higher exit rates among low-quality (and low-priced) providers, combined with higher prices at least in some facilities, hurt the poor even though prices for the average patient did not increase? To test for this possibility, we assess the impact on the distribution of patients by socioeconomic status. We construct a wealth index using exit surveys of 11,098 outpatients based on asset ownership following the Demographic and Health Survey (DHS) in Kenya (see variable construction in Section 6 in the Supplemental Material). If care seeking had declined among the poor, we should have seen a mean increase in wealth among those visiting facilities in treated areas and lower densities at lower wealth levels. In fact, as Figure 2 shows, we cannot reject the hypothesis that the distribution of the wealth index is identical among patients in treatment and control markets (Kolmogorov–Smirnov test p-value = 0.325). Table A6 in the Appendix presents further robustness checks confirming that there is no treatment effect, either for the mean or for different quantiles of the wealth index. We can thus confirm that access to health care among poorer patients was not reduced by the intervention, suggesting an overall improvement in their quality of care.

Figure 2: Distribution of Patients by Wealth Index and Treatment Status



Notes. Kolmogorov-Smirnov test p-value = 0.325. Index range is observed range. Wealth index is estimated following the methodology of the Kenya Demographic and Health Survey based on household ownership of selected assets.

IV.3 Decomposition of JHIC improvements and the role of closures

One simple way to assess whether our impacts are driven by the increased closure of lower-quality facilities is to evaluate the treatment effect among the (selected) sample of facilities that were always open. This is shown in Table A5, where we find that the treatment effect on the JHIC score is *higher* in this sample, a first indication that exits are not the main reason for our observed improvements. In our next exercise, we now present a fuller accounting of the different channels of improvement by first decomposing the observed average gains for patients in the JHIC score into its separate components of facility improvements, exits, entries and patient reallocation, stressing that this is an accounting decomposition. We then leverage the market-level randomized design to estimate the impact of the regulation on the different components.

Following Chandra et al. (2016), Foster et al. (2001), and Foster et al. (2008) we write the change in average market quality for patients as:

$$\begin{aligned}\Delta\bar{q}_m = &\underbrace{\sum_{h\in C_m}\theta_{h,0}\Delta q_h}_{\text{within}} + \underbrace{\sum_{h\in C_m}\Delta\theta_h\left(q_{h,0}-\bar{q}_{m,0}\right)}_{\text{between}} + \underbrace{\sum_{h\in C_m}\Delta\theta_h\Delta q_h}_{\text{cross}} \\ &+ \underbrace{\sum_{h\in M_m}\theta_{h,1}\left(q_{h,1}-\bar{q}_{m,0}\right)}_{\text{entry}} - \underbrace{\sum_{h\in X_m}\theta_{h,0}\left(q_{h,0}-\bar{q}_{m,0}\right)}_{\text{exit}}\end{aligned} \tag{3}$$

where q_h indicates patient safety defined as the facility JHIC score of health facility h in market m and θ_h is its market share in terms of outpatients. We look at two periods: the endline period (period 1) and the baseline period (period 0). $\bar{q}_m$ is the market-share-weighted average JHIC score in market m (at period 0 or 1), and Δ is the difference operator, applied between endline and baseline (or in actual notation, between period 1 and 0). $\Delta\bar{q}_m$ is then the change in the market weighted average JHIC score between baseline and endline for market m. C_m is the set of health facilities in each market which were open both at baseline and at endline. M_m is the set of health facilities which did not exist at baseline but were active at endline. X_m is the set of health facilities which were active at baseline but inactive at endline.

This decomposition divides the weighted change in patient safety into five terms. The first term, "within," captures the change due to health facilities improving while keeping their baseline market share constant. The second "between," reflects the change due to patients reallocating (at endline) to health facilities with baseline JHIC score above the weighted baseline mean of their market. The third, "cross," shows the covariance between changes in market share and changes in patient safety between baseline and endline for facilities active at baseline and endline. The "cross" term can be interpreted as whether changes of facilities' JHIC score were accompanied by changes in market shares. The final two terms, "entry" and "exit" are, respectively, the change due to facilities entering each market with patient safety scores above the market weighted mean

at baseline and facilities exiting the market with patient safety scores below the weighted baseline mean of their market.[22] Having computed the decomposition for each market, we then compare treatment and control markets to estimate the impact of the intervention on each component.

Table 6, column 1 shows that 87% of the total increase in the (patient-weighted) JHIC score of 3.6 percentage points (p-value < 0.010), is driven by "within" health facilities changes.[23] The exit of facilities with quality below the market baseline mean contributes only 5% of total impact (p-value $= 0.013$) with reallocation of patients across facilities barely contributing to the overall improvement. Therefore, gains in the JHIC score for the average patient was primarily due to improvements within facilities, rather than reallocation, exits or entries. This reflects the fact that entering facilities account for less than 12% of market share and exiting facilities less than 3% of market share and that patient reallocations are among facilities with similar quality, as we would expect if movers are "marginal." Figure A6 presents robustness checks, which do not change the main results presented here.

Table 6: Treatment Effects on Weighted JHIC Score and Decomposition Components
Percentage Points of Maximum JHIC Score

		Contribution				
	Total Impact	Within	Between	Cross	Entry	Exit
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	3.559***	3.080***	0.298	−0.046	0.044	0.182**
	(0.933)	(0.876)	(0.309)	(0.277)	(0.159)	(0.073)
	[0.000]	[0.001]	[0.335]	[0.869]	[0.782]	[0.013]
Control Mean	−0.314	−0.331	0.047	0.065	−0.294	0.200
Observations (Markets)	252	252	252	252	252	252
Observations (Facilities)	1303	1303	1303	1303	1303	1303

Notes. Robust standard errors are reported in parentheses and p-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. Regressions include randomization strata controls (by county and market size) and control for the percentage of health facilities of each level in the market.

IV.4 Cost Effectiveness

The operational cost of this intervention during the pilot phase was USD 165 per visit, which includes inspections and visits for the enforcement of warnings and sanctions, as well as closures of facilities and/or departments within facilities. Multiple factors would allow us to reduce the costs in a scaled-up version to USD 95 per visit. With an average of 3 visits (2 inspections) per

[22] This analysis includes 92% of the markets identified at randomization. We restrict the sample to markets that were active at both baseline and endline and exclude markets where missing data accounts for more than 30% of the share in the market at any period. We also exclude facilities with missing data. These restrictions reduce the total sample by 15% of all facilities (11% of facilities active at baseline and 10% of facilities active at endline), which account for 3.0% of patients in the baseline and 4.8% in the endline.

[23] The difference with the weighted impact presented in Table 3 stems from a slightly different sample due to the restriction to markets open at baseline and endline as explained in the previous footnote.

treated facility, we estimated the operational cost per facility for the pilot to be USD 495, which could be reduced to USD 285 for the scaled-up model.[24] Supplemental Material Section 8 presents a snapshot of the costs with further details in Bedoya et al. (2020). To ensure the validity of our estimates, we also provided data to an independent team to complete a third-party costing of our intervention. Including the fixed cost, they computed a per-visit estimate of USD 103 or USD 309 for a full cycle in the scaled-up model, which is only slightly higher than our estimates (Chege et al., 2022). This compares to a cost of $8,000 per facility reported by King et al. (2021) for a similar standards-based approach intervention for private facilities, and much higher than costs of between $8,900 and $108,000 for results-based financing interventions, which have become one important mechanism for quality improvement in this region.[25]

What about benefits? Although we do not have data on health outcomes, we can interpret the increase in quality as an equivalent decline in price and use the (back-of-the) envelope theorem via Roy's identity to compute the gain in consumer surplus as the decline in price multiplied by the total number of patients.[26] Based on Table 2 (column 2) we estimate that patients are willing to pay USD 0.15 PPP (USD 0.075 nominal) for one additional percentage point JHIC score in a facility at baseline, after controlling for patient-level characteristics. Facilities in treatment markets receive 8.1 million outpatient visits each year and the impact of the intervention on the JHIC score, weighted by patient load, is 3.93 percentage points. This yields annual estimated gains in consumer surplus of USD 2.4 million in nominal exchange rates, compared to an operational cost of USD 242,000 for scaled-up program per year. This gain in consumer surplus is 10 times the cost of the program but it may still be underestimated, both because we have assumed it accrues for only one year and because we have excluded inpatients, who may value quality even more, from this computation.

IV.5 Additional Results

Having demonstrated the impacts of the regulation on the JHIC score and the market structure, we now present three additional results before turning to potential mechanisms. Specifically, we assess

[24] Costs were higher in the pilot because of a single office in the county headquarters, which increased travel costs and the fact that inspectors were seconded from different government institutions and transferred from other regions, resulting in a salary supplement. In a scaled-up version, the number and location of inspectors can be flexibly determined to minimize costs.

[25] See examples cited in Chege et al. (2022) such as De Allegri et al. (2019), Zeng et al. (2018), and Borghi et al. (2015).

[26] There are no systems for measuring nosocomial infections, vital statistics are incomplete and are not linked either to facilities or to geographical areas at a sufficiently granular level. Even if mortality data were available, the sample size requirements for sufficient power are exceedingly large. If we use a value of a statistical life of USD 50,000, which may be relevant for very poor populations, the intervention would have to save an additional 9 lives over 8 million outpatient visits for benefits to exceed costs (Li, 2020). To be well powered, this requires samples of more than 1 billion patients in each treatment arm, assuming mortality rates common to the literature (National Academies of Sciences, 2018; World Bank, 2020). Using a VSL of USD 200,000, or 100 times the Kenyan nominal per capita GDP (based on the World Bank's Indicators), would imply that the intervention is cost effective even if it saves two additional lives which requires even larger sample sizes to detect (World Bank, 2020).

cross-market spillovers, the impact of program duration, and spillovers on other quality measures that were not part of the inspection process. The estimating equations and accompanying tables are detailed in Sections 7 and 9 of the Supplemental Material.

Cross-Market Externalities: Cross-market externalities, whereby control health facilities in markets located near treatment facilities are affected by the treatment, may bias our estimates of the impact of the regulation. We identify cross-market externalities using exogenous variation in the local density of facilities induced by the stratified market-level randomization, following a method similar to Miguel & Kremer (2004). We find no significant cross-market externalities in the JHIC score, patient load, OOP payments, exit and entry of new facilities (Table S12 in the Supplemental Material).

Program Duration: Next, we exploit variation in the timing of the inspections and the endline to examine the impact of program duration, which captures both the fact that facilities that were in the program longer will have been inspected more often (2.4 times versus 1.6 to 2.0 times for other groups) and that program impacts can fade out over time. Our main identifying assumption, which we verify in Section 7.2 in the Supplemental Material, is that conditional on the controls, the variation in the date of first inspection and the date of the endline are not correlated with the JHIC score.

In markets where the time elapsed from first (last) inspection to endline was 15 (10) months, the JHIC score increased by 7 percentage points (0.65 SD, p-value < 0.01), compared to 4 percentage points for treated markets where the time from first (last) inspection to endline was 11 (7) months. This suggests little "fade-out" and potentially larger effects as the model scales up (Figure S4 and Table S14 in the Supplemental Material).

Impacts on non-incentivized outcomes: One concern with regulations on specific inputs is that they can reduce quality along non-incentivized dimensions (Blau, 2003, 2007). We were particularly concerned about this possibility given the results presented by Contreras-Loya et al. (2021), who find that structural improvements are accompanied by declines in the quality of clinical processes in private facilities. We therefore estimated the impact of the regulation on multiple process and structural measures of quality that were not part of the JHIC instrument. These include: (a) compliance with infection prevention and control practices across 19,178 observations of clinical interactions; (b) quality indicators reported by patients in 11,098 exit surveys and; (c) healthcare staff composition and remuneration for 7,663 staff.[27]

Fortunately, we do not find significant negative changes along any of these dimensions, with small typically positive effect sizes and statistically insignificant after correcting for multiple hypothesis. To the extent that we can interpret the individually significant estimates, in public facilities, we find an *increase* in consultation length, which has shown to be positively correlated with clinical accuracy, as well as an increase in the ratio of healthcare workers to total staff and total

[27] In the 13% of health facilities with more than 15 staff, we chose a random sample stratified by cadre.

staff compensation. These results show that across multiple dimensions of quality the intervention does not lead to negative spillover effects. In fact, there is a suggestion of improvements in some non-incentivized dimensions of quality in the public sector (Table A7 to A9).

V A Discussion of Possible Mechanisms

In order to understand the mechanisms at play it is worth emphasizing, first, that if there is no market failure, minimum quality standards are welfare decreasing. Facilities below the minimum quality are eliminated (they either improve or shut down)—but this increases prices and decreases use for those with lower willingness-to-pay. For MQS to improve welfare therefore requires a market failure—and the distributional impact depends on the source and extent of this failure.

Two canonical sources of market failure have been extensively studied. In Shapiro's model, the source of the market failure is asymmetric information (Shapiro, 1986). Firms choose to invest in quality but consumers cannot initially distinguish high from low quality, so firms are in a pooling equilibrium. In a second period, quality is revealed and higher quality firms charge higher prices. For a firm to invest in quality, it therefore requires a rent in the second period to compensate for the lower price in the initial period. An MQS increases the average quality in the (pooling) first period and therefore increases prices; in the second period, it decreases the rent necessary for firms to invest in high quality. These changes benefit consumers with higher willingness-to-pay and hurt consumers with lower willingness-to-pay as facilities close down and prices increase for low quality facilities.[28]

In contrast, in Ronnen's formulation, the inefficiency arises from market power due to vertical differentiation in oligopolies (Ronnen, 1991). In a model where firms choose quality and then price, the choice of vertical differentiation trades-off market access and market power. MQS increases the quality of the lowest firm—but by decreasing the market power of the higher quality firm, it also puts pressure on the high-quality firm to improve. The equilibrium is similar to what would obtain in a Stackelberg rather than Nash Equilibrium—lower quality firms would like to be able to commit to a higher quality, but cannot do so because it is not subgame perfect. The MQS allows them to achieve this higher quality equilibrium. Consumers in this model are strictly better off because overall market power is reduced. The distributional implications of this channel are thus very different from those in Shapiro's model.

Although formal tests of these models are difficult to execute in our broad-ranging experiment, a set of ancillary results help us disentangle these forces.[29] Interestingly, the results elevate the

[28] Multiple models since Shapiro (1986) confirm the basic intuition that for a separating equilibrium to emerge in markets with asymmetric information, there must be an informational 'rent' for high quality firms. It is this rent that provides the leverage for consumers to punish the firm in case they choose to lie about their quality.

[29] Formal tests of these models require the emergence of sharp cut-offs, which we do not see in our data, and at least some subset of facilities to be unaffected by the regulation. Given the ambitious standards, 97% of facilities could have been subjected to some sort of sanctions—and therefore beliefs over the regulation determine investments, as do

importance of the market-power channel although alternate interpretations, which we discuss, may also be consistent with the findings.

Result 1: Facilities improved in ways that went beyond the "letter of the law:" We first looked for strategic behavior among facilities with respect to the regulation, which would suggest that it was the regulation itself that led to the changes we observed. A facility interested in minimizing the cost of complying with the regulatory requirements would have (a) started with the lowest-cost items and (b) undertaken changes that were just sufficient to meet the compliance threshold. Indeed, a striking consequence of the scoring rubric in the JHIC was that if facilities had complied with all items in the lowest-cost category, their score would have increased by 34 percentage points or 3.2 SD, placing the average facility well above the 60% compliance score that would have staved off future warnings or sanctions. Instead, consistent with our previous results, we find that the impact of the intervention was 3.4, 7.4 and 6.3 percentage points (all p-values < 0.01) on compliance with the lowest, medium and high cost items (Table A10). An alternative classification by items that affected the marginal versus the fixed-cost again yielded similarly sized impacts on both, despite the fact that all the items in the lowest-cost category were fixed-cost items that are therefore independent of the number of patients (Table A10).

We also do not find evidence that facilities focused on "just" meeting the compliance threshold. For instance, 66% of facilities had a JHIC score lower than 40% at baseline, implying they faced the most frequent follow-ups (every three months) and risk of closure if the facility did not move to the next category by the third visit. Facilities closest to this cutoff-point could have strategically moved to the next higher compliance category (41-60%), with more lenient warnings and sanctions. Figure A7 shows evidence of lack of strategic behavior on this front; using a McCrary-type density test we cannot reject the null hypothesis of continuity of the density of the JHIC score for treatment facilities around 40% of the maximum score (p-value = 0.246).

In contrast to the regulatory-driven incentives, we find some evidence that market-based incentives played a role among private facilities. Table A3 shows that price-quality correlations are statistically significant and higher for infrastructure, equipment and medium-cost supplies, compared to SOPs and these correlations are robust to the inclusion of machine-selected controls. This is consistent with our finding in Table 4 of little improvement in compliance with SOP standards, despite strong regulatory incentives and very low costs of doing so. Further, impacts for private facilities are higher in markets where there were more public facilities, suggesting an important role of public facilities in the market (Table A11).

We emphasize that these results do not imply a zero role for regulatory incentives, but rather that facilities invested in ways that went *beyond* the regulatory incentives, potentially driven by

beliefs over other facility's beliefs. These models also do not include the public sector, which accounts for 71% of the market share in our setting. The improvement in the public sector can be modelled as "exogenous" with implications for other private facilities, but this does not address the question of why the public sector improved in the first place.

market rewards in the case of private facilities.

Result 2: No impact of additional information: We have shown in Table 2 that facilities with low JHIC scores have lower prices, lower market shares and are more likely to exit the market. This already suggests that there must be some information in the market regarding the quality of health facilities. We now provide additional evidence that the impacts we observe on quality were not driven by additional patient information.

Recall that our intervention divided treatment markets into those who received inspections only and those who received inspections and information. In the second arm, inspectors posted a scorecard with the result of the inspection, while the first kept the results private. If the source of the market failure was a lack of patient information that allowed the community to hold health workers accountable (like in Björkman Nyqvist et al. (2017)), we should find that the impact is driven by the arm with the scorecard. In fact we find exactly the same treatment effects across both arms (Table 7).

Perhaps the information treatment did not have any additional impact because the report cards did not improve patient information—Table A12 shows, for instance, that even though patients in treatment markets understood the scoring system, only 8 percentage points more patients actually noticed the scorecard (versus control) despite a fairly extensive dissemination effort. However, it is then difficult to ascribe the impact of the intervention as a whole to an improvement in information because the arm with *less* information saw just as much of an improvement as the arm with the report cards. Further, the report card intervention did improve the awareness of the scorecards by 58 percentage points in T2 (p-value < 0.01) among facility in-charges. If information was indeed a binding constraint, an external, verifiable certification should have provided sufficient incentive for facilities to improve quality and advertise their services. This did not happen.

Result 3: Heterogeneity by market size and across the quality distribution: Our final set of results explores further potential heterogeneity across the outcome distribution in patient safety using quantile treatment effects. Appendix Figure A8 shows the distribution of the (endline) JHIC score in private and public facilities in treated and control markets. In both public and private facilities, there is a clear shift of the distribution towards higher quality and an equally clear decline in the fraction of facilities with very low JHIC scores. This is consistent with the aims of the regulation. What is striking though, is the increase in the fraction of facilities with very high scores relative to control; for the private sector, it appears that the increases in the JHIC score are just as marked at the top of the distribution as at the bottom.

Figure A9 investigates this formally using unconditional quantile treatment effects and confirms that there are significant impacts across the entire distribution of JHIC score, but higher impacts on the top part of the distribution. Figure 3 then shows conditional quantile treatment effects by market size group (1-2, 3-10 or 11+ health facilities) at percentiles 10th, 25th, 50th, 75th and 90th. Again, the intervention increased JHIC scores at the upper quantiles of the safety distribution

more than the lower quantiles within each market size group, and particularly so for markets with greater competition, as measured by the number of facilities. Interestingly, the differences between the lowest and highest quantiles are larger and more precisely estimated for private facilities.[30]

Table 7: Treatment Effects on JHIC Score, OOP, Outpatients, and Entry by Treatment Groups

	Unweighted		Weighted			
	JHIC Score (pp of max) (1)	OOP (USD PPP) (2)	JHIC Score (pp of max) (3)	OOP (USD PPP) (4)	Daily Outpatients (5)	New (6)
Inspection Only (T1)	5.435***	0.917**	4.193***	0.174	1.421	-0.024
	(1.112)	(0.438)	(1.582)	(0.607)	(2.180)	(0.025)
	[0.000]	[0.037]	[0.009]	[0.775]	[0.515]	[0.328]
Inspections plus Information (T2)	4.924***	1.020**	3.686***	0.106	1.537	0.032
	(0.858)	(0.491)	(1.245)	(0.556)	(1.886)	(0.025)
	[0.000]	[0.039]	[0.003]	[0.849]	[0.416]	[0.200]
Observations	1285	1285	1285	1285	1285	1319
R^2	0.317	0.127	0.517	0.178	0.247	0.054
Control Mean	35.493	4.069	42.526	3.136	20.793	0.133
T1 Impact: {%; SD}	{15%; 0.51}	{23%; 0.19}	{10%; 0.35}	{6%; 0.04}	{7%; 0.05}	{-18%; -0.07}
T2 Impact: {%; SD}	{14%; 0.46}	{25%; 0.21}	{9%; 0.31}	{3%; 0.02}	{7%; 0.05}	{24%; 0.09}
Test (T1)=(T2) (p-value)	0.629	0.805	0.633	0.849	0.955	0.021

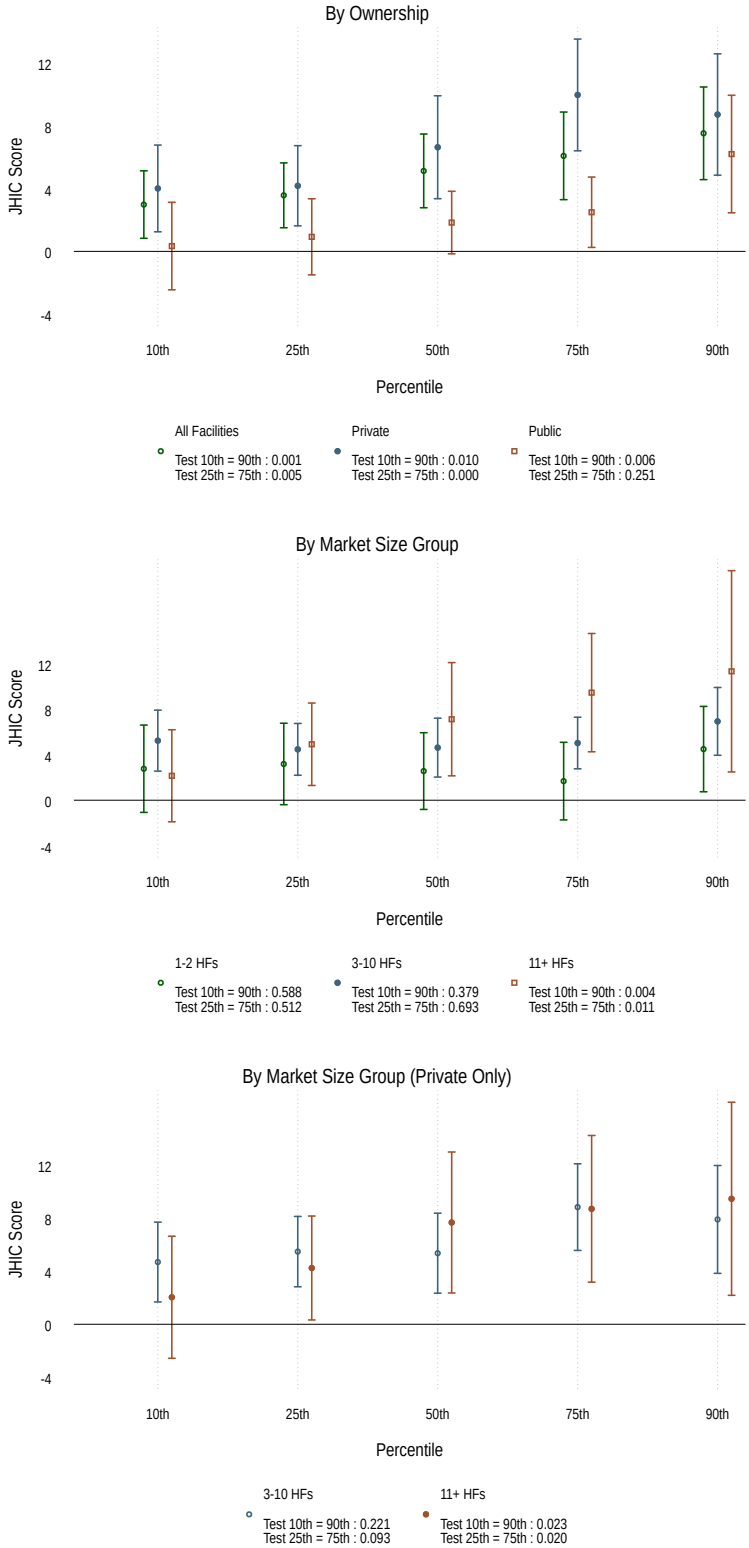
Notes. Robust standard errors are reported in parentheses and clustered at the market level. P-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. Missing values for OOP in 5.8% of observations are imputed using means defined by level, ownership, treatment, license status at randomization, and daily outpatients. Regressions include randomization strata controls (by county and market size) and health facility level controls. JHIC = Joint Health Inspection Checklist; OOP = out-of-pocket payments; PPP = purchasing power parity.

The quantile treatment effects indicate improvements at the higher end of the quality distribution, but because they do not tell us which facilities improved, we cannot link facility improvements to regulatory incentives. Our second exercise therefore assesses the importance of the threat of government closures as a channel for our results. The idea is that if facilities have those characteristics that make them more likely to be closed by the government, a regulatory channel would suggest that they should also have more incentive to improve.

To assess this possibility, we use a logit model to predict a facility's likelihood of closure by the government as a function of pre-treatment or fixed characteristics for all private facilities at randomization (see Supplemental Material Section 6 for variable construction details). We then classify facilities into three groups based on their predicted probability of closure: (i) 57% (536 facilities) are classified as "Low" with closure probability equal or less than 0.4; (ii) 20% (187 facilities) as "Mid" with closure probability greater than 0.4 and less 0.6 and; (iii) 23% (219 facilities) as "High" with closure probability equal or greater than 0.6. The mean predicted probability of

[30]Table S16 in the Supplemental Material shows similar analyses using unconditional quantile treatment effects, with similar qualitative results. Table S8 in the Supplemental Material also shows that there is no statistically significant correlation between the market size, and the average JHIC score at the market level at baseline, and, for treatment markets, there is no significant correlation between market size at randomization and the month of first inspection visit in the market, or the average number of inspections per facility in the market.

Figure 3: Conditional QTE on JHIC Score, by Ownership and Market Size



Notes. Vertical lines correspond to 95% confidence intervals. In the third panel, we exclude the estimates for private facilities of sizes 1-2 which are based on too small a sample to allow for convergence. Regressions include controls for the county and health facility level.

closure is 11% in the Low group, followed by 50% and 64% in the Mid and High groups. However, because many facilities that were closed by the government reopened subsequently, we have endline data for 453 facilities in the Low group (85% of those listed at randomization), 120 (64%) in the Mid and 145 (66%) in High groups. Having classified facilities by their propensity to be closed by the government, we then assess how the treatment effects vary by the this propensity, following Equation 2 and using a leave-one-out estimator as in Abadie et al. (2018) to reduce the risk of over-fitting.

Figure 4 and Table S17 in the Supplemental Material show that "Low" treated facilities report the *largest* increase in the JHIC score by 7.6 percentage points (0.72 SD, p-value < 0.010), while "High" facilities reported an increase of 5.5 percentage points (0.63 SD, p-value < 0.010). Observed gains are only for surviving facilities. Since surviving facilities are those that improved the *most*, especially for the "High" group, even this smaller improvement is likely overestimated so that the actual differences are even starker.[31] As with the quantile treatment effects, it is the facilities with the lowest probability of government closure–who are also those with high JHIC scores in the baseline–that improve the most.

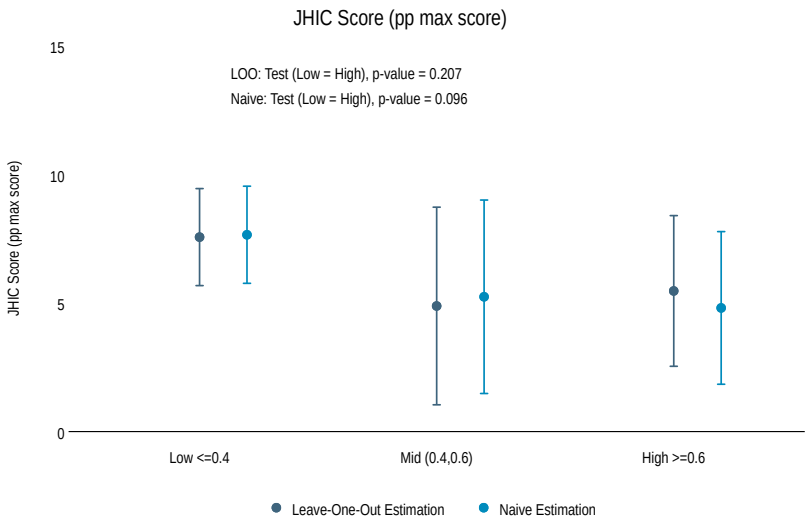
Discussion: Our results are consistent with the idea that MQS drives improvement across the range of quality with private clinics readjusting their positions to maintain market power in response to improvements in other facilities. An important alternate explanation is that the checklist provided feedback that led to improvements, with greater improvements among altruistic providers who were better to begin with. This explanation leaves unchanged our finding that MQS leads to changes throughout the quality distribution, including in facilities that faced little regulatory incentives.

Nevertheless, feedback alone seems to be insufficient to explain our results. Previous evaluations that provided feedback showed null to relatively small improvements and, in our case, both control and treatment facilities had access to the checklist (King et al., 2021; Dunsch et al., 2022). Our treatment effects thus net out the effects of giving facilities in the control group the necessary materials for improvement.

It is still possible that it was individual in-person feedback that mattered and in fact, Brock et al. (2016) and Leonard & Masatu (2017) demonstrate that such feedback can improve clinical processes even 18 months after it was given. Importantly, they also show that improvement is not correlated with altruism as measured by performance in a dictator game (although clinical quality is) and that it requires multiple visits for feedback to impact performance. A single visit and

[31] Another way to see the same result is to focus on a group of facilities with very small likelihood of closure. For instance, we see only 5 closures among facilities with JHIC scores above 40 and if we were to use rational expectations, this group would have faced virtually zero regulatory incentives to improve. Nevertheless, we again see large improvements of 6 percentage points in the JHIC score for this group, compared to control. Similarly, we see 3 closures of facilities with JHIC scores above 50 and we see improvements of 6.9 percentage points in the JHIC score for this group, compared to control (Tables S18a and S18b in the Supplemental Material).

Figure 4: Treatment Effects on JHIC Score
by Closure Probability Group at Randomization



Notes. Vertical lines correspond to 95% confidence intervals. Regressions include controls for the 16 strata included in the randomization (by county and market size), health facility level, and baseline market controls for JHIC, OOP, and outpatients. The table corresponding to this figure can be found in Supplemental Material Table S17.

several hours of observation leads to an immediate improvement and an equally rapid fade-out of performance (Leonard & Masatu, 2006). Our result that the top of the distribution improved the most thus seems inconsistent with the existing literature on the links between altruism, feedback and clinical performance, although we caution that we cannot fully rule-out this alternate channel.

Our results also provide the first evidence that bringing public sector facilities under a uniform government regulation can lead to quality improvements without any further investments. There is little previous evidence on this in the health literature; farther afield, the education literature has posited a positive role for school inspections (Muralidharan et al., 2017; Ehren et al., 2013), but again, with little experimental evidence in support.

One potential reason for the improvement we see in public clinics may be linked to the devolution of responsibilities under Kenya's 2010 constitution, under which each of the counties became responsible for the functioning of their public clinics. Multiple studies show that counties improved access to healthcare and infrastructure in public clinics after devolution (Masaba et al., 2020). Formal models of bureaucracy take seriously the problems of communication within hierarchies with results showing how inefficient outcomes may obtain, for instance, due to the emergence of cheap-talk equilibrium (Gailmard & Patty (2012) present an overview). Inspections in this context present verifiable information to the politician by a third party—the federal government—rather than the facility that requires the resources and may have thus helped alleviate the concerns arising

from strategic communications.

VI Conclusion

Health markets in Kenya are characterized by a public sector with 70% market share and a private sector that is highly varied in quality, with some very low-quality and unlicensed providers who enter and exit the market frequently. This group accounts for 12% of facilities but a small share of patients (3%). The ubiquity of these clinics prompted an important regulatory reform, establishing a minimum quality standards (MQS) that was uniformly implemented for both public and private sector health facilities. We draw three overarching conclusions from the experimental evaluation of this reform.

First, regulation and inspections without additional resources can lead to improvements, establishing a positive role for MQS within the health sector. Second, improvements for the average patient are driven by within-facility changes rather than re-allocation of patients across facilities or the exit of low-quality facilities. Third, we find a diminished role for information as a market failure, which is consistent with baseline patterns showing that quality is rewarded through higher prices and market share.

Coupled with improvements in the public sector, this opens up the possibility that MQS can lead to improvements in quality across the distribution, which are critical because the market share of the lowest quality facilities is very low and low entry costs imply that the costs of regulation among this group are very high. If quality improvements had occurred only among the lowest performing facilities, the impacts on patients would have been quite limited. Instead, bringing the public sector into the regulatory framework and allowing for the possibility that regulation can affect the entire market could lead to significant improvements.

References

Abadie, A., Chingos, M. M., & West, M. R. (2018, October). Endogenous Stratification in Randomized Experiments. *The Review of Economics and Statistics*, *100*(4), 567–580.

Anderson, M. L. (2008, December). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, *103*(484), 1481–1495.

Andrabi, T., Das, J., & Khwaja, A. I. (2017, June). Report cards: the impact of providing school and child test scores on educational markets. *American Economic Review*, *107*(6), 1535–1563.

Andrabi, T., Das, J., Khwaja, A. I., Ozyurt, S., & Singh, N. (2020, October). Upping the ante: the equilibrium effects of unconditional grants to private schools. *American Economic Review*, *110*(10), 3315–3349.

Arudo, J., Gimnig, J. E., ter Kuile, F. O., Kachur, S. P., Slutsker, L., Kolczak, M. S., ... Phillips-Howard, P. A. (2003, April). Comparison of government statistics and demographic surveillance to monitor mortality in children less than five years old in rural western Kenya. *The American Journal of Tropical Medicine and Hygiene*, *68*(4), 30–37.

Ashraf, N., Bandiera, O., Davenport, E., & Lee, S. S. (2020, May). Losing prosociality in the quest for talent? Sorting, selection, and productivity in the delivery of public services. *American Economic Review*, *110*(5), 1355–1394.

Banerjee, A., Das, J., Hammer, J., Hussam, R., & Mohpal, A. (2020). *The Market for Healthcare in Low Income Countries - Working Paper - Faculty & Research - Harvard Business School.*

Bedoya, G., Bittarello, L., Davis, J., & Mittag, N. (2017, July). *Distributional Impact Analysis: Toolkit and Illustrations of Impacts Beyond the Average Treatment Effect* (Working Paper). Washington, DC: World Bank.

Bedoya, G., Das, J., Dolinger, A., De Guttry, R., Hur, Y. S., & Lee, J. Y. (2020). Regulation for Safety and Quality of Care : A Process Evaluation of the Health Inspection Pilots of the Kenya Patient Safety Impact Evaluation. *World Bank*.

Bedoya, G., Dolinger, A., Rogo, K., Mwaura, N., Wafula, F., Coarasa, J., ... Das, J. (2017, July). Observations of infection prevention and control practices in primary health care, Kenya. *Bulletin of the World Health Organization*, *95*(7), 503–516.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006, September). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507.

Bennett, D., & Yin, W. (2019, March). The Market for High-Quality Medicine: Retail Chain Entry and Drug Quality in India. *The Review of Economics and Statistics*, *101*(1), 76–90.

Björkman Nyqvist, M., de Walque, D., & Svensson, J. (2017, January). Experimental Evidence on the Long-Run Impact of Community-Based Monitoring. *American Economic Journal: Applied Economics*, *9*(1), 33–69.

Blau, D. M. (2003). Do child care regulations affect the child care and labor markets? *Journal of Policy Analysis and Management*, *22*(3), 443–465.

Blau, D. M. (2007, June). Unintended consequences of child care regulations. *Labour Economics*, *14*(3), 513–538.

Borghi, J., Little, R., Binyaruka, P., Patouillard, E., & Kuwawenaruwa, A. (2015, March). In Tanzania, The Many Costs Of Pay-For-Performance Leave Open To Debate Whether The Strategy Is Cost-Effective. *Health Affairs*, *34*(3), 406–414.

Bosk, C. L., Dixon-Woods, M., Goeschel, C. A., & Pronovost, P. J. (2009, August). Reality check for checklists. *The Lancet*, *374*(9688), 444–445.

Brock, J. M., Lange, A., & Leonard, K. L. (2016, January). Generosity and Prosocial Behavior in Healthcare Provision: Evidence from the Laboratory and Field. *Journal of Human Resources*, *51*(1), 133–162.

Bruhn, M., & McKenzie, D. (2008, October). In Pursuit of Balance : Randomization in Practice in Development Field Experiments. *World Bank*.

Chandra, A., Finkelstein, A., Sacarny, A., & Syverson, C. (2016, August). Health Care Exceptionalism? Performance and Allocation in the US Health Care Sector. *American Economic Review*, *106*(8), 2110–2144.

Chege, T., Wafula, F., Tama, E., Khayoni, I., Ogira, D., Gitau, N., & Goodman, C. (2022, March). *How much does effective health facility inspection cost? An analysis of the economic costs of Kenya's Joint Health Inspection innovations.*

Chen, B. K., Gertler, P. J., & Yang, C.-Y. (2016, December). Physician ownership of complementary medical services. *Journal of Public Economics*, *144*, 27–39.

Chipty, T., & Witte, A. D. (1997, July). *An empirical investigation of firms' responses to minimum standards regulations.* National Bureau of Economic Research.

Contreras-Loya, D., Gertler, P., & Kwan, A. (2021). Managerial Practices and Altruism in Health Care Delivery. *NBER Working Paper*.

Dafny, L., Ho, K., & Lee, R. S. (2019, June). The price effects of cross-market mergers: theory and evidence from the hospital industry. *The RAND Journal of Economics*, *50*(2), 286–325.

Daniels, B., Dolinger, A., Bedoya, G., Rogo, K., Goicoechea, A., Coarasa, J., ... Das, J. (2017, June). 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).

Das, J., Daniels, B., Ashok, M., Shim, E.-Y., & Muralidharan, K. (2022, May). Two Indias: The structure of primary health care markets in rural Indian villages with implications for policy. *Social Science & Medicine*, *301*, 112799.

Das, J., Holla, A., Das, V., Mohanan, M., Tabak, D., & Chan, B. (2012, December). In Urban And Rural India, A Standardized Patient Study Showed Low Levels Of Provider Training And Huge Quality Gaps. *Health Affairs*, *31*(12), 2774–2784.

De Allegri, M., Makwero, C., & Torbica, A. (2019, May). At what cost is performance-based financing implemented? Novel evidence from Malawi. *Health Policy and Planning*, *34*(4).

Dizon-Ross, R., Dupas, P., & Robinson, J. (2017, December). Governance and the effectiveness of public health subsidies: Evidence from Ghana, Kenya and Uganda. *Journal of Public Economics*, *156*, 150–169.

Duflo, E., Greenstone, M., Pande, R., & Ryan, N. (2018). The value of regulatory discretion: estimates from environmental inspections in india. *Econometrica*, *86*(6), 2123–2160.

Dunsch, F. A., Evans, D. K., Eze-Ajoku, E., & Macis, M. (2022, February). Management, supervision, and healthcare: A field experiment. *Journal of Economics & Management Strategy*.

Ehren, M. C. M., Altrichter, H., McNamara, G., & O'Hara, J. (2013, February). Impact of school inspections on improvement of schools—describing assumptions on causal mechanisms in six European countries. *Educational Assessment, Evaluation and Accountability*, *25*(1), 3–43.

Flodgren, G., Gonçalves-Bradley, D. C., & Pomey, M.-P. (2016, December). External inspection of compliance with standards for improved healthcare outcomes. *Cochrane Database of Systematic Reviews*.

Flodgren, G., Pomey, M.-P., Taber, S. A., & Eccles, M. P. (2011, November). Effectiveness of external inspection of compliance with standards in improving healthcare organisation behaviour, healthcare professional behaviour or patient outcomes. *Cochrane Database of Systematic Reviews*.

Foster, L., Haltiwanger, J., & Syverson, C. (2008, March). Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability? *American Economic Review*, *98*(1).

Foster, L., Haltiwanger, J. C., & Krizan, C. J. (2001, January). Aggregate Productivity Growth: Lessons from Microeconomic Evidence. In *New Developments in Productivity Analysis* (pp. 303–372). University of Chicago Press.

Gailmard, S., & Patty, J. W. (2012, June). Formal Models of Bureaucracy. *Annual Review of Political Science*, *15*(1), 353–377.

Gatti, R., Andrews, K., Avitabile, C., Conner, R., Sharma, J., & Yi Chang, A. (2021). *The Quality of Health and Education Systems Across Africa: Evidence from a Decade of Service Delivery Indicators Surveys*. Washington, DC: World Bank.

Hotz, V. J., & Xiao, M. (2011, August). The Impact of Regulations on the Supply and Quality of Care in Child Care Markets. *American Economic Review*, *101*(5), 1775–1805.

Jain, R. (2022, October). Private hospital behavior under government insurance:evidence from reimbursement changes in india. *Forthcoming*.

Jain, R., & Dupas, P. (2022, October). Can beneficiary information improve hospitalaccountability? experimental evidence from a publichealth insurance scheme in india. *Forthcoming*.

King, Powell-Jackson, T., Makungu, C., Spieker, N., Risha, P., Mkopi, A., & Goodman, C. (2021, September). Effect of a multifaceted intervention to improve clinical quality of care through stepwise certification (SafeCare) in health-care facilities in Tanzania: a cluster-randomised controlled trial. *The Lancet Global Health*, *9*(9), e1262–e1272.

King, J. J. C., Powell-Jackson, T., Makungu, C., Hargreaves, J., & Goodman, C. (2021, June). How much healthcare is wasted? A cross-sectional study of outpatient overprovision in private-for-profit and faith-based health facilities in Tanzania. *Health Policy and Planning*, *36*(5), 695–706.

Klein, B., & Leffler, K. B. (1981, August). The Role of Market Forces in Assuring Contractual Performance. *Journal of Political Economy*, *89*(4), 615–641.

Kovacs, R. J., Lagarde, M., & Cairns, J. (2022, February). Can patients improve the quality of care they receive? Experimental evidence from Senegal. *World Development*, *150*, 105740.

Kwan, A., Boone, C. E., Sulis, G., & Gertler, P. J. (2022, March). Do private providers give patients what they demand, even if it is inappropriate? A randomised study using unannounced standardised patients in Kenya. *BMJ Open*, *12*(3), e058746.

Leland, H. E. (1979, December). Quacks, Lemons, and Licensing: A Theory of Minimum Quality Standards. *Journal of Political Economy*, *87*(6), 1328–1346.

Leonard, & Masatu, M. C. (2006, November). Outpatient process quality evaluation and the Hawthorne Effect. *Social Science & Medicine*, *63*(9), 2330–2340.

Leonard, & Masatu, M. C. (2017, May). Changing health care provider performance through measurement. *Social Science & Medicine*, *181*, 54–65.

Li, S. (2020). *How we measured the value of a statistical life in kenya and ghana.*

Masaba, B., Moturi, J., Taiswa, J., & Mmusi-Phetoe, R. (2020, December). Devolution of healthcare system in Kenya: progress and challenges. *Public Health*, *189*, 135–140.

McKenzie, D., & Paffhausen, A. L. (2019, October). Small Firm Death in Developing Countries. *The Review of Economics and Statistics*, *101*(4), 645–657.

Miguel, E., & Kremer, M. (2004, January). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica*, *72*(1), 159–217.

Ministry of Health. (2015). *Ministry of Health Implementation Guidelines for the Joint Health Inspection Checklist.*

Mohanan, M., Vera-Hernández, M., Das, V., Giardili, S., Goldhaber-Fiebert, J. D., Rabin, T. L., ... Seth, A. (2015, April). The Know-Do Gap in Quality of Health Care for Childhood Diarrhea and Pneumonia in Rural India. *JAMA Pediatrics*, *169*(4), 349.

Muralidharan, K., Das, J., Holla, A., & Mohpal, A. (2017, January). The fiscal cost of weak governance: Evidence from teacher absence in India. *Journal of Public Economics*, *145*, 116–135.

National Academies of Sciences. (2018). *Crossing the Global Quality Chasm: Improving Health Care Worldwide*. Washington, D.C.: National Academies Press.

Pittet, D., Hugonnet, S., Harbarth, S., Mourouga, P., Sauvan, V., Touveneau, S., & Perneger, T. V. (2000, October). Effectiveness of a hospital-wide programme to improve compliance with hand hygiene. *The Lancet*, *356*(9238), 1307–1312.

Romano, J. P., & Wolf, M. (2010, February). Balanced control of generalized error rates. *The Annals of Statistics*, *38*(1).

Ronnen, U. (1991). Minimum Quality Standards, Fixed Costs, and Competition. *The RAND Journal of Economics*, *22*(4), 490.

Shapiro, C. (1986, October). Investment, Moral Hazard, and Occupational Licensing. *The Review of Economic Studies*, *53*(5), 843.

Siam, Z. A., McConnell, M., Golub, G., Nyakora, G., Rothschild, C., & Cohen, J. (2019, July). Accuracy of patient perceptions of maternity facility quality and the choice of providers in Nairobi, Kenya: a cohort study. *BMJ Open*, *9*(7), e029486.

Tama, E., Khayoni, I., Goodman, C., Ogira, D., Chege, T., Gitau, N., & Wafula, F. (2021, August). What Lies Behind Successful Regulation? A Qualitative Evaluation of Pilot Implementation of Kenya's Health Facility Inspection Reforms. *International Journal of Health Policy and Management*.

WHO. (2006). *The world health report : 2006 : working together for health.*

WHO. (2011). *Report on the Burden of Endemic Health Care-Associated Infection Worldwide.*

WHO. (2021). *WHO SCORE for health data kenya assessment summary.*

World Bank. (2020). *Population, total - lower middle income.*

Yi, H., Miller, G., Zhang, L., Li, S., & Rozelle, S. (2015, August). Intended And Unintended Consequences Of China's Zero Markup Drug Policy. *Health Affairs*, *34*(8), 1391–1398.

Zeng, W., Shepard, D. S., Nguyen, H., Chansa, C., Das, A. K., Qamruddin, J., & Friedman, J. (2018, November). Cost–effectiveness of results-based financing, Zambia: a cluster randomized trial. *Bulletin of the World Health Organization*, *96*(11), 760–771.