Randomized Regulation:
The impact of minimum quality standards on health markets*

Guadalupe Bedoya[†] Jishnu Das[‡] Amy Dolinger[†]

Version: September 2022

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, the census of 1348 health facilities in three Kenyan counties were classified into 273 markets, and the markets were then randomly allocated to treatment and control groups. Government inspectors visited health facilities in treated markets 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.

JEL Codes: H75, I11, I18, 017

*Bedoya: e-mail: gbedoya@worldbank.org; Das: e-mail: jishnu.das@georgetown.edu; Dolinger: e-mail: adolinger@worldbank.org. 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, Ana Goicoechea, Arianna Legovini, Nikolas Mittag, Edit Velenyi, as well as seminar participants at UCLA, Georgetown University, Warwick University, 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.

[†]Development Impact Evaluation (DIME) department, World Bank

[‡]McCourt School of Public Policy and Walsh School of Foreign Service, Georgetown University

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. 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) and changed the content, frequency and consequences of facility inspections. It first established a standardized inspection protocol through the creation of a "joint" health inspection checklist or JHIC that all boards and councils agreed to use. The JHIC was used to assess the facility's compliance with patient safety protocols. Previously, inspections had not used a standardized scored rubric. Second, it replaced an earlier system of infrequent and ad-hoc inspections implemented by regular members of each regulatory board with a specially trained cadre of full-time health inspectors who carried out periodic inspections. Finally, the scores generated through the JHIC triggered well-defined warnings and sanctions according to a facility's level of risk. These ranged from immediate closure for unlicensed or very low scoring facilities to less frequent inspections for those with higher scores.

[1] See, for instance, World Health Organization (2006). Two systematic reviews on the impacts of healthcare regulation (Flodgren et al., 2011, 2016), found only two studies that met the eligibility criteria. 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, Gertler, & Yang (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; 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 LIC. See Das et al. (2012), Banerjee et al. (2020), Mohanan et al. (2015), Daniels et al. (2017), Kovacs et al. (2022), J. J. C. 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.

With cabinet approval, we implemented this new regulation in an experimental manner in three counties across the country (Meru in the center, Kakamega in the lakes region and Kilifi in the East coast). These counties 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.

We coupled the experimental allocation of the MQS regulation with a market-level randomization, where we first allocated all 1348 health facilities in the census of the 3 counties, including unlicensed providers, 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. A red closure scorecard was used in both treatment arms (T1 and T2) in cases where the facility did not have appropriate licenses, scored between 0-10% on the JHIC or failed to comply with minimum standards according to the regulation after three visits. This market-level allocation of experimental treatments allows us to estimate the causal effects of the regulation on 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 JHIC score as well as patient volume, and prices.

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 facilities (0.80 SD vs. 0.50 SD for unlicensed) and for facilities that had been in the program longer (0.50-0.65 SD). We highlight that improvements of 0.31 SD in the public sector were also substantial and an important first 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 actually in facility infrastructure, equipment, and supplies—all of which required substantial investments.

We then show that the intervention meaningfully altered the market structure. In treated markets, private facilities that were unlicensed at baseline were 10.5 percentage points more likely to exit, and visits to public facilities increased by 18% at the expense of private facilities. 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 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 our market-level randomization show that within-facility improvements accounted for 92% of the improvement, with another 6% attributable to the exit of facilities with lower than the mean market quality. We thus conclude that the 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 discuss potential mechanisms 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). We establish that (a) facilities invested in improvements that were (far) more costly than what was required and were not optimizing decisions to meet compliance thresholds; (b) there was no difference in treatment outcomes between the inspection only and the inspection + information arms and; (c) quantile treatment effects by market density show that impacts were highest at the *top-end* of the distribution of patient safety, which was plausibly least affected by regulatory requirements, and in markets with greater competition.

These results have important implications for theory and policy at the intersection of regulation and market structure in LLMICs. In terms of the theory, our results suggest that facilities' investment decisions went beyond the letter of the law, potentially responding to (derived) demand in markets with multiple facilities.[4] Surprisingly, given that fundamental problems of healthcare are often tied to a poor informational environment, we are unable to find a clear informational channel; our results instead elevate the relative importance of a market-power based explanation, like in Ronnen (1991). This could be because patient safety is one of the few dimensions of quality that is broadly observable and unidirectional—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.

In terms of policy, our results 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

[4] Our results on the mechanism are more speculative as 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. 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.

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 alone can improve patient safety without decreasing utilization. This contrasts with more common and expensive models of mentoring and financial assistance in the health sector that surprisingly yield worse results. Two previous experimental evaluations sought to improve patient safety measures similar to ours using mentoring and supervision. One evaluation of primary public facilities in Nigeria finds no impacts on quality of care one year after the intervention (Dunsch et al., 2022) compared to a 3.1 percentage point increase (or 8% versus control) in our case. The other targets private formal facilities in Tanzania, reporting a 4.4 pp increase or 8.5% over control facilities (King et al., 2021) vs. 8.8 pp (or 23%) 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 over $8000, which is multiple times what our intervention cost (King et al., 2021).

Third, the study allays the concern that even if MQS regulation improves quality, it does so by hurting the poor as the cost of care, either in terms of distance or price, increases (Leland, 1979; Shapiro, 1986; Klein & Leffler, 1981). We find that quality increased across the board without increases in prices for the average patient or declines in utilization. This is consistent with theoretical

[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 low- and lower-middle-income countries, 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 randomization.

[7]We compare the same types of facilities (public or licensed private/NGOs), surviving at endline, since both studies randomized at the facility level. In another intervention to improve quality in Kenya's private sector, Contreras-Loya et al. (2021) also find small effects of a large and costly intervention designed to improve business management and care delivery on healthcare quality, although it increases facility investments.

predictions from the literature on vertically differentiated oligopolies, mediated in our case by the presence of the public sector.

While we thus make substantial progress in understanding the impacts of regulation, an important limitation of our study is the lack of data on health outcomes or iatrogenic diseases. These require expensive patient follow-ups, biological tests and/or tests of surface contamination in health facilities. While the measures we use were developed by organizations because they are believed to be correlated with outcome measures, we do not have the data from this study to support that hypothesis. With that limitation in mind, 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.

II Setting and Context

Healthcare in Kenya is delivered through public (61%) and private (39%) facilities, with public facilities managed independently by each of 47 counties following a process of devolution of responsibilities in 2010.[8] These facilities are further divided into levels; Levels 2 and 3 are primary care facilities and Level 4 hospitals and above offer increasingly advanced care. Patients are free to choose what facility they visit; prices in public facilities are substantially lower and private facilities can set their own prices.

Most health facilities operate in a setting with some competition. In our study, 79% of all health facilities are in markets with 4 or more facilities (we define "market" more precisely in Section III); 15% in markets with 2 to 3 facilities, and 7% are the only ones in their market. These "singleton" facilities tend to be publicly-owned and located in rural areas. Consequently, 88% of markets have at least one public provider, implying that even if all private sector facilities are closed, a public sector option will be available in virtually all cases. Mirroring the market structure, 70% of patients in the 3 study counties seek care in markets with 4 or more providers and 11% seek care from singleton facilities. The distribution of markets in the 3 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 Online Appendix).[9]

The regulation of patient safety is a responsibility of the national government carried out by nine "Boards and Councils," each responsible for a different facet of healthcare delivery (for instance, the Kenya Medical Practitioners and Dentists Board deals with licensing of the majority of 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

[8]Figures correspond to data from the Kenya Master Facility List before the intervention.

[9]We are comparing data that we collected in the study counties with administrative data in other counties. Therefore, differences could also reflect the undercounting of unlicensed providers in the administrative data.

basis in small geographic regions based on the quota for the inspection period, or by individual boards and councils, usually following a complaint or a serious adverse event: 4% of facilities were inspected annually and the likelihood of two inspections in one year was zero.[10] This process led to considerable arbitrariness in the implementation of inspections and the resulting sanctions; Section 2 of the Online Appendix provides details on the inspection system under the old regime.

Finally, the quality of care in Kenyan facilities is among the best in the region.[11] Nevertheless, concerns around patient safety were raised after a national survey in 2012 found that only 2% of health facilities were compliant with minimum patient safety protocols and systems (IFC & WHO, 2012). Actual patient observations thankfully suggest a more nuanced situation with variation across specific tasks. For instance, compliance is 87% with safe injections and blood draw practices but only 2% for hand-hygiene. Even then, outpatients faced on average 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).

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, including a Joint Inspection Health Checklist (JHIC) to be used in each visit 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 briefly 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 Online Appendix, Section 2.

The JHIC instrument: The JHIC focuses on input-driven measures of patient safety with 471 individual items across 14 sections (Online Appendix Section 3).[12] Scores are computed by equally weighting each section of the checklist, certain subsections, and components within subsections,

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

[11] Kenya ranks among the best in the region on indicators such as health care worker knowledge of clinical guidelines and health facility equipment availability as per the World Bank's Service Deliver Indicators (Gatti et al., 2021) and providers perform better for select tracer conditions in comparisons with India and China (Daniels et al., 2017).

[12] 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.

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 the 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 weights just as much as introducing a costly waste management system.

Sanction and Warning System: Following an inspection, sanctions and warnings are graded by the JHIC score. Facilities scoring less than 10% or those without a valid license to operate are categorized as "non-compliant" and recommended for immediate closure. Facilities scoring 11-40% are considered "minimally compliant" and receive a 3-month notice for correction and re-inspection, while facilities with scores between 41-60% are classified as "partially compliant" and receive a 6-month notice for correction and re-inspection. For these two categories, facilities are closed if they do not improve to a higher category by the third inspection. Once they reach the next category, the cycle resets and they start a new notice for correction and re-inspection. Facilities that score above 60% do not any face 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 be minimally compliant 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 with the top 12 candidates selected for the intervention. 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). Indeed, a third-party qualitative assessment, separate from our team, showed that the inspectors were highly valued by the facilities they inspected as a source of important feedback (Tama et al., 2021). 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.[13] 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

[13]76% of facility in-charges in treated markets provided comments on their experience with inspections in the endline survey. Of these, only 2% of the comments were related to corruption. In Addition, random inspection quality checks performed during the implementation, show small discrepancies with inspectors results.

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

The experimental design had three components: The construction of markets, the allocation of treatment and control arms and the timing of inspections. We describe each briefly.

Construction of Markets: We started with a census of 1,258 facilities that we could locate in the 3 countries between January and September 2015, and updated the census between October and November 2016 (Section 1 of the Online Appendix).[14] We then defined a market using a "z-center" clustering algorithm and the fact that in the baseline 70% of respondents lived within 4 km of the health facility. The algorithm assigned facilities to markets such that no facility was more than 4km from the centroid of its assigned market, where, the centroid was computed recursively from the location of all facilities mapped to the market. 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 S1 in the Online Appendix), 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 inspections in the year of the intervention.

[14] Between implementation and endline, we identified an additional 90 facilities that were missed at randomization or were temporarily or permanently closed when the census was ongoing, and were listed later. They were assigned to the closest market using a closest-neighbor algorithm preserving the 4 km clustering rule. Therefore, in total, there were 1348 facilities in the 273 markets at randomization.

[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] Table S4 in Section 4 of the Online Appendix presents details by treatment arm and county at randomization and endline.

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 A3 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 A4). 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 A4).

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 federal team and county health officials (Panel B in Figure A3 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.

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 fully baseline survey. These are the facilities we used for the randomization.

The endline was completed between March and August 2018. We counted 1,319 facilities and completed the endline in 1,285 facilities for a 97% response rate.[17] During the endline we located 173 new facilities, which we allocated to existing markets using a nearest-neighbor algorithm. 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. Consequently, 90 facilities that opened at some point during the census updates were missed till the endline; these facilities accounted for 2.7% to 4.5% of 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 1348 facilities, which includes the baseline census and the pre-randomization update (1258 facilities) and the facilities listed later that were operational before the intervention (90 facilities).[19] When we estimate impacts on exit/entry, we use all facilities operational at randomization (1348) and/or endline (1319) regardless

[17] At endline all facilities in 5 markets had closed, reducing the total number of markets at endline to 268. We also exclude three facilities that were part of the census but are greater than 4km from our existing markets.

[18] Of these, 23 facilities were temporarily or permanently closed during the initial census listing exercises and found operational later with starting years prior to the intervention based on self-reports by facility in-charges. The range in market share stems from whether or not those facilities are considered operational at baseline.

[19] Section 5 of the Online Appendix describes details of ethical considerations and approvals of the trial.

of whether they have a completed survey. The average time 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 will turn to when we examine the impact of program duration on impact.

III.3 Data Sources and Description of Main Outcomes

Section 6 in the Online Appendix lists the outcome variables and key covariates, along with how they were constructed. In brief, our primary data sources are based on surveys of health facilities and their staff, exit surveys of patients, and direct clinical observations. Sample sizes across endline (baseline) were 1,285 (1,027) health facilities, 11,098 (8,577) patient exit surveys, 2,098 (1,625) healthcare workers, and 19,178 (18,558) clinical observations. We augment these survey data with additional administrative information on licensing status.

Summary statistics, presented in Table 1 show that 70% of facilities in our data were private and 30% public, although higher patient volumes of 49 patients per day in the public facilities implied that they accounted for 71% of all outpatient visits at baseline. Among private providers, 53% either did not have a valid operating license or were operating with an expired license before the intervention.[20] Out-of-pocket expenditures were lower in public facilities at USD 0.7 PPP (USD 0.3 nominal) compared to private facilities at USD 8.4 PPP (USD 3.8 nominal) and a wealth index for patients visiting private facilities is 0.6 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. At baseline, JHIC scores did not differ by market size (Table S17 of the Online Appendix), although public facilities scored 0.67 SD higher.

One aspect of these markets that we had not appreciated fully was the significant churn in the private sector. Of the 301 private facilities in the control group operational at randomization, 57 (19%) had closed 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 low-income countries. 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.

Another 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 as shown in Table A3. In the private sector at baseline, the JHIC score for unlicensed relative to licensed providers was 21% lower. Further, a 1 SD increase in the JHIC score (11.3 percentage points) was correlated with an increase of USD 2 (PPP) in out-of-pocket (OOP) payments, a correlation that remains robust to the inclusion of machine-selected controls (Table 2). Next, JHIC scores and licensing status were

[20] Within the private facilities, we include NGOs (3% of private facilities) and faith-based organizations (13% of private facilities).

strongly correlated with facility exits in the control group.

Finally, vertical differentiation requires a positive price-quality correlation (which we find) but not necessarily a quantity-quality correlation (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. In the public sector, where prices are administratively set and therefore uncorrelated with the JHIC score, we indeed find a strong quantity response. One consequence of this pattern of correlations is that facilities that exit the market by endline represent only a small 2% of all patients in the data (Table S16 of the Online Appendix).

These results strongly suggest that (a) consumers placed a premium on safety as measured by 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 IV, 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 1.561 (p=0.131) (Table A2 of the Appendix). Response rates were 93% at baseline and 97.5% at endline and non-response is balanced between treatment and control at endline with an estimated null difference (p-value = 0.974). At baseline, facilities in treatment markets report a 4 percentage points higher response rate (p-value < 0.001), Table S6 of the Online Appendix Section 4), however, as Table A2 shows, there are no systematic or significant differences between treatment and control groups in the main outcome variables or key covariates.

III.4.2 Compliance with Treatment

Table S5 (Panel A) in the Online Appendix shows that we reached 90% of active facilities in the T1 arm, 85% in the T2 arm and 97% of facilities in the control group did not receive the intervention (3% contamination), compared to our listing at randomization.[21] The reason why some facilities in the treatment arms did not receive an inspection was because they were found (or opened) at some point after the randomization. This is a plausible reflection of how an actual inspection process would work in markets with considerable churn.

[21] By the first inspection, 5% of T1 facilities and 10% of T2 facilities were out of business.

Bedoya, Das, & Dolinger (2020) show that 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 A1). Random quality checks showed that 89% of facilities left the scorecards displayed after the inspection (in T2) and 52% of facilities closed by the government re-opened within a few months, the majority without obtaining a valid license.

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 (on average, inspector absences implied that an equivalent of 6 full-time inspectors conducted the inspections during 13 months of intervention despite a team size of 9), vehicle breakdowns and general strikes (Figure A2).[22] 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, we had to stop just short of the full cycle of interventions, which would have required 2 follow-up visits every 3 months after the first inspection for those in the "minimally-compliant" category. We are able to eventually observe the impact of a full cycle of 3 inspections for 6% of treated facilities.

Second, most closures occurred because facilities did not have operating licenses, rather than a lack of improvement and the time elapsed between the report for closure and enforcement of closure (which required the arrival of a national team) was on average 70 days (vs. a 1-day protocol). Facility in-charges may have realized that enforcement capacity was a constraint, 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:

$$Y_{im} = \alpha + \delta T_m + \sum_{j=1}^{n} \emptyset_j V_j + \gamma X_{im} + \epsilon_{ijm} \tag{1}$$

Here, Y_{im} indicates the outcome of interest for facility i in market m at endline and T_m is an indicator that equals one for facilities in treated markets. 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.[23] X_{im} are facility or

[22]Due to capacity constraints by the B&Cs, while the top 12 inspectors were selected for the intervention only 10 inspectors were seconded and at most a team of 9 conducted inspections at any given time during the intervention.

[23]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 (97.5% of the census of facilities), they are unlikely to differ largely from treatment-on-the-treated effects.

market-level covariates, and ϵ_{im} are time-varying unobserved characteristics. Since we stratified by county-market level groups, we follow Bruhn & McKenzie (2008) and include V_j, which is a dummy variable equal to one if the facility or market is in one of the 16 randomization strata. Standard errors are clustered at the market level, unless otherwise stated. To account for multiple hypothesis testing, we also report sharpened q-values for the main outcomes of interest in braces, following 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 also estimate the heterogeneity of impacts, using the following specification:

$$Y_{ik} = \alpha + \beta_k T_k + \gamma_k T_k W_{ik} + \delta_k W_{ik} + \sum_{j=1}^{n} \emptyset_j V_j + \gamma X_{im} + \epsilon_{ijk} \tag{2}$$

Here, W_{ik} is a binary variable, indicating whether the observation belongs to one of the subgroups over which we are running the heterogeneity analysis, for instance, whether a facility is private or unlicensed. All other notations are similar as in 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 happens to the patient load of unlicensed facilities in regulated 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, as the value for closed facilities can be (correctly) coded as zero. For other characteristics, such as the JHIC score, we will have missing data for the 16% of facilities in the census at randomization that closed 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.

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 arm into a single treatment allocation. Health facilities in the treated markets improve their JHIC score by 5.2 percentage points (0.49 SD, q-value = 0.002), a 15% increase from 36% of the maximum score in the control group (Column 1). There is no significant change in the daily outpatients (increased by 7% compared to 21 patients in the control group;

q-value = 0.602) and no increase in the entry of new facilities (Columns 5 and 6). At the facility level, fees increased by USD 0.93 PPP or 24% (q-value = 0.022) (Column 2). However, when weighted by patient load (Column 4), fee increases are negligible and never statistically significant, suggesting that they were concentrated among facilities with low patient loads. The impact on the weighted JHIC score is also smaller (Column 3), again 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, the JHIC score for private facilities in treated markets is 6.3 percentage points higher (0.58 SD, p-value $<$ 0.010) compared to 2.8 percentage points (0.31 SD, p-value $<$ 0.010) for public facilities, versus their control counterparts (Column 1). Second, the intervention increases daily outpatients in public facilities by 7.8 patients (0.25 SD, p-value = 0.027), which is a 19% increase compared to 41 daily mean outpatients in control, while it decreased daily outpatients in private facilities by 1.5 patients (0.06 SD, p-value = 0.436), a 13% decrease compared to 11.2 in control (Column 5). Again, weighted impacts on OOP are statistically insignificant for patients attending both public and private facilities (Column 4).

Panel C, Table 3 then shows that licensed and unlicensed private facilities both had similarly higher JHIC scores in treated markets relative to control (Column 1). Further, there is no significant difference in the patient load of licensed versus 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.

Finally, 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 the intervention change?

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 Online Appendix shows, several checklist items could be fulfilled simply by printing 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 on item compliance 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.[24]

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 and supplies; for medium cost supplies there was a 44% increase relative to a baseline of 27.6% 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 A5 in the Appendix shows that baseline quality-price correlations by functional category are statistically significant and higher for infrastructure, equipment and supplies, compared to SOPs, and remain robust to the inclusion of machine-selected control.

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 an quality in the control markets, Table A3 now shows private facility exits in treated markets, again by licensing status and by quintiles of JHIC score. We also include an additional column showing the facilities that were closed by the government. We note first that 24% of all private facilities that were open at our 2015 baseline were closed by the government at some time. These facilities were mostly unlicensed (45% versus 7%), 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 11% in the top quintile. 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 are smaller than the closure

[24] 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.

rate: this is because many facilities reopen 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, obtained from regular monitoring data.

In Table 5, we use the census of facilities at randomization and code all outpatients as zero for inactive facilities to estimate a 3.4-percentage point increase (p-value = 0.246) in exits (inactivity) among treated private facilities. This impact is not statistically significant and is zero for public facilities. It is only when we look at licensing status that significant differences arise, with unlicensed facilities 8.8 percentage points (36%, p-value = 0.049), more likely to be inactive in treated compared to control markets. The increased rate of inactivity among facilities that were unlicensed at randomization also appears in a decrease in their outpatient load of 3.1 patients (p-value < 0.001) or 44% compared to an average of 7.1 in control, while we observe a null impact on 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 closed were replaced by new unlicensed facilities or simply reopened without obtaining their licenses.

Table A4 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 < 0.001) and even more so for licensed facilities that report an increase in the JHIC score of 8.8 percentage points (p-value < 0.001), 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 of the mechanisms below.

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 conditions. We construct a wealth index using exit surveys of 11,100 outpatients based on asset ownership following the Demographic and Health (DHS) Survey in Kenya (see variable construction in Section 6 in the Online Appendix). 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 > 0.99). Table A6 in the Appendix present 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 patient safety improved for all types of providers, including unlicensed providers, some of whom become formal as part the intervention. Further, there was a reallocation of patients towards the public sector, which could have mitigated the impact on health facility exits. Access to health care among poorer patients was not reduced by the intervention, suggesting an overall improvement in their quality of care. The results also point to several nuances, the most important of which is that the government closed very few higher quality facilities, especially if they were licensed and that even with closures, outpatient load among unlicensed facilities in treated markets at endline did not decline.

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

The impacts in Table 3 reflect the composite effect of facility improvements, exits, entries and patient reallocation. We now present two further exercises. To understand the relative importance of each of these, we first decompose the observed average gains for patients in the JHIC score into its separate components, stressing that this is an accounting decomposition, and leverage the market-level random design to estimate the impact of the regulation on the different components. Then, we focus further on closures and how they affect relative patient loads, taking selection into exits seriously.

IV.3.1 Decomposition of Patient Safety Changes

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

$$\begin{aligned}\Delta\bar{q}_{mt} = {} & \underbrace{\Sigma_{mh\in C_t}\theta_{mh,t-1}\Delta q_{mh,t}}_{\text{within}} + \underbrace{\Sigma_{mh\in C_t}\Delta\theta_{mh,t}\left(q_{mh,t-1}-\bar{q}_{mt-1}\right)}_{\text{between}} + \\ & \underbrace{\Sigma_{mh\in C_t}\Delta\theta_{mh,t}\Delta q_{mh,t}}_{\text{cross}} + \underbrace{\Sigma_{mh\in M_t}\theta_{mh,t}\left(q_{mh,t}-\bar{q}_{mt-1}\right)}_{\text{entry}} - \\ & \underbrace{\Sigma_{mh\in M_t}\theta_{mh,t-1}\left(q_{mh,t-1}-\bar{q}_{mt-1}\right)}_{\text{exit}}\end{aligned} \tag{3}$$

where $q_{mh,t}$ indicates patient safety defined as the facility JHIC score of health facility h in market m at period t and $\theta_{mh,t}$ is its market share in terms of outpatients. We look at two periods:

the endline period t and the baseline period *t-1*. $\bar{q}_{mt}$ is the market-share-weighted average JHIC score in market m in period t, and Δ is the difference operator, applied between t and *t-1*. $\Delta\bar{q}_{mt}$ is then the change in the market weighted average JHIC score between baseline and endline for market m. C_t is the set of health facilities which were open both at baseline and at endline. M_t is the set of health facilities which did not exist at baseline but were active at endline. X_t is the set of health facilities which were active at baseline but closed 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.[25] 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 92% of the total increase in the (patient-weighted) JHIC score of 3 percentage points (p-value = 0.002), is driven by "within" health facilities changes.[26] The exit of facilities with quality below the market baseline mean contributes 6% of total impact (p-value = 0.027) 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 exiting and entering facilities account for less than 20% of market share for entry and less than 7% of market share for exit 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.

IV.3.2 How do closures drive the results

As a consequence of the regulation, 29% of private facilities (174) in treated markets operational at randomization were closed by the government. Given the importance of closures, our second

[25] 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. There 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.

[26] The differences 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.

exercise assesses the importance of government closures under the regulation as a channel for our results. The problem that we face is that in order to estimate the impact of government closures, we cannot just compare facilities or markets with closures to control facilities or markets as exits in control markets were also quite frequent and the same types of facilities (unlicensed and low quality) were also more likely to exit in control markets.

We therefore proceed as follows. We first use a random-forest model to predict a facility's likelihood of closure by the government as a function of pre-treatment (or fixed) characteristics. We then aggregate this likelihood to the market-level, yielding the the number of facilities in the market that are predicted to face closure. Finally, we assess how the treatment effects vary by these characteristics and by the interaction of treatment and number of closure-type facilities per market at randomization. Table A7 shows that we are able to predict government closures with a balanced accuracy rate of 79%, compared to 50% from a random classification.[27] Using the prediction, we classify markets into those with 0,1,2,3+ facilities at high risk of closure; these account for 44%, 20%, 13%, and 23% (17%, 15%, 13%, and 55%) of all markets (facilities) in our study respectively.

At the facility level, Figure 3 shows that declines in outpatients at private facilities seem to be important only in markets with 3+ predicted closure-type facilities, with the largest and more precisely estimated drop reported for outpatients in unlicensed facilities at randomization.

At the facility level, a second approach instead uses mediation analysis. Specifically, we use a two-stage regression estimator, the *sequential g-estimator*, following Acharya et al. (2016) which, under certain (fairly strong) assumptions, allows us to estimate the direct effect of the intervention, fixing the hypothesized mediator (closures) at zero, to assess the importance of "number of closures in a market" as a mechanism. This estimator answers the question "does the regulatory reform causes a shift of patients from private to public facilities independent of the mediator, in this case, closures?" Table S7 of the Online Appendix shows that in public facilities, the impact on daily outpatients decreases from +7.2 (p-value = 0.03) in our base specification to a *demediated* impact of +4.4 (p-value = 0.19), suggesting that the effect of the intervention is reduced considerably if physical closures are set to zero. For private facilities, although less precisely measured, the impact on outpatients increases from -1.5 (p-value = 0.17) to +0.2 (p-value = 0.87 in the *demediated* impact, suggesting again an important role of closures in the shift of patients from the private to the public sector. Consistent with previous results, the drop in outpatients for private providers is larger and more precisely estimated for unlicensed facilities at randomization, where the impact on outpatients' increases from -2.8 (p-value < 0.001) in the baseline estimate to -1.7 (p-value = 0.17) in the *demediated* impact (Section 7 in the Online Appendix).

These results suggest a role of government closures mediating the switch of patients from the

[27] The balanced accuracy rate is a performance indicator that weights equally the true positive and true negative rates. For binary classification and imbalanced classes, when both errors matter, balanced accuracy tends to be more appropriate than other performance indicators. A balanced accuracy rate between 0.70 and 0.90 is considered good, and above 0.90 very good.

private to the public sector.

IV.4 Cost Effectiveness

Bedoya, Das, Dolinger, De Guttry, et al. (2020) estimate the cost of this intervention focusing on the variable cost of conducting the routine operations (excluding the fixed cost to set up the system). It is worth emphasizing that for a number of reasons the costs for this pilot are likely higher than what they would be in a scaled up version.[28] The authors report that a visit during the pilot costs USD 165 in operational costs, which includes inspections and visits for the enforcement of warnings and sanctions as inspection and follow-ups, as well as closures of facilities and/or departments. They also estimate this cost could be reduced to USD 95 per visit in an scaled-up model. With an average 3 visits (2 inspections) per treated facility, the operational cost per facility estimated for the pilot is USD 495 per visit and USD 285 for the scaled-up model (See Section 8 of the Online Appendix for further details).[29] This compares to a cost of $8000 per facility reported by King et al. (2021) for an 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.[30]

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 discussing potential mechanisms. Specifically, we assess cross-market externalities, the impact of program duration and we examine spillovers on other quality measures that were not part of the inspection process. The estimating equations and accompanying tables are detailed in Section 9 of the Online Appendix.

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 S10 in the Online Appendix).

[28] In our study, inspectors were located in county headquarters and used vehicles provided by central and regional governments to visit health facilities. They were seconded from different government institutions and most transferred from other regions, resulting in a salary supplement. Facility closures required staff from the central government to travel to the regions. In a scaled-up version, the number and location of inspectors can be flexibly determined to minimize costs.

[29] We also provided data to an independent team to complete a third-party an economic costing for our intervention. Including the fixed cost, they emerge at a per-visit estimate of USD 103 and a per-completed-inspection estimate of USD 155 for a scaled-up inspection model (Chege et al., 2022).

[30] See for example De Allegri et al. (2019), Zeng et al. (2018), and Borghi et al. (2015).

Program Duration: Next, we exploit variation in the rollout of the reform and the timing of 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 9.3 in the Online Appendix, is that conditional on the controls, the variation in the market date of first inspection and endline date are exogenous.

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 S2 and Table S13 in the Online Appendix).

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 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; (b) quality indicators reported by patients in 11,098 exit surveys and; (c) healthcare staff composition and remuneration for 7,663 staff.

In general, we do not find significant negative changes along any of these dimensions, with small and statistically insignificant effect sizes, especially given the large number of outcome variables. Important exceptions are in public facilities, where we find an increase in consultation length, which has shown to be positively correlated with clinical accuracy as well as the ratio of healthcare workers to total staff and total staff compensation. These results show that across multiple dimensions of quality the intervention does not lead to negative spillover effects. In fact, they are consistent with improvements in some non-incentivized dimensions of quality in the public sector indicating that at least some public sector facilities are able to negotiate and bargain for more resources after being inspected (Table A8 to A10).

V A Discussion of Possible Mechanisms

In order to understand the mechanisms at play, we now turn to the theoretical literature on minimum quality standards. Consider first, the case of a government imposing MQS on a market with only private firms. If there is no market failure, 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. Minimum standards are strictly welfare decreasing. For MQS to improve welfare therefore requires a market failure—and the impact depends on the source and extent of this failure.

Two canonical sources of market failure have been extensively studied in the literature. In

Shapiro's (1986) model, the source of the market failure is asymmetric information. 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 at the bottom of the market.[31]

In contrast, in Ronnen's (1991) formulation, the inefficiency arises from market power due to vertical differentiation in oligopolies. 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 usually 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 the overall inefficiency in the market is reduced.

Understanding the importance of each of these two channels matters for healthcare regulation in LLMICs because of the different distributional implications. As our intervention was designed to allow for a range of possibilities rather than to test a theory, formal tests of these models are difficult to execute.[32] Nevertheless, guided by the theory, we present a set of ancillary results to disentangle these forces. Interestingly, the results elevate the importance of the market-power channel, although alternate interpretations are also consistent with the findings.

Result 1: Facilities improved in ways that went beyond the "letter of the law:" We first look 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 in the JHIC was that if facilities had complied with all items in the lowest-cost category, their JHIC score would have increased by 34

[31] 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.

[32] 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 beliefs over other facility's beliefs. These models also do not include the public sector, which accounts for 70% 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.

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 A11). An alternative classification by items that affected the marginal versus the fixed-cost again yielded similar impacts, 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 A11).

We also do not find any 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. The largest improvements happened in functional categories that were rewarded in the market as we show in Table 4. Table A5 shows that quality-price correlations by functional category at baseline are statistically significant and higher for infrastructure, equipment and supplies, compared to SOPs (and remain robust to the inclusion of machine-selected control), which sheds light on why the gains in compliance were the lowest for improvements in SOPs, the least costly items for compliance. We also find that impacts for private facilities are higher in markets where there were more public facilities, which suggest an important role of public facilities in the market (Table A12).

Result 2: No impact of additional information: Our second result investigates the role of 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, when analyzed separately, we find exactly the same treatment effects across both arms (Table 7).

It could be that the information treatment did not have any additional impact because the report cards did not improve patient information—Table A13 shows, for instance, that even though patients understood the scoring system, only 8 percentage points more patients actually noticed the scorecard despite a fairly extensive dissemination effort. However, such an explanation makes it then harder for information to have had an impact in the inspection only arm: it is 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 cumulative distribution functions 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 4 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.

Panel B shows, for instance, that for markets of 1-2 facilities, the intervention increased the JHIC score at the 90th quantile by 4.5 percentage points or 3 times the impact in the 10th quantile (Test 10th=90th p-value = 0.365), while for markets of 3-10 facilities the impacts on safety for the 90th and 10th quantiles are 7.0 and 4.0 percentage points, respectively (Test 10th=90th p-value = 0.095), and for markets of 11+ facilities, the intervention increases safety at the 90th quantile by 9.7 percentage points or 5.6 times the impact in the 10th quantile (Test 10th=90th p-value < 0.010). When only private facilities are analyzed, the differences between the lowest and highest quantile are larger and more precisely estimated (Panel C).[33] Table S14 in the Online Appendix shows

[33] Dividing the sample for public facilities leaves very small groups and lack of common support for this type of analysis.

similar analyses using unconditional quantile treatment effects, with similar qualitative results.[34]

Discussion: The results are consistent with the idea that MQS drives improvement across the range of quality. One possibility is that firms in the private sector readjust their positions to maintain market power in response to improvements in other facilities. While a formal test would require us to posit beliefs over closures, we only see three government closures of facilities with JHIC scores above 50 and if we were to use rational expectations, this group would have faced zero regulatory incentives to improve. Nevertheless, a market-power based explanation is not the only possibility. Explanations whereby the cost of investments is negative correlated with quality will lead to similar patterns, as will models with different beliefs and knowledge regarding the regulation. For instance, if the checklist provided feedback to providers and the providers who were altruistic also had the lowest cost of improvement, similar patterns may emerge.

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. The positive results that we observe for the JHIC score is accompanied with higher budgets and a larger workforce in the public sector, suggesting that these facilities used the inspections to bargain for better resources from the county government.

One hypothesis is 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 3rd 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 a large share of facilities (21%) but

[34] Table S17 in the Online Appendix 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.

a small share of patients (2%). The ubiquity of these clinics prompted an important regulatory reform, establishing 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, MQS increased quality as measured by the JHIC score in both public and private facilities without any decline in utilization. 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. This is not because exits are infrequent or there was no reallocation. Rather it reflects the fact that exits are also quite frequent among control facilities and that the market share of facilities that exit are small. Similarly, reallocations were likely between facilities with similar qualities, which is consistent with economic theory.

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 are very high. 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

Acharya, A., Blackwell, M., & Sen, M. (2016, August). Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects. *American Political Science Review*, *110*(3), 512–529. Retrieved 2022-09-10, from https://www.cambridge.org/core/product/identifier/S0003055416000216/type/journal_article doi: 10.1017/S0003055416000216

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. Retrieved 2022-09-10, from https://doi.org/10.1198/016214508000000841 doi: 10.1198/016214508000000841

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. Retrieved 2022-09-10, from https://www.aeaweb.org/articles?id=10.1257/aer.20140774 doi: 10.1257/aer.20140774

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. Retrieved 2022-09-10, from https://www.aeaweb.org/articles?id=10.1257/aer.20180924 doi: 10.1257/aer.20180924

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. Retrieved 2022-09-10, from https://www.aeaweb.org/articles?id=10.1257/aer.20180326 doi: 10.1257/aer.20180326

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.* Retrieved 2022-07-24, from https://www.hbs.edu/faculty/Pages/item.aspx?num=59594

Bedoya, G., Das, J., & Dolinger, A. (2020). *Safety First : Improving Access to Quality Health Services in Kenya, Expanding Global Knowledge on Disease Prevention* [Text/HTML]. Retrieved 2022-09-15, from https://documents.worldbank.org/en/publication/documents-reports/documentdetail/751991605198328124/Safety-First-Improving-Access-to-Quality-Health-Services-in-Kenya-Expanding-Global-Knowledge-on-Disease-Prevention

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 [Text/HTML]. *World Bank*. Retrieved

2022-09-15, from https://documents.worldbank.org/en/publication/documents-reports/documentdetail/560481605198769107/Regulation-for-Safety-and-Quality-of-Care-A-Process-Evaluation-of-the-Health-Inspection-Pilots-of-the-Kenya-Patient-Safety-Impact-Evaluation

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. Retrieved 2022-07-28, from https://apps.who.int/iris/handle/10665/272121 doi: 10.2471/BLT.16.179499

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. Retrieved 2022-07-24, from https://direct.mit.edu/rest/article/101/1/76/58661/The-Market-for-High-Quality-Medicine-Retail-Chain doi: 10.1162/rest_a_00758

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. Retrieved 2022-09-10, from https://pubs.aeaweb.org/doi/10.1257/app.20150027 doi: 10.1257/app.20150027

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. Retrieved 2022-09-10, from https://onlinelibrary.wiley.com/doi/10.1002/pam.10140 doi: 10.1002/pam.10140

Blau, D. M. (2007, June). Unintended consequences of child care regulations. *Labour Economics*, *14*(3), 513–538. Retrieved 2022-09-10, from https://www.sciencedirect.com/science/article/pii/S0927537106000042 doi: 10.1016/j.labeco.2006.01.003

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. Retrieved 2022-09-15, from https://www.healthaffairs.org/doi/10.1377/hlthaff.2014.0608 doi: 10.1377/hlthaff.2014.0608

Bosk, C. L., Dixon-Woods, M., Goeschel, C. A., & Pronovost, P. J. (2009, August). Reality check for checklists. *The Lancet*, *374*(9688), 444–445. Retrieved 2022-07-24, from https://linkinghub.elsevier.com/retrieve/pii/S0140673609614409 doi: 10.1016/S0140-6736(09)61440-9

Bruhn, M., & McKenzie, D. (2008, October). *In Pursuit of Balance : Randomization in Practice in Development Field Experiments* (Tech. Rep.). Washington, DC: World Bank. Retrieved 2022-09-10, from https://openknowledge.worldbank.org/handle/10986/6910 doi: 10.1596/1813-9450-4752

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. Retrieved 2022-07-24, from https://www.aeaweb.org/articles?id=10.1257/aer.20151080 doi: 10.1257/aer.20151080

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.* Retrieved 2022-09-15, from https://www.researchsquare.com doi: 10.21203/rs.3.rs-1359824/v1

Chen, B. K., Gertler, P. J., & Yang, C.-Y. (2016, December). Physician ownership of complementary medical services. *Journal of Public Economics*, *144*, 27–39. Retrieved 2022-07-23, from https://www.sciencedirect.com/science/article/pii/S0047272716301311 doi: 10.1016/j.jpubeco.2016.09.005

Chipty, T., & Witte, A. D. (1997, July). *An empirical investigation of firms' responses to minimum standards regulations* [Working Paper]. National Bureau of Economic Research. Retrieved 2022-09-13, from https://www.nber.org/papers/w6104 doi: 10.3386/w6104

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. Retrieved 2022-07-24, from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.12270 doi: 10.1111/1756-2171.12270

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), e000333. Retrieved 2022-07-24, from https://gh.bmj.com/lookup/doi/10.1136/bmjgh-2017-000333 doi: 10.1136/bmjgh-2017-000333

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. Retrieved 2022-07-24, from https://linkinghub.elsevier.com/retrieve/pii/S0277953620300186 doi: 10.1016/j.socscimed.2020.112799

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. Retrieved 2022-07-24, from http://www.healthaffairs.org/doi/10.1377/hlthaff.2011.1356 doi: 10.1377/hlthaff.2011.1356

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), 282–288. Retrieved 2022-09-15, from https://academic.oup.com/heapol/article/34/4/282/5491586 doi: 10.1093/heapol/czz030

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. Retrieved 2022-07-24, from https://linkinghub.elsevier.com/retrieve/pii/S0047272717301482 doi: 10.1016/j.jpubeco.2017.09.005

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. Retrieved 2022-09-06, from https://www.econometricsociety.org/doi/10.3982/ECTA12876 doi: 10.3982/ECTA12876

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*, jems.12471. Retrieved 2022-07-24, from https://onlinelibrary.wiley.com/doi/10.1111/jems.12471 doi: 10.1111/jems.12471

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. Retrieved 2022-07-24, from http://link.springer.com/10.1007/s11092-012-9156-4 doi: 10.1007/s11092-012-9156-4

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*. Retrieved 2022-07-28, from https://doi.wiley.com/10.1002/14651858.CD008992.pub3 doi: 10.1002/14651858.CD008992.pub3

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. In The Cochrane Collaboration (Ed.), *Cochrane Database of Systematic Reviews* (p. CD008992.pub2). Chichester, UK: John Wiley & Sons, Ltd. Retrieved 2022-07-28, from https://doi.wiley.com/10.1002/14651858.CD008992.pub2 doi: 10.1002/14651858.CD008992.pub2

Foster, L., Haltiwanger, J., & Syverson, C. (2008, March). Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability? *American Economic Review*, *98*(1), 394–425. Retrieved 2022-09-10, from https://www.aeaweb.org/articles?id=10.1257/aer.98.1.394 doi: 10.1257/aer.98.1.394

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. Retrieved 2022-09-10, from https://www.nber.org/books-and-chapters/new-developments-productivity-analysis/aggregate-productivity-growth-lessons-microeconomic-evidence

Gailmard, S., & Patty, J. W. (2012, June). Formal Models of Bureaucracy. *Annual Review of Political Science*, *15*(1), 353–377. Retrieved 2022-07-24, from https://www.annualreviews.org/doi/10.1146/annurev-polisci-031710-103314 doi: 10.1146/annurev-polisci-031710-103314

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. Retrieved 2022-08-27, from https://openknowledge.worldbank.org/handle/10986/36234 doi: 10.1596/978-1-4648-1675-8

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. Retrieved 2022-07-24, from https://www.aeaweb.org/articles?id=10.1257/aer.101.5.1775 doi: 10.1257/aer.101.5.1775

IFC, & WHO. (2012). *Minimum patient safety standards compliance kenya survey dataset* (Tech. Rep.). International Finance Corporation and World Health Organization.

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. Retrieved 2022-07-24, from https://linkinghub.elsevier.com/retrieve/pii/S2214109X2100228X doi: 10.1016/S2214-109X(21)00228-X

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. Retrieved 2022-07-24, from https://academic.oup.com/heapol/article/36/5/695/6225718 doi: 10.1093/heapol/czab039

Klein, B., & Leffler, K. B. (1981, August). The Role of Market Forces in Assuring Contractual Performance. *Journal of Political Economy*, *89*(4), 615–641. Retrieved 2022-07-28, from https://www.journals.uchicago.edu/doi/10.1086/260996 doi: 10.1086/260996

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. Retrieved 2022-07-24, from https://linkinghub.elsevier.com/retrieve/pii/S0305750X21003557 doi: 10.1016/j.worlddev.2021.105740

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. Retrieved 2022-07-24, from https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2021-058746 doi: 10.1136/bmjopen-2021-058746

Leland, H. E. (1979, December). Quacks, Lemons, and Licensing: A Theory of Minimum Quality Standards. *Journal of Political Economy*, *87*(6), 1328–1346. Retrieved 2022-07-28, from https://www.journals.uchicago.edu/doi/10.1086/260838 doi: 10.1086/260838

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. Retrieved 2022-07-24, from https://linkinghub.elsevier.com/retrieve/pii/S0033350620304431 doi: 10.1016/j.puhe.2020.10.001

McKenzie, D., & Paffhausen, A. L. (2019, October). Small Firm Death in Developing Countries. *The Review of Economics and Statistics*, *101*(4), 645–657. Retrieved 2022-09-11, from https://direct.mit.edu/rest/article/101/4/645/58570/Small-Firm-Death-in-Developing-Countries doi: 10.1162/rest_a_00798

Miguel, E., & Kremer, M. (2004, January). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica*, *72*(1), 159–217. Retrieved 2022-09-10, from http://doi.wiley.com/10.1111/j.1468-0262.2004.00481.x doi: 10.1111/j.1468-0262.2004.00481.x

Ministry of Health Implementation Guidelines for the Joint Health Inspection Checklist (Tech. Rep.). (2015). Kenya Ministry of Health: Joint Health Inspections Technical Working Group.

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. Retrieved 2022-07-24, from http://archpedi.jamanetwork.com/article.aspx?doi=10.1001/jamapediatrics.2014.3445 doi: 10.1001/jamapediatrics.2014.3445

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. Retrieved 2022-07-24, from https://www.sciencedirect.com/science/article/pii/S0047272716301621 doi: 10.1016/j.jpubeco.2016.11.005

Ronnen, U. (1991). Minimum Quality Standards, Fixed Costs, and Competition. *The RAND Journal of Economics*, *22*(4), 490. Retrieved 2022-07-24, from http://doi.wiley.com/10.2307/2600984 doi: 10.2307/2600984

Shapiro, C. (1986, October). Investment, Moral Hazard, and Occupational Licensing. *The Review of Economic Studies*, *53*(5), 843. Retrieved 2022-07-24, from https://academic.oup.com/restud/article-lookup/doi/10.2307/2297722 doi: 10.2307/2297722

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. Retrieved 2022-07-24, from https://bmjopen.bmj.com/content/9/7/e029486 doi: 10.1136/bmjopen-2019-029486

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*, 1. Retrieved 2022-07-24, from http://www.ijhpm.com/article_4103.html doi: 10.34172/ijhpm.2021.90

World Health Organization. (2006). *The world health report : 2006 : working together for health* (Tech. Rep.). World Health Organization. Retrieved 2022-07-24, from https://apps.who.int/iris/handle/10665/43432 (number-of-pages: xxvi, 209)

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. Retrieved 2022-07-24, from http://www.healthaffairs.org/doi/10.1377/hlthaff.2014.1114 doi: 10.1377/hlthaff.2014.1114

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. Retrieved 2022-09-15, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6239017/pdf/BLT.17.207100.pdf/ doi: 10.2471/BLT.17.207100

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	943
Daily outpatients, mean [SD] [1]	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 [2]	5.47 [8.50]	0.70 [0.98]	8.39 [9.67]	958
Patients' OOP, mean [SD] USD nominal	2.45 [3.80]	0.31 [0.44]	3.75 [4.33]	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	0.49	0.65	0.23	8523 (958 HFs)
Patients reporting facility distance from home <=4km [3]	0.73	0.72	0.75	8116 (966 HFs)
Patient's wealth index is, mean [SD] (-4 to 12) [4]	0.87 [2.09]	0.34 [1.71]	1.70 [2.35]	8477 (960 HFs)
Violations of infection prevention and control (IPC) practice during outpatient visit, mean [SD] [5]	5.11 [3.33]	4.85 [3.18]	5.72 [3.58]	14108 (926 HFs)
Indications during outpatient visit, mean [SD] [6]	7.50 [5.61]	7.18 [5.46]	8.28 [5.90]	14108 (926 HFs)
Panel C: Indication-level indicators from patient-HCW interactions				
Compliance with all IPC practices measured, mean [SD] [7]	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. HFs = Health facilities. All statistics are constructed using active facilities at randomization in the three study counties and a sample of outpatients that were surveyed in each facility for patient-level indicators. Patient-level statistics are unweighted. [1] Daily outpatients from facility records. Data are available for 1025 facilities (658 private and 367 public). [2] Out-of-pocket (OOP) expenses as reported by patients is used as a proxy for price and is given in nominal and PPP-adjusted USD, set at 2015 prices using the Kenya CPI and PPP conversion factor from the IMF. [3] Distance in kilometers from the facility where the survey is conducted to the patient's household. [4] Index range is observed range. To assess the socioeconomic status of our sample and how it may compare to the population at the national level we created a Proxy Wealth Index following the methodology of the 2014 DHS Wealth Index (WI). This WI is a measure of a household's cumulative living standard based on easy-to-collect data on a household's ownership of selected assets. [5] A safety violation occurred when the required infection prevention and control action was not taken during patient-health care worker (HCW) interactions, following Bedoya et al. (2017). [6] An indication refers to a situation in which an infection prevention and control practice must be undertaken to prevent the risk of a pathogen being transmitted from one surface to another. [7] The compliance is the proportion of indications for an infection prevention and control practice for which the corresponding action was taken.

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

	Private			Public	
	OOP (USD PPP) at Baseline	Market Share at Baseline	Exit by Endline (Control Facilities)	OOP (USD PPP) at Baseline	Market Share at Baseline
	(1)	(2)	(3)	(4)	(5)
JHIC Score at Baseline	0.180***	0.040	-0.009**	0.007	0.294*
	(0.055)	(0.044)	(0.004)	(0.010)	(0.155)
Unlicensed at Baseline	-0.748***	0.863	0.125**	0.000	0.000*
	(0.701)	(0.866)	(0.049)	(0.000)	(0.000)
Observations	586	648	189	364	367
R^2	0.15	0.77	0.08	0.55	0.79
Dependent Variable Mean	8.37	9.66	0.15	0.70	53.05
Mean (SD) JHIC Score at Baseline	34.07 (11.25)	33.55 (11.28)	32.66 (9.56)	41.19 (10.13)	41.18 (10.20)
Total Controls Selected by PDF (out of 26)	9	14	4	8	11

Notes. Robust standard errors are reported in parentheses and p-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. Controls selected by PDSLASSO out of a list of 26 variables. Unlicense status, Facility Levels and strata FE at baseline are partialled out (imposed as controls in the regression) so not included in the list of 26 variables. Out-of-pocket (OOP) expenses as reported by patients is used as a proxy for price and is given in PPP-adjusted USD, set at 2018 prices using the Kenya CPI and PPP conversion factor from the IMF. OOP is computed as unweighted averages at the facility level. Market share is defined as the percentage of daily outpatients in the market coming from the facility. Exit correspond to facilities inactive at endline. There were no public facilities inactive at endline, we then present results only for private facilities. JHIC score ranges from 0 to 100 (% of maximum score x 100). There are 8% of facilities missing in these estimates (74 private facilities and 3 public facilities). Baseline data includes 660 private facilities (191 in the control group) and 367 public facilities. For estimates with private facilities, we lose 12 facilities (2 in the control group) for which one of the control variable is missing and we lose 62 facilities for which we did not have OOP due to lack of patient flow at the time of the survey. For estimates with public facilities, we lose 3 facilities for which we did not have OOP for the same reason.

Table 3: Treatment Effects on JHIC Score, OOP, Outpatients, and Entry:
Overall and Interacted with Ownership and License Status 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*						
All Inspections (T)	5.159***	0.973**	3.926***	0.138	1.484	0.004
	(0.836)	(0.419)	(1.319)	(0.553)	(1.741)	(0.022)
	{0.001}***	{0.022}**	{0.007}***	{0.474}	{0.247}	[0.853]
Observations	1285	1285	1285	1285	1285	1319
R^2	0.317	0.126	0.517	0.178	0.247	0.050
Control Mean	35.493	4.069	42.526	3.136	20.793	0.135
Impact: {%; SD}	{15%; 0.49}	{24%; 0.20}	{9%; 0.33}	{4%; 0.03}	{7%; 0.05}	{3%; 0.01}
Panel B: *Interaction with Private*						
All Inspections (T)	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]	[0.351]
Private HF	-5.929***	4.373***	-0.038	5.486***	-28.353***	0.148***
	(1.011)	(0.377)	(2.364)	(1.012)	(2.989)	(0.029)
	[0.000]	[0.000]	[0.987]	[0.000]	[0.000]	[0.000]
Private HF x T	3.498***	1.509***	3.091	0.047	-9.303**	-0.016
	(1.176)	(0.569)	(2.505)	(1.072)	(4.117)	(0.035)
	[0.003]	[0.009]	[0.218]	[0.965]	[0.025]	[0.656]
Observations	1285	1285	1285	1285	1285	1319
R^2	0.337	0.219	0.524	0.350	0.409	0.079
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	0.187
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}	{-0%; -0.00}
Test T + Private x T = 0 (p-value)	0.000	0.007	0.003	0.713	0.436	0.978
Panel C: *Interaction with Unlicensed (Private and active at endline only)*						
All Inspections (T)	6.766***	1.093	7.712***	-0.136	-1.986	-0.042
	(1.222)	(0.762)	(1.956)	(1.418)	(2.786)	(0.037)
	[0.000]	[0.153]	[0.000]	[0.924]	[0.477]	[0.255]
Unlicensed at Endline	-3.815***	-1.297**	-2.148	-3.497***	-3.859	0.007
	(1.050)	(0.568)	(2.155)	(1.088)	(2.641)	(0.057)
	[0.000]	[0.023]	[0.320]	[0.002]	[0.146]	[0.903]
Unlicensed at Endline x T	-1.906	0.504	-4.303	1.516	1.515	0.102
	(1.427)	(0.832)	(2.904)	(1.517)	(3.221)	(0.064)
	[0.183]	[0.546]	[0.140]	[0.319]	[0.639]	[0.116]
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	6.394	45.718	8.083	15.821	0.168
Control Mean Unlicensed	30.086	4.973	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}	{-25%; -0.11}
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	0.245

Notes. *** (**) (*) denotes significance at 1% (5%) (10%) level. "Naive" p-values are reported in brackets. Sharpened q-values are reported in braces accounting for false discovery rate given that we jointly test five hypotheses (impacts on JHIC score and OOP, both weighted and unweighted, and outpatients). When sharpened q-values are applied, stars reported next to the estimated coefficients denote significance related to the "naive" p-value. This table reports impact estimates for patient safety as measured by the JHIC score (ranges from 0 to 100 of maximum score) as per regulation; out-of-pocket expenses (OOP) as reported by patients; daily outpatients from facility records; and facility entry. Missing values for OOP are imputed using means defined by level, ownership, treatment, license status at randomization, and daily outpatients. All regressions include 16 randomization strata controls (by county and market size) and health facility level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided). Level 2 is the excluded category. The difference in the total number of observations in column (6) is 34 facilities (of which, 33 are private) that did not consent to endline data collection. All monetary amounts are PPP-adjusted USD terms, set at 2018 prices using the Kenya CPI and PPP conversion factor from the IMF, unless otherwise stated. Details with outcome definitions and calculations are in Online Appendix Section 6.

Table 4: Treatment Effects on JHIC Item Compliance by Functional Categories:
Overall and Interacted with Ownership and License Status 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*							
All Inspections (T)	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
	50927	16726	53711	2892	56321	6337	29617
	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*							
All Inspections (T)	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]
	50927	16726	53711	2892	56321	6337	29617
	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)*							
All Inspections (T)	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 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 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]
	33125	10700	33929	1752	36640	3646	18352
	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 p-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. Regressions include controls for the 16 strata included in the randomization (by county and market size) and health facility level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided). Items are categorized in 7 functional groups: supplies (low cost), supplies (medium cost), equipment, management, medical records and SOPs, and infrastructure.

Table 5: Treatment Effects on Outpatients and Inactivity:
Overall and Interacted with Ownership and License Status at Randomization

	Daily Outpatients	Inactive
	(1)	(2)
Panel A: *Overall Impact*		
All Inspections (T)	0.672	0.027
	(1.636)	(0.021)
	{0.733}	{0.335}
Observations	1321	1347
R^2	0.253	0.042
Control Mean	20.150	0.131
Impact: {%; SD}	{3%; 0.02}	{20%; 0.08}
Panel B: *Interaction with Private*		
All Inspections (T)	7.620**	0.003
	(3.449)	(0.009)
	[0.028]	[0.708]
Private HF at Randomization	-29.211***	0.170***
	(3.088)	(0.025)
	[0.000]	[0.000]
Private HF at Randomization x T	-9.325**	0.031
	(4.243)	(0.030)
	[0.029]	[0.308]
Observations	1321	1347
R^2	0.419	0.090
Control Mean Public	41.424	0.000
Control Mean Private	10.286	0.190
Impact Public: {%; SD}	{18%; 0.24}	{ .%; .}
Impact Private: {%; SD}	{-17%; -0.07}	{18%; 0.09}
Test T + Private x T = 0 (p-value)	0.366	0.246
Panel C: *Interaction with Unlicensed (Private and active at randomization only)*		
All Inspections (T)	0.006	-0.015
	(2.663)	(0.034)
	[0.998]	[0.659]
Unlicensed at Randomization	-0.606	0.093*
	(1.998)	(0.050)
	[0.762]	[0.062]
Unlicensed at Randomization x T	-3.111	0.103*
	(2.645)	(0.057)
	[0.241]	[0.074]
Observations	918	943
R^2	0.311	0.080
Control Mean Licensed	14.378	0.124
Control Mean Unlicensed	7.123	0.240
Impact Licensed: {%; SD}	{0%; 0.00}	{-12%; -0.05}
Impact Unlicensed: {%; SD}	{-44%; -0.23}	{36%; 0.20}
Test T + Unlicensed x T = 0 (p-value)	0.002	0.049

Notes. *** (**) (*) denotes significance at 1% (5%) (10%) level. "Naive" p-values are reported in brackets. Sharpened q-values are reported in braces accounting for false discovery rate given that we jointly test five hypotheses (impacts on JHIC score and OOP, both weighted and unweighted, and outpatients). When sharpened q-values are applied, stars reported next to the estimated coefficients denote significance related to the "naive" p-value. This table reports impact estimates for daily outpatients from facility records and facility inactivity, or exits. All regressions include 16 randomization strata controls (by county and market size) and health facility level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided). Level 2 is the excluded category. The difference in the total number of observations in column (2) is 26 facilities (of which, 25 are private) that did not consent to endline data collection. Details with outcome definitions and calculations are in the Online Appendix Section 6.

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)
All Inspections (T)	3.044***	2.792***	0.110	0.014	−0.041	0.170**
	(0.971)	(0.915)	(0.321)	(0.290)	(0.166)	(0.076)
	[0.002]	[0.003]	[0.733]	[0.961]	[0.803]	[0.027]
Control Mean	−0.314	−0.331	0.047	0.065	−0.294	0.200
Observations (Mkts)	252	252	252	252	252	252
Observations (HFs)	1300	1300	1300	1300	1300	1300

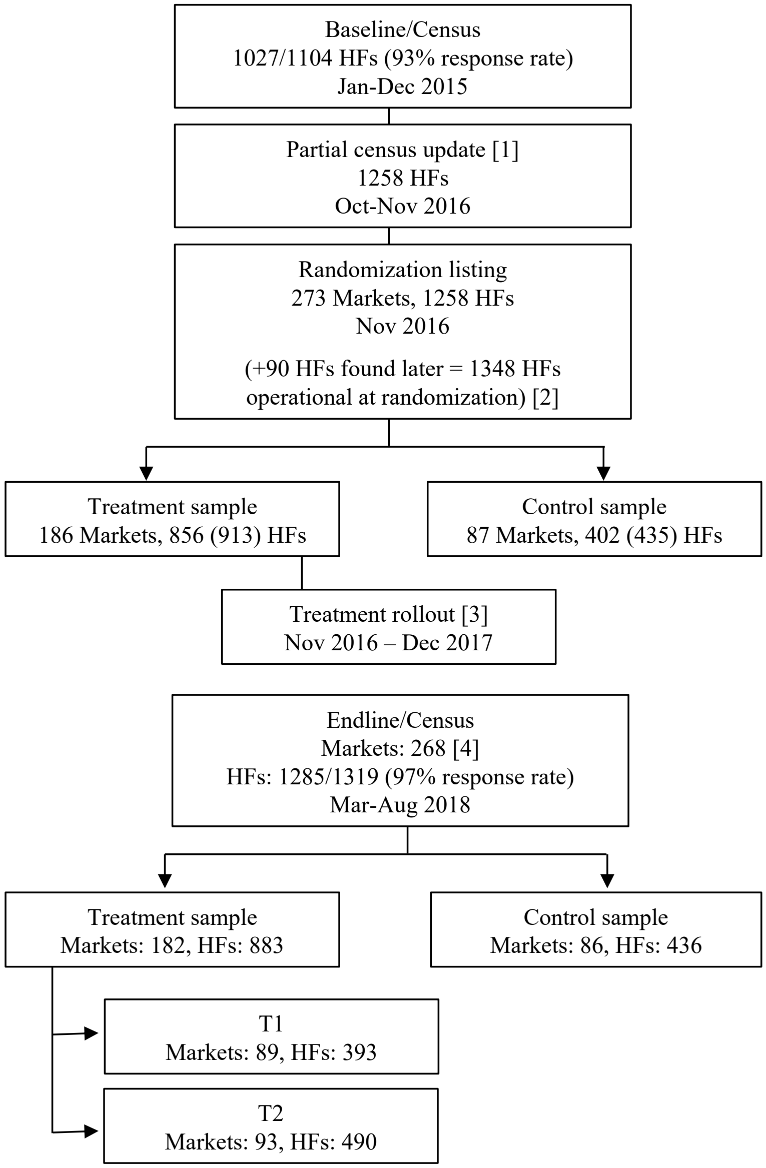
Notes. Robust standard errors are reported in parentheses and p-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. This table reports impact estimates on the decomposition of the change of weighted average patient safety score in a market between baseline and endline using the decomposition formula given in Equation 3. Patient safety is defined as the weighted average JHIC score. Column 1 reports the total change of weighted average JHIC score between baseline and endline. Columns 2 to 6 include the impact on each of the five terms of the decomposition. All regressions are at the market level and weighted by the market size at baseline. All regressions include 16 randomization strata controls (by county and market size) and control for the percentage of health facilities of each level in the market (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided).

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

	Unweighted		Weighted			
	JHIC Score (pp of max)	OOP (USD PPP)	JHIC Score (pp of max)	OOP (USD PPP)	Daily Outpatients	New
	(1)	(2)	(3)	(4)	(5)	(6)
Inspections (T1)	5.435***	0.917**	4.193***	0.173	1.421	-0.027
	(1.112)	(0.438)	(1.582)	(0.607)	(2.180)	(0.024)
	[0.000]	[0.037]	[0.009]	[0.775]	[0.515]	[0.279]
Inspections with Scorecards (T2)	4.924***	1.019**	3.686***	0.106	1.537	0.030
	(0.858)	(0.491)	(1.245)	(0.556)	(1.886)	(0.024)
	[0.000]	[0.039]	[0.003]	[0.849]	[0.416]	[0.217]
Observations	1285	1285	1285	1285	1285	1319
R^2	0.317	0.127	0.517	0.178	0.247	0.055
Control Mean	35.493	4.069	42.526	3.136	20.793	0.135
T1 Impact: {%; SD}	{15%; 0.51}	{23%; 0.19}	{10%; 0.35}	{6%; 0.04}	{7%; 0.05}	{-20%; -0.08}
T2 Impact: {%; SD}	{14%; 0.46}	{25%; 0.21}	{9%; 0.31}	{3%; 0.02}	{7%; 0.05}	{22%; 0.09}
Test (T1)=(T2) (p-value)	0.629	0.805	0.633	0.849	0.955	0.021

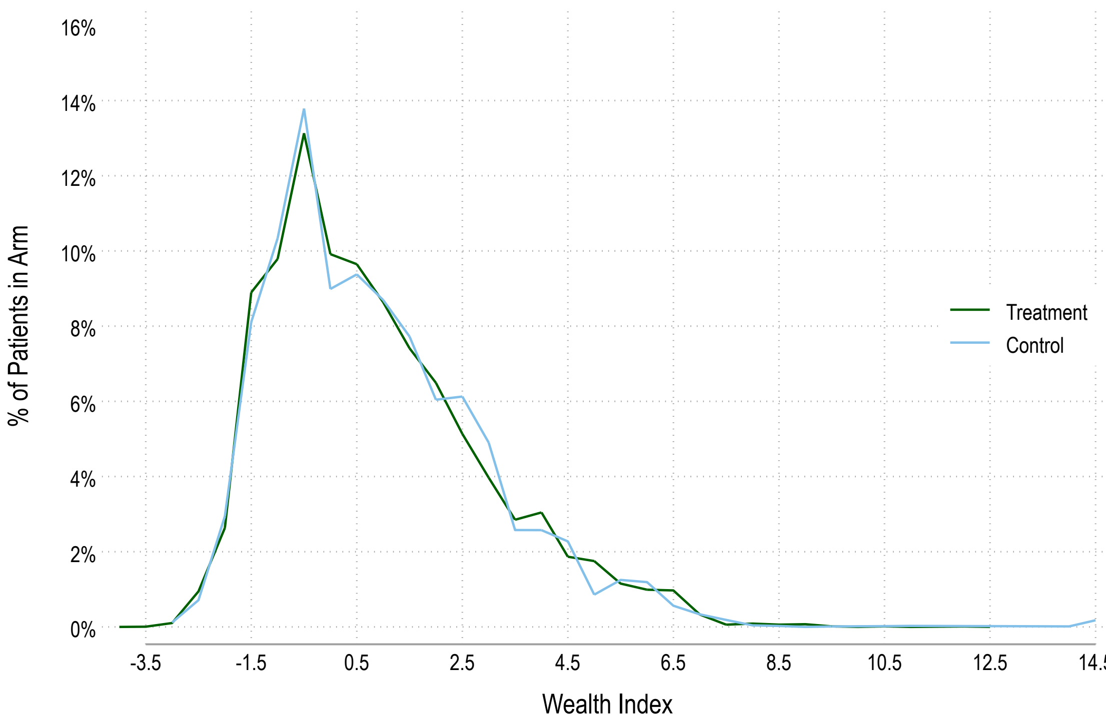
Notes. Robust standard errors are reported in parentheses and p-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. This table reports impact estimates for patient safety as measured by the JHIC score (ranges from 0 to 100 of maximum score) as per regulation; out-of-pocket expenses (OOP) as reported by patients; daily outpatients from facility records; and facility entry. Missing values for OOP are imputed using means defined by level, ownership, treatment, license status at randomization, and daily outpatients. All regressions include 16 randomization strata controls (by county and market size) and health facility level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided). Level 2 is the excluded category. The difference in the total number of observations in column (6) is 34 facilities that did not consent to endline data collection. All monetary amounts are PPP-adjusted USD terms, set at 2018 prices using the Kenya CPI and PPP conversion factor from the IMF, unless otherwise stated.

Figure 1: Timeline of Study



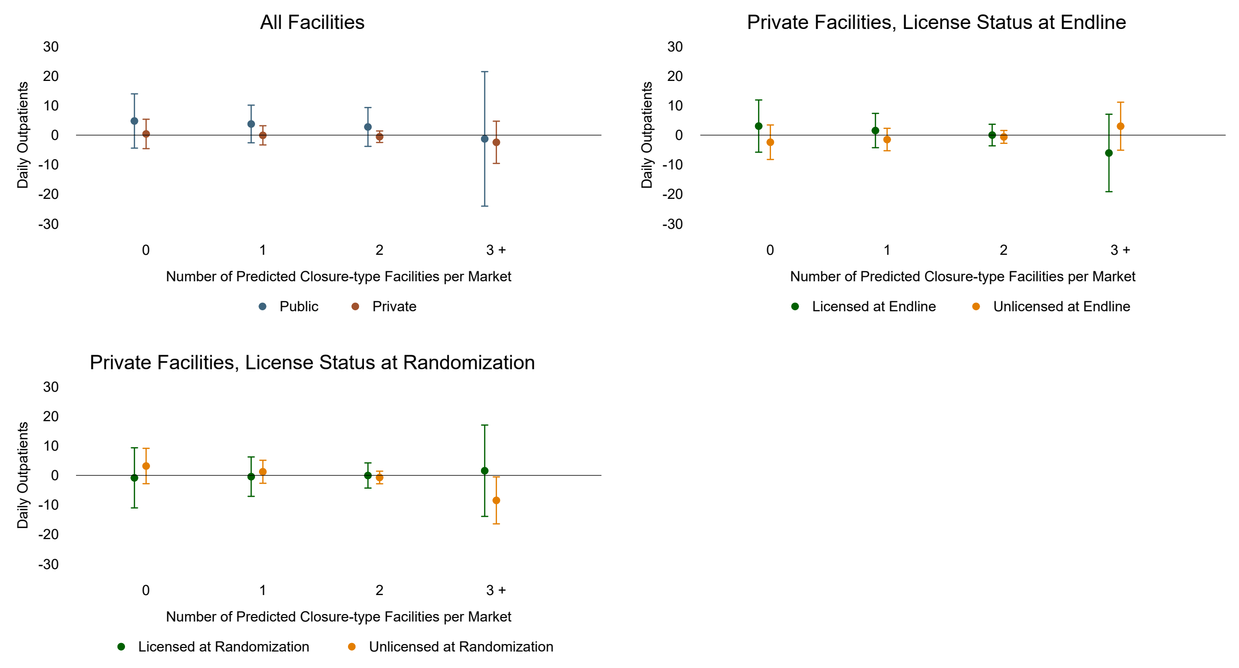
Notes. Timeline and treatment arms. HFs = health facilities. [1] Due to the high turnover of facilities and delay in the implementation, we conducted a partial update of the census in markets of size 1, 2, and 3 between October and November 2016. We used this partial update of the census of 1,258 facilities 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] A 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 the HFs permanently closed or went out of business.

Figure 2: Distribution of Patients by Socioeconomic Status and Treatment Status



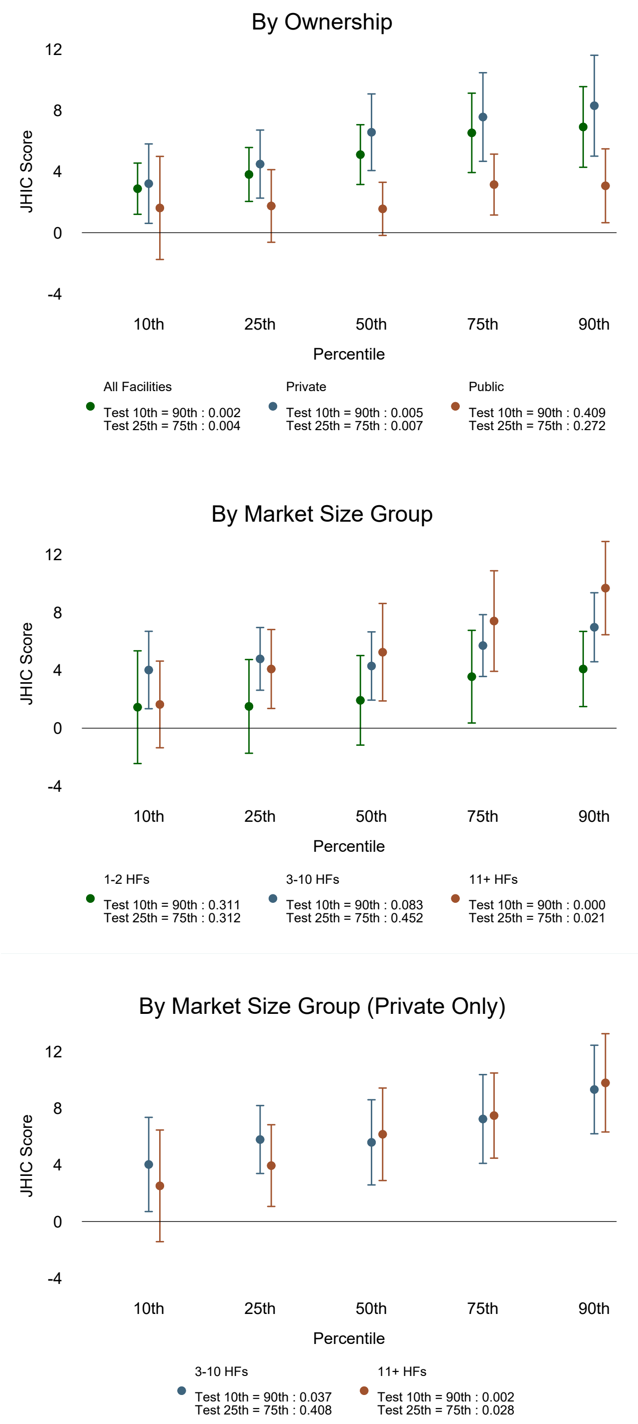
Notes. Kolmogorov-Smirnov test p-value > 0.99. Index range is observed range. To assess the socioeconomic status of our sample and how it may compare to the population at the national level we created a Proxy Wealth Index following the methodology of the 2014 DHS Wealth Index (WI). This WI is a measure of a household's cumulative living standard based on ownership of selected assets.

Figure 3: Treatment Effects on Daily Outpatient interacted with the Number of Predicted Closure-type Facilities per Market Evaluated at Different Values of the Interaction Term, by Ownership Type and License Status at Randomization and Endline



Notes. Vertical lines correspond to 95% confidence intervals. To ensure common support, we code all markets with more than 3 predicted closure-type facilities in the same category, labelled "3 +". For that category, we evaluate the coefficient at the median number of predicted closure-type facilities in markets with three or more predicted closure-type facilities. For the other categories, we evaluate the coefficient at the number of predicted closure-type facilities mentioned. In addition to controlling for the facility level and the 16 randomization strata (combination of county and market size), we control for facility daily outpatients at baseline to adjust for facility outpatient imbalances at baseline. To include facilities missing at baseline, we set baseline outpatient as equal to 0 and include a dummy identifying these facilities.

Figure 4: Conditional QTE on JHIC Score, by Ownership and Market Size



Notes. Vertical lines correspond to 95% confidence intervals. This figure presents the conditional quantile treatment effect estimates on patient safety measured by the JHIC score as per regulation, by ownership (A) and market size (B and C).In graph C, we exclude the private 1-2 HFs estimates which are based on a too small sample (66 health facilities) to allow for convergence. Regressions include controls for the county, health facility level (2, 3, and 4 and 5 merged together, which are positively related to the amount, and specialization of services provided. Level 2 is the excluded category. Regression also controls for JHIC score quartile dummies at baseline, where missing values are replaced by zero and an indicator for missing JHIC score at baseline is included. Conditional quantile treatment effect are estimated using the qreg2 stata command, which allows to estimate robust standard errors clustered at the market level but does not allow to test the equality of the 10th and 90th coefficient. For the test, we run simultaneously the five quantile regressions using the stata command sqreg, which allows to test the equality of percentile coefficients but does not estimate robust standard errors accounting for intra cluster correlation. In this case, standard errors are estimated after performing a 2,000 bootstrap replications.

Appendix

Table A1: JHIC Compliance Categories with Warnings and Sanctions as per 2016 Regulation

Facility score (% of maximum score)	Compliance category	Warning/sanction/reward
0-10% *(or absence of licenses)	Non-Compliant	**Close facility.** Recommend prosecution where requisite registration and licensing is absent (as specified by the respective legislation) Closed facilities that have a license, may ask for a re-inspection after corrections for purposes of reopening and re-categorization, subject to the criteria established by the competent authority.
11-40%	Minimally compliant	**3-month notice for correction + re-inspection** Facility will be closed if it does not score over 40% of the maximum score in the 3rd inspection. Facility may ask for re-inspection for purposes of re-categorization after corrections
Facility score (% of maximum score)	Compliance category	Warning/sanction/reward
41-60%	Partially Compliant	**6-month notice for correction + re-inspection** Facility will be closed if it does not score over 60% of the maximum score in the 3rd inspection. Facility may ask for re-inspection for purposes of re-categorization after corrections
61-75%	Substantially Compliant	**12-month notice for correction + re-inspection** Facility may ask for re-inspection for purposes of re-categorization after corrections
>75%	Fully Compliant	No re-inspection for two years

* Other incentives under consideration include client signaling (quality mark and/or SMS feedback mechanism) and insurance recommendation

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

Table A2: Balance Checks

	Unweighted		Weighted		
	(C) Control Mean	(T-C) Adj. Diff.	(C) Control Mean	(T-C) Adj. Diff.	Obs.
	(1)	(2)	(3)	(4)	(5)
Panel A: *Balance using baseline sample*					
JHIC Score (% of max)	35.539 (10.412)	0.631 (0.947) [0.506]	42.781 (11.906)	0.436 (1.775) [0.806]	1027
OOP	4.525 (7.136)	1.023* (0.545) [0.062]	3.493 (7.171)	-0.072 (0.578) [0.901]	958
Daily Outpatients	24.817 (30.961)	-0.397 (2.041) [0.846]	24.817 (30.961)	-0.397 (2.041) [0.846]	1025
Compliance with IPC Practices (Patient-HCW indication level)	0.318 (0.466)	-0.001 (0.010) [0.900]	0.198 (0.399)	-0.009 (0.010) [0.367]	105876 (929 HFs)
IPC Knowledge (HCW level)	0.735 (0.098)	0.017*** (0.007) [0.010]	0.732 (0.082)	0.023** (0.009) [0.012]	1625 (973 HFs)
IPC Supplies (Site level)	0.639 (0.188)	-0.003 (0.012) [0.834]	0.621 (0.204)	0.005 (0.014) [0.741]	1886 (1006 HFs)
Public	0.350 (0.478)	-0.024 (0.028) [0.405]	0.350 (0.478)	-0.024 (0.028) [0.405]	1104
Level 2	0.824 (0.382)	0.020 (0.023) [0.389]	0.824 (0.382)	0.020 (0.023) [0.389]	1104
Level 3	0.133 (0.340)	-0.023 (0.022) [0.290]	0.133 (0.340)	-0.023 (0.022) [0.290]	1104
F-test from regression of treatment on all outcome variables listed above	1.561 [0.131]				
Panel B: *Balance using randomization sample (select variables)*					
Public	0.308 (0.462)	-0.013 (0.024) [0.607]	0.308 (0.462)	-0.013 (0.024) [0.607]	1348
Level 2	0.855 (0.352)	0.007 (0.019) [0.729]	0.855 (0.352)	0.007 (0.019) [0.729]	1348
Level 3	0.108 (0.311)	-0.005 (0.019) [0.811]	0.108 (0.311)	-0.005 (0.019) [0.811]	1348
Unlicensed (Private only)	0.570 (0.496)	-0.054 (0.034) [0.109]	0.570 (0.496)	-0.054 (0.034) [0.109]	943

Notes. Robust standard errors are reported in parentheses and p-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. Columns (1) and (3) include all information from facilities that are located in the markets assigned to the control group. Columns (2) and (4) present adjusted differences between the means for the treatment markets and the control group. These differences include controls for the 16 strata included in the randomization (by county and market size). The indicators include patient safety as measured by the JHIC score (ranges from 0 to 100 of maximum score) as per regulation; out-of-pocket expenses (OOP) as reported by patients; daily outpatients from facility records; IPC compliance, knowledge, and supplies; public facility ownership; facility level; and license status at randomization. Details with outcome definitions and calculations are in the Online Appendix Section 6.

Table A3: Government Closures During Implementation and Inactivity at Endline
by Baseline JHIC Score Quintile and License Status at Randomization
(Private Facilities)

	Licensed			Unlicensed				All Private			
	Control	Treatment		Control		Treatment		Control		Treatment	
	Inactive	Closed	Inactive	Closed	Inactive	Closed	Inactive	Closed	Inactive	Closed	Inactive
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
JHIC Quintile											
Lowest	0.27	0.21	0.14	0.04	0.32	0.61	0.45	0.03	0.31	0.49	0.36
2nd	0.08	0.11	0.19	0.00	0.36	0.47	0.33	0.00	0.26	0.33	0.28
3rd	0.04	0.10	0.04	0.00	0.14	0.33	0.29	0.00	0.08	0.21	0.16
4th	0.00	0.03	0.09	0.00	0.11	0.44	0.33	0.00	0.05	0.15	0.16
Top	0.05	0.01	0.04	0.00	0.06	0.11	0.11	0.00	0.05	0.03	0.05
All	0.07	0.07	0.09	0.01	0.23	0.45	0.34	0.01	0.15	0.24	0.20

Notes. JHIC quintile estimated using baseline JHIC scores by treatment group within private facilities. License status as per randomization. The estimates only include facilities for which baseline JHIC score is available. Closure indicates government enforcement of facility closure during the implementation. For all columns, the denominator is the number of private facilities per quintile, treatment group, and license status. There was one facility in the control group that was closed due to contamination as per Table S5 in the Online Appendix.

Table A4: Treatment Effects on JHIC Score and OOP:
Overall and Interacted with Ownership and License Status at Randomization
(Sample: Facilities in census at randomization and surviving by endline)

	Unweighted		Weighted	
	JHIC Score (pp of max) (1)	OOP (USD PPP) (2)	JHIC Score (pp of max) (3)	OOP (USD PPP) (4)
Panel A: *Overall Impact*				
All Inspections (T)	5.708***	0.936**	4.177***	0.131
	(0.773)	(0.459)	(1.258)	(0.579)
	{0.000}***	{0.076}*	{0.000}***	{0.861}
Observations	1120	1120	1120	1120
R^2	0.315	0.137	0.524	0.186
Control Mean	36.329	3.993	42.769	3.158
Impact: {%; SD}	{16%; 0.54}	{23%; 0.19}	{10%; 0.36}	{4%; 0.03}
Panel B: *Interaction with Private*				
All Inspections (T)	3.111***	-0.066	3.242**	0.388
	(1.080)	(0.272)	(1.600)	(0.270)
	[0.004]	[0.809]	[0.044]	[0.152]
Private HF at Randomization	-5.686***	4.497***	0.135	5.690***
	(1.089)	(0.455)	(2.350)	(1.080)
	[0.000]	[0.000]	[0.954]	[0.000]
Private HF at Randomization x T	4.120***	1.500**	3.182	-0.044
	(1.232)	(0.675)	(2.507)	(1.193)
	[0.001]	[0.027]	[0.205]	[0.970]
Observations	1120	1120	1120	1120
R^2	0.333	0.232	0.532	0.362
Control Mean Public	39.946	0.658	42.300	0.819
Control Mean Private	34.240	5.920	43.645	7.525
Impact Public: {%; SD}	{8%; 0.35}	{-10%; -0.07}	{8%; 0.35}	{47%; 0.34}
Impact Private: {%; SD}	{21%; 0.67}	{24%; 0.28}	{15%; 0.42}	{5%; 0.06}
Test T + Private x T = 0 (p-value)	0.000	0.023	0.001	0.781
Panel C: *Interaction with Unlicensed (Private and active at randomization only)*				
All Inspections (T)	8.782***	1.504*	8.452***	0.039
	(1.218)	(0.812)	(2.104)	(1.735)
	[0.000]	[0.066]	[0.000]	[0.982]
Unlicensed at Randomization	-2.217*	-0.175	-1.503	-1.929*
	(1.242)	(0.651)	(2.545)	(1.157)
	[0.076]	[0.789]	[0.555]	[0.097]
Unlicensed at Randomization x T	-3.803***	-0.355	-5.457*	0.620
	(1.415)	(0.745)	(3.070)	(1.751)
	[0.008]	[0.635]	[0.077]	[0.724]
Observations	719	719	719	719
R^2	0.381	0.097	0.628	0.082
Control Mean Licensed	37.471	6.652	48.285	8.408
Control Mean Unlicensed	31.326	5.259	36.404	6.146
Impact Licensed: {%; SD}	{23%; 0.80}	{23%; 0.26}	{18%; 0.52}	{0%; 0.01}
Impact Unlicensed: {%; SD}	{16%; 0.50}	{22%; 0.26}	{8%; 0.28}	{11%; 0.15}
Test T + Unlicensed x T = 0 (p-value)	0.000	0.081	0.077	0.455

Notes. *** (**) (*) denotes significance at 1% (5%) (10%) level. "Naive" p-values are reported in brackets. Sharpened q-values are reported in braces accounting for false discovery rate given that we jointly test five hypotheses (impacts on JHIC score and OOP, both weighted and unweighted, and outpatients). When sharpened q-values are applied, stars reported next to the estimated coefficients denote significance related to the "naive" p-value. This table reports impact estimates for patient safety as measured by the JHIC score (ranges from 0 to 100 of maximum score) as per regulation and out-of-pocket expenses (OOP) as reported by patients. Missing values for OOP are imputed using means defined by level, ownership, treatment, license status at randomization, and daily outpatients. All regressions include 16 randomization strata controls (by county and market size) and health facility level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided). Level 2 is the excluded category. All monetary amounts are PPP-adjusted USD terms, set at 2018 prices using the Kenya CPI and PPP conversion factor from the IMF, unless otherwise stated. Details with outcome definitions and calculations are in the Online Appendix Section 6.

Table A5: Quality-Price Gradient per JHIC Functional Category
(Private Facilities)

	OOP (USD PPP)						
	Infrastructure	Equipment	Supplies (Low Cost)	Suppplies (Medium Cost)	Management	Medical Records	SOPs
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mean Adherence per JHIC Category (in Percentage)	0.144***	0.055*	0.051*	0.042**	0.132**	0.023**	0.056
	(0.041)	(0.030)	(0.028)	(0.019)	(0.057)	(0.011)	(0.035)
Unlicensed	0.107***	-0.827*	-1.143*	-0.711**	-0.729**	-0.847**	-0.963
	(0.685)	(0.666)	(0.616)	(0.639)	(0.647)	(0.495)	(0.623)
Average Mean Adherence in Percentage (SD)	43.29 (14.22)	27.94 (16.06)	39.58 (14.23)	16.84 (29.20)	22.90 (13.35)	45.90 (37.31)	6.48 (10.34)
Observations	586	586	586	586	586	586	586
R^2	0.145	0.135	0.128	0.124	0.132	0.103	0.131
Total Controls Selected by PDS (out of 32)	8	10	6	6	11	1	7

Notes. Robust standard errors are reported in parentheses and p-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. Out-of-pocket (OOP) expenses as reported by patients is used as a proxy for price and is given in PPP-adjusted USD, set at 2018 prices using the Kenya CPI and PPP conversion factor from the IMF. OOP is computed as unweighted averages at the facility level. JHIC score ranges from 0 to 100 (% of maximum score x 100). OLS estimates include all 32 variables used in the PDSLasso and controls for strata and facility level, and ownership when including all facilities (private/public). There are 8% of facilities missing in these estimates (74 private facilities and 3 public facilities). For 7% of facilities, we did not have OOP due to lack of patient flow at the time of the survey, the 1% remaining is lost due to missing one of the 32 controls included.

Table A6: Average and Quantile Treatment Effects on Outpatients Wealth Index

	Wealth Index (1)	Observations (2)
All Inspections (T)	−0.036	10957
	(0.171)	
	[0.834]	
QTE		
20th	0.018	10957
	(0.040)	
	[0.650]	
40th	0.028	10957
	(0.055)	
	[0.605]	
60th	−0.080	10957
	(0.072)	
	[0.265]	
80th	−0.009	10957
	(0.083)	
	[0.915]	

Notes. *** p<0.01, ** p<0.05, * p<0.1. Wealth Index is constructed with a subset of variables taken from DHS. Adjusted with sampling patients weights. Regression controls for facility levels and strata. QTE estimates using qreg command.

Table A7: Out-of-Sample Performance Indicators for Closure-type Prediction Models

	Random Forest	Logit	Unlicensed	Random
	Leave-One-Out Cross-Validation			
	(1)	(2)	(3)	(4)
True Positive Rate	0.83	0.65	0.84	0.5
False Positive Rate	0.25	0.15	0.29	0.5
Precision	0.58	0.63	0.55	0.3
Balanced Accuracy	0.79	0.75	0.78	0.5

Notes. The table presents the performance metrics for predicting closure using different prediction methods. The first column uses a Random Forest model, the second column uses a logit model, the third column uses the variable unlicensed (defined as having no valid license) as the predictor of closure and the fourth column is estimating the convergence of using a 50-50 percent chance model that the facility was closed. Both the random forest and the logit model use the same predictor variables: not having a license at randomization, having an expired license at randomization, the facility level, a categorical variable corresponding to when the facility was open, the type of private and the county. The sample is restricted to private facilities operating at randomization.True Positive Rate = True Positive / Positive, False Positive Rate = False Positive / Negative, Precision = True Positive / (True Positive + False Positive), Balanced Accuracy = 1/2 (True Positive / Positive + True Negative / Negative).

Table A8: Treatment Effects on Quality Indicators Not Included in the JHIC (Infection Prevention and Control (IPC)): Overall and Interacted with Ownership and License Status at Endline

	Practice, Knowledge, and Supplies in IPC		
	Practice (Patient-HCW indication level)	Knowledge (HCW level)	Supplies (Site level)
	(1)	(2)	(3)
Panel A: *Overall Impact*			
All Inspections (T)	-0.006	-0.001	0.001
	(0.010)	(0.001)	(0.002)
Observations	104565	2098	2532
R^2	0.011	0.383	0.229
Control Mean	0.336	0.764	0.709
Impact: {%; SD}	{-2%; -0.01}	{-0%; -0.04}	{0%; 0.02}
Panel B: *Interaction with Private*			
All Inspections (T)	-0.012	-0.002	-0.002
	(0.013)	(0.001)	(0.003)
	[0.374]	[0.107]	[0.404]
Private HF	0.014	-0.007***	-0.023***
	(0.017)	(0.001)	(0.003)
	[0.423]	[0.000]	[0.000]
Private HF x T	0.015	0.002	0.004
	(0.018)	(0.002)	(0.004)
	[0.415]	[0.293]	[0.309]
Observations	104565	2098	2532
R^2	0.012	0.418	0.298
Control Mean Public	0.336	0.769	0.728
Control Mean Private	0.335	0.761	0.700
Impact Public: {%; SD}	{-3%; -0.02}	{-0%; -0.10}	{-0%; -0.08}
Impact Private: {%; SD}	{1%; 0.01}	{0%; 0.00}	{0%; 0.05}
Test T + Private x T = 0 (p-value)	0.814	0.989	0.560
Panel C: *Interaction with Unlicensed (Private and active at endline only)*			
All Inspections (T)	-0.000	0.000	0.003
	(0.015)	(0.002)	(0.005)
	[0.978]	[0.815]	[0.524]
Unlicensed at Endline	-0.022	-0.001	-0.002
	(0.017)	(0.001)	(0.005)
	[0.193]	[0.535]	[0.602]
Unlicensed at Endline x T	0.014	-0.001	-0.002
	(0.020)	(0.002)	(0.006)
	[0.467]	[0.688]	[0.670]
Observations	42497	1302	1665
R^2	0.018	0.254	0.205
Control Mean Licensed	0.351	0.762	0.705
Control Mean Unlicensed	0.310	0.759	0.694
Impact Licensed: {%; SD}	{-0%; -0.00}	{0%; 0.03}	{0%; 0.09}
Impact Unlicensed: {%; SD}	{5%; 0.03}	{-0%; -0.04}	{0%; 0.02}
Test T + Unlicensed x T = 0 (p-value)	0.400	0.711	0.880

Notes. Robust standard errors are reported in parentheses and p-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. Regressions include controls for the 16 strata included in the randomization (by county and market size) and health facility level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided). Compliance means are estimated at the indication level over 104,565 indications that required an action by the healthcare workers (HCWs) in terms of an IPC practices. An indication refers to a situation in which an infection prevention and control practice must be undertaken to prevent the risk of a pathogen being transmitted from one surface to another. Knowledge surveys were administered to HCWs observed. Supplies were assessed through structured observation in consultation rooms, laboratories, and injection rooms during HCW-patient interactions, and only for HCWs who had the relevant indications e.g., the supply of puncture-resistant sharps containers was only checked for the health care workers that had to conduct injections. Details with variable definitions and calculations are in Online Appendix Section 6.

Table A9: Treatment Effects on Quality Indicators Not Included in the JHIC (Reported by Patients): Overall and Interacted with Ownership and License Status at Endline

	Time with HCW and Waiting Time		Patient Satisfaction	Provider Consultation Practices		
	Minutes spent with HCW in examination	Minutes waiting before examination, laboratory, and pharmacy	Patient is satisfied or very satisfied (1-5 scale)	Physical examination (PCA index)	Prescribed or gave medicines	Referred to another HF
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: *Unweighted*						
All Inspections (T)	0.423*	1.933	-0.013**	0.021	-0.004	-0.005
	(0.246)	(1.945)	(0.006)	(0.061)	(0.015)	(0.005)
Observations	9634	11098	11098	9649	9737	9736
R^2	0.013	0.045	0.011	0.017	0.007	0.007
Control Mean	7.760	31.992	0.926	-0.028	0.810	0.054
Impact: {%; SD}	{5%; 0.06}	{6%; 0.04}	{-1%; -0.05}	{-75%; 0.01}	{-1%; -0.01}	{-9%; -0.02}
Panel B: *Weighted*						
All Inspections (T)	0.562**	3.090	-0.019*	0.025	-0.000	-0.009
	(0.263)	(2.715)	(0.010)	(0.064)	(0.020)	(0.007)
	[0.034]	[0.256]	[0.051]	[0.695]	[0.997]	[0.218]
Private HF	2.119***	-20.770***	0.045***	0.780***	-0.051**	-0.013
	(0.388)	(2.811)	(0.010)	(0.105)	(0.022)	(0.009)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.024]	[0.140]
Private HF x T	-0.185	-4.330	0.019	0.062	-0.017	0.011
	(0.475)	(3.522)	(0.012)	(0.117)	(0.027)	(0.011)
	[0.697]	[0.220]	[0.120]	[0.596]	[0.540]	[0.315]
Observations	9634	11098	11098	9649	9737	9736
R^2	0.024	0.092	0.020	0.069	0.011	0.007
Control Mean Public	6.897	40.711	0.906	-0.329	0.829	0.056
Control Mean Private	9.133	17.841	0.958	0.449	0.781	0.051
Impact Public: {%; SD}	{8%; 0.09}	{8%; 0.06}	{-2%; -0.06}	{-8%; 0.02}	{-0%; -0.00}	{-16%; -0.04}
Impact Private: {%; SD}	{4%; 0.05}	{-7%; -0.05}	{0%; 0.00}	{19%; 0.05}	{-2%; -0.04}	{5%; 0.01}
Test T + Private x T = 0 (p-value)	0.357	0.569	0.942	0.392	0.424	0.780
Panel C: *Interaction with Unlicensed (Private and active at endline only)*						
All Inspections (T)	0.653	-1.320	0.010	0.225**	-0.002	0.009
	(0.425)	(2.371)	(0.010)	(0.106)	(0.026)	(0.012)
	[0.126]	[0.579]	[0.325]	[0.034]	[0.937]	[0.445]
Unlicensed at Endline	0.791	-2.521	0.021*	0.140	0.035	0.016
	(0.760)	(3.082)	(0.012)	(0.187)	(0.039)	(0.016)
	[0.299]	[0.414]	[0.075]	[0.454]	[0.364]	[0.310]
Unlicensed at Endline x T	-0.608	-1.574	-0.026*	-0.194	-0.070	-0.008
	(0.900)	(3.468)	(0.015)	(0.216)	(0.045)	(0.021)
	[0.501]	[0.650]	[0.080]	[0.370]	[0.123]	[0.705]
Observations	3127	3681	3681	3132	3166	3165
R^2	0.013	0.085	0.011	0.024	0.009	0.020
Control Mean Licensed	8.944	21.443	0.947	0.390	0.772	0.046
Control Mean Unlicensed	9.765	14.722	0.973	0.512	0.808	0.064
Impact Licensed: {%; SD}	{7%; 0.10}	{-6%; -0.04}	{1%; 0.04}	{58%; 0.13}	{-0%; -0.00}	{19%; 0.04}
Impact Unlicensed: {%; SD}	{0%; 0.01}	{-20%; -0.12}	{-2%; -0.10}	{6%; 0.02}	{-9%; -0.18}	{2%; 0.00}
Test T + Unlicensed x T = 0 (p-value)	0.957	0.202	0.130	0.877	0.048	0.951

Notes. Robust standard errors are reported in parentheses and p-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. Regressions include controls for the 16 strata included in the randomization (by county and market size) and health facility level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided). Details with variable definitions and calculations are in Online Appendix Section 6.

Table A10: Treatment Effects on Quality Indicators Not Included in the JHIC (Healthcare Staff): Overall and Interacted with Ownership and License Status at Endline

	Ratio of healthcare workers to total staff	Healthcare workers per outpatient [1]	Monthly total staff cost (USD PPP)	Monthly staff cost per staff (USD PPP)	Staff cost per out-patient [1] (USD PPP)
	(1)	(2)	(3)	(4)	(5)
Panel A: *Overall Impact*					
All Inspections (T)	0.006	0.009	1416.664**	23.945	6.085
	(0.011)	(0.008)	(707.763)	(15.347)	(3.834)
Observations	1284	1273	1284	1284	1273
R^2	0.090	0.011	0.363	0.087	0.014
Control Mean	0.660	0.031	4022.330	379.387	13.595
Impact: {%; SD}	{1%; 0.03}	{29%; 0.09}	{35%; 0.15}	{6%; 0.10}	{45%; 0.21}
Panel B: *Interaction with Private*					
All Inspections (T)	0.035*	-0.001	3393.735*	-1.334	-0.595
	(0.020)	(0.002)	(1891.131)	(17.738)	(1.288)
	[0.080]	[0.753]	[0.074]	[0.940]	[0.645]
Private HF	0.075***	0.022**	-3191.873***	-34.757	6.692**
	(0.022)	(0.009)	(1159.038)	(21.102)	(2.989)
	[0.001]	[0.015]	[0.006]	[0.101]	[0.026]
Private HF x T	-0.043	0.014	-2923.509	37.416	9.928*
	(0.026)	(0.013)	(2411.402)	(24.448)	(5.420)
	[0.102]	[0.278]	[0.226]	[0.127]	[0.068]
Observations	1284	1273	1284	1284	1273
R^2	0.100	0.018	0.375	0.088	0.019
Control Mean Public	0.600	0.012	7340.894	385.531	6.483
Control Mean Private	0.689	0.040	2443.708	376.465	17.038
Impact Public: {%; SD}	{6%; 0.18}	{-6%; -0.05}	{46%; 0.25}	{-0%; -0.01}	{-9%; -0.08}
Impact Private: {%; SD}	{-1%; -0.04}	{34%; 0.11}	{19%; 0.07}	{10%; 0.14}	{55%; 0.27}
Test T + Private x T = 0 (p-value)	0.575	0.281	0.629	0.066	0.087
Panel C: *Interaction with Unlicensed (Private and active at endline only)*					
All Inspections (T)	0.004	0.039**	1627.777	36.110	19.612*
	(0.022)	(0.020)	(992.866)	(22.164)	(10.312)
	[0.870]	[0.049]	[0.103]	[0.105]	[0.059]
Unlicensed at Endline	0.034	0.036*	430.217	17.981	12.032**
	(0.027)	(0.018)	(889.601)	(38.166)	(5.674)
	[0.220]	[0.053]	[0.629]	[0.638]	[0.035]
Unlicensed at Endline x T	-0.042	-0.059**	-1562.430	8.990	-22.993*
	(0.031)	(0.028)	(1113.349)	(53.461)	(12.634)
	[0.178]	[0.036]	[0.162]	[.867]	[0.070]
Observations	718	714	718	718	714
R^2	0.085	0.015	0.426	0.067	0.015
Control Mean Licensed	0.644	0.021	4177.144	384.911	12.117
Control Mean Unlicensed	0.713	0.058	1190.724	391.635	21.393
Impact Licensed: {%; SD}	{1%; 0.02}	{182%; 1.48}	{39%; 0.17}	{9%; 0.20}	{162%; 1.40}
Impact Unlicensed: {%; SD}	{-5%; -0.16}	{-35%; -0.11}	{5%; 0.05}	{12%; 0.13}	{-16%; -0.07}
Test T + Unlicensed x T = 0 (p-value)	0.089	0.307	0.810	0.326	0.619

Notes. [1] Monthly outpatients. Robust standard errors are reported in parentheses. P-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. For health facility-level variables, the standard errors are clustered at the market level. Regressions include controls for the 16 strata included in the randomization (by county and market size) and health facility level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided). Details with variable definitions and calculations are in Online Appendix Section 6.

Table A11: Treatment Effects on JHIC Item Compliance by Cost Categories:
Overall and Interacted with Ownership and License Status at Endline

	Cost Groups				Marginal and Fixed Costs	
	Lowest Cost	Low Cost	Medium Cost	High Cost	Marginal	Fixed
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: *Overall Impact*						
All Inspections (T)	0.034***	0.061***	0.074***	0.063***	0.064***	0.047***
	(0.008)	(0.010)	(0.010)	(0.012)	(0.009)	(0.008)
	0.000	0.000	0.000	0.000	0.000	0.000
Observations	82979	62872	19618	51062	73329	143202
R^2	0.036	0.018	0.052	0.045	0.023	0.032
Control Mean	0.199	0.407	0.290	0.409	0.359	0.295
Impact: {%; SD}	{17%; 0.08}	{15%; 0.12}	{25%; 0.16}	{15%; 0.13}	{18%; 0.13}	{16%; 0.10}
Panel B: *Interaction with Private*						
All Inspections (T)	0.027**	0.031**	0.052***	0.035**	0.040***	0.028**
	(0.012)	(0.013)	(0.015)	(0.015)	(0.013)	(0.011)
	[0.021]	[0.013]	[0.001]	[0.020]	[0.002]	[0.014]
Private HF	-0.093***	-0.030**	-0.037**	-0.093***	-0.027**	-0.089***
	(0.012)	(0.014)	(0.015)	(0.014)	(0.013)	(0.012)
	[0.000]	[0.037]	[0.012]	[0.000]	[0.047]	[0.000]
Private HF x T	0.009	0.046***	0.034**	0.043***	0.038**	0.028**
	(0.014)	(0.016)	(0.017)	(0.017)	(0.015)	(0.013)
	[0.508]	[0.004]	[0.045]	[0.010]	[0.012]	[0.034]
Observations	82979	62872	19618	51062	73329	143202
R^2	0.045	0.019	0.052	0.048	0.023	0.036
Control Mean Public	0.271	0.425	0.322	0.482	0.377	0.363
Control Mean Private	0.157	0.397	0.272	0.370	0.348	0.258
Impact Public: {%; SD}	{10%; 0.06}	{7%; 0.06}	{16%; 0.11}	{7%; 0.07}	{11%; 0.08}	{8%; 0.06}
Impact Private: {%; SD}	{23%; 0.10}	{20%; 0.16}	{32%; 0.19}	{21%; 0.16}	{22%; 0.16}	{22%; 0.13}
Test T + Private x T = 0 (p-value)	0.000	0.000	0.000	0.000	0.000	0.000
Panel C: *Interaction with Unlicensed (Private and active at endline only)*						
All Inspections (T)	0.047***	0.072***	0.103***	0.079***	0.078***	0.063***
	(0.013)	(0.015)	(0.015)	(0.015)	(0.014)	(0.013)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Unlicensed at Endline	-0.023***	-0.061***	-0.043***	-0.068***	-0.053***	-0.039***
	(0.008)	(0.018)	(0.013)	(0.015)	(0.015)	(0.011)
	[0.007]	[0.001]	[0.001]	[0.000]	[0.001]	[0.000]
Unlicensed at Endline x T	-0.031**	0.003	-0.052***	-0.016	-0.011	-0.024
	(0.015)	(0.024)	(0.020)	(0.019)	(0.021)	(0.015)
	[0.036]	[0.917]	[0.010]	[0.393]	[0.614]	[.109]
Observations	52378	40101	12452	33213	46381	91763
R^2	0.039	0.027	0.061	0.051	0.028	0.036
Control Mean Licensed	0.185	0.430	0.315	0.424	0.381	0.294
Control Mean Unlicensed	0.117	0.348	0.215	0.304	0.300	0.209
Impact Licensed: {%; SD}	{25%; 0.12}	{17%; 0.15}	{33%; 0.22}	{19%; 0.16}	{20%; 0.16}	{22%; 0.14}
Impact Unlicensed: {%; SD}	{14%; 0.05}	{21%; 0.16}	{24%; 0.13}	{21%; 0.14}	{22%; 0.15}	{19%; 0.10}
Test T + Unlicensed x T = 0 (p-value)	0.023	0.000	0.000	0.000	0.000	0.000

Notes. Robust standard errors are reported in parentheses and p-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. Regressions include controls for the 16 strata included in the randomization (by county and market size) and health facility level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided). Details with variable definitions and calculations are in Online Appendix Section 6.

Table A12: Treatment Effects on JHIC Score for Private Facilities:
Interacted with Number of Public Facilities in Market

	JHIC Score (Private Facilities) (1)
All Inspections (T)	2.425*
	(1.250)
	[0.054]
No. Public Facilities in Market	-0.890***
	(0.333)
	[0.008]
No. Public Facilities in Market x T	1.780***
	(0.427)
	[0.000]
Observations	872
R^2	0.354
Control Mean	33.463
Mean No. of Public in Markets	2.234
Impact Evaluated at Mean No. of Public {%; SD}	{19%; 0.59}
T + No. of Public x T = 0 (p-value)	0.000

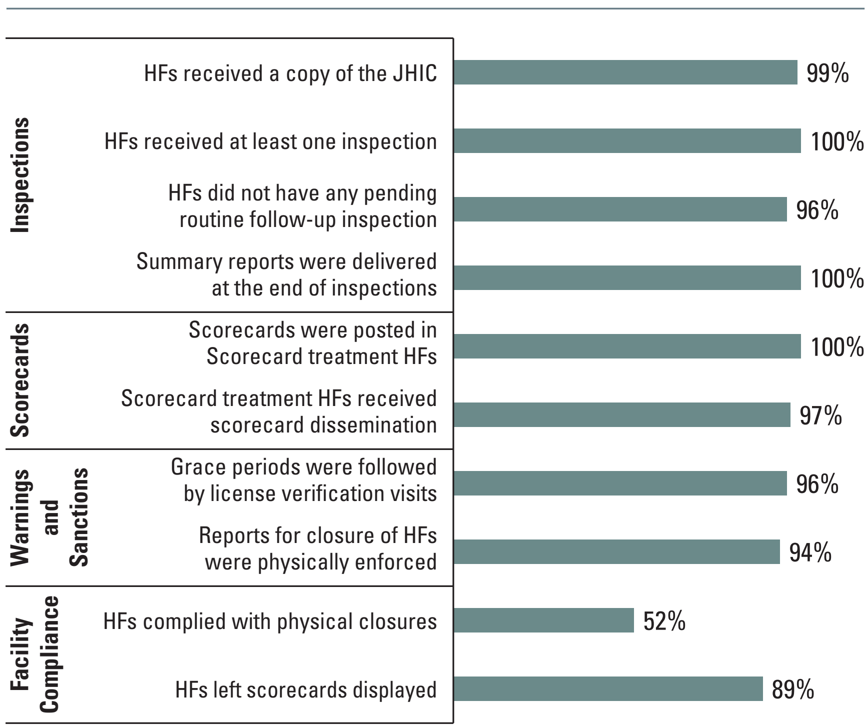
Notes. Robust standard errors are reported in parentheses and p-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. Regressions include controls for the 16 strata included in the randomization (by county and market size) and health facility level controls health facility level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided).

Table A13: Treatment Effects on Select Intermediate Outcomes:
Intervention Awareness, Knowledge, and Perceptions

	I. In-Charge Level		II. Patient Level			
	Familiar with the New Legislation JHIC (Awareness)	Ever Noticed a Scorecard (Awareness)	Know Scorecards' Letter Ranking (Vignette: A vs C vs D)	Ever Noticed a Scorecard (Awareness)	Perceive Improvement in 2017 in HF's Quality (If opened before 2018)	Perceive Recent Government Inspection
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: *Unweighted*						
Inspections (T1)	0.280***	0.090***	0.024	0.009	−0.019	0.012
	(0.039)	(0.032)	(0.019)	(0.012)	(0.029)	(0.020)
	[0.000]	[0.006]	[0.204]	[0.430]	[0.518]	[0.557]
Scorecards with Inspections (T2)	0.321***	0.576***	0.026	0.082***	−0.024	0.025
	(0.033)	(0.031)	(0.018)	(0.012)	(0.026)	(0.017)
	[0.000]	[0.000]	[0.145]	[0.000]	[0.358]	[0.148]
Observations						
Facilities/ (Patients)	1285	1285	1210 (11095)	1210 (11095)	1142 (10162)	1210 (11095)
R^2	0.107	0.287	0.012	0.015	0.012	0.010
Control Mean	0.306	0.233	0.727	0.127	0.667	0.375
T1 (SD Control)	0.608	0.213	0.053	0.028	−0.039	0.025
T1 (% Control Mean)	92%	39%	3%	7%	−3%	3%
T2 (SD Control)	0.695	1.361	0.059	0.246	−0.051	0.052
T2 (% Control Mean)	105%	247%	4%	65%	−4%	7%
Test T1 = T2 (p-value)	0.309	0.000	0.881	0.000	0.834	0.493
Panel B: *Weighted*						
Inspections (T1)	0.239***	0.176***	−0.011	0.020	0.028	0.002
	(0.051)	(0.062)	(0.020)	(0.014)	(0.021)	(0.021)
	[0.000]	[0.005]	[0.582]	[0.143]	[0.182]	[0.943]
Scorecards with Inspections (T2)	0.337***	0.651***	0.006	0.091***	0.009	0.026
	(0.048)	(0.046)	(0.017)	(0.014)	(0.021)	(0.023)
	[0.000]	[0.000]	[0.705]	[0.000]	[0.677]	[0.259]
Observations						
Facilities/ (Patients)	1285	1285	1210 (11095)	1210 (11095)	1142 (10162)	1210 (11095)
R^2	0.165	0.347	0.017	0.015	0.011	0.016
Control Mean	0.338	0.219	0.749	0.130	0.642	0.404
T1 (SD Control)	0.504	0.426	−0.025	0.061	0.059	0.003
T1 (% Control Mean)	71%	80%	−1%	16%	4%	0%
T2 (SD Control)	0.711	1.571	0.015	0.269	0.019	0.052
T2 (% Control Mean)	100%	297%	1%	70%	1%	6%
Test T1 = T2 (p-value)	0.057	0.000	0.302	0.000	0.317	0.253

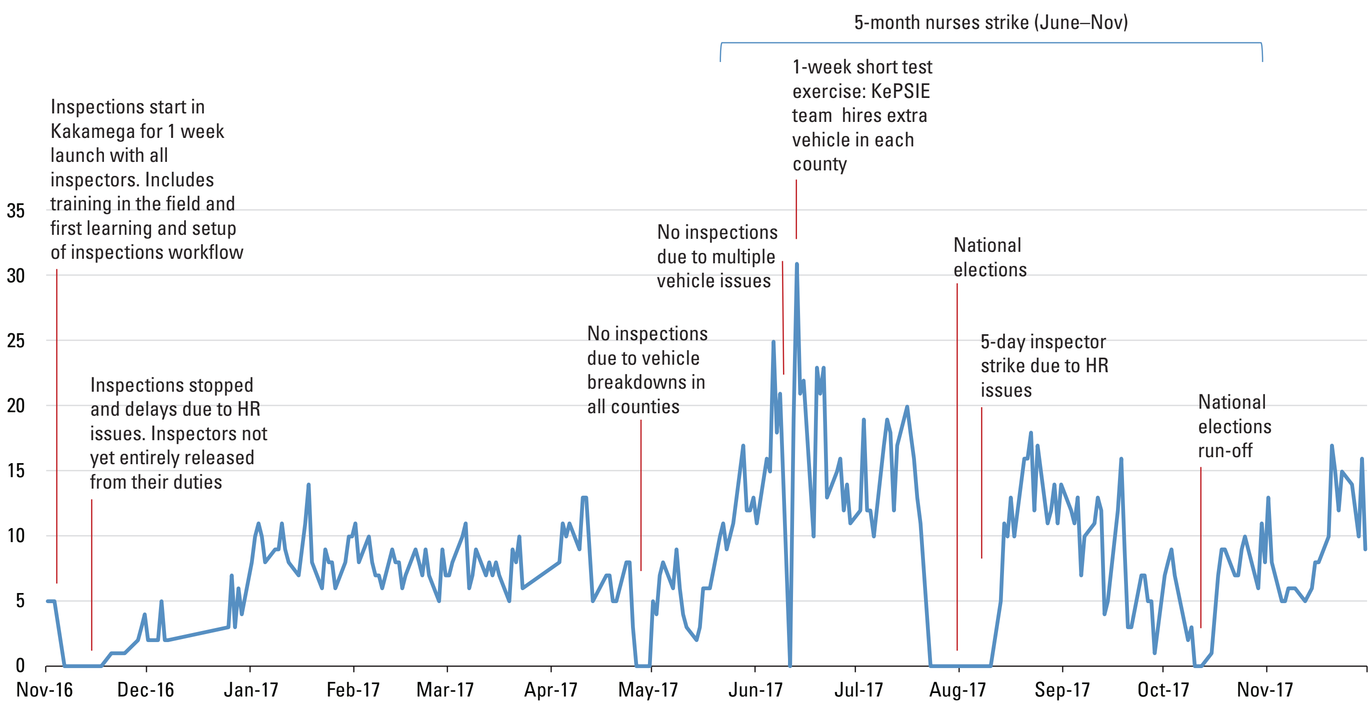
Notes. Robust standard errors are reported in parentheses and p-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. This table reports impact estimates on the knowledge of the new JHIC regulation, the awareness of scorecards, the understanding of scorecards theoretically and practically and the perception of improvement and recent inspection. Column (1) and (2) are at the health facility in-charge level and column (3) to (7) are at the patient level. Regressions include controls for the 16 strata included in the randomization (by county and market size) and health facility level controls health facility level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided). Level 2 is the excluded category. Panel A includes unweighted estimates. Panel B includes weights constructed using average facility outpatients. Details with variable definitions and calculations are in Online Appendix Section 6.

Figure A1: Compliance with Intervention Components



Notes. Source: Bedoya, Das, & Dolinger (2020). HFs = health facilities. Facility compliance with physical closures is based on quality checks on average 3 months after the inspection. Facility compliance with scorecards displayed is based on quality checks on average 2 months after closure.

Figure A2: Inspection Visits by Day and Select Events



Notes. Source: Bedoya, Das, & Dolinger (2020). Vehicle issues include breakdowns/maintenance, no fuel due to payment delays, and vehicles being used by county government.

Figure A3: Scorecards

A. Scorecards for information arm

B. Scorecards for closures in all treatment arms

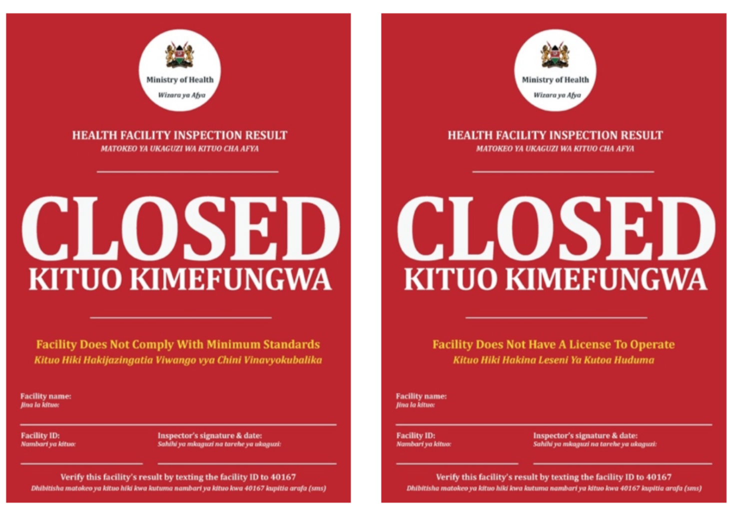
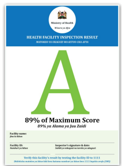
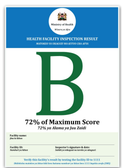
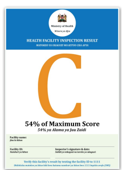
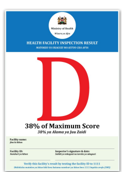


Figure A4: Scorecard Dissemination Materials

A. Description Sheet

	COMPLIANCE CATEGORY	FACILITY SCORE (% OF MAX SCORE)
	• Fully Compliant	• >75%
	• Substantially Compliant	• 61- 75%
	• Partially Compliant	• 41- 60%
	• Minimally Compliant	• 11- 40%
	• Non-Compliant	• 0 - 10% (or absence of licenses)

B. Dissemination Flyer

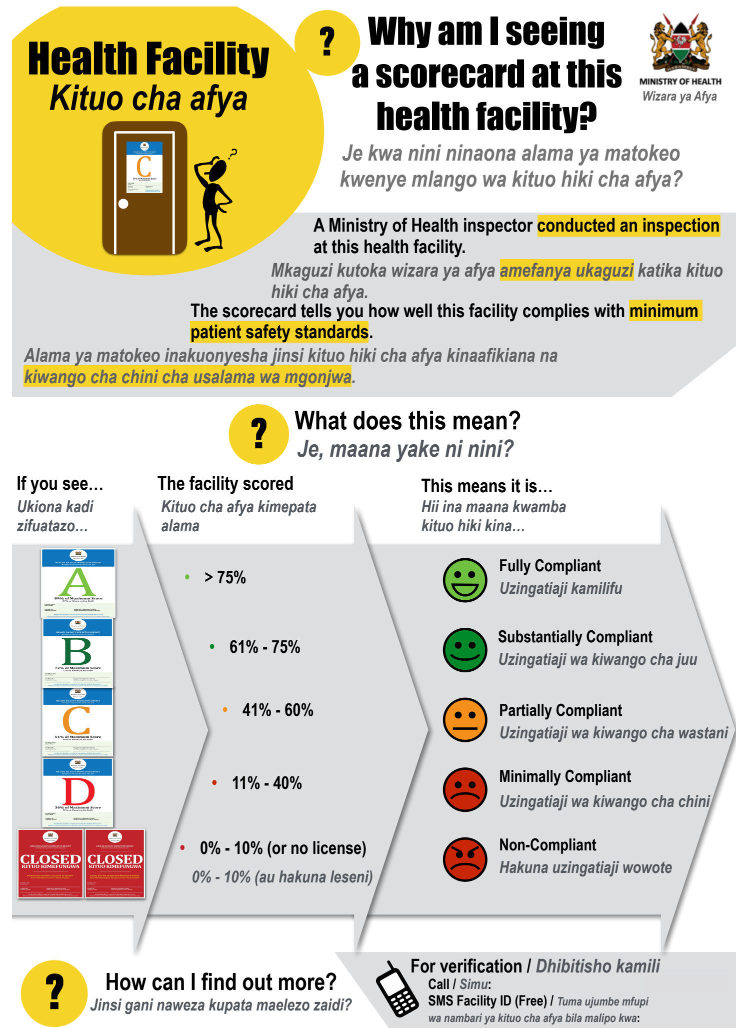
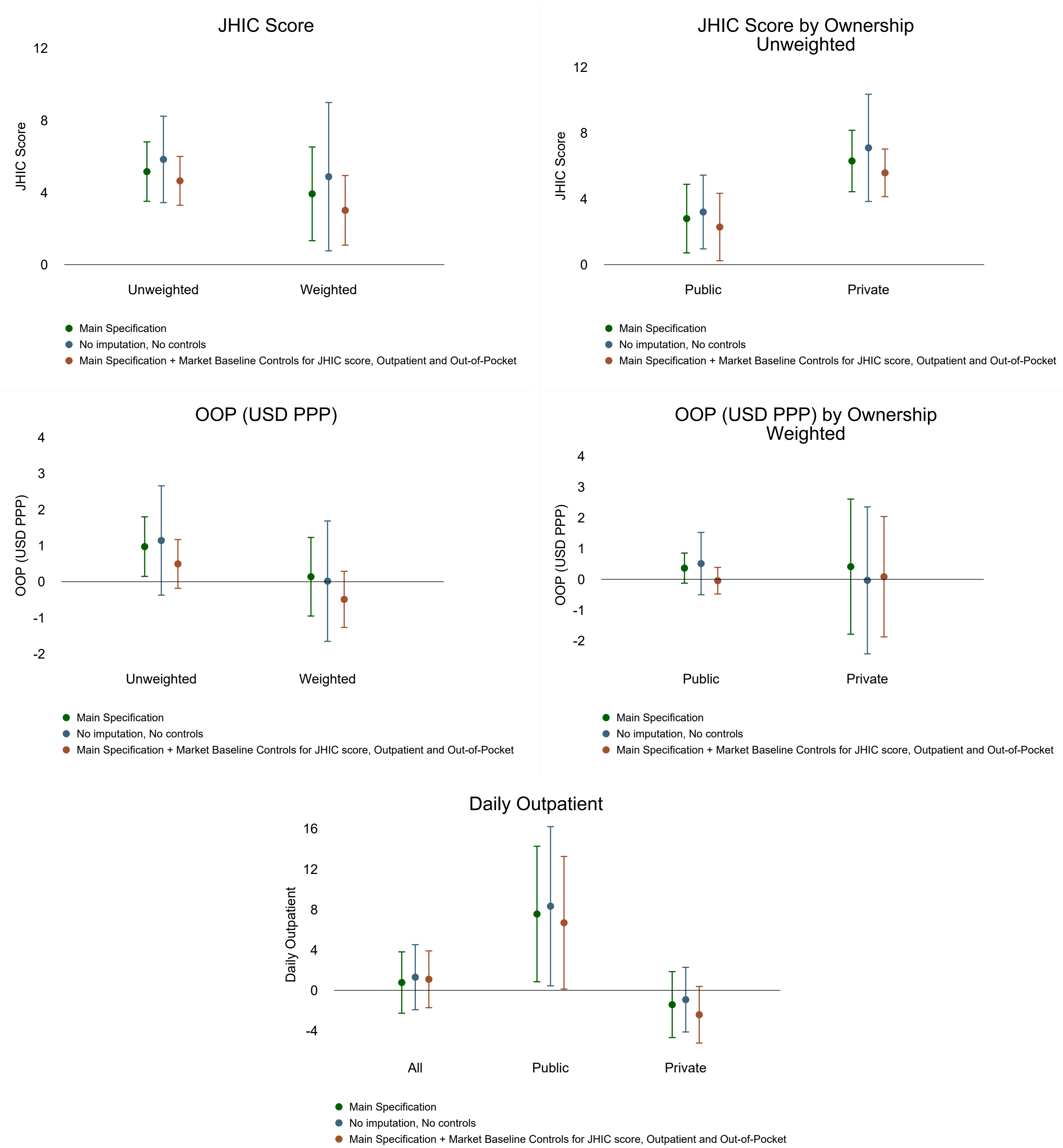
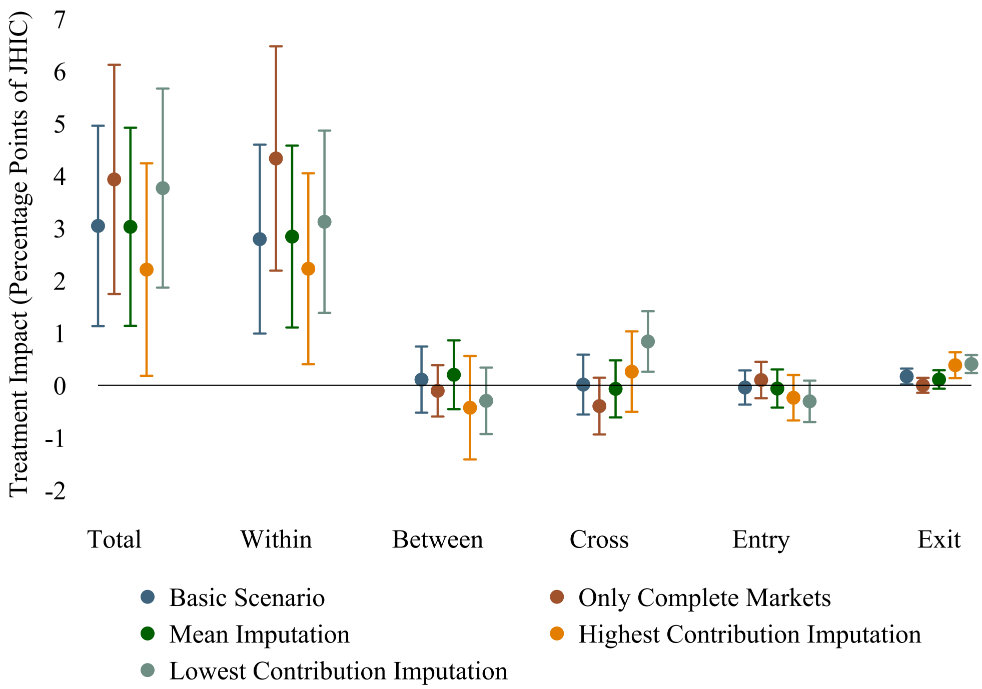


Figure A5: Main Outcomes: Robustness Checks



Notes. Vertical lines correspond to 95% confidence intervals.

Figure A6: Treatment Effects on Weighted JHIC Score and Decomposition Components
Robustness to Different Scenarios



Notes. The dots are the coefficient estimates and the range represents the 95% confidence interval. This graph presents impact estimates on the decomposition of the change of weighted average patient safety score in a market between baseline and endline using the decomposition formula given in equation (1) for different scenarios. Patient safety is defined as the weighted average JHIC score. 'Total' reports the total change of weighted average JHIC score between baseline and endline and is followed by the impact on each of the five terms of the decomposition. All regressions are at the market level and weighted by the market size at baseline. All regressions include 16 randomization strata controls (by county and market size) and control for the percentage of health facilities of each level in the market (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided).

Figure A7: McCrary Test of Density Discontinuity of JHIC Score at Endline, by Treatment Group

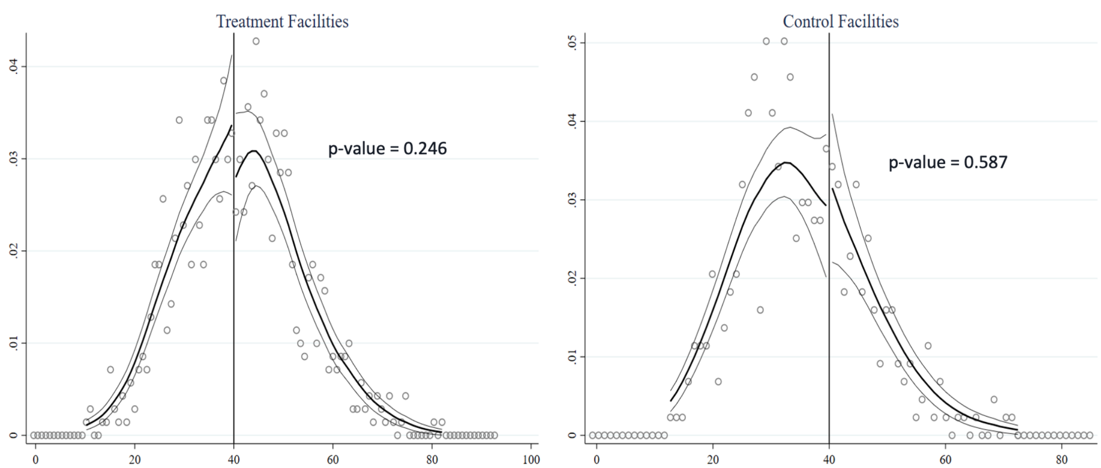
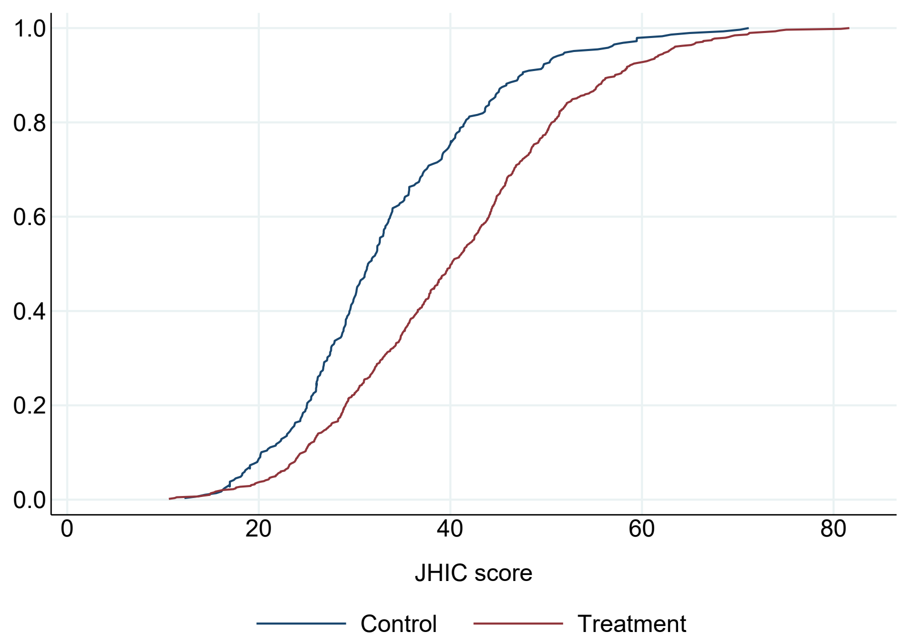


Figure A8: JHIC Score Density by Treatment Status
By Ownership

(a) Private Facilities



(b) Public Facilities

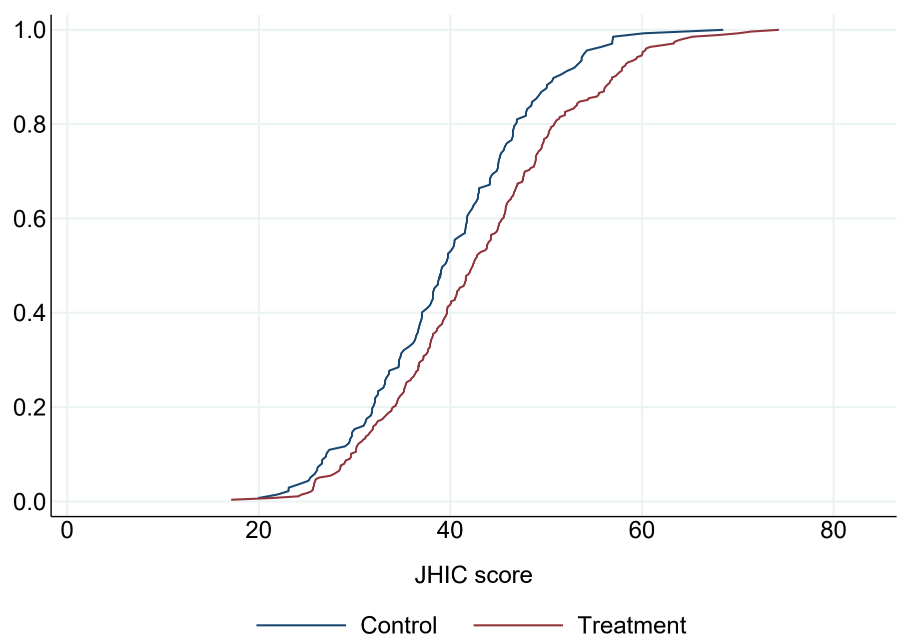
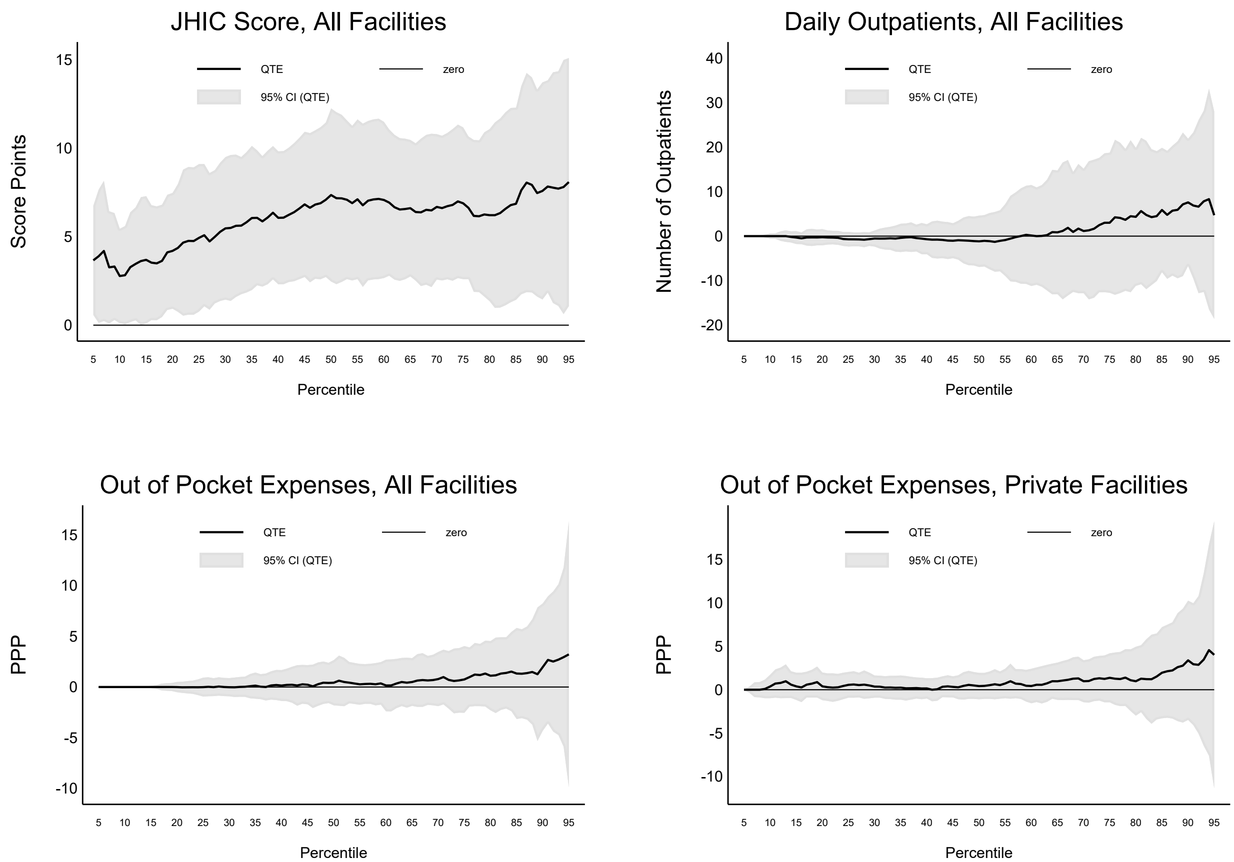


Figure A9: Unconditional Quantile Treatment Effects on JHIC Score, Outpatients and OOP by Percentiles



Notes. Unconditional quantile treatment effect estimates of the outcomes at endline. Estimation is made for every percentile between the 5th and 95th. Bootstrapped 95% confidence intervals (2,000 replications). The confidence intervals control for the 1-family-wise error rates (probability of at least one false rejection across tests), following Romano and Wolf (2010), using codes from Bedoya et al. (2017). Impact estimates are presented for patient safety as measured by the JHIC score, as per regulation, excluding licensing; daily average of out-patients for the month of January as recorded in the health facility's books; and OOP prices in USD PPP as reported by patients, for all facilities and for private facilities only. JHIC score range from 0 to 100 and indicate score as percentage of maximum score. JHIC = Joint Health Inspection Checklist; OOP = Out-of-Pocket ; PPP = purchasing power parity

SUPPLEMENTARY APPENDIX
Randomized Regulation:
The Impact of Minimum Quality Standards on Health Markets

Version: September 2022

1 Health Markets Construction and Characteristics

Markets are defined by geographic delimitations. In the 2015 baseline with 8,598 patients surveyed we find that 73% of the respondents live 4 km or closer from the health facility they visited.[35] Using this distance, we apply a z-center clustering algorithm to the facilities. First, a randomly-chosen facility is defined as a "cluster center." Next, every facility is matched to the market defined by the nearest cluster center. If any facility is more than 4 km from its corresponding cluster center, the facility which is farthest from its cluster center becomes a new cluster center. Then the process is repeated. Our algorithm therefore stops creating new clusters when all facilities are within the predefined distance (4 km) from the cluster center which defines their market.

In our three study counties, 273 markets were identified through the clustering algorithm and were randomized into the treatment and control groups. The three counties have diverse market sizes ranging from singletons to markets with 2, 3, 4, 5, or large markets with 30, 40, or more facilities. Figure S1 shows examples of typical markets in these areas: a singleton, a small market, and a densely populated market. At randomization, 1,258 health facilities in those markets were assigned the same treatment.[36] By endline, 90 facilities that were missed or classified as temporarily or permanently closed during the census were found operational and added to the nearest market for a total of 1,348 facilities operational at randomization. We also applied the market clustering algorithm to the government data as of August 2014 to assess how they compare to the three study countries and compare it with the national figures. We identified 2,795 markets for the 8,776 facilities listed, of which 863 facilities were located in our three counties.[37]

Table S1 shows market characteristics in Kenya and our three study counties using government data (Columns 1 and 2) and for our three counties using our census and survey data (columns 3 and 4). We find the government data over-represent public facilities and singletons with a large overlap between these groups (76% of singleton markets are public) and are missing mostly small, private facilities.[38] Columns 1 and 2 show that there are more private facilities in the 3 countries (48% vs. 39%), more markets of sized 4 or more (65% vs. 57%) and less smaller markets (11% vs. 17% singletons), compared to the national level.

[35] We define a health facility as any premise offering outpatient care or primary care services. Level 2 includes Dispensaries and Clinics; Level 3 consists of Health Centers and Maternity and Nursing homes; Level 4 includes Primary Hospitals; Level 5 includes Secondary Hospitals, and Level 6 consists of Tertiary Hospitals.

[36] This excludes 14 facilities originally listed in the 273 randomized markets were identified later as duplicates.

[37] A feature important to highlight is that we found more facilities in our 2015 census than registered in government records in February of the same year (1,115 vs. 940, or 19% more): We identified 379 new facilities (40%) that were not included the MFL, and we also identified that 202 (21%) from the MFL have gone out of business.

[38] Within the private facilities, we include NGOs and faith-based organizations.

Table S1: Facility Market Characteristics in KePSIE Counties and at the National Level Before the Intervention

	Government Data		KePSIE Data		
	National	3 Counties	3 Counties		
	% of facilities		% of facilities	% of patients [1]	% of patients [2]
	(1)	(2)	(3)	(4)	(5)
Panel A. Facility level					
Facility ownership					
Public	0.61	0.52	0.30	0.71	0.67
Private	0.39	0.48	0.70	0.29	0.33
Facility level					
Level 2	0.80	0.83	0.85	0.62	0.66
Level 3	0.15	0.12	0.11	0.21	0.19
Levels 4 and 5	0.05	0.05	0.04	0.17	0.15
Proportion by market size					
1 facility	0.17	0.11	0.07	0.11	0.11
2-3 facilities	0.25	0.24	0.15	0.19	0.18
4+ facilities	0.57	0.65	0.79	0.70	0.71
Observations	8776	863	1348	1025	1348
Panel B. Market level	% of markets		% of markets	% of patients	% of patients
Singleton markets	0.50	0.38	0.30	0.11	0.11
Only public	0.47	0.32	0.23	0.09	0.09
Only private	0.04	0.07	0.08	0.01	0.01
Markets with at least 1 public facility	0.95	0.93	0.88	0.98	0.97
Observations	2795	240	273	273	273

Notes. Panel A reports statistics from the randomization of study facilities as of 2016. Panel B is estimated using government data provided by Kenya's Ministry of Health in a master health facility list (kmhfl.health.go.ke) available as of August 2014. Markets are generated using a z-center clustering algorithm with a 4 km radius for all health facilities and applied to the government data for comparable estimates at the study and national levels. [1] Outpatient caseload is based on the daily outpatients from facility records for the month of January 2015. [2] Missing values for OOP are imputed using means defined by level, ownership, treatment, license status at randomization, and daily outpatients.

Figure S1: Illustration of Markets

A. Singleton market B. Small market (2-3 facilities) C. Large market (4 or more facilities)

Health facilities

Notes. Markets are defined by geographic delimitations with a 4 km. radius using a z-center clustering algorithm. Panel A is an example from Kakamega County of a singleton market (around 30% of markets are singletons). Panel B is an example from Meru County of a small to medium-size market (around 45% of markets have between 2 and 5 facilities). Panel C is an example from Kilifi County of a large market (around 25% of markets have more than 5 facilities, and where the largest randomized market has 45 facilities).

2 Regulatory Reform Background

Kenya's government initiated a series of reforms to improve their health inspections system in the past decade. In 2012, a Joint Health Inspections Checklist (JHIC) was published in the official gazette in an effort to improve fairness, coordination, and transparency. This JHIC replaced multiple individual inspections from Kenya's regulatory "Boards and Councils" (B&Cs), improved standardization, and reduced costs to health facilities. There are three main types of inspections conducted: (1) inspections of new facilities before issuing licenses; (2) inspections of operating facilities due to complaints received, for instance, due to malpractice cases (i.e., reactive inspections), and (3) inspections to verify the quality of health providers. Before the intervention we evaluate, inspections in (3) which are closest to the subject of this study, occurred with low probability in small geographic regions (inspectors would go to these regions and inspect all health facilities in the surrounding areas based on the quota for the inspection period). The boards and councils and the Ministry of Health that are responsible for the inspections must supervise them from their headquarters in Nairobi, since there was no regional network of inspectors established. At the moment of starting this study, there were nine regulatory boards and councils: Medical Practitioners and Dentists Board, Clinical Officers Council, Nursing Council of Kenya, Radiation Protection Board, Pharmacy and Poisons Board, Council of the Institute of Nutritionists and Dieticians, Kenya Medical Laboratory Technicians and Technologists Board, and Public Health Officers and Technicians Council.

In October 2013, all stakeholders led by the Ministry of Health (MOH) held a two-day workshop to assess progress and signed the "Windsor Agreement," embarking on an ambitious health inspections reform to improve patient safety and health care in Kenya. This agreement became the basis for this study.[39] The assessment identified the following challenges with the JHIC and the inspection framework:

1. Unclear and discretionary rules of the game (e.g., rating in the inspections was not consistent due to lack of definitions on specific items to be inspected): The 2012 version of the JHIC did not adequately define the concept of patient safety, and failed to specify consequences of inspections. The definition of patient safety was 'fuzzy', both in terms of concepts being measured, and how they were measured. For example, the 2012 JHIC asked inspectors to rank from 1 to 5 whether there was an emergency tray in the facility, without listing specific components of the tray that guarantee minimum patient safety, or what to do in cases where some components were present. This gave inspectors a certain degree of discretion, thereby weakening the regulatory agencies' ability to monitor facility performance and improvement, and prioritize improving elements that matter the most for patient safety. Lack of clarity on

[39] The stakeholders and signatories of the Windsor Agreement included the MOH, the regulatory boards and councils, and representatives from the private health sector.

regulation has also been linked to problems of bribery and corruption in the Kenyan health sector. Another weakness was poor dissemination of the 2012 JHIC. While the checklist was gazetted and made publicly available online, many health facilities reported that they were unaware of the current regulation, and could not, as a result, conduct self-assessments and improve voluntarily.

2. Lack of incentives to improve patient safety at different levels of compliance with the standards, and with unclear and virtually nonexistent sanctions and weak enforcement except for extreme cases of malpractice: There were no clear follow-up actions to take based on inspection results. The lack of clear consequences, combined with the lack of information on part of health facilities on the consequences of inspections, weakened the capacity of the then current inspections regime to encourage improvement or enforce sanctions at the facility level.

3. Inadequate capacity to conduct inspections: Only about 4% of health facilities in Kenya were inspected each year by the B&Cs and virtually none were re-inspected (as of July 2015). Reasons for the low inspection coverage include limited numbers of trained inspectors and inadequate resources to support field visits (there's only a small pool of joint inspectors, all of whom are based in Nairobi). Furthermore, the B&Cs had limited capacity to follow up on warnings and sanctions imposed by inspectors. There was also variation in the background of inspectors representing the different boards and councils. This is likely to result in some variability in the outcome of inspections.

4. Lack of an integrated information system: The 2012 JHIC was not linked to an integrated information system, therefore limiting the ability of actors to use inspections results to strategically manage inspection process and prioritize follow-up efforts. As of July 2015, the results of the inspections were physically archived in some B&C, with very few manually typing reports for electronic archiving. Not having an electronic system limits the regulators' ability to, for example, identify priority locations to support, or reference results from previous inspections of the same facility, or compare results across facilities to identify shared constraints to inform better targeting of policies and remedial interventions.

Stakeholders judged this as a sufficient basis for action and constituted a technical working group led by the MOH to further the reform of a regulatory framework, and committed to conduct an evaluation to provide evidence on the impact of the inspection reform on patient safety and quality of care. The task force was appointed with authority and mandate to oversee all aspects of the design and implementation of the pilot inspections regimes, provide support on key decisions, deal with any issues that arise during the implementation, and to play a critical role after the intervention including dissemination of results and decision-making for national policy. The task force consisted of members from the MOH, the regulatory boards councils, representatives from the private sector, and county health teams, with the aim for involvement from public and private sectors as well as from county health teams responsible for providing health services and the central government, which is responsible for the regulatory function.

In March 2016, an enhanced regulatory framework was published in the official gazette based on the inspection reform efforts of the task force. The main elements of the new regulatory framework defined by Legal Notice No. 46 of the Public Health Act and the implementation guidelines prepared by the technical working group include a refined JHIC focused on fundamentals of patient safety, a scoring system that allows facilities to be categorized according to the level of risk presented to patients, and warnings and sanctions that are triggered by the scores according to a facility's level of risk. The interventions evaluated as a part of the study, which are based on the new regulatory framework, as well as the instruments, and evaluation design, were designed through a participatory approach over a three-year process. The three counties where the study takes place, for instance, were selected by the health management representatives of the 47 Kenyan counties to represent different conditions and markets in the country. The inspection checklist was designed by the regulators and tested before application in the three study counties. The efforts were the result of commitment by the Kenyan government and stakeholders to conduct joint health inspections and continue improving the inspections reform process by testing alternative regimes through the study.

3 Joint Health Inspection Checklist (JHIC) Instrument

The standardized regulatory checklist largely focuses on structural measures of quality of care and patient safety. Specific elements within the structural category look at standard procedures and policies, infrastructure, equipment and supplies, and medical records. For instance, inspectors confirm that basic medicines are available in an emergency tray, or that the facility has adequate storage for safe water supply. In addition to these measures, there are a few process measures of quality of care and patient safety. This may include, for example, reviewing a sample of patient charts for appropriate care, such as whether a completed partograph accurately monitors TPR/BP and codes drugs. Overall, most sections are very much a combination of these elements within structural measures of quality of care. The JHIC scores for each item are defined in the regulation using an equal-weight scoring system with total values per section ranging from 0 to 100. The score represents the score as a percentage of maximum score. The new regulatory framework defines all items to be inspected with a specific score for each item. Therefore, any facility can predict with precision what score and in which category they would fall using the gazette checklist which includes the table with scores and corresponding warnings, sanctions, and rewards.

There are five general sections that represent minimum standards for all facilities, such as measures of infrastructure, management, and infection prevention and control. The remaining eight sections are connected to particular services that the facilities offer, including labor ward, medical and pediatric wards, theatre, pharmacy, laboratory, radiology, nutrition and kitchen services, and mortuary.[40]

Examples of the minimum standards for all facilities include items across administration (e.g., having licenses for the facility, departments, and staff), health facility infrastructure (e.g., having a stable source of power, and safe, clean water available from a tap or container with sufficient storage, and the presence of an emergency signage), general management and recording of information (e.g., having patient records with names and unique patient numbers, collecting and regularly reporting performance indicators on immunization and notifiable diseases, and having service contracts for maintenance and calibration of medical equipment), infection prevention and control (e.g., evidence of disinfectant use, availability of soap at hand-washing areas, and protocols for sterilization of equipment), and medical consultation (e.g., access to essential medicines including glucose, adrenaline, sodium bicarbonate, diazepam, and phenobarbitone in an emergency tray, and a system for coding patients in triage at the first point of contact with patients).

Further sections are activated for facilities that provide additional services including labor ward (e.g., access to a sterile delivery set and resuscitation equipment with oxygen, suction machine, and ambu bags, evidence of monitoring labor with partographs where contraction, cervical dilation, color

[40]In our sample, health facilities offer the following services: consultation (100%); labor ward (31%); medical and pediatric wards (11%), theater (4%); pharmacy (59%); laboratory (58%); radiology (3%); and nutrition (5%).

coding, TPR/BP, urine input/output, and drugs are properly charted and recorded, and availability of procedures for obstetrics emergencies, such as obstructed labor fetal distress, eclampsia, and APH/PPH/HELLP), medical and pediatric wards (e.g., having regular ward rounds, standardized procedures for patient admission as well as handover and discharge reports, and access to basic monitoring equipment and supplies such as pulse oximeter, BP machine, thermometer, and suction machine), theatre (e.g., presence of an operating area with adequate space, lighting, and equipment such as anesthetic machine, monitors, laryngoscope, endotracheal tubes, and a back-up oxygen source, and a scrubbing area with clean clothing, masks, caps, theatre shoes, and hand-washing area), pharmacy (e.g., having adequate storage conditions for drugs, having a system for feeding back medication errors, documentation showing where medicines are procured, and staff that are familiar with reference materials), laboratory (e.g., having procedures for handling, labeling, and storage of specimens, having quality controls for internal and external review and improvement of lab tests, and having documentation for equipment registration, validation, and calibration), radiology (e.g., having radiation badges for monitoring, a radiation safety service provider, and safety measures with waste management), nutrition and dietetics (e.g., availability of nutrition equipment, formulations, and supplements, such as a weighing scale, glucometer, reference charts, MUAC tape, vitamin A, iron and folate and zinc, and F75 and F100 diet milks), and mortuary (e.g., having standard operating procedures for receiving, identifying, storage, release, and disposal of bodies, adequate infrastructure such as a working drainage system, and adequate security measures).

4 Census, Response Rates, Attrition and Accretion

We conducted two complete censuses that coincide with baseline (2015) and endline (2018) data collection. The sample is very close to the census and therefore all our statistics can be interpreted as population statistics. We will refer to 1,319 as the 2018 census of facilities in this study at follow-up, and similarly to 1,104 as the 2015 census of facilities at baseline. The response rate was 97% at endline and 93% at baseline. **??** shows the response rates at baseline and endline and by survey instruments. The randomization was conducted in December 2016, more than one year after the baseline census due to delays in the start of the implementation. With high turnover in facilities, a partial update of the census was conducted between October and November 2016 in markets of size 1, 2, and 3.[41] At randomization, 1,258 were listed after the census update.Throughout the project we continuously conducted listing updates on a rolling basis. By endline, 90 facilities that were missed or classified as temporarily or permanently closed during the census were found operational and added to the nearest market for a total of 1,348 facilities operational at randomization. In total, between randomization and endline, 1,537 facilities are listed in the study as operational at some point, of which 1319 are active at endline (Table S2).

As Table S4 indicates, we allocated all new facilities that were found by endline census to one of the 273 fixed markets that were randomized into the treatment and control groups. These new facilities were matched using an algorithm to the nearest randomized market, according to their proximity (if they are within the established proximity). Three of these new facilities are excluded from the analysis as they are beyond a defined distance (5km) from any randomized market. Five markets disappear because they are permanently closed or out of business, for a total of 268 markets at endline. We show the facilities that were active at randomization based on the new facilities that were added to the census in the continuous updates, and given we know when they started operations. **??** shows treatment status as per fixed randomization market versus treatment received. Note that most new facilities and the facilities missed at randomization that were allocated into the fixed treatment markets did not received the treatment, as the majority were added during endline. That is the reason why the proportion of treatment effective for the census at endline is lower than for the randomization census.

Table S6 shows the response rate by data collection wave.

[41]Given the financial and logistical constraints, the rationale for this partial update came from these being the market sizes more potentially affected by the intervention by having one more or one less facility around. However, this partial update was limited, in particular for updating facilities that were out of business (or inactive).

Table S2: Response Rates at Baseline and Endline

	Baseline	Endline
	(1)	(2)
Panel A. Facility level indicators		
Active facilities [1]	1104	1319
Consent rate	0.93	0.97
Facilities surveyed [2]	1027	1285
Panel B. Surveys completed [3]		
Facility level surveys	1027	1285
Patient exit interviews	8577	11098
HCW-patient interactions (est.) [4]	18558	19178
Healthcare worker surveys	1625	2098
Site observations	1886	2534
Panel C. Response rates		
Patient exit interview	88%	85%
Healthcare worker surveys	97%	98%
Site observations	100%	100%
HCW-patient interactions (patient consent to observation)	99%	100%

Notes. Non-response includes direct refusals, which explicitly request right to not participate in the surveys, and indirect refusals, where surveyors unable to complete surveys after multiple visits and attempts following a visits protocol. [1] Active facilities include all facilities found operational as per the last KePSIE visit, and excludes those ever listed but that have permanently closed (out of business), are duplicate, only offer specialized services, cannot be located, or have relocated. [2] At endline, three facilities were part of the census and consented but are excluded as they are beyond a defined distance (5km) from any market defined by our clustering algorithm. [3] Include all surveys consented and completed in their entirety. The totals reported here refer exclusively to facilities where data collection is complete. [4] Surveyors conducted the patient-provider observations for a minimum of 3 hours in the baseline data collection, and for a minimum of 2 hours in the follow-up.

Table S3: Census at Randomization and Endline

	All	Public	Private
	(1)	(2)	(3)
a. Randomization census (Dec 2016)	1258	383	875
b. Facilities active at randomization but missed in census [1]	90	21	69
c. Facilities active at randomization (a+b) [2]	1348	404	944
d. New facilities after randomization [3]	189	12	177
e. Inactive facilities at endline	218	2	216
f. Total in study from randomization to endline (c+d)	1537	416	1121
g. Of which, active at endline (f-e)	1319	414	905

Notes. [1] Based on updated census data with facility self-reported year of establishment. [2] This is our estimate of active facilities at randomization and may be inflated as new facilities were listed in a partial census update in markets of sizes 1, 2, and 3, but out of business (inactive) facilities were not able to be confirmed. [3] Active at endline but were not active at randomization.

Table S4: Randomized Markets and Corresponding Facilities by County

	Markets				Facilities			
	Kakamega	Kilifi	Meru	Total	Kakamega	Kilifi	Meru	Total
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. At Randomization [1]								
Treatment 1	30	31	29	90	125	82	202	409
Treatment 2 (Scorecards)	30	33	33	96	139	146	219	504
Control	28	31	28	87	113	91	231	435
Total	88	95	90	273	377	319	652	1348
Panel B. At Endline [2]								
Treatment 1	30	30	29	89	119	76	198	393
Treatment 2 (Scorecards)	30	33	30	93	131	140	219	490
Control	28	30	28	86	106	85	245	436
Total	88	93	87	268	356	301	662	1319

Notes. [1] Includes baseline census plus a partial census update that was conducted during the implementation. 273 markets were identified through the clustering algorithm and were randomized into the treatment and control groups. Health facilities in those markets were assigned the same treatment. [2] Includes the randomized markets that were still active at endline, or those with at least one health facility found in the market. 268 markets still remain. Five markets disappear because they permanently closed or out of business. The endline census includes 213 new facilities that were not in the randomization stage and excludes 189 facilities that were out of business. New facilities identified were matched using an algorithm to the nearest randomized market. Three of these new facilities are excluded from the analysis as they are beyond a defined distance (5km) from any randomized market.

Table S5: Treatment Status vs. Treatment Received

	Treatment Received			Out of business at 1st inspection	Total
	Treatment 1	Treatment 2 (Scorecards)	No Treatment		
	(1)	(2)	(3)	(4)	(5)
Panel A. Active at Randomization [1]					
Treatment 1	0.90	0.02	0.03	0.05	409
Treatment 2 (Scorecards)	0.00	0.85	0.05	0.10	504
Control	0.03	0.00	0.97	NA	435
All					1348
Panel B. Active at Endline [2]					
Treatment 1	0.87	0.02	0.11	NA	393
Treatment 2 (Scorecards)	0.00	0.82	0.17	NA	490
Control	0.03	0.00	0.97	NA	436
All					1319

Notes. [1] Effective treatment differs from treatment by taking into account (i) re-allocation of some markets due to contamination (19 facilities) and (ii) the fact that some facilities included in the randomization became inactive (went out of business, were found duplicate, etc.) during implementation before they were ever inspected (66 facilities). Column (3) includes facilities that were reported by inspectors as out of business but found to be operating during the endline census and therefore did not receive treatment (5 facilities). [2] The endline census includes 213 new facilities that were not in the randomization. New facilities are matched to the nearest randomized market using a clustering algorithm. Three of the new facilities are excluded from the analysis as they are beyond a defined distance (5km) from any randomized market. Column (3) includes facilities that were reported by inspectors as out of business but found to be operating during the endline census and thereforer did not receive treatment (5 facilities).

Table S6: Response Rates by Waves

	(C) Control Mean	(T-C) Adj. Diff.	Obs.
	(1)	(2)	(3)
Response Rate at Baseline	0.902 (0.298)	0.040*** (0.015) [0.008]	1104
Surveyed Facilities at Baseline (Randomization sample)	0.717 (0.451)	0.060** (0.027) [0.025]	1348
Response Rate at Endline	0.975 (0.157)	-0.000 (0.010) [0.974]	1319

Notes. Standard errors are clustered at the market level and reported in parentheses. *** (**) (*) denotes significance at 1% (5%) (10%) level. Column (2) presents adjusted differences between the means for the treatment markets and the control group. These differences include controls for the 16 strata included in the randomization (by county and market size). This table shows the response rate of the targeted facilities listed in the census at baseline and endline. Non-response includes direct refusals, which explicitly request right to not participate in the surveys, and indirect refusals, where surveyors unable to complete surveys after multiple visits and attempts following a visits protocol.

5 Ethical Considerations

The study was approved 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. We took care to ensure that our research satisfied the 3 basic principles advanced in the Belmont Declaration (1979) of Respect for Persons, Beneficence and Justice. We sought consent from every facility and patient who participated in the study and by working with every facility in the 3 counties (which were themselves chosen to represent the diversity of counties in the country) we ensured that the principle of justice was fulfilled. The question of whether facilities who do not have a license to practice and/or do not meet minimum patient safety requirements should be closed down, as required by the regulators, satisfied the principle of equipoise, with uncertain potential for harm or benefit. Although there is a strong belief in the global health community and among policymakers that unlicensed practitioners (who are therefore practicing illegally) should not be allowed to function, there is no prior evidence on the costs and benefits of actually doing so. If such clinics are the only option for local populations, shutting them down could decrease health care provision. Prior to the intervention, we used the full census data to plot each market and found that in Kenya, clinics that are stand-alone (that is, do not have any other clinics within 4Km) were predominantly public. This meant that were unlicensed clinics, which are all private, to be shutdown, it would still allow for multiple other options in the vicinity. In our data, 97% of unlicensed facilities have at least another facility in the market and 93% have 2 or more. In cases where clinics were closed by federal regulators, the closures were conducted with the full cooperation of the county authorities and the local community.

6 Variable Definition and Construction

There are various units of analysis for our outcomes including health facility, health care worker (HCW), patient, HCW-patient interaction, and health facility market. Markets are defined by geographical clustering as described in Text 1. For instance, the Joint Health Inspection Checklist (JHIC) score, and the outpatient caseload are collected at the health facility level. Questions relating to out-of-pocket expenses are asked to a sample of outpatients in exit interviews. Measures of Infection Prevention and Control (IPC) practices (adherence, knowledge and supplies) are collected through direct observation of healthcare worker-patient interactions, while IPC knowledge surveys were applied to the healthcare workers observed, and observation of the availability of supplies required for the practices at the sites were taken while observation took place. We use weights to adjust for sampling variation and estimate population statistics at the patient, healthcare worker and site level, based on the proportion of surveyed outpatients, healthcare workers and observation sites, with respect to the total outpatient caseload, medical staff and observation sites of the census of facilities.

6.1 Main Outcomes

JHIC Score

This indicator measures patient safety as per adherence to Kenya's Joint Health Inspection Checklist developed as part of the intervention and gazetted in March 2016 under the Public Health Act. The JHIC includes indicators of patient safety related to protocols, infrastructure, and equipment at the facility level. We included 11 JHIC sections in the survey: (1) health facility infrastructure; (2) general management and recording of information; (3) infection prevention and control; (4) medical consultation; (5) labor ward; (6) medical and pediatric wards; (7) theatre; (8) pharmacy; (9) laboratory; (10) radiology, and (11) nutrition and dietetics. The first 4 sections are common to all facilities. License information and the mortuary section were excluded for difficulties to implement them using non-government surveyors. The JHIC scores and its aggregation are defined in the regulation using a scoring system, we follow exactly this aggregation. The values range from 0 to 100 and represents the score as a percentage of maximum score.

Outpatient Caseload

Equals the daily outpatient caseload per facility. We use three main sources of data starting with monthly records and, when no records are kept, self-reported weekly and daily data. All data were collected as part of the health facility surveys. The correlation between daily outpatients from monthly records and weekly self-reported data, when data are available for both, is 0.7.

The monthly records checked at the facility level include (1) Ministry of Health (MOH) official reporting books for aggregate totals (MOH Form 717); (2) daily records by patient age (MOH under and over 5); and (3) any other available facility-based monthly patient registers. Sources (2) and (3) were only asked if (1) was not available. We also collect the number of days missing from the registers, if any. We collect monthly data for 4 different months: November 2017 to February 2018 from any facility records in endline, and November 2014 to February 2015 in baseline. For facilities that started operating in 2018, we also collect data for March and April of that year during endline. Self-reported weekly outpatients for the previous week is collected for all facilities. Daily outpatients for the previous day and two days before the survey are collected for facilities that do not have monthly records.

For the main analysis we use the data from the month of January (2015 for baseline and 2018 for follow-up) as this month has the largest proportion of monthly data. We conduct robustness checks with other months, as well as the average across all months available and results are not affected by the selection of month.

Out-of-pocket Expenses

Total amount paid for the facility visit as self-reported by patients for all services the day of the visit. Monetary amounts are reported in nominal and purchasing power parity (PPP)-adjusted USD terms. The latter is set at 2015 prices using the Kenyan CPI and PPP conversion factor from the IMF, unless otherwise stated. Figures in current USD are converted at the exchange rate for the year the data were collected with the IMF exchange rates for the corresponding year: 2015 (for baseline data). The exchange rates used are: 1 USD = 98.18 KES (2015).

6.2 Secondary Outcomes

Inactivity/Exits

Equals 1 if a facility is no longer operating as a health facility. This includes those facilities that have permanently closed (out of business) or have changed their operations and only offer specialized services (e.g., standalone laboratories or pharmacies) and no longer provide general outpatient services.

New/Entry

Equals 1 if a facility opened after randomization.

Formality

Equals 1 if a private facility has a valid license as per Boards and Councils registries.

Wealth Index

Constructed asset index based on ownership of key assets based on the socioeconomic questions from the Kenya Demographics and Health Survey (DHS) survey to permit calibration against the national distribution. For the correlation of the DHS wealth index with selected variables, all variables are highly statistically significant and explain 80.7% of DHS wealth index variables, and there is high correlation between DHS wealth index and the wealth index with study variables on the DHS sample ($R^2 = 0.76$).

Infection Prevention and Control

Outcomes from this set of outcomes follow definitions as per Bedoya et al. (2017).

(a) Compliance with IPC Practices. Direct observation of IPC practices is based on patient-HCW interactions observed in the consultation room, laboratory, and injection room, where most outpatient care occurs. Compliance indicators are constructed based on indications and actions. Indications are the principal rationale for performing a safety action and occur when there is a risk of transmitting pathogens from one surface to another. For example, "before patient contact" is an indication as there is a risk of microbial transmission to the patient. Actions occur in response to the indications and therefore define whether the safety practice took place, for instance, "washing hands with water and soap prior to patient contact." Compliance is estimated as the proportion of indications accompanied by the appropriate IPC action. Compliance is estimated individually for each of 20 indications that we measure across the five domains and is defined as:

$$Compliance_{iw} = Action_{iw}/Indication_{iw} \tag{4}$$

where i denotes indication (of 20 indications), and w denotes healthcare workers. Compliance equals 1 if the action corresponds to the right action for each indication, and zero otherwise.

We analyze compliance with safety practices at the indication level: as long as an indication is observed, the action taken in response by the healthcare worker is recorded, regardless of whether the indication occurred at the same time as another indication. We present the mean of the individual indicators as well as the means aggregated by domain and for all safety measures. This implies that the weight of each indicator is its frequency. Weighting across domains/indicators may rely on several possible schemes. For instance, if we wanted to compute the costs of compliance, we could weight these by the cost of the supplies for each; alternatively, if we focused on the benefits, we would weight by the mapping from compliance to the likelihood of Healthcare Acquired Infections. Unfortunately, the literature is far from establishing these likelihoods even in high-income countries, and particularly for primary care.

(b) HWC IPC Knowledge. Equal to 1 if the HCW answers correctly each question related to every indication and domain measured. To develop the measures of knowledge of healthcare workers, we selected questions that map into the practices observed in each of the five domains. We use a combination of open-ended questions such as “Can you name the most important indications where hand hygiene is recommended?” as well as questions such as “Do you agree or disagree with the following statement: Washing hands before putting on gloves to examine a patient is not necessary,” to elicit knowledge of appropriate use of gloves. We further measure knowledge on characteristics of the practice that may or may not be observed such as “Do you know what is the minimum time recommended to perform hand washing to kill most germs on the hands?” which is not observed if the health care worker never washes his hands. We triangulate some of the questions (e.g., asking a similar concept with different wording and changing the answer between “agree” and “disagree”) and find overall agreement for the questions tested.

(c) Compliance with IPC Supplies. Equal to 1 if the supplies are available based on supplies required to perform the practice for every indication and domain measured. We measure the availability of supplies, policies, and practices necessary for compliance with IPC standard procedures. Supplies were assessed through structured observation during HCW-patient interactions, and only for healthcare workers who had the relevant indications e.g., the supply of puncture-resistant sharps containers was only checked for the health care workers that had to conduct injections. Data are collected at the level of the observation site since their presence may be critical at more than one site in the facility, and as some of them are particular to the procedures performed at that site. Only a selection of practices could be mapped into this set due to difficulties in observing the availability of some supplies without asking for them if they are not visible and/or due to the HCW not performing the safety practice at the moment of observation.

Facility In-charge awareness and knowledge of the regulation and the scorecard

(a) Familiar with the new legislation JHIC (awareness)

Equal 1 if facility in-charge answered “Yes” to the following question, 0 otherwise: “Are you familiar with the new legislation, gazetted in 2016, that uses a new Joint Health Inspection Checklist to inspect public and private health facilities?“

(b) Ever noticed a scorecard (awareness)

Equal 1 if in-charge answered “Yes” to the following question, 0 otherwise: “Have you noticed a scorecard such as this one posted on any health facility?”

Patient awareness and knowledge of scorecards, perceptions about quality and government inspections

(a) Ever noticed a scorecard (awareness)

Equal 1 if patient answered "Yes" to the following question, 0 otherwise: "The government has introduced scorecards in some health facilities. Have you ever noticed a scorecard similar to this one posted on a facility? (Show the image to the respondent/patient until he/she give you an answer.)"

(b) Know scorecards' letter ranking (A vs C)

Equal 1 if patient answered "A" to the following question, 0 otherwise: "Imagine there are two health facilities with two different scorecards (Note: one facility has an "A" scorecard, the other facility has a "C" scorecard), as shown below. Which one do you think is better? (A vs C)"

(c) Know scorecards' letter ranking (vignette: A vs C vs D)

Equal 1 if patient answered "I would go to Clinic Belief" to the following question, 0 otherwise: "Imagine that there are three clinics in your village/city/town: Faith, Hope, and Belief. You have always gone to Clinic Faith. Yesterday you found out that the government gave a scorecard D to your facility, a C to Clinic Hope, and an A to Clinic Belief. Based on this information, would you change the health facility you go to? (Show the image to the patient/respondent.)" (Note: An image is given to help understanding)

(d) Perceive improvement in 2017 in facility's quality (If opened before 2018)

Equal 1 if patient answered "improved" to the following question, 0 otherwise: "Since 2017, have you perceived any change in the quality of services provided by this facility?"

(e) Perceive recent government inspection

Equal 1 if patient answered "Yes" to the following question, 0 otherwise: "Do you think this facility has been inspected recently by the government to verify the quality of the services it provides?"

6.3 Other Key Variables

Closure-type Facilities

In the experiment on health facility inspections, an important number of facilities in the treatment group were closed by the government (mostly due to the lack of a license). We want to assess whether the intervention had differential impacts in markets with "closure-type" facilities, which could have acted as a signal for the population due to their prominence. Since facilities were only closed in the treatment group, to run the analysis, we need to estimate a counterfactual, or in other words, which facilities in the control group would have been closed by the government had they been in the treatment group.

The end goal is to estimate the number of closure-type facilities per market, which we will use for heterogeneity analysis. To predict whether a facility is a closure-type, we use a random forest algorithm applied to a limited number of "time-fixed" variables (including the license status). The variables used for the prediction are: not having a license at randomization, having an expired license at randomization, the facility level, a categorical variable corresponding to when the facility was open, the type of private and the county. The sample is restricted to private facilities operating at randomization.

We also run two robustness checks. The first one uses a logit model instead of the random forest algorithm to make the predictions. For the second robustness check, we use whether a facility has a license as a proxy for closure, since most of the facilities were closed because they had no license. However, a number of unlicensed facilities were not closed by the government, creating false positives. Note: Even though we have an extensive baseline on facility characteristics, there was a significant time between the baseline and the start of the intervention and the market is characterized by a high churn rate, therefore, we lose important information if we were to use baseline data on precisely facilities that are likely to be closed, limiting the number of variables we can use in this model.

7 Closures and Mediation Analysis

Closures Analysis. To estimate how the impact varies by number of predicted closure-type facilities per market we use a machine-learning prediction of closure-type facilities using pre-treatment or fixed variables, which allows us to construct a counterfactual for control markets where facilities do not face any closure event. Then we assess how treatment effects vary with the interaction of treatment and number of closure-type facilities per market at randomization, following Equation 5:

$$Y^j_{m,t} = \alpha + \delta_1 T_m + \beta_2 NCT_{m,t-1} + \delta_2 T_m \cdot NCT_{m,t-1} + \sum_{j=1}^{n} \emptyset_j V_j + \gamma X_m + \theta Z_{i,t-1} + \epsilon_{m,t} \tag{5}$$

where $Y^j_{m,t}$ represents the outcome of interest, in this case outpatient caseload, for type of facility j, for instance private or public, in market m at endline, and the variable $NCT_{m,t-1}$ represents the number of predicted closure-type facilities in market m at randomization.[42] As before, T_m is a treatment indicator assigned to market m, X_m represents market level characteristics such as percentage of facilities per level, V_j indicates the randomization strata, and $\epsilon_{m,t}$ represents a set of random time-varying unobserved characteristics. Controls for JHIC score and outpatients at baseline $Z_{i,t-1}$ are included in all specifications.[43] We use a machine learning Random Forest algorithm to predict closure-type facilities, including variables such as unlicensed status, type of private facility, level and years of operation. Table A7 presents performance indicators for different prediction models. We focus on fixed characteristics for the prediction due to the high churn of facilities in the market, which restrict our sample size if time-varying characteristics are included. The number of predicted closure-type facilities is balanced at randomization at the market level (not shown). Figure 3 shows the results. **Mediation Analysis**. We use a two-stage regression estimator, the *sequential g-estimator*, following Acharya et al. (2016) which, under certain assumptions, allows us to estimate the controlled direct effect of the intervention, fixing the hypothesized mediator (closures) at zero, to assess in this way the importance of "number of closures in a market" as a mechanism. Under certain conditions, to which our market randomization design and high compliance with treatment status help considerably, this estimator answers the question "does the regulatory reform causes a shift of patients from private to public facilities independent of the mediator, in this case, closures?" The control direct effect estimate helps rule out the presence of bias due to intermediate confounders, a type of post-treatment bias due to the fact that closures occur post-treatment and there may be variables that are also affected by treatment and affect the

[42] We use the predicted variable both in treatment and control markets. Predictors include not having a license at randomization, having an expired license at randomization, the facility level, a categorical variable corresponding to when the facility was open, the type of private and the county. The sample is restricted to private facilities operating at randomization.

[43] Missing observations per market are replaced with zero and an indicator for missing observations is also included in the regressions.

mediator (closures) and the outcome. This method uses a *demediated* outcome (switching closures to zero) to remove the causal effect of closures.

We could identify three potential intermediate confounders including the number of inactive facilities in the market, the level of formality and the level of competition. Inactivity is affected by treatment and could have potentially affected closures, since facilities that find it difficult to comply with the regulation may have become inactive before the government attempted to close them. However, many facilities became inactive after the closure event, therefore, including this as a confounder would potentially reduce part of the mediator direct effect. We choose to be conservative and perform a robustness check later. Another potential confounder is changes in formality (or in the number of unlicensed facilities), since many facilities got the license before they were physically closed (they received a report for closure on average 70 days before the closure visit). Finally, competition may also be a confounder. Larger markets were more affected with closures due to the higher concentration of closure-type facilities in dense markets, and the size of the market could be affected by the treatment and affect closures. Therefore, we include the continuous endline market size indicator. Because of the randomized designed, pre-treatment confounders are less of a concern–and, as expected, our results are not sensitive to the inclusion of these variables, including facility level, distance from the main county town, baseline market patient safety score and the type of private facility (if private). At the second stage, randomization strata that includes county and market size group are included. Standard errors are bootstrapped and clustered at the market level.

Table S7 shows the comparison of the control direct effect (CDE) when closures are set at zero (g-estimator) and the baseline impacts that control only for pre-treatment variables for private and public facilities.

Table S7: Treatment Effects on Daily Outpatients
Baseline estimate vs. Sequential g-estimates (Controlled Direct Effects),
by Ownership Type and License Status at Randomization and Endline

	Main Specification (Baseline)		Sequential-g (ACDE)		
	Est. Coefficient (1)	% of Control Mean (2)	Est. Coefficient (3)	% of Control Mean (4)	N (5)
A. By Ownership					
Public	7.22**	17%	4.42	11%	410
	[0.03]		[0.19]		
Private	-1.53	−16%	0.20	2%	1070
	[0.17]		[0.87]		
B. Private by License Status at Endline					
Licensed	-2.29	−20%	-0.62	−5%	704
	[0.21]		[0.76]		
Unlicensed	-0.14	−2%	1.97	31%	366
	[0.87]		[0.19]		
C. Private by License Status at Randomization					
Licensed	-1.71	−12%	1.77	12%	430
	[0.47]		[0.59]		
Unlicensed	-2.79***	−39%	-1.65	−23%	471
	[0.00]		[0.17]		

Notes. P-values reported in brackets. Regressions include controls for the 16 randomization strata (by county and market size). Standard errors are clustered at the market level. Sequential G estimation uses a bootstrap procedure with 2000 repetitions. Pretreatment confounders are facility levels, distance from market center to nearest economic market, distance from market center to nearest town, market average JHIC score and private type. Intermediate confounders are the number of facilities at the market level which are inactive, the number of facilities at the market level which are unlicensed or have a missing license status at randomization, and market size at endline. The mediator is the number of facilities closed by the government. The baseline regression is similar to the main regression presented in table 3. The difference comes from that we control for all the pretreatment confounders presented above and not only facility levels and strata as in Table 3.

8 Costing

There are two important components of the cost of the intervention. First, the investment to set up the system. Second, once the system is in place, the cost of conducting routine operations. We focus on the cost of routine operation activities in this section, which helps us assess the average operational cost per visit to make the system work as intended by the regulation.

The pilot followed a particular model of inspections where all inspectors were located in the county headquarters and used vehicles provided by central and regional governments to visit health facilities. Inspectors were seconded from different government institutions, and most transferred from other regions. Facility closure visits required staff from the central government to travel to the regions. Additional external support was required for implementation and monitoring of the operation. This is a poor model for costs in a fully scaled-up version, where the number and location of inspectors can be flexibly determined and external support is minimized. Nevertheless, the routine pilot costs help provide a benchmark that can be improved upon using standard tools from operational research.

On average, a visit to a health facility during the pilot cost around USD 165 in operational costs. Of this, USD 54 (33%) were inspector costs, including salaries, allowances and compensations for being outside of their duty station; USD 17 (10%) were transportation costs to visit each facility; and USD 13 (8%) included other costs related to office, supplies and technology. The remaining USD 81 (49%) of the total, included government management (USD 22 per visit) and external World Bank support for implementation, MIS management and inspection quality assurance (USD 59 per visit). However, several factors complicate the interpretation of this cost. First, for 28% of visits the inspector could not start the inspection and the facility required multiple visits.[44] Second, there were days when vehicles were used for other government activities or were not functioning. Third, there were days when vehicles were available, but inspectors were absent. Therefore, we view this cost per visit as an upper-bound, since at least three of these problems–unsuccessful visits, non-functioning vehicles and inspector absence (during which we paid for the vehicles)–can be sharply reduced in subsequent years with more experience.[45] For instance, at best, a team of two inspectors could complete 6 inspection visits in a day (versus 3.5 during the pilot) with variation across regions based on market structures. Additionally, the World Bank support management valued at local government costs would be reduced considerably. These two actions would imply a per-facility cost of USD 95 per visit. Further, alternate models where (for instance) inspectors are either located in multiple cities in the county or have multiple bases within which they travel will

[44] Visits did not result in an inspection because the in-charges were absent or had run away when the inspector arrived (likely due to lack of a license). Due to vehicle constraints, inspectors in these cases were required to wait for a shared vehicle to come back after taking other inspectors to separate (sometimes distant) facilities, before proceeding to the next facility. Waiting times could be up to several hours.

[45] For example, the government established a new protocol that would lead to closure reports for multiple unsuccessful visits. These visits are expected to decrease further as the system matures.

further decrease transport costs, and a larger scale of inspections will also decrease costs per visit related to office, supplies, technology and management.[46]

[46]We exclude from these costs the fixed costs of building the inspection system, which included the development of the enhanced regulatory framework, implementation protocols, training materials, and the electronic inspection system.

9 Additional Results

9.1 Robustness Checks

Table S8: Treatment Effects on JHIC Score, OOP, and Outpatients
with Baseline Market Level Controls:
Overall and Interacted with Ownership and License Status 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)
Panel A: *Overall Impact*					
All Inspections (T)	4.640***	0.495	3.675***	0.209	1.789
	(0.685)	(0.343)	(1.007)	(0.349)	(1.573)
	{0.000}***	{0.279}	{0.004}***	{0.243}	{0.283}
Observations	1285	1285	1285	1285	1285
R^2	0.358	0.150	0.567	0.252	0.262
Control Mean	35.493	4.069	42.526	3.136	20.793
Impact: {%; SD}	{13%; 0.44}	{12%; 0.10}	{9%; 0.31}	{7%; 0.04}	{9%; 0.06}
Panel B: *Interaction with Private*					
All Inspections (T)	2.303**	-0.367	2.475	0.160	6.797**
	(1.040)	(0.263)	(1.534)	(0.267)	(3.268)
	[0.028]	[0.164]	[0.108]	[0.549]	[0.039]
Private HF	-6.058***	4.210***	-1.541	4.612***	-28.898***
	(1.014)	(0.307)	(2.130)	(0.523)	(2.916)
	[0.000]	[0.000]	[0.470]	[0.000]	[0.000]
Private HF x T	3.252***	1.553***	3.637*	0.525	-9.260**
	(1.120)	(0.507)	(2.118)	(0.764)	(4.066)
	[0.004]	[0.002]	[0.087]	[0.492]	[0.024]
Observations	1285	1285	1285	1285	1285
R^2	0.379	0.234	0.572	0.378	0.420
Control Mean Public	39.760	0.643	42.236	0.808	41.060
Control Mean Private	33.463	5.698	43.033	7.211	11.151
Impact Public: {%; SD}	{6%; 0.26}	{-57%; -0.41}	{6%; 0.27}	{20%; 0.14}	{17%; 0.22}
Impact Private: {%; SD}	{17%; 0.52}	{21%; 0.23}	{14%; 0.39}	{10%; 0.12}	{-22%; -0.10}
Test T + Private x T = 0 (p-value)	0.000	0.010	0.000	0.386	0.130
Panel C: *Interaction with Unlicensed (Private and active at endline only)*					
All Inspections (T)	6.030***	0.901	8.110***	0.574	-2.697
	(0.933)	(0.719)	(0.895)	(0.918)	(2.369)
	[0.000]	[0.212]	[0.000]	[0.533]	[0.256]
Unlicensed at Endline	-3.280***	-0.953	-0.137	-2.156***	-3.276
	(0.960)	(0.631)	(1.329)	(0.764)	(2.267)
	[0.001]	[0.132]	[0.918]	[0.005]	[0.150]
Unlicensed at Endline x T	-1.976	0.228	-5.447**	0.497	1.263
	(1.316)	(0.848)	(2.186)	(1.135)	(2.904)
	[0.135]	[0.788]	[0.014]	[0.662]	[0.664]
Observations	872	872	872	872	872
R^2	0.418	0.111	0.661	0.138	0.317
Control Mean Licensed	36.703	6.394	45.718	8.083	15.821
Control Mean Unlicensed	30.086	4.973	35.991	4.924	6.283
Impact Licensed: {%; SD}	{16%; 0.54}	{14%; 0.16}	{18%; 0.51}	{7%; 0.09}	{-17%; -0.08}
Impact Unlicensed: {%; SD}	{13%; 0.43}	{23%; 0.25}	{7%; 0.23}	{22%; 0.27}	{-23%; -0.13}
Test T + Unlicensed x T = 0 (p-value)	0.000	0.032	0.146	0.213	0.178

Notes. *** (**) (*) denotes significance at 1% (5%) (10%) level. "Naive" p-values are reported in brackets. Sharpened q-values are reported in braces accounting for false discovery rate given that we jointly test five hypotheses (impacts on JHIC score and OOP, both weighted and unweighted, and outpatients). When sharpened q-values are applied, stars reported next to the estimated coefficients denote significance related to the "naive" p-value. This table reports impact estimates for patient safety as measured by the JHIC score (ranges from 0 to 100 of maximum score) as per regulation; out-of-pocket expenses (OOP) as reported by patients; and daily outpatients from facility records. Missing values for OOP are imputed using means defined by level, ownership, treatment, license status at randomization, and daily outpatients. All regressions include 16 randomization strata controls (by county and market size), baseline JHIC score, OOP, and outpatient controls at the market level with missing values replaced by zero and with a dummy indicator for whether values are missing, and health facility level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided). Level 2 is the excluded category. All monetary amounts are PPP-adjusted USD terms, set at 2018 prices using the Kenya CPI and PPP conversion factor from the IMF, unless otherwise stated. Details with outcome definitions and calculations are in Section 6.

Table S9: Treatment Effects on JHIC Score, OOP, and Outpatients without Imputation and Including Strata Controls Only: Overall and Interacted with Ownership and License Status 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)
Panel A: *Overall Impact*					
All Inspections (T)	5.090***	0.975**	4.217**	0.186	1.252
	(0.838)	(0.423)	(1.808)	(0.594)	(1.601)
	{0.000}***	{0.048}**	{0.044}**	{0.705}	{0.402}
Observations	1285	1285	1285	1285	1285
R^2	0.193	0.125	0.235	0.162	0.034
Control Mean	35.493	4.069	42.526	3.136	20.793
Impact: {%; SD}	{14%; 0.48}	{24%; 0.20}	{10%; 0.35}	{6%; 0.04}	{6%; 0.04}
Panel B: *Interaction with Private*					
All Inspections (T)	2.728**	-0.040	3.464**	0.483	8.282**
	(1.088)	(0.243)	(1.683)	(0.301)	(3.478)
	[0.013]	[0.868]	[0.041]	[0.110]	[0.018]
Private HF	-7.782***	4.209***	-1.525	5.317***	-33.463***
	(1.069)	(0.389)	(3.475)	(1.243)	(3.049)
	[0.000]	[0.000]	[0.661]	[0.000]	[0.000]
Private HF x T	3.515***	1.484**	2.227	-0.154	-10.252**
	(1.262)	(0.584)	(3.867)	(1.317)	(4.418)
	[0.006]	[0.012]	[0.565]	[0.907]	[0.021]
Observations	1285	1285	1285	1285	1285
R^2	0.235	0.216	0.236	0.318	0.263
Control Mean Public	39.760	0.643	42.236	0.808	41.060
Control Mean Private	33.463	5.698	43.033	7.211	11.151
Impact Public: {%; SD}	{7%; 0.31}	{-6%; -0.05}	{8%; 0.38}	{60%; 0.42}	{20%; 0.27}
Impact Private: {%; SD}	{19%; 0.58}	{25%; 0.28}	{13%; 0.37}	{5%; 0.06}	{-18%; -0.08}
Test T + Private x T = 0 (p-value)	0.000	0.009	0.126	0.804	0.320
Panel C: *Interaction with Unlicensed (Private and active at endline only)*					
All Inspections (T)	6.292***	1.062	5.586	-0.164	-3.633
	(1.266)	(0.767)	(3.975)	(1.506)	(2.806)
	[0.000]	[0.168]	[0.162]	[0.913]	[0.197]
Unlicensed at Endline	-5.527***	-1.348**	-10.237**	-3.380**	-9.189***
	(1.153)	(0.543)	(3.983)	(1.327)	(3.056)
	[0.000]	[0.014]	[0.011]	[0.012]	[0.003]
Unlicensed at Endline x T	-1.358	0.526	-2.449	1.477	3.279
	(1.484)	(0.824)	(4.704)	(1.573)	(3.234)
	[0.361]	[0.524]	[0.603]	[0.349]	[0.312]
Observations	872	872	872	872	872
R^2	0.310	0.089	0.284	0.075	0.065
Control Mean Licensed	36.703	6.394	45.718	8.083	15.821
Control Mean Unlicensed	30.086	4.973	35.991	4.924	6.283
Impact Licensed: {%; SD}	{17%; 0.57}	{17%; 0.19}	{12%; 0.35}	{-2%; -0.03}	{-23%; -0.11}
Impact Unlicensed: {%; SD}	{16%; 0.53}	{32%; 0.35}	{9%; 0.27}	{27%; 0.34}	{-6%; -0.03}
Test T + Unlicensed x T = 0 (p-value)	0.000	0.009	0.136	0.056	0.748

Notes. *** (**) (*) denotes significance at 1% (5%) (10%) level. "Naive" p-values are reported in brackets. Sharpened q-values are reported in braces accounting for false discovery rate given that we jointly test five hypotheses (impacts on JHIC score and OOP, both weighted and unweighted, and outpatients). When sharpened q-values are applied, stars reported next to the estimated coefficients denote significance related to the "naive" p-value. This table reports impact estimates for patient safety as measured by the JHIC score (ranges from 0 to 100 of maximum score) as per regulation; out-of-pocket expenses (OOP) as reported by patients; and daily outpatients from facility records. Missing values for OOP are imputed using means defined by level, ownership, treatment, license status at randomization, and daily outpatients. All regressions include 16 randomization strata controls (by county and market size). All monetary amounts are PPP-adjusted USD terms, set at 2018 prices using the Kenya CPI and PPP conversion factor from the IMF, unless otherwise stated. Details with outcome definitions and calculations are in Section 6..

9.2 Cross-Market Externalities

A potential spillover effect of interest in this setting results from cross-market externalities which may lead to impacts on control health facilities in markets located near treatment facilities. If these externalities are present, impact estimates comparing health facilities in treated and untreated markets will underestimate the benefits of the treatment. We identify cross-market externalities using exogenous variation in the local density of facilities induced by the stratified market-level randomization, following a similar method as Miguel & Kremer (2004), but focusing exclusively on the externalities for control facilities. In particular, although all treatment markets, defined as per our z-center clustering algorithm (z=4 km) are treated, control facility x could still be located within 4 kms of treatment facility w, but would not be included in w's market because x is more than 4 kms away from the center market of facility w. Therefore, we leverage the random variation in treatment facilities close to control facilities for this estimation.[47] Equation 6 below presents the basic specification we use to estimate cross-market externalities, or the impact of the program on facilities in the control group:

$$Y_{im}^{C} = \alpha + \delta_{04} \cdot N_{04i}^{T} + \lambda_{04} \cdot N_{04i} + \sum_{j=1}^{n} \emptyset_j V_j + \gamma X_i + \theta Z_{m,t-1} + \epsilon_{ijm} \tag{6}$$

where Y_{im}^{C} indicates any of the outcomes of interest for facility i in market m for control markets C; N_{04i}^{T} is the number of facilities randomly assigned to treatment markets T within a 4 kms from facility i and N_{04i} is the number of all facilities (treatment and control) within distance d from facility i; X_i is the level of facility i; $Z_{m,t-1}$ is a set of market baseline controls, and V_j are the randomization strata; and δ_{04} is our coefficient of interest measuring the regulation treatment externalities across markets. In this specification $\delta_{04} \cdot \bar{N}_{04i}^{T}$ is the average externality of inspections on control facilities, where $\bar{N}_{04i}^{T}$ is the average number of treatment facilities located within 4 km from treatment facility i. We are controlling for the total density of health facilities since it may affect the outcomes of interest or the impact of the intervention, for instance, through higher competition. Therefore, we argue that conditional on the total number of facilities within certain distance from the facility, the number of these treatment facilities assigned to treatment is exogenous. Table S11 shows that for control facilities the main outcomes of interest at baseline are not significantly associated with the local density of treatment facilities within 4km, conditional on the total local density of health facilities.

We find no statistically significant spillover impacts on patient safety in control facilities close to treatment facilities due to cross-facility externalities. Table S10 shows that the JHIC score improved by 0.21 percentage points (p-value = 0.50) in control health facilities within four kilometers of

[47] For instance, taking 4 kms as our benchmark, 31% of control facilities have no treatment facility within 4 kms but 44% have 1-3 treatment facilities within this distance and 25% have 4 or more treatment facilities within the same distance.

treatment facilities. To see this, note that the average spillover impact on control facilities is the average number of treatment facilities located within four kilometers $\bar{N}^T_{04i}$ times the average estimated effect of an additional facility in the treatment group located within four kilometers γ_{04}. Based on estimates in Table S10 this implies that the estimated cross-facility externality in patient safety is $[\hat{\delta}_{04} \cdot \bar{N}^T_{04i}] = [0.063*3.3] = 0.21$ percentage points. Even if we lack power to detect this impact, it represents less than 1% of the mean JHIC score in the control group of 35%, and just over 4% the impact on treatment facilities of 5.2 percentage points (Table 3).

Table S10: Spillover Effects on JHIC Score, OOP, Outpatients, Facility Inactivity, and Entry for Control Facilities: with Baseline Market Level Controls

	JHIC Score (pp of max) (1)	OOP (USD PPP) (2)	Daily Outpatients (3)	Inactive (4)	New (5)
Number of T HFs within 4 kms	0.063	0.044	0.097	0.006	-0.006
	(0.115)	(0.055)	(0.322)	(0.005)	(0.004)
	[0.583]	[0.427]	[0.764]	[0.182]	[0.129]
Number of HFs within 4 kms	-0.095	0.002	-0.086	-0.002	0.004
	(0.076)	(0.037)	(0.166)	(0.002)	(0.002)
	[0.213]	[0.950]	[0.607]	[0.282]	[0.110]
Observations	421	390	481	491	431
R^2	0.184	0.209	0.072	0.094	0.099
Control Mean	35.379	3.955	18.212	0.122	0.135
Mean No. T HFs in 4kms	3.3	3.3	3.3	3.3	3.3

Notes. Robust standard errors are clustered at the market level for facility-level outcomes and reported in parentheses. P-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. All regressions include 16 randomization strata controls (by county and market size), proportion of health facility level at market-level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided), and baseline JHIC score, OOP, and outpatient controls at the market level applied to all columns, with missing values replaced by zero and with a dummy indicator for whether values are missing values at the market level. 5 outliers are dropped for facilities with number of T facilities within 4 kms >=40. This table includes unweighted estimates and the sample is for health facilities at endline.

Table S11: Spillovers: Balance on Main Outcomes

	JHIC Score at Baseline (pp of max) (1)	OOP at Baseline (USD PPP) (2)	Daily Outpatients at Baseline (3)
Number of T HFs within 4 kmxs	0.004	-0.020	0.764
	(0.139)	(0.122)	(0.613)
	[0.979]	[0.869]	[0.216]
Number of HFs within 4 kms	-0.001	0.053	-0.132
	(0.085)	(0.069)	(0.249)
	[0.988]	[0.448]	[0.598]
Observations	310	287	310
R^2	0.103	0.163	0.079
Control Mean	35.479	4.516	24.881
Mean No. T HFs in 4kms	3.3	3.3	3.3

Notes. Robust standard errors are clustered at the market level for facility-level outcomes and reported in parentheses. P-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. All regressions include 16 randomization strata controls (by county and market size). 5 outliers are dropped for facilities with number of T facilities within 4 kms >=40. This table includes unweighted estimates and the sample is for health facilities at endline.

9.3 Program Duration

We exploit exogenous variation in the implementation of the regulatory reform rollout to assess if impacts vary depending on whether markets were inspected earlier or later in the cycle. The combination of the time-bounded implementation of the regulation in the treatment markets of 13 months and delays caused by internal and external exogenous factors such as the government release of the inspectors, national healthcare worker strikes, a national presidential re-election, and unavailability of vehicles during the implementation, led to markets being exposed to different times where inspections were rolled out first. We combine this with time of endline data collection. Facilities inspected earlier and with endline data collected later, will have been in the program longer, so this both captures (potential) fade-out and the fact that earlier facilities will have received more inspections. We use the following specification:

$$\begin{aligned} JHIC\ Score_{im} \quad &= \alpha + \delta_1 T_{m123} + \delta_2 T_{m456} + \beta_1 E_{i12} + \delta_3 T_{m123} * E_{i12} + \\ \delta_4 T_{m456} \quad &* E_{i12} + \theta MainTownDistance_i + \gamma X_i + \textstyle\sum_{k=1}^{n} \emptyset_k V_k + \varepsilon_{im} \end{aligned} \tag{7}$$

where $JHIC\ Score_{im}$ represents the JHIC score for facility i in market m at endline; T_{mj} is an indicator that equals one if the first inspection in treatment market m is in one for two groups j=123 (months 1 to 3), 456 (months 4 to 6), and zero otherwise;[48] E_{i12} is an indicator that equals one if facility i has endline data collection in months 1 or 2, and zero otherwise;[49] δ_l are the parameters of interest. As before, X_i represents the facility level, V_j indicates the randomization strata, and ε_{im} represents a set of random time-varying unobserved characteristics. We control for the distance to the main county town, $MainTownDistance_i$, to account for potential unbalances due to the starting rollout of inspections and data collection of facilities closer to the main county town, which tend to have higher average JHIC than facilities in more rural areas. Our main identifying assumption is that conditional on the controls, the variation in the market date of first inspection (T_{mj}) and endline date (E_{i12}) are exogenous. Table S12 reports a test for this assumption, showing balance of baseline JHIC score across the groups of interest for these estimates.

Table S13 and Figure S2 show that there is no evidence of fade-out. In treatment markets where the first inspection occurred in the first three months and endline occurred in the last two months, the intervention increases JHIC score between 7 percentage points (0.65 SD, p-value $<$ 0.005), compared to control. For this group, the average time between first inspection (last inspection) and endline is the largest at 15 (10) months. In treatment markets with first inspection in the first

[48] The first inspection happened in 6 months between mid-January and mid-July 2017 for the majority of the markets (85%). Therefore, we use this as the main benchmark for this analysis. However, the intervention started slowly between mid-November 2016 until mid-January 2017 (7% of the markets). We include these markets in the 1-to-3-month group. For the remaining markets (8%, 2% facilities), the first inspection was spread between mid-July and November 2017. To avoid outliers, we decided to include these markets in the 4-to-6-month group.

[49] Most of the endline data collection happened in 4 months between March and June 2018 (98% of facilities). The endline for the remaining facilities (2%) happened between July and August, so these are included in the second group with 3-to-4-month endline.

three months and endline in the last two months, the JHIC score increases by 5.3 percentages points (0.5 SD, p-value < 0.005), compared to control. For treatment markets with first inspection in the last three months, the JHIC score impact ranges between 4 percentage points (0.37 SD, p-value $<$ 0.005) for those with endline in the first two months, and 3.7 percentage points (0.35 SD, p-value $= 0.004$) for those with endline in the last two months, compared to control. The markets with the largest time between first inspection and endline are also the markets where more of the process could be implemented (receiving 2.4 inspection vs. between 1.6 and 2.0 inspections for the other groups, not shown), highlighting the importance of continuous supervision and completing the full cycle of the regulatory schedule.[50]

[50]Given the most facilities fell within the two lowest categories of compliance (aside from those who lack a license), the full cycle of the new regulation would give time to facilities to improve to the minimum required by the regulation (60% of the maximum score).

Table S12: Balance on JHIC Score at Baseline by Time of First Market
Inspection with Endline Time Interactions

	JHIC Score at Baseline (pp of max) (1)
First Market Inspection in Months 1-3 (Ta)	0.153
	(1.284)
	[0.905]
First Market Inspection in Months 4-6 (Tb)	1.447
	(0.966)
	[0.135]
Endline in Months 1-2	3.314***
	(1.101)
	[0.003]
Endline in Months 1-2 x Ta	1.216
	(1.702)
	[0.476]
Endline in Months 1-2 x Tb	-1.686
	(1.708)
	[0.325]
Observations	889
R^2	0.382
Control Mean	37.461
Estimated Coefficients for Endline in Months 3-4	
Inspection in Months 1-3	1.369
P-value	[0.200]
Inspection in Months 4-6	-0.238
P-value	[0.872]

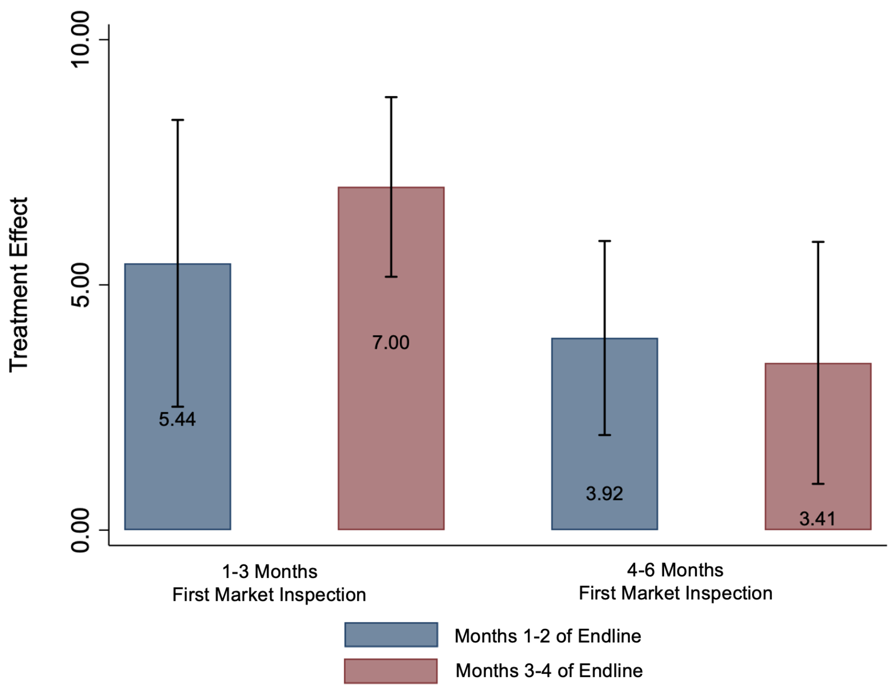
Notes. Robust standard errors in parentheses. P-values reported in brackets below standard errors. *** p<0.01, ** p<0.05, * p<0.1. For health facility-level variables, the standard errors are clustered at the market level. Regressions include controls for the 16 strata included in the randomization (by county and market size), health facility level controls (levels are 2, 3, and 4 and 5 merged together), and distance to the main town in each county. The start of the intervention was interrupted over the first two months. The first inspection happened in 6 months between mid-January and mid-July 2017 for the majority of the markets (85%). Therefore, we use this as the main benchmark for this analysis. However, the intervention started slowly between mid-November 2016 until mid-January 2017 (7% of the markets). We include these markets in the 1-to-3-month group. For the remaining markets (8%, 2% facilities), the first inspection was spread between mid-July and November 2017. To avoid outliers, we decided to include these markets in the 4-to-6-month group. Most of the endline data collection happened in 4 months between March and June 2018 (98% of facilities). The endline for the remaining facilities (2%) happened between July and August, so they are included in the second group 3-to-4-month endline.

Table S13: Impact on JHIC Score by Time of First Market Inspection with Endline Time Interactions

	JHIC Score (pp of max) (1)
First Market Inspection in Months 1-3 (Ta)	5.341***
	(1.473)
	[0.000]
First Market Inspection in Months 4-6 (Tb)	3.961***
	(0.999)
	[0.000]
Endline in Months 1-2	2.730***
	(0.991)
	[0.006]
Endline in Months 1-2 x Ta	1.616
	(1.673)
	[0.335]
Endline in Months 1-2 x Tb	-0.267
	(1.598)
	[0.867]
Observations	1285
R^2	0.354
Control Mean	39.399
Estimated Coefficients for Endline in Months 3-4	
Inspection in Months 1-3	6.957***
P-value	[0.000]
Inspection in Months 4-6	3.694***
P-value	[0.004]
Impact {%; SD}	
Inspection in Months 1-3 and Endline in Months 1-2	{15%; 0.50}
Inspection in Months 1-3 and Endline in Months 3-4	{20%; 0.65}
Inspection in Months 4-6 and Endline in Months 1-2	{11%; 0.37}
Inspection in Months 4-6 and Endline in Months 3-4	{10%; 0.35}

Notes. Robust standard errors are reported in parentheses. P-values are reported in brackets. *** (**) (*) denotes significance at 1% (5%) (10%) level. For health facility-level variables, the standard errors are clustered at the market level. Regressions include controls for the 16 strata included in the randomization (by county and market size), health facility level controls (levels are 2, 3, and 4 and 5 merged together), and distance to the main town in each county. The start of the intervention was interrupted over the first two months. The first inspection happened in 6 months between mid-January and mid-July 2017 for the majority of the markets (85%). Therefore, we use this as the main benchmark for this analysis. However, the intervention started slowly between mid-November 2016 until mid-January 2017 (7% of the markets). We include these markets in the 1-to-3-month group. For the remaining markets (8%, 2% facilities), the first inspection was spread between mid-July and November 2017. To avoid outliers, we decided to include these markets in the 4-to-6-month group. Most of the endline data collection happened in 4 months between March and June 2018 (98% of facilities). The endline for the remaining facilities (2%) happened between July and August, so they are included in the second group 3-to-4-month endline.

Figure S2: Impact on JHIC Score by Time of Market's First Inspection
With Endline Time Interactions



Notes. For health facility-level variables, the standard errors are clustered at the market level. Regressions include controls for the 16 strata included in the randomization (by county and market size), health facility level controls (levels are 2, 3, and 4 and 5 merged together). The start of the intervention was interrupted over the first two months. These facilities are included in Months 1-3. 18 facilities had their first market inspection after month 6 (including delays in the first two months of the intervention). These facilities are included in Months 4-6.

9.4 Quantile Treatment Effects

9.4.1 Unconditional Quantile Treatment Effects Estimates

To estimate unconditional quantile treatment effects in Table S14, we use the following specification:

$$\Delta_{QTT}(\tau) = q_1(\tau|R=1) - q_0(\tau|R=0) \tag{8}$$

where $q_D(\tau|R=1)$ is the τ-th quantile of potential outcomes Y_D under treatment. This specification assumes full compliance with the random assignment to estimate the treatment-on-the-treated effects. Given the nearly universal program compliance measured for the policy, this approximation is justifiable to simplify analysis without risk of bias.

9.4.2 Conditional Quantile Treatment Effects Estimates

To estimate conditional quantile treatment effects in Figure 4, we use conditional quantile regressions with the following specification:

$$Y^q_{i,m} = \alpha + \beta_1 T_m + \beta_2 Z^k_i + Level_i + FE_{counties} + \epsilon_{i,m} \tag{9}$$

where $Y^q_{i,m}$ is the outcome of interest (patient safety as measured by the JHIC score, as per regulation, excluding licensing) at quantile q for health facility i in market m at follow-up; T_m is a treatment indicator at the market level that equals one for facilities belonging to the treatment group and zero for facilities in the control group; Z^k_i is a vector of dummy variables equal to one if facility i belongs to quartile k for the outcome at baseline and an indicator for missing baseline outcome, whenever it was not available; $Level_i$ and $FE_{counties}$ are two vectors of dummy variables for the health facility level and county; and $\epsilon_{i,m}$ represents a set of random time-varying unobserved characteristics. Standard errors are robust and clustered at the market level. In Figure 4, we run the previous regression first on the whole sample and by ownership, (part A) and second by market density (1-2 health facilities, 3-10 health facilities and 11 and more health facilities in the market) in part B and C of the table. Note that for consistency across part A and parts B and C, we do not control for market size in the quantile regression.

Table S14: Unconditional QTE on JHIC Score, by Ownership and Market Size

	By Ownership			By Market Size Group			By Market Size Group (Private)		
	All	Private	Public	1-2 HFs	3-10 HFs	11+ HFs	1-2 HFs	3-10 HFs	11+ HFs
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Percentile									
10th	2.780**	4.450***	2.560	2.610	4.430***	2.970	4.670	4.780***	2.380
	(1.162)	(1.569)	(2.022)	(3.029)	(1.460)	(2.140)	(7.550)	(1.539)	(2.093)
25th	4.970***	4.860***	2.330	3.160	4.580***	7.340***	5.610	3.950**	6.760***
	(1.284)	(1.542)	(1.515)	(2.803)	(1.647)	(2.361)	(5.332)	(1.676)	(2.406)
50th	7.400***	8.620***	2.910*	1.300	6.200***	10.860***	10.320**	7.180***	10.390***
	(1.376)	(1.790)	(1.523)	(2.377)	(1.552)	(2.831)	(4.678)	(1.998)	(3.284)
75th	6.860***	8.760***	3.840***	1.650	6.520***	11.540***	10.110**	8.090***	9.740***
	(1.320)	(2.194)	(1.428)	(2.098)	(1.380)	(3.227)	(4.505)	(1.960)	(3.136)
90th	7.550***	9.640***	5.710***	3.320	4.820**	11.150***	4.040	7.440***	11.150**
	(2.006)	(3.019)	(1.861)	(2.155)	(2.068)	(4.060)	(5.600)	(2.654)	(4.633)
Observations	1285	872	413	191	638	456	66	413	393
Test 10th = 90th	0.001	0.018	0.169	0.833	0.869	0.010	0.927	0.319	0.012

Notes. Robust standard errors are clustered at the market level and reported in parentheses. *** (**) (*) denotes significance at 1% (5%) (10%) level. This table presents the conditional quantile treatment effect estimates on patient safety measured by the JHIC score as per regulation, by ownership (A) and market size (B and C). Regressions include controls for the county, health facility level (2, 3, and 4 and 5 merged together, which are positively related to the amount, and specialization of services provided. Level 2 is the excluded category. Regression also controls for JHIC score quartile dummies at baseline, where missing values are replaced by zero and an indicator for missing JHIC score at baseline is included. Conditional quantile treatment effect are estimated using the qreg2 stata command, which allows to estimate robust standard errors clustered at the market level but does not allow to test the equality of the 10th and 90th coefficient. For the test, we run simultaneously the five quantile regressions using the stata command sqreg, which allows to test the equality of percentile coefficients but does not estimate robust standard errors accounting for intra cluster correlation. In this case, standard errors are estimated after performing a 1,000 bootstrap replications.

9.5 Additional Tables

Table S15: Share of Total Patients and Facilities, by Facility Stability and Ownership Type

	Patients				Facilities			
		Private				Private		
	Public	Stable	Unstable	All	Public	Stable	Unstable	All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Baseline	0.72	0.22	0.06	1.00	0.36	0.43	0.22	1.00
Endline	0.67	0.27	0.06	1.00	0.28	0.53	0.19	1.00
Endline (control only)	0.64	0.31	0.05	1.00	0.28	0.52	0.20	1.00

Notes. The table presents the share of patients (columns 1 to 4) and facilities (columns 5 to 8) by facility ownership type at baseline and endline. Private stable facilities are facilities which are active at baseline and at endline. Private unstable facilities are facilities which were either not open at baseline or closed at endline.

Table S16: Share of Total Patients and Facilities,
By Facilities that Exited and Entered by Endline

	Share of patients when operational	Share of facilities when operational
	(1)	(2)
Exited by endline (active at baseline)	0.02	0.12
Entered by endline	0.12	0.24
Entered by endline (control only)	0.12	0.26

Notes. Exit equals the share of patients that the facilities that exited the market by endline were responsible for in the baseline. Enter equals the share of patients that the facilities that entered the market after baseline were responsible for in the endline. Share of facilities when operational includes those with patient data available. This table does not include faciliies that exited by endline and were active after the baseline.

Table S17: Correlations of Select Indicators with Market Size at Randomization

	JHIC Score at Baseline at the Market Level (All Markets)	Market Size at Randomization (Treated Markets Only)	
	(1)	(2)	(3)
Market Size at Randomization	0.311		
	(0.212)		
Month of First Inspection Visit in Market		-0.079	
		(0.079)	
Average Inspections in the Market			-0.083
			(0.352)
Observations	259	178	178
R^2	0.284	0.849	0.848

Notes. Robust standard errors are reported in parentheses. *** (**) (*) denotes significance at 1% (5%) (10%) level. Regressions include controls for the 16 strata included in the randomization (by county and market size) and health facility level controls (levels are 2, 3, and 4 and 5 merged together, and are positively related to the amount, and specialization of services provided) as percentage in each market. Columns (2) and (3) include controls for mean baseline JHIC score at the market level.