Regular Article

Data and policy decisions: Experimental evidence from Pakistan☆

Michael Callen a,b,*, Saad Gulzar c, Ali Hasanain d, Muhammad Yasir Khan e, Arman Rezaee f

a London School of Economics, UK 
b NBER, USA 
c Stanford University, USA 
d Lahore University of Management Sciences, Pakistan 
e University of California, Berkeley, USA 
f University of California, Davis, USA

ARTICLE INFO

Keywords: Absenteeism Bureaucracies Data-informed policy Information communication technology Health

ABSTRACT

We evaluate a program in Pakistan that equips government health inspectors with a smartphone app which channels data on rural clinics to senior policy makers. The system led to rural clinics being inspected 104% more often after 6 months, but only 43.8% more often after a year, with the latter estimate not attaining significance at conventional levels. There is also no clear evidence that the increase in inspections led to increases in general staff attendance. In addition, we test whether senior officials act on the information provided by the system. Focusing only on districts where the app is deployed, we find that highlighting poorly performing facilities on a dashboard viewed by supervisors raises doctor attendance by 75%. Our results indicate that technology may be able to mobilize data to useful effect, even in low capacity settings.

1. Introduction

Information technology is providing governments across the globe with greater access to data to inform policymaking. Technologies like smartphones and tablets make it simple and cheap to collect, compile, and visualize specialized and timely information relevant to a range of decisions. Increasingly, policymakers no longer need to rely on the collection and aggregation of geographically disparate paper records to understand how their government is operating. They can have this information instantly, and presented in the form that best suits their needs.

But will policymakers use these data? Will activating such technologies improve the quality of service provision? These are complex questions where the capabilities of government personnel, the specific government organization, and the broader political and institutional environment all potentially impact the answers.

To provide evidence on these questions, we conduct a randomized controlled evaluation of a smartphone monitoring program in Punjab, Pakistan. The program—officially termed ‘Monitoring the Monitors’—equips government inspectors with a smartphone application that collects data and feeds it to an online dashboard system. This

☆ Authors’ Note: This paper combines two previous papers circulated with the titles “The Political Economy of Public Sector Absence: Experimental Evidence from Pakistan” and “Personalities and Public Sector Absence: Evidence from a Health Experiment in Pakistan”. We thank Farasat Iqbal for championing and implementing the project and Asim Fayaz and Zubair Bhatti for designing the smartphone monitoring program. Support is generously provided by the International Growth Centre (IGC) political economy program, the IGC Pakistan Country Office, and the University of California Office of the President Lab Fees Research Program Grant #235855. Callen was supported by grant #FA9550-09-1-0314 from the Air Force Office of Scientific Research. We thank Erlend Berg, Eli Berman, Leonardo Bursztyn, Ali Cheema, Melissa Dell, Ruben Enikolopov, Barbara Geddes, Naved Hamid, Gordon Hanson, Michael Kremer, Asim Ijaz Khwaja, Craig McIntosh, Ijaz Nabi, Aprajit Mahajan, Monica Martinez-Bravo, Benjamin A. Olken, Gerard Padro-i-Miquel, Karthik Muralidharan, Rohini Pande, Daniel N. Posner, Ronald Rogowski, Jacob N. Shapiro, Christopher Woodruff, Oliver Vanden Eynde, David Yanagizawa-Drott, Ekaterina Zhuravskaya and various seminar participants for insightful comments. Excellent research assistance was provided by Muhammad Zia Mehmood and Haseeb Ali. We thank Ali Cheema and Farooq Naseer for kindly sharing their data on election outcomes.

* Corresponding author. London School of Economics, UK.

E-mail addresses: m.j.callen@lse.ac.uk (M. Callen), gulzar@stanford.edu (S. Gulzar), hasanain@lums.edu.pk (A. Hasanain), yasir.khan@berkeley.edu (M.Y. Khan), abrezaee@ucdavis.edu (A. Rezaee).

https://doi.org/10.1016/j.jdeveco.2020.102523

Received 3 September 2019; Received in revised form 8 June 2020; Accepted 13 June 2020

Available online 2 July 2020

0304-3878/© 2020 Elsevier B.V. All rights reserved.
provides real-time information on rural public health clinics in Punjab, Pakistan, aggregated into simple charts and tables for the review of senior health officials. It also includes several fail-safes to ensure accurate reporting: reports are geo-stamped and time-stamped and all staff reported present must be photographed with the inspector. In this environment, irregular inspections (in our baseline, only 23% of facilities had received their required monthly inspection) and doctor absence (doctors were present at 24% of facilities during regular operating hours in our baseline) are serious issues. The smartphone system supplanted the previous paper-based system for collecting operational data on public health facilities, which rarely functioned.

The evaluation spans 35 of 36 districts in Punjab. Punjab is a province of 100 million people, with many citizens utilizing public health services. The experiment involved 117 inspectors, 35 senior officers, 2496 rural health clinics (of which, we sample 850), and took place across 240 different parliamentary assembly constituencies. This scale allows us to cluster randomize at a relevant policy unit – the district. This setting also provides variation to begin to speculateively examine whether individual and institutional constraints are relevant to the performance of the system.

Beyond improving the flow of data, the system also changed the behavior of inspectors, at least temporarily. In independent audits conducted six months after the survey, the inspection rate increased from 25.5% to 51.9% (\(p < 0.01\)). After a year of operation, inspection rates were 33.8% in the treatment districts and 23.5% in control districts, with this difference not being significant at conventional levels.\(^5\)

Theoretically, if clinic staff base their beliefs about the likelihood they will be inspected on how often they are actually inspected, this increase in inspections could change staff attendance. However, the data generally do not indicate that increases in inspections resulted in better staff attendance. Of the seven categories of clinic staff, we only find highly speculative evidence that doctors may have increased attendance. However, this result is highly sensitive to the choice of specification, and should be interpreted accordingly.

We also built into the experiment a feature that allows us to examine whether providing senior officials with data changes their behavior.\(^6\) Specifically, if more than three of the seven health workers that are supposed to staff a rural clinic are absent during a health inspection, we ‘flagged’ a facility as underperforming by highlighting it in red on the dashboard.\(^7\) We test for effects on officials’ behavior by examining whether doctor attendance increases in flagged facilities.\(^8\) We find that flagging increases doctor attendance from 23.6% to 41.3% (standard error of difference = 8.2 percentage points), while it has no effect on the attendance of other, less senior, clinic staff. We conceive of this as creating a regression discontinuity that allows us to study whether data changes the behavior of senior policymakers. We discuss the identifying assumptions required for this to be causal and subject it to an extensive set of validity tests. We interpret this as evidence that policymakers use data when making decisions. We report results from a battery of robustness checking the causal interpretation of this result in section 4.4.1.

Last, we investigate two dimensions of heterogeneity for flagging effects. First, we find that attendance responds more to flagging when the information is channeled to senior health officials with normatively better personality measures as measured using the Big Five Personality Test and the Perry Public Sector Motivation battery. Second, we find smaller effects in less politically competitive parliamentary constituencies. While speculative, these results suggest that both intrinsic and extrinsic forces are important factors in determining the extent to which senior bureaucrats might act on new information.

Our study principally concerns the potential for information technology and data to increase accountability and improve policy (Duflo et al., 2012; Blum and Pande, 2015; Callen and Long, 2015; Callen et al., 2016; Nealer et al., 2017), but relates to three additional literatures.

The first regards whether monitoring of government workers improves attendance (Banerjee and Duflo, 2006; Banerjee et al., 2008; Chaudhury et al., 2006; Daliwal and Hanna, 2017; Muralidharan et al., 2019, 2020). The second broadly studies incentives in bureaucracies in developing countries (Ashraf et al., 2014, 2015; Bertrand et al., 2017; Dal Bó et al., 2013; Deserranno, 2017; Finan et al., 2017; Gulyar and Pasquale, 2017; Khan et al., 2016; Rasul and Rogger, 2018; Xu, 2017; Habyrimanana et al., 2020). The final involves experiments at scale (Muralidharan et al., 2016; Muralidharan and Niehaus, 2017).

The paper proceeds as follows. Section 2 provides background on the ‘Monitoring the Monitors’ program. Section 3 describes the data and the experiment. Section 4 provides results, and section 5 concludes.

2. Background

2.1. Public health services in Punjab

In Punjab, public health services are provided by the Department of Health, which is headed by the Secretary of Health. The smartphone intervention we describe in this paper works by facilitating flows of information through the existing chain of command, so describing that chain is fundamental to characterizing the reform.

This provincial Department comprises 36 District Health Departments, each headed by an Executive District Officer, hereafter referred to as a ‘senior health official’. Senior health officials report directly to the Secretary. Performance by senior health officials is commonly rewarded with appointment to a higher office. Senior health officials are, in turn, each supported by several Deputy District Officers, hereafter referred to as ‘health inspectors’, typically one for each sub-district (there are, on average, 3.4 sub-districts per district). Fig. 1 depicts this administrative hierarchy.

Health inspectors are charged with inspecting all of the health facilities in their sub-district at least once every month (see
there are several ways that senior officials can compel or encourage inspectors to do their jobs. Conversations with several senior officials reveal that they typically begin simply by having a conversation with a problematic inspector. The next step is to refer the matter to a senior provincial-level official in charge of general administration. This can result in a formal inquiry and ultimately in pay cuts. Recent research in the same context provides direct evidence of how structured interactions between senior health officials and clinic staff can improve performance (Khan, 2020).  

While their primary job is to collect data, health inspectors do have the authority to directly punish absent clinic staff by issuing a ‘show-cause notice’, which requires staff to explain their absence to senior health officials. They can also suspend and deny pay to contract staff, including doctors. In severe cases of persistent absence, health inspectors can transfer staff to less-desirable locations, but they cannot terminate employment. Unlike their superiors, health inspectors rarely ascend to higher leadership positions. 

Senior health officials do not typically sanction facility-level staff directly. Rather, they will send verbal communications through the chain of command via the inspector. The next step is to call for a formal explanation for absence. After that, the matter is referred to a provincial-level official who can recommend pay cuts or transfers. 

There are five classifications of health facilities; we focus on the frontline tier, called Basic Health Units, hereafter referred to as ‘health clinics’. Each health clinic is headed by a Medical Officer, henceforth ‘doctor’. These doctors are of particular interest for this study. Doctors are general practitioners who have completed five years of medical school, and are therefore the most trained health professionals in rural areas (see Appendix Section C for details on doctor hiring practices). While more senior doctors are paid more, they have essentially the same portfolio of responsibilities. Very few doctors rise through the ranks to become health inspectors: compared to the 2496 Medical Officer posts in clinics, there are only 123 such senior positions.

2.2. Pre-existing paper-based monitoring system

During their required monthly inspections, health inspectors are required to collect information on a standard paper form. This form records utilization, resource availability, and worker absence. We provide this form in Appendix E. Once collected, forms are brought to a central district facility, manually entered into a spreadsheet, and aggregated into a monthly report for senior health officials. 

This inspection system affords only limited visibility into inspectors’ activities to senior officials. Compounding this problem, senior health officials have only two weak means of sanctioning an inspector: issuing a verbal reprimand or, in serious cases, sending a written request for investigation to provincial authorities. The investigation process is long, highly bureaucratic, and, anecdotally, prone to interference by elected politicians.

2.3. ‘Monitoring the Monitors’ smartphone monitoring program

We partnered with the Department of Health to design and experimentally evaluate the ‘Monitoring the Monitors’ program. This program replaced the existing paper-based monitoring system with an Android-based smartphone application, which collects the same data as the paper forms and transmits them instantly to a central online dashboard for the Secretary of Health and senior health officials. The dashboard provides summary statistics, charts, and graphs in a format designed in collaboration with senior health officials. Inspections are geotagged, timestamped, and require photos of the inspector and all health clinic staff marked present to check for reliability. The geotagging and time-stamping features are designed to ensure inspectors visit health clinics, while the staff photos are intended to ensure that the digital reports of staff attendance are accurate.

Fig. 2, Panel A, depicts the view of the dashboard that the Secretary of Health sees when first logging on. It presents a bar chart giving the number of health clinic inspections conducted in a district as a proportion of number of inspections assigned that month, allowing the Secretary to compare performance across districts. Panel B provides an alternate view available to senior health officials—a summary spreadsheet where each row corresponds to a different health clinic inspection that occurred in a senior health official’s district.

3. Data and experiment

3.1. Data

To measure the impacts of our smartphone monitoring on health clinic inspections and doctor attendance, we collected primary data on a representative sample of 850 (34%) of the 2496 health clinics in Punjab. All districts in Punjab except Khanewal are represented in our data. To our knowledge, this is the first representative survey of health clinics in Punjab. Fig. 3 provides a map of the health clinics in our experimental sample along with district boundaries.

Enumerators made three unannounced visits to these 850 health clinics: one before smartphone monitoring began, in November 2011, and two after smartphone monitoring began (in treatment districts), in June and October 2012. 

During these unannounced visits, enumerators collected the same information that health inspectors record—information on health clinic utilization, resource availability, and worker presence—as well as information on the occurrence of health clinic inspections themselves. Enumerators physically verified health clinic staff presence, filling out an attendance sheet at the end of their visit and in private for doctors as well as dispensers, lady health workers, health/medical technicians, school health and nutrition supervisors, and midwives. Summary statistics from unannounced visits at baseline are presented in Appendix Table A1.

Health inspectors record visits by signing paper registers maintained at the health facility. Enumerators measured whether a health inspection occurred in the prior month by interviewing facility staff and verifying the register record. In some cases enumerators were unable to confidently verify whether or not an inspection had occurred in the prior month. We treat such cases as missing data for analysis and verify in Appendix Table A1 (considering a dummy variable equal to one if the enumerator could not verify the last health inspection) that such

---

9 The exercise in this paper encourages frontline community health workers to watch a video of the senior health official describing the mission of the Health Department and requires them to take part in sessions reflecting on that mission. This is cross-randomized with a performance bonus. Both increase performance, though financial incentives are less effective in the presence of the mission treatment.

13 Doctors are officially required to be present and see patients at the health clinic. An unannounced visit therefore captures the official work assigned to doctors. We did not capture data on computer technicians as they are rarely assigned.

---
cases do not correlate with treatment assignment at baseline and in Appendix Table A2 that we do not have treatment-driven attrition wave by wave. Additional information about data collection can be found in Appendix Section D.

We also conducted face-to-face time use surveys with all health inspectors in Punjab between February and March 2013.
3.2. Experiment

Our experimental sample comprises 35 of the 36 districts in Punjab. We randomly implemented the smartphone program in 18 of the 35 remaining districts. We randomized at the district level for two reasons. First, the intervention channels information about health inspections to district-level senior health officials. Second, all inspectors in a district are required to attend monthly meetings and so interact frequently, while these relations are much weaker across districts. District-level randomization therefore makes sense in terms of the design of program and also reduces concerns about contamination.

We stratified treatment on baseline health clinic staff attendance, the number of clinics in a district (to ensure a roughly even number of clinics in treatment and control), and whether or not the district was being run by the World Bank-led Public Sector Reform Program (PSRP). Fig. 3 depicts control and treatment districts. We then re-randomized to achieve balance on a set of variables related to health clinic staff attendance, the frequency of health inspections, and the quality of service provision. We randomly implemented the smartphone program in 18 of the 35 clinics, it did not do so for doctors. We therefore use a difference-in-differences specification to estimate impacts of the program on doctors.16

4. Results

We now present results from our experimental evaluation of the ‘Monitoring the Monitors’ program. We estimate treatment effects on health inspection rates and on health inspector time use with the following specification:

\[ Y_{itd} = \alpha + \beta_{Treatment} + \delta_{i} + \gamma_{s} + \epsilon_{itd}. \]  

where \( Y_{itd} \) is either a dummy equal to one for health clinics inspected in the previous month or a measure of health inspector time use, \( t \) refers to the clinic, \( t \) to the survey wave, \( d \) refers to the district, and \( s \) to the randomization stratum that the district was in. \( Treatment \) is a dummy variable equal to 1 for treated districts, \( \delta_{i} \) are survey wave fixed effects, and \( \gamma_{s} \) are randomization strata fixed effects. These regressions use only post-treatment data (survey waves 2 and 3). We cluster all standard errors at the district level.

We estimate treatment effects on health clinic staff attendance and assignment using three specifications, all meant to account for baseline imbalances in our primary outcomes (which are only present and statistically significant for our measure of doctor absence). First, we use a difference-in-differences specification:

\[ Y_{itd} = \alpha + \beta_{1} Treatment + \beta_{2} Post + \beta_{3} Treatment \times Post + \delta_{i} + \gamma_{s} + \epsilon_{itd}. \]  

where \( Y_{itd} \) is now either a dummy for whether a health clinic staff member is present during our announced visits or a dummy for whether a staff member is currently assigned to work at a health clinic at the point of our announced visits, and \( \delta_{i} \) are survey wave fixed effects, \( \gamma_{s} \) are randomization strata fixed effects and \( \theta_{s} \) are clinic fixed effects. Second, we use our primary treatment effect specification, equation (1), with adding \( Y_{itd} \) the baseline level of the outcome, to the set of covariates in the regression. Third, we estimate:

\[ Y_{itd} = \alpha + \beta_{Inspected} + \delta_{i} + \gamma_{s} + \epsilon_{itd}. \]  

14 Specifically, we randomized using the ‘big stick’ approach, whereby we redrew our treatment assignment until the minimum p-value from difference in means tests between health clinics in treatment and control districts for a pre-specified set of balance variables was greater than some threshold. In our case we selected a threshold of 0.21. We balanced on the share of assigned health clinic staff who were present during the baseline unannounced visit, whether the health clinic had been inspected in the previous month by its health inspector, whether the health clinic had been inspected in the previous month by its senior health official, the number of antenatal visits recorded on the health clinic register in the previous month, the log of the number of polio vaccines administered at the health clinic in the previous month, whether the health clinic’s doctor claimed a connection to their local parliamentarian, the tenure of the health clinic’s doctor, the log of the population in the health clinic’s catchment area, and the log of the distance of the health clinic to the district’s headquarters.

15 Treatment districts at baseline have 17.1 percentage point higher doctor attendance (p-value = 0.003).

16 While it would be ideal, we cannot check for balance in pre-trends in doctor attendance, a necessary assumption for our difference-in-differences estimation to be unbiased, because we only have one pre-treatment observation for each clinic.
where we instrument for whether a facility was inspected in the prior month using our treatment assignment.\textsuperscript{17}

For all of our primary treatment effects, we also present results from regressions with additional controls and/or sample selection criteria as robustness checks. We point out any cases where results are sensitive to such decisions.

### 4.1. Approach to inference

With only 35 districts in our sample, the asymptotic reference distributions for our test statistics may be invalid. We therefore report Fisher exact p-values (Fisher, 1935) which do not require a limiting distribution (Gerber and Green, 2012). This test assumes a null of no treatment effect for any unit. We perform this test by creating a set of artificial treatment assignments that satisfy the balancing requirements of the assignment protocol for actual treatment. For each treatment assignment, a corresponding artificial treatment effect is generated. The effect estimated using the actual treatment assignment is then compared against the 1000 artificial treatment effect estimates. The p-value is the share of artificial treatment effects that have a larger magnitude than the actual treatment effect.

### 4.2. Impact on health inspectors

Table 1 presents estimates of the program’s impact on the rate of inspections. We find that health clinics in treatment districts were 18.1 percentage points more likely to be inspected in the previous month during the treatment period. This represents a 74 percent increase in inspection rates in treatment districts relative to control districts. Breaking this up into the two waves of post-treatment data collection, we find comparable effects, though there is evidence that the effect of treatment had attenuated by October 2012, a year after the introduction of the program.\textsuperscript{18} In Table 1 we report simple means to facilitate inference on both means directly and on comparisons both between groups and across time. Appendix Tables A4 and A5 report corresponding treatment effect estimates from a range of specifications (including with strata fixed effects) which are consistent with the raw means, except, importantly, the wave three treatment effect is not significant at conventional levels with both randomization strata and wave fixed effects included (exact test p-value of 0.19).

We examine whether the additional time for inspections come at the costs of other tasks. Appendix Table A3 reports the monitoring program’s impact on the time use of health inspectors. We do not find significant evidence that health inspectors in treatment districts are spending less time on other tasks after treatment, while we do find significant increases in time inspecting health clinics. However, we treat these results speculatively, as we only have a sample of 117 inspectors and a noisy measure of time use. We do, for example, find a negative but insignificant coefficient on time spent monitoring hospitals that could account for 60 percent of the increased time monitoring clinics.

#### 4.3. The impact on health clinic staff

Our estimates indicate that the program increased health inspections; this could increase health clinic staff attendance (an explicit goal of the program). We report estimates of program impact on staff attendance in Table 2. Panels A and B report impacts using a difference-in-differences specification. Panel C reports results using an average treatment effects regression controlling for baseline attendance.\textsuperscript{19} Panel D reports a two-stage least squares estimate of the impact of a health facility being inspected on staff attendance, instrumenting for inspection with treatment.\textsuperscript{20} We find mixed results here—we find no impacts of monitoring on staff assignment in Panels A and B and we find large, positive, and significant impacts on doctor attendance in Panels C and D. These results are not robust across specifications, and should be viewed accordingly.\textsuperscript{21}

\textsuperscript{17} Tables A4 and A5 therefore report the first stage of this regression.

\textsuperscript{18} The p-value corresponding to the test of equal treatment effects in wave 2 and in wave 3 is 0.08.

\textsuperscript{19} We do not present results not controlling for baseline attendance as we have a large imbalance in doctor attendance between treatment and control districts at baseline, as reported in Appendix Table A1.

\textsuperscript{20} In this setting, inspections could impact staff attendance through the following channels. First, staff see an inspection take place directly, and therefore shift their beliefs about the probability of an inspection in the future. Second, they learn that an inspection took place through colleagues indirectly, and shift their beliefs. Third, they learn that an inspection took place because they are contacted about their absence during the inspection by a supervisor. There was no effort to publicize the reform to clinic staff and we expect that most staff learned about the reform when the inspector arrived at their facility with a smartphone and compelled all present staff to take a picture in front of the facility to confirm the attendance data. If the program affects attendance through channels other than inspections, then the exclusion restriction is not satisfied.

\textsuperscript{21} Appendix Tables A6 and A7 present results on attendance by survey wave. While our results become more imprecise, we do not see qualitatively different results across waves. This loss of precision also means our first-stage becomes too weak to conduct IV analysis by wave.
In contrast to doctors, other staff do not appear to respond to treatment regardless of specification. Speculatively, there are at least three reasons for this. First, attendance rates for clinic staff who are not doctors, with the exception of school and health nutrition supervisors, are substantially higher, and so there is less room for improvement. Second, doctors carry out most of their work outside of clinics, and so may be less affected by these reports. Note our measures of staff attendance do not condition on staff assignment to facilities as staff assignment itself could be impacted by treatment. We test for treatment effects on doctor assignment directly. We test for treatment effects on doctor assignment directly.

### Table 2

| Doctor Dispenser Lady Health Worker Health Tech School Health Nutri. Supervisor Midwife |
|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|
| #Observations                          | 2408                                          | 2408                                          | 2408                                          | 2408                                          |
| #Districts                            | 35                                            | 35                                            | 35                                            | 35                                            |
| Mean in Controls                      | 0.227                                         | 0.736                                         | 0.594                                         | 0.413                                         |
| R-Squared                             | 0.533                                         | 0.464                                         | 0.447                                         | 0.552                                         |

Notes: This table reports average treatment effects of the 'Monitoring the Monitors' program on the staff attendance. The unit of observation is the clinic, and data come from primary unannounced surveys after the treatment was launched. The dependent variable is an indicator for whether a doctor was present at the clinic during an announced visit. All models include randomization block fixed effects. Standard errors clustered at the district level are in parentheses. Square brackets report the statistical power.

### 4.4. Do policymakers use data on staff attendance?

In the 18 districts of Punjab that received the 'Monitoring the Monitors' treatment, the system aggregated and presented inspection report data to senior health officials through an online dashboard. This dashboard was also visible to the Health Secretary and the Director General of Health for Punjab.

To test whether senior health officials would act on these data, we introduced a manipulation to the dashboard that made certain health inspection reports salient. Specifically, we highlighted in red (hereon ‘flagged’) inspection reports that reported three or more staff (of seven generally) as absent during a health inspector’s visit to the clinic. We deliberately selected three as the cutoff for flagging as the majority of reports indicated that two or three staff were absent, affording the greatest statistical power.

---

22 We attempted to obtain detailed doctor assignment records from the Health Department on multiple occasions and it was clear there was no regular system for keeping track of doctor assignment during the period of our experiment. Hence, we are not able to conclusively identify what is driving the increased assignment of doctors. It is possible that this increase is achieved by moving doctors from control to treatment districts. However, as described in Appendix Section C, it is unlikely that doctors are moved across districts under ordinary circumstances.

23 We deliberately selected three as the cutoff for flagging as the majority of reports indicated that two or three staff were absent, affording the greatest statistical power.
some facilities flagged in red. We then test for changes in senior health official action through subsequent staff attendance.

The exact formula for this arbitrary threshold was not known to anyone but the research team. This approach creates a sharp discontinuity and permits measurement of the impacts on subsequent staff attendance. Our identifying assumption requires that facilities just below the cutoff (those with two staff absent during a health inspector’s visit) and facilities just above the cutoff (those with three staff absent) share potential outcomes in the absence of the flagging. There are several reasons that this identifying assumption may not hold. For example, there could be mean reversion. If facilities with two staff absent revert to the mean of having three staff absent, or if facilities having three staff absent revert to a mean of having two staff absent (or both), our specification would pick this up as an effect of flagging. In Section 4.4.1 we report four tests of this identifying assumption, all of which are consistent with a causal interpretation of our estimates of the impact of flagging on subsequent attendance.

Our data have limitations, however, which we also acknowledge in Section 4.4.1. Perhaps the greatest limitation is the fact that the data are limited to treatment districts and furthermore to facilities that had a health inspection (and thus flagging or not flagging) within a window before one of our survey visits to measure attendance. Thus, while we can conduct some placebo tests to test whether our flagging was “as-if” random for those facilities right below and above the cutoff, we are limited in how well we can verify parallel pre-trends and control for facility absence history. Results should be interpreted accordingly.

More specifically, we examine whether this manipulation within treated districts affected subsequent doctor absence in our primary data with the following specification:

\[
\text{Present Survey}_{jt} = \alpha + \beta_1 \text{Flagged}_{jt-1} + \delta t + \eta_{jt} \tag{4}
\]

Present Survey\(_{jt}\) is a dummy variable equal to 1 if the doctor \(j\) was absent during an unannounced visit by our enumerator in wave \(t\), Flagged\(_{jt-1}\) is a dummy variable that equals 1 if the facility was flagged in red on the dashboard in a window of time prior to the primary survey wave \(t\). For our primary analysis, we restrict to those facilities with either two or three staff absent, those on either side of the sharp discontinuity, though we report all possible bandwidths in Appendix Table A9. The estimate of the flagging effect is not significant at conventional levels for bandwidths that include data away from the discontinuity.

Selecting the window for which Flagged\(_{jt-1}\) is non-missing (i.e. 0 or 1) involves trade-offs and provides us substantial discretion in the analysis. We expect it would take at least a few days for senior officials to acknowledge and act on data from the dashboard. Practically, they need to receive the report that a clinic is underperforming, and then communicate their dissatisfaction to clinic personnel. Senior officials have a number of tools to reprimand clinic staff, with verbal warnings being by far the most common. Staff then need to react and change their behavior. This process will require at least a few days to play out, but should not take more than a few weeks. Also, practically, if the window for which Flagged\(_{jt-1}\) is defined is too long, virtually every facility will become flagged, limiting the variation with which to estimate effects. For transparency, we present estimates of the effect on doctor attendance for a broad number of potential time windows.

Specifically, we run the regression from equation (4) 750 times, varying the window for which we define a clinic as flagged prior to a primary unannounced visit to a clinic along two dimensions—we vary the length of the window being used along the x-axis and the delay from when a clinic is highlighted in red to when the window begins along the y-axis (so for example, a length of 25 and delay of 15 corresponds to considering a clinic as flagged if it was highlighted in red anytime 15–40 days prior to an unannounced visit).\(^{24}\)

We observe a positive, robust, and significant treatment effect of flagging on doctor attendance across a wide range of windows, depicted in the top left panel of Appendix Fig. A1. At our preferred window,\(^{25}\) reported in Table 3, doctor attendance subsequent to flagging on the dashboard increases by 17.7 percentage points or about 75 percent.\(^{26}\) Conversations with government partners suggest that the most likely driver of this effect is verbal reprimands from senior health officials to doctors in charge of clinics.

Table 3

<table>
<thead>
<tr>
<th></th>
<th>Doctor (1)</th>
<th>Dispenser (2)</th>
<th>LHW (3)</th>
<th>Health Tech (4)</th>
<th>SHNS (5)</th>
<th>Midwife (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flagged</td>
<td>0.177**</td>
<td>0.076</td>
<td>−0.115</td>
<td>−0.066</td>
<td>−0.037</td>
<td>−0.125</td>
</tr>
<tr>
<td></td>
<td>(0.082)</td>
<td>(0.083)</td>
<td>(0.090)</td>
<td>(0.092)</td>
<td>(0.093)</td>
<td>(0.091)</td>
</tr>
<tr>
<td>Unflagged Mean</td>
<td>0.236</td>
<td>0.676</td>
<td>0.703</td>
<td>0.405</td>
<td>0.324</td>
<td>0.568</td>
</tr>
<tr>
<td># Clinics</td>
<td>112</td>
<td>116</td>
<td>116</td>
<td>116</td>
<td>116</td>
<td>116</td>
</tr>
<tr>
<td># Reports</td>
<td>139</td>
<td>136</td>
<td>136</td>
<td>136</td>
<td>136</td>
<td>136</td>
</tr>
<tr>
<td>R-Squared</td>
<td>0.298</td>
<td>0.258</td>
<td>0.190</td>
<td>0.264</td>
<td>0.176</td>
<td>0.242</td>
</tr>
<tr>
<td>District Fixed Effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Notes: This table reports on the effect on subsequent doctor attendance of flagging on an online dashboard the fact that a clinic had three or more staff absent to a senior policymaker. Clinics were flagged in red on an online dashboard if three or more of the seven staff were absent in one or more health inspections of the clinic 11–25 days prior to an unannounced visit by our survey enumerators. The data sample limits to facility reports in which either two or three staff were absent (the threshold to trigger the underreporting red flag). In addition, the sample in all columns is limited to ‘Monitoring the Monitors’ treatment districts due to the necessity of the web dashboard for flagging clinics. All regressions include survey wave fixed effects. LHWs are Lady Health Workers and SHNS are School Health and Nutrition Supervisors. Standard errors clustered at the clinic level are reported in parentheses. *\(p < 0.1\), **\(p < 0.05\), ***\(p < 0.01\).

\(^{24}\) This analysis necessarily involves testing the same null of \(\beta_1 = 0\) in many closely related specifications. We do not adjust inference for multiple tests.

\(^{25}\) We chose this window to allow for the interpretation of a specific point estimate. It is a window we find plausible and it has a coefficient and significance level that is in the middle of those reported in Appendix Fig. A1.

\(^{26}\) Note that this positive result cannot be directly compared to the estimated average treatment effect of ‘Monitoring the Monitors’ on doctor attendance. To detect the effect of flagging we are limiting our sample to treatment facilities only (those that could be flagged on the dashboard) that were right below or above the staff absence cutoff for flagging.

4.4.1. Validity tests for ‘flagging’ results

This section presents four validity tests for the flagging results. First, an alternative explanation for the effect of the dashboard could be simple mean reversion. If absence is unusually high in one month, it might drop in the next month because of mean reversion, whether or not it was flagged. We check this in Appendix Fig. A2 where we use placebo thresholds between one and two and three and four absences. We find no evidence of mean reversion at placebo thresholds. Note that the first placebo test compares facilities recording two absences (the actual control group) against the group recording one absence (who are also not flagged on the dashboard). The second placebo test involves comparing facilities recording three absences (the actual ‘treatment’ group) against those with four recorded absences (who are also flagged on the dashboard). If a tendency for facilities with either two or three absences to revert to the mean were driving the results, with no role for flagging, we would expect the first placebo test to yield a negative estimate, and the second placebo test to yield a positive estimate. In the small number of cases where our placebo regressions find significance, the signs are the opposite of what mean reversion would predict (and consistent with a true positive effect at the 2/3 cutoff).

Second, we present in Appendix Fig. A3 a placebo test in which we test whether flagging on the dashboard predicts increases in attendance recorded in survey visits prior to the flag. That is, we define a facility as flagged if it was highlighted red on the online dashboard in varying windows after our survey visits. It is not possible for dashboard flags to causally impact outcomes in the past. This exercise therefore tests for a time series process whereby either facilities with three absences tend toward a mean of having only two absences, or for a process whereby facilities with two absences tend toward a mean of having three absences (or both). Again, we find no clear evidence of this pattern of mean reversion. An important caveat, however, is that these tests have low power. Using the analogous placebo window to our preferred specification above, for example, while the coefficient on the flagging placebo is −0.004, its 95% confidence interval is from −0.25 to 0.25.

Third, Appendix Table A10 presents two results controlling for the possible time dynamics of inspections, flaggings, and doctor attendance. For example, we might think a facility that is flagged multiple times in a row sees different outcomes than one that is flagged and then not flagged or flagged and then not inspected. Column 1 repeats our main flagging result. Then, in column 2, we restrict the sample to cases in which the flagging that occurred prior to our survey visit was the first time a facility was ever flagged. Treatment in this sample is straightforward to interpret since it does not require accounting for prior flags. Taking this approach shrinks the sample as it removes facilities that were flagged multiple times before our survey visit. Estimates in this much smaller sample comprising only 130 reports are similar in magnitude to our result in column 1, suggesting doctors respond to flagging even when it is the first time. To additionally control for dynamic effects, in column 3, we add fixed effects for the number of prior inspections in our treatment period (all of which must have been cases where two or less staff was absence for the flagging visit to be the first). Results are again similar. While this does not rule out dynamic effects, it suggests flagging can have an effect for facilities with a range of inspection histories.

Fourth, a standard practice to check the validity of regression discontinuity designs is to examine whether fixed or predetermined variables are smooth across the discontinuity (Calonico et al., 2014). The analogue here is to check for balance on predetermined variables between facilities with either three absences or two absences. Appendix Table A11 presents balance across facilities in our main specification in Table 3 coded as flagged versus facilities coded as unflagged. We do find imbalance on four of 18 variables at the 10% level, though not in baseline doctor attendance, our primary outcome variable. Finally, in Appendix Table A12 we re-run our primary flagging result controlling for all variables for which there is baseline imbalance, and find that results are similar.

Fifth, it is worth noting that flagging effects materialize approximately 10–20 days after the reports are first filed on the online dashboard. Throughout the project, we interacted with bureaucrats using the dashboard. They indicated very consistently that they would review the dashboard approximately once a week. A 10 day response is consistent with a manager checking about once a week and then demanding action from their subordinates. We therefore did not expect that attendance would respond immediately. That we only find effects in an intermediate window of a reasonable length is consistent with our explanation of how the dashboard was used: actions by senior officials and subsequent responses by doctors require time rather than through simple mean reversion. These results can be found on the top left panel of Appendix Fig. A1.

4.5. Heterogeneity of results

Whether increased monitoring and data affect service delivery reasonably depends not only on the abilities of government personnel but also on the broader political environment. We now examine two sources of heterogeneity, taking advantage of the scale of our experiment.

First, a growing literature documents the role of individuals in public service delivery across the developing world, summarized in Finan et al. (2017). We examine whether personality characteristics of the health workers in our study predict their response to monitoring and to information. To do this, we measured personality characteristics—the Big Five Personality Index and the Perry Public Service Motivation Index—of all of the doctors in our sample clinics and the universe of health inspectors and senior health officials in Punjab. In Callen et al. (2017a) we explore this dimension of heterogeneity in great detail. As an example of the role of personality characteristics in this setting, we present heterogeneity of our flaggings results in Table A13. We find that personality characteristics systematically predict responses by senior health officials to our dashboard experiment, as measured by future doctor attendance in flagged facilities: a one standard deviation increase in the Big 5 Personality Index of senior bureaucrats increases the effect of flagging on the likelihood that doctor is present at the clinic during a subsequent unannounced visit by 28 percentage points. We see a similar, though smaller and statistically insignificant, point estimate for the interaction with Public Service Motivation. Of course, these results are merely predictive as we do not have identifying variation in personality traits.

Second, we explore whether measures of political competition predict flagging effects. The bureaucrats we worked with to create the program felt strongly that the program would break down when politicians interfered with senior officials’ attempts to sanction their subordinates. Indeed, in our surveys, senior health officials report that politicians routinely interfere in this way. In Callen et al. (2017b) we match each clinic in our sample to a provincial assembly constituency and examine

---

27 In general, at any given point in time, different facilities will have been flagged different numbers of times. With sufficient data, flexibly controlling for the comprehensive treatment history would allow recovery of the causal (and recovery of heterogeneous treatment effects given different histories). With only three survey visits and being restricted to the smartphone monitoring districts, we lack the data to pursue this approach.

28 Identification, in this case, only requires parallel trends prior to flagging. If we had reliable time series data on attendance prior to flagging, we could test this directly. Unfortunately, we only have at most two pre-flagging survey audits.

29 Based on interviews with all senior health officials in Punjab, we find that 44% report a politician interfering in their decision to sanction an underperforming employee during the previous year.
in detail the extent to which these political moderators affect the efficiency of the Monitoring the Monitors program. As an example, we find that the treatment effect of flagging on doctor attendance varies by the degree of competitiveness in the previous election. Table A14 presents the results. We find that while flagging increases subsequent attendance by 35.9 percentage points in the most politically competitive third of constituencies, flagging has no apparent effect in the least competitive third. In addition, flagging works better on doctors who do not report a direct connection with a local politician. Indeed, the point estimates, though noisy, suggest the program may have negatively affected attendance of connected doctors.

5. Discussion and conclusion

A fundamental objective of policy research is to convey facts and data to policy makers. We find that providing senior health officials with information on low staff attendance causes them to take corrective action, indicating that data can change how policy makers behave, even in settings with weak institutions. This suggests two additional general lines of inquiry.

First, our test of whether information affects policy decisions is direct. For our senior officials, who are tasked with responsibilities that affect millions, ensuring the functioning of frontline service facilities is a priority. A natural question is whether we would see the same response to a more complicated object, like causal estimates of program effect, cost-benefit analyses, or more general forms of research evidence. The ‘Monitoring the Monitors’ program provides proof-of-concept that technology can mobilize data to real effect. This also suggests, but by no means proves conclusively, that the massive investments being made by governments, technologists, researchers, philanthropists, and aid organizations to promote evidence-based policy can make a difference in developing countries.

Second, whether data and evidence impact policy likely depends on the characteristics of policy makers, and the political and institutional environment in which they operate. The scale of our experiment provides enough variation to examine preliminarily whether these factors matter. We find some evidence, albeit highly speculative, that these considerations are relevant. The personalities of both inspectors and senior health officials predict how the ‘Monitoring the Monitors’ program impacted them. Similarly, the degree of local political competition predicts where the program will work best. These results are correlational, and should be treated with appropriate skepticism, but do suggest that the effect of data on policy decision might depend critically on context.

The ‘Monitoring the Monitors’ program cost 17,800 USD to set up and 510 USD per month to operate. While the results of the program are mixed, given this low cost, we would expect it to pass a cost-effectiveness test. The government of Punjab scaled the program up to cover the entire province at the conclusion of the study. This investment by Punjab and others like it have driven a revolution in the amount of data that can quickly and cheaply be accessed for policy decisions. This trend is only likely to accelerate with the rise of remote sensing, digital trace (e.g., cell phone call and mobile money transaction records), smartphones, and other research innovations. A key lesson from this exercise is that, appropriately channeled, these data streams can improve policy outcomes.

Author statement

All authors contributed equally to the research project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jdeveco.2020.102523.

References

Blum, Florian, Pande, Rohini, July 2015. Data Poverty Makes it Harder to Fix Real Poverty. That’s Why the UN Should Push Countries to Gather and Share Data.
Cullen, Michael, Gibson, Clark, Jung, Danielle, Long, James, 2016. Improving electoral integrity with information and communications technology. J. Exp. Polit. Sci. 3 (1), 4–17.
Dhalaiwal, Ishbal, Hanna, Rema, 2017. Deal with the devil: the successes and limitations of bureaucratic reform in India. J. Dev. Econ. 124 (C), 1–21.

30 Recent research finds that providing information to politicians can similarly impact their decisions (Hjort et al., 2019).
31 The set up costs included 4470 USD to develop the app and 13,330 USD for smartphones.