Official Attendance Data for Politically Connected Bureaucrats Are Less Accurate*

Michael Callen†
Saad Gulzar‡
Ali Hasanain§
Yasir Khan¶

April 11, 2016

Abstract

Research shows that official data can often deviate from the truth. This paper shows that the absence of bureaucrats is underreported when they are politically connected. We compare absence reports in the universe of government inspections of public clinics in Punjab, Pakistan (N=79,318), with independent unannounced inspections of a representative sample of 850 clinics. We present robust evidence that government data underreport doctor and staff absence by up to 12.9 percentage points. Importantly, we show that doctors who personally know the local politician are less likely to be reported absent in official data. Our results signal caution in the use of official data as incentives to misrepresent data may be correlated with political objectives.

Authors' Note: We thank Farasat Iqbal (Punjab Health Sector Reforms Project), Asim Fayaz, and Zubair Bhatti (World Bank) for assistance in obtaining data. Support is provided by the International Growth Centre Pakistan Country Office. Excellent research assistance was provided by Jawad Karim, Zia Mehmood, and Arman Rezaee.

†Harvard Kennedy School. michael_callen@hks.harvard.edu

‡New York University. saad.gulzar@nyu.edu

§Lahore University of Management Sciences. hasanain@lums.edu.pk

¶University of California Berkeley. yasir.khan@berkeley.edu

Data generated by official agencies is critical for well-functioning governments. Government agencies rely on these data for process evaluations, as well as for the efficient allocation of scare public sector resources. This is perhaps why aid agencies and developing country governments are devoting considerable resources to improving government reporting systems and statistics agencies.

Recent research suggests that the quality of data generated by official sources may be severely lacking, particularly in the developing world (Sandefur and Glassman, 2015). This paper aims to show that there may be political reasons for this. We evaluate the quality of government data in comparison to independently collected primary data, and then show that these differences are predicted by political connections.[1]

We focus on the public health sector in Pakistan, and test for differences between unannounced government health inspection reports, and data from independent visits to a representative sample of 850 rural clinics collected during December 2011 in Punjab, Pakistan. We find that government inspectors systematically underreport health worker absence. The absence rate for doctors is 67.6 percent in the primary data and 63.5 percent in the government data. All paramedic staff (Medical/Health Technicians, and Dispensers) are absent in 18.2 percent of visits in the primary data but only 10.1 percent of government inspections. We reject that both these differences are equal to zero with 95 percent confidence. Next, we evaluate a political economy reason for this difference. In patronage and clientelistic systems, where public sectors jobs are often given to favored personnel, We find that for the *same* politician, government data underreports absence when doctors know local politicians personally.

Besides policy, our results are important for social scientists who routinely rely on official datasets. In 2015, 56 percent of all articles published in top political science journals

[1]Our study resembles de Mel, McKenzie and Woodruff (2009) in that we deploy independent surveyors to evaluate the quality of commonly available data critical to development economics research.

made use of some official dataset – a figure that jumps to 70 percent if attention is restricted to papers employing quantitative analysis. Conditional on quantitative analysis, articles use official data as follows: 50 percent in the American Political Science Review, 61 percent in the American Journal of Political Science, and 88 percent in the Journal of Politics. Table 1 provides a detailed breakdown. There may be reasons for caution if inaccuracies in official data are correlated with political outcomes under study in these papers.

Table 1 about here.

Background

In 2005, the government of Punjab, Pakistan launched a program to use Monitoring and Evaluation Agents (MEAs) to inspect government schools under the Punjab Education Sector Reform Program. This was later expanded to include health facilities. The program established a system whereby MEAs provide monthly information on education and health facilities. The program sought to establish a system of monitoring that remained within the ambit of the Government of Punjab, but that was not subordinate to the departments being monitored. As of 2013, these data are reported directly to the most senior official in the District,[2] as well as to province-level agencies like the Punjab Health Sector Reform Program (PHSRP). The PHSRP is in charge of utilizing these incoming data to improve health management in the province.

Our focus is on Basic Health Units (BHUs) ('clinics' here onwards) which are the most local-level public health care units in Pakistan. They are designed to be the first stop for patients seeking medical treatment in government facilities. Services provided at clinics include out-patient services, ante-natal and reproductive healthcare, as well as vaccinations against diseases. Each facility is headed by a doctor who is supported by a Dispenser, and

[2]Punjab, a province of 100 million, is divided into 36 administrative districts.

a Health/Medical Technician.[3]

Description of Data

Government Data: MEAs visit each clinic every month to conduct a surprise inspection. They gather an extensive log of service delivery using a standardized form developed by the Health Department. Each visit typically lasts around 30 minutes. The forms record information on: (i) attendance for approximately 15 clinic staff in total including doctors, paramedics, preventive and administrative staff, (ii) medicine availability, (iii) cleanliness and infrastructure quality, (iv) facility usage, and (v) outreach activities being conducted by the clinic.

The paper forms filled by the MEAs at the facility are digitized with a lag of one month. At PHSRP, these data are collated for the entire province to construct a provincial database. PHSRP then aggregates these data and ranks each district on their performance on various indicators. We use 79,318 health inspection visits from this database over the period January 7, 2008 - February 29, 2012. These data represent the universe of clinic inspections by MEAs in Punjab province for this period.

Primary Data: We collect primary data on a representative sample of clinics in Punjab in December of 2011. Clinics were selected randomly using an Equal Probability of Selection (EPS) design, stratified on district and distance between the district headquarters and the clinic. Because of the EPS design, our estimates of absence are self-weighting, and so no sampling corrections are used in our analysis. We sampled 850 (34 percent) of the 2,496 clinics. All districts in Punjab except Khanewal are represented in our data.[4] To our

[3]Other staff including a Lady Health Visitor, a School Health and Nutrition Supervisor, and a Mid-wife, are associated with clincs but perform roving duties in surrounding areas.

[4]Khanewal served as the pilot district in a randomized control evaluation of a government program for which this dataset was a baseline. Figure A1 locates the clinics in our sample

knowledge, this is the first representative survey of clinics in Punjab.

Our data are comparable to official data. As in the case of MEA data, our team collected information on staff absence, availability of medicine, and information on facility usage. We restrict our comparison to doctor and paramedic staff posting and attendance, as these are the only variables measured using identical protocols across both data sets. When considering comparisons, it is important to note that the median difference between when our surveyors visit and when the MEAs visit a clinic is nine days. However, as both types of visits are supposed to be random, these differences should not systematically affect absence reports.

Our enumerators interviewed the Doctor, and other staff, before physically verifying the attendance of the Mid-Wife and the School Health and Nutrition Specialist. The attendance sheet for the staff was filled out at the end of the interviews and in private. We present summary statistics from the data in Table A1. Of the 850 clinics in our primary data, we have no government data on 3 clinics. In addition 25 clinics that were found closed during our primary visits are not considered in the analysis.

Results

We compare absence levels across our two datasets. We begin our analysis using the same definition for absence as Muralidharan et al. (2011)—service providers are defined as absent if enumerators cannot find them in the health facility at a time when they should be on duty. According to this definition, health workers will be counted as absent even if they are away from the clinic for sanctioned reasons or if the position is not filled. We term this *unconditional absence*. Ignoring the reason for absence allows for the most direct comparison between the two data sets. If reporting on the reasons for absence is subject to different biases in the two data sets, then comparisons between justified absence will suffer from an additional source of error.

---

across Punjab.

However, filling staff positions and preventing worker absence are distinct policy challenges. We attempt to focus narrowly on the latter issue by analyzing absence among providers in positions that are reported to be filled. We term this *conditional absence*. As in the previous analysis, however, we do not attempt to account for sanctioned absence in the analysis.

Unconditional Absence

We create two indicator variables for absence of staff whose duty is primarily at the clinic: *Doctor Absent*, and *Paramedic Staff Absent*, where the latter is equal to one if the Medical/Health Technician, and Dispenser are both absent. This definition is consistent with the MEA data set.

Table 2, Panel A, reports means of health worker absence collected by our enumerators and reported by MEAs in 847 facilities. No MEA visits are recorded for 3 of the 850 BHUs in our primary sample, while another 25 clinics are not considered in the analysis as the clinics were found closed during our survey. Means from the primary data are reported in Column (1) and means from the government data are reported in Column (2). We also report differences between these means and p-values corresponding to a t-test that these differences are equal to zero. In Column (1), we see that absence rates during unannounced visits are 67.6 percent for doctors and 18.2 percent for paramedic staff. In Column (2), we see that absence rates are 63.5 percent for doctors and ten percent for paramedic staff. The final column of the table shows that the difference in means for both absence measures is statistically different from zero.

Table 2 about here.

While these differences are large, there is still a considerable amount of absence reported in the MEA data. This may be in part due to the fact that MEAs have some institutional independence from the Health Department. In any case, these absence rates of roughly 68 percent for doctors are stark and comparable to Bihar, the worst state reported in India in

2003 in Muralidharan et al. (2011).

Conditional Absence

Next, we analyze differences in attendance for staff positions that have someone actually posted to the clinic. We define three variables: *Assigned Doctors Absent*, *Share of Paramedic Staff Absent*, and *Share of Measured Staff Absent*. *Share of Paramedic Staff Absent* is the share of all paramedic staff (Medical/Health Technicians, and Dispensers) assigned to the clinic who are absent. *Share of Measured Staff Absent* is the share of paramedic staff and doctors assigned to the facility who are absent.

Table 2 Panel B reports the results. Again, we find that absence is higher in primary data than in the government data. Doctors in filled positions are absent during 48 percent of surveyor visits and during 40.6 percent of MEA visits. 32.4 percent of the assigned paramedic staff are absent when our enumerators visit and and only 21 percent of assigned paramedic staff are absent in the government data. Both differences as well as the the difference for all measured staff are statistically different from zero.

Figure 1 about here.

Figure 1 plots the estimated empirical densities of the share of assigned staff absent in the primary data, as well as the government data for December 2011. We also plot the clinic average absence over the period that ranges from January 2008 to December 2011. It is clear that absence is underreported in the government data, and that there is less noise in the government data taking averages over a longer period.

Doctor Connections

There can be several reasons for the differences we observe between the primary and government data. A critical question for our analysis is the extent to which these differences are

attributable to deliberate misreporting of staff presence.[5]

A literature shows the positive benefits that accrue to firms from their political connectedness (Fisman, 2001; Khwaja and Mian, 2005). We study whether the discrepancies in absence reporting are explained by political connections of doctors. If doctors rely on their political connections to shirk from work, they should be absent more often in the primary data than in the government data. During fieldwork we collect self-reported data on whether doctors know the local politician personally. We collect additional primary data in two waves: in June 2012 and in October 2012. Our measure of connections is an indicator variable that equals 0 unless doctors report they know the local Member of the Provincial Assembly (MPA) in all three waves where this question is answered, in which case, it is coded as 1. Of the 534 doctors assigned in the primary data, we have political connections information on 441 doctors.

To measure the discrepancy across the two datasets, we create a new indicator variable that equals 1 every time a doctor is marked absent in the primary data but present in the government data. In Table 3, we regress this outcome on doctor connections with politicians. We can see a positive relationship between the two variables, even when comparing differences between doctors who know and do not know the *same* politician (Columns (3) and (4)). This suggests that when doctors know the politician, they are more likely to be marked absent in the primary data than in government data.[6]

Table 3 about here.

[5]In Appendix B we test and do not find support for measurement error explaining the results.

[6]Table A2 shows that doctor connections with politicians negatively, though not robustly, predict an outcome that measure instances where doctors are marked absent in the primary data, but present in the government data.

Discussion

Policy makers and researchers need reliable data for their work. During our conversations with senior health bureaucrats in Pakistan, the unreliability of government data was frequently connected with political economy and administrative problems.

We show that the reliability of government data may be suspect because of political reasons. We compare government data on the absence of health staff in rural clinics with independent surveyor visits and visits by government health inspectors. We find a range of evidence consistent with government underreporting absence. In addition, doctors who personally know the local politician are more likely to be reported absent in the primary data than in government data. This is consistent with the findings of a growing literature on the value political of connections.

Our results carry implications for policy and political science research. Official data is produced by agents who act under rational incentives. If the use of official data is not related to the incentives for misreporting, statistics should only include random bias. However, as we demonstrate here, if misreporting is correlated precisely with the kinds of political interference researchers will want to identify, and policy makers will try to rectify, then the presence of such behavior will be biased towards zero. Our results therefore recommend caution when official data is being used to study political phenomenon.

References

de Mel, Suresh, David J. McKenzie and Christopher Woodruff. 2009. "Measuring microenterprise profits: Must We Ask How the Sausage is Made?" *Journal of Development Economics* 88(1):19 – 31.

Fisman, Raymond. 2001. "Estimating the Value of Political Connections." *American Economic Review* 91(4):1095–1102.

Khwaja, Asim Ijaz and Atif Mian. 2005. "Do Lenders Favor Politically Connected Firms? Rent Provision in an Emerging Financial Market." *Quarterly Journal of Economics* 120(4):1371–1411.

Muralidharan, Karthik, Nazmul Chaudhury, Jeffrey Hammer, Michael Kremer and F. Halsey Rogers. 2011. Is There a Doctor in the House? Medical Worker Absence in India. Working paper Unviersity of California San Diego.

Sandefur, Justin and Amanda Glassman. 2015. "The Political Economy of Bad Data: Evidence from African Survey and Administrative Statistics." *Journal of Development Studies* 51(2):116–132.

Figures and Tables

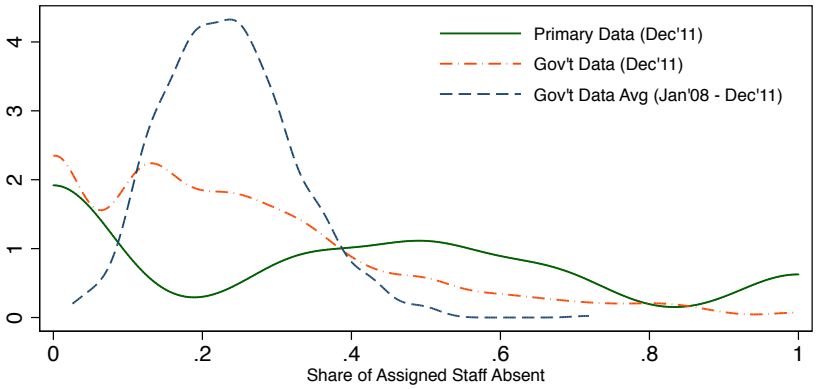

Figure 1: Absence Densities

Table 1: Use of Official Data in Political Science Research 2015

	Articles Considered	Data Analysis	Official Data
American Political Science Review	48	28	14
American Journal of Political Science	66	65	40
Journal of Politics	82	64	56
Total	196	157	110

Table 2: Absence in Primary and Government Data

	Primary Data	Gov't Data	Diff	t-test (p-value)	N
	(1)	(2)	(1) - (2)	(1) - (2)	
Panel A: Unconditional Absence					
Doctor Absent (=1)	0.676	0.635	0.041	0.022	822
	[0.468]	[0.482]	(0.018)	.	.
Paramedic Staff Absent (=1)	0.182	0.101	0.080	0.000	820
	[0.386]	[0.302]	(0.016)	.	.
Panel B: Conditional Absence					
Assigned Doctors Absent	0.480	0.406	0.073	0.009	465
	[0.500]	[0.491]	(0.028)	.	.
Share of Paramedic Staff Absent	0.324	0.210	0.114	0.000	803
	[0.376]	[0.322]	(0.016)	.	.
Share of Measured Staff Absent	0.363	0.234	0.129	0.000	818
	[0.345]	[0.215]	(0.014)	.	.

Notes: Standard deviations in brackets and standard errors reported in parentheses. Data are from 847 of the 850 clinics in the primary sample for which government data are also recorded. 25 of these clinics are not considered in the analysis as they were found closed during primary visits staff assignment could not be verified. Observations in Panel B are restricted to facilities for which staff in the respective category are assigned in both the primary and government data.

Table 3: The Value of Political Connections

Outcome: Doctor Absent in Primary Data and Present in Gov't Data (=1)				
	Unconditional		Conditional	
	(1)	(2)	(3)	(4)
Knows Local Politician Personally	0.152**	0.162	0.246***	0.270*
	(0.063)	(0.126)	(0.071)	(0.153)
Constant	0.176***	0.175***	0.168***	0.165***
	(0.018)	(0.014)	(0.020)	(0.016)
Observations	519	519	440	440
Politician Fixed Effects	No	Yes	No	Yes

Notes: Standard errors clustered at the politician level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

ONLINE APPENDIX

A Additional Tables and Figures

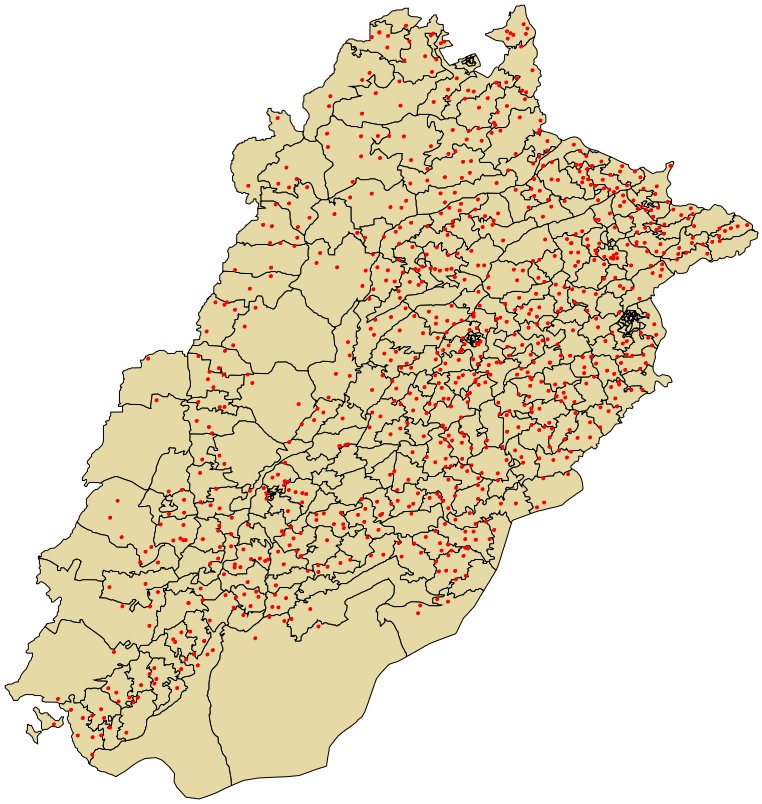

Figure A1: Clinic Locations in Political Constituencies in Primary Data

Table A1: Summary statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
Clinics in Primary Sample					850
Clinics Not in Gov't Sample					3
Clinics Found Closed in Primary Sample					25
Total Clinics in Analysis $(850 - 3 - 25)$					822
Doctor Assigned, Primary (=1)	0.65	0.48	0	1	822
Paramedic Staff Assigned, Primary (=1)	0.99	0.1	0	1	820
Doctor Assigned, Gov't (=1)	0.63	0.48	0	1	847
Paramedic Staff Assigned, Gov't (=1)	0.98	0.13	0	1	847
Doctor Absent, Primary (=1)	0.68	0.47	0	1	822
Paramedic Staff Absent, Primary (=1)	0.18	0.39	0	1	820
Doctor Absent, Gov't (=1)	0.63	0.48	0	1	847
Paramedic Staff Absent, Gov't (=1)	0.1	0.3	0	1	847
Doctor Absent Conditional, Primary (=1)	0.5	0.5	0	1	534
Average Doctor Absence Conditional, Gov't (Dec' 11)	0.42	0.49	0	1	530
Paramedic Staff Absent Conditional, Primary (=1)	0.33	0.38	0	1	812
Average Paramedic Staff Absence Conditional, Gov't	0.21	0.32	0	1	833
Share of Assigned Staff Present, Primary	0.36	0.34	0	1	818
Share of Assigned Staff Present, Gov't	0.23	0.21	0	1	847
Doctor Knows Local Politician Personally (=1)	0.11	0.31	0	1	535
Doctor Absent in Primary Data, Present in Gov't Data (=1)	0.16	0.36	0	1	822
Doctor Present in Primary Data, Absent in Gov't Data (=1)	0.11	0.32	0	1	822

Table A2: Additional Results on Political Connections

Outcome: Doctor Present in Primary Data and Absent in Gov't Data (=1)				
	Unconditional		Conditional	
	(1)	(2)	(3)	(4)
Knows Local Politician Personally	-0.120***	-0.109	-0.134***	-0.092
	(0.040)	(0.101)	(0.047)	(0.104)
Constant	0.189***	0.188***	0.221***	0.216***
	(0.020)	(0.011)	(0.023)	(0.011)
Observations	519	519	440	440
Observations	519	519	440	440
Politician Fixed Effects	No	Yes	No	Yes

Notes: Standard errors clustered at the politician level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B Measurement Error

A concern for our analysis is that our surveyors and MEAs are not visiting the same facilities on the same days. While both visits occur during periods when doctors should be at the clinic, it is possible that absence varies considerably across days. If our surveyors systematically visit on days where health staff are less likely to be present, then the difference we document is not a difference in true average absence. We think it is unlikely that this could generate our result, as our visits are random and MEA reports should also be random. We nevertheless, present a tests for this.

First, we constrain our sample to visits that have happened within 30 days of one another in Table A3 and find that results are virtually identical to our core results.

Second, we show in Figure A2 that the difference between primary data absence reporting and the government data always remains positive no matter how apart the primary data collection is from the MEA inspector's visit.[7] This increases confidence that our results are not driven by measurement error.

Table A3: Health Worker Absence in Primary and Government Data

	Primary Data	Gov't Data	Diff	t-test (1) - (2)	N
	(1)	(2)	(1) - (2)	(p-value)	
Doctor Absent (=1)	0.672	0.634	0.039	0.040	822
	0.470	0.482	0.019	.	.
Paramedic Staff Absent (=1)	0.183	0.095	0.087	0.000	820
	0.387	0.294	0.016	.	.

Notes: Standard deviations in brackets and standard errors reported in parentheses. Data are from 847 of the 850 Basic Health Units in the primary sample for which government data are also recorded. Doctors are Medical Officers (MOs). Paramedic staff are Medical Technicians (MTs), Health Technicians (HTs), and Dispensers.

[7]The size of the circle shows the amount of data in each bin.

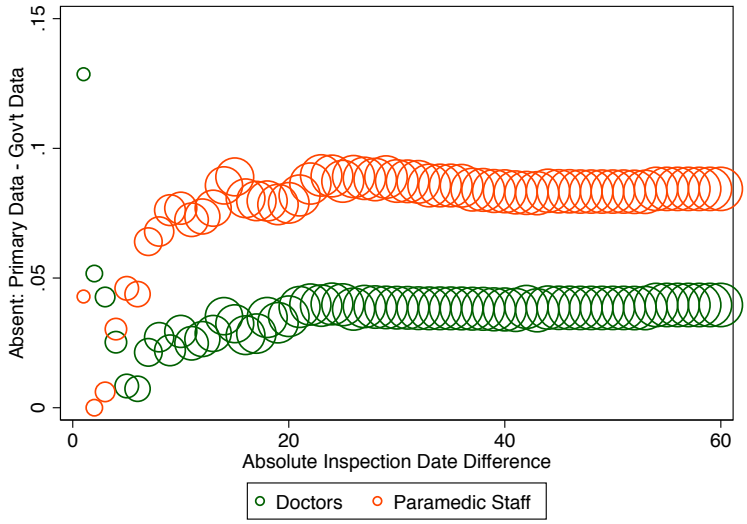

Figure A2: Absence Densities