Policy Research Working Paper 8951

Incentivizing Quantity and Quality of Care

Evidence from an Impact Evaluation of Performance-Based Financing in the Health Sector in Tajikistan

Tashrik Ahmed
Aneesa Arur
Damien de Walque
Gil Shapira

WORLD BANK GROUP
Development Economics
Development Research Group
&
Health, Nutrition and Population Global Practice
July 2019

Policy Research Working Paper 8951

Abstract

This paper presents the results of an impact evaluation of a performance-based financing pilot in rural areas of two regions of Tajikistan. Primary care facilities were given financial incentives conditional on general quality and the quantity provided of selected services related to reproductive, maternal and child health, and hypertension-related services. The study relies on a difference-in-difference design and large-scale household and facility-based surveys conducted before the launch of the pilot in 2015 and after three years of implementation. The performance-based financing pilot had positive impacts on quality of care. Significant impacts are measured on facility infrastructure, infection prevention and control standards, availability of equipment and medical supplies, provider competency, provider satisfaction, and even some elements of the content of care, measured through direct observations of provider-patient interactions. While the communities in the performance-based financing districts reported higher satisfaction with the local primary care facilities, and despite the improvements in quality, the findings suggest moderate effects on utilization: among the incentivized utilization indicators, only timely postnatal care and blood pressure measurements for adults were significantly impacted.

This paper is a product of the Development Research Group, Development Economics and the Health, Nutrition and Population Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at gshapira@worldbank.org.

Incentivizing Quantity and Quality of Care: Evidence from an Impact Evaluation of Performance-Based Financing in the Health Sector in Tajikistan*

Tashrik Ahmed, Health Nutrition and Population, The World Bank and Johns Hopkins School of Public Health

Aneesa Arur, Health Nutrition and Population, The World Bank

Damien de Walque, Development Research Group, The World Bank

Gil Shapira, Development Research Group, The World Bank

Keywords: Performance-based Financing; Health; Quality of care

JEL classification: I15; J13; O15.

* We are extremely grateful for a fruitful collaboration with the Ministry of Health of Tajikistan. The evaluation was supported by the project implementation unit of the Tajikistan Health Services Improvement Project (HSIP), under the leadership of Mahbuba Mustafaevna. In particular, the team would like to acknowledge the work of Fatima Gaibova and Saodat Mirsaburova. Zerkalo Analytics led the data collection. For support and inputs during the study design and implementation, the team would like to thank Sarvinoz Barfieva, Mutriba Latypova, Rouselle Lavado, Kate Mandeville, Wezi Msisha, Ha Nguyen, Sabrina Qandenova, Jeanette Walldorf and Huihui Wang. This impact evaluation is funded by the Health Results Innovation Trust Funds (HRITF) 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 views of the World Bank, its Executive Directors, or the countries they represent.

1. Introduction

Substantial disparities in health care coverage persist in low- and middle-income countries despite the progress towards the Millennium Development Goals (Wagstaff, Bredenkamp and Buisman, 2014). These gaps exist for basic maternal and child health services, as well as for services aimed at preventing and treating non-communicable diseases. This assessment is starker if *effective* coverage is measured, i.e., coverage of high-quality services (Shengelia et al. 2005). There is substantial evidence that the quality of care in many low- and middle-income countries is low. Health conditions are often misdiagnosed and even when correctly diagnosed, the appropriate treatment or interventions may not be prescribed or available (Das, Hammer and Leonard 2008, Das and Hammer 2014, Kruk et al. 2018). Because of these gaps in quality of care, health outcomes improve at a slower pace than coverage rates.

Confronted with limited progress, low- and middle-income countries have experimented in incentives involving a mix of salaries, budgets and bonuses linked to performance (Witters et al. 2012; Miller and Barbiarz 2013; Peabody et al. 2014; Yip et al. 2014; Sun et al. 2016). Performance-based financing (PBF) is a pay-for-performance mechanism aimed at improving effective coverage through financial incentives to health providers which reward both quantity and quality of health services delivered (Fritsche et al. 2014; Friedman and Scheffler 2016). In recent years, a growing number of prospective impact evaluations of PBF programs have been completed (Basinga et al. 2011; Bonfrer et al. 2014a; Bonfrer et al. 2014b; Gertler et al. 2014; Celhay et al. 2015; de Walque et al. 2015; Engineer et al. 2016; Friedman et al. 2016; de Walque et al. 2017; Van de Poel et al. 2016; Shapira et al. 2018). These programs have had mixed success in improving the coverage of health services. Most studied interventions managed to positively improve coverage of only some of the incentivized services and a few programs failed to show any significant improvements on coverage. There is solid evidence that PBF positively impacts quality of care (Kandpal 2017). Almost all programs were successful in improving indicators of structural quality such as facility infrastructure and availability of medical supplies. A smaller number of studies have also shown positive impacts on the content of care (Basinga et al. 2011; Engineer et al. 2016; Friedman et al. 2016).

This paper presents an evaluation of a PBF pilot in Tajikistan, launched in January 2015 in selected rural districts in Khatlon and Sughd regions. Rural health centers (RHCs) and health houses receive

financial incentives conditional on the quantity and quality of provided services, including family planning, antenatal and postnatal care, child vaccination and growth monitoring, and hypertension. Up to 70 percent of the quarterly performance payments can be distributed as bonuses to the clinical staff, while at least 30 percent had to be reinvested in the facility. The program included multiple layers of supervision and verification by regional and national agencies and a third-party counter-verification of the performance indicators.

This impact evaluation relies on a difference-in-difference design, which compared changes in outcomes between the districts implementing PBF and control districts within the same regions. Coverage rates were measured at the population level by household surveys conducted before the launch of the program and after three years of implementation. Detailed data on quality of care was collected through extensive facility-based surveys that included general facility assessments, interviews with health providers, and direct clinical observations of patient-provider interactions.

The results provide evidence of substantial positive impacts of the PBF reform on many dimensions of quality of care. It increased the availability of equipment and supplies at the primary health centers. It had positive impacts on infrastructure and infection prevention and control standards, such as the availability of containers for sharps and needles in consultation rooms. We also find positive impacts on provider competency, measured through clinical vignettes. Finally, we find evidence that the improvements in structural quality and provider knowledge also translated into better content of care. For example, providers in the PBF facilities are more likely to perform key physical exams such as to measure the height and weight of children under 5.

The PBF pilot had positive impacts on health providers. Their income increased by about two-thirds due to performance bonuses, and they reported higher satisfaction, especially concerning the working conditions in the facilities. Community perceptions of health providers were also positively impacted. During the follow-up survey, individuals living in PBF districts reported significantly higher perceived competency of providers and that the staff work closely with and listens to the community. Additionally, individuals reported that during the three years of PBF implementation, the attitude of health providers improved.

While we find strong evidence of improved quality at the primary level and observe that the communities noticed the change, we find more modest impacts on the utilization of health services by the community. Concerning adult health, we find a positive impact of 3 percentage points on

the likelihood of adults aged 40 and above had their blood pressure measured by a health professional in the preceding year. With respect to maternal and child health services, we find an 18 percentage-point increase in the rate of women who received timely postnatal care. However, we do not find statistically significant impacts on timing and number of antenatal consultations, or coverage rates of child growth monitoring and vaccination. We also do not find an impact on family planning coverage, but it must be stressed that our sample of households with recently pregnant women is not optimal for measuring this outcome.

This study contributes to the literature on pay for performance in the health sector in low- and middle-income countries by presenting the first evidence from a Central Asian country with a post-Soviet health system. The political, economic, and epidemiological contexts in Tajikistan are distinct from those in Sub-Saharan Africa, where most of the previous PBF evaluations have been conducted. Given the dual burden of disease in the country, the intervention in Tajikistan is unique by incentivizing hypertension-related services for adults and not only focusing on maternal and child health.

The results of this study are overall in line with the global evidence about PBF. In many countries, PBF reforms have had mixed effects on utilization, while often finding positive impacts for quality of care. The quality of care impacts measured in Tajikistan are noteworthy in the existing literature because they extend beyond structural quality (infrastructure and equipment) and include instances of improvements in the content of care, such as greater percentage of patient consultations following recommended examinations.

2. The Tajikistan Health Services Improvement Project

Tajikistan is a mountainous and landlocked former Soviet republic in Central Asia. The 2017 population is an estimated 8.93 million persons with about three-quarters of the country living in rural areas. Tajikistan has seen significant advances since the conclusion of the civil war in 1997, although it remains the poorest among the former USSR states. Industrialization has driven high rates of economic growth over the past decade, and literacy rates are nearly 100%. Life expectancy

increased to 71.1 years and under-5 mortality per 1,000 live births reduced to 33.6 by 2017.[1] Despite these significant improvements, the burden of chronic diseases is growing.

Tajikistan's health system is heavily centralized. Health services are provided overwhelmingly within the public sector, focusing historically on hospital-based curative care. The share of public health expenditure is only 2% of GDP, among the lowest of countries in the Europe and Central Asia regional group.[2] In rural areas, rural health centers (RHCs) provide primary care services that are matched against the available resources and infrastructure. A rural health center typically has at least one physician in its staff, but most centers do not provide laboratory and inpatient services. The RHCs often manage affiliate health houses (HHs), which provide more basic services (immunization, first aid, referrals) in remote areas. Services of both RHCs and HHs are often provided through home visits.

The Tajikistan Health Services Improvement Project (HSIP), financially supported by the World Bank, and the Health Results Innovation Trust Fund were designed to improve the coverage and quality of basic primary health care services in selected rural districts in the regions of Sughd and Khatlon. The main component of the project was a pilot of Performance-Based Financing (PBF) at the RHC and HH levels, launched in January 2015. Additionally, the project financed trainings of primary health care providers in family medicine, maternal and child health, and noncommunicable diseases.

Performance-based financing in Tajikistan incentivizes maternal, child, and hypertension health services, and general quality of care. RHCs and HHs receive quarterly financial rewards according to the volume of services delivered and an overall quality score. Quantity and quality are linked to prevent providers from providing higher volumes of low-quality services. Quantity indicators are reported by health facilities and verified by district health teams and the State Health Activities Supervision Services (SHASS) agency by review of registers and patient records. Quality of care is scored by a quality checklist conducted during the SHASS agency visits. The UNICEF country office was contracted to conduct additional external independent counter-verification to confirm performance indicators. A maximum of 70% of the quarterly PBF payments to the facilities can

[1] World Bank, World Development Indicators.
[2] World Bank, World Development Indicators, the latest estimation available is for 2016.

be distributed as bonuses to clinical staff, with 30% allocated towards reinvestment in the facility. District health teams approve action plans developed by the facilities for these reinvestments.

Rewards for the quantity indicators are on a fee-for-service basis. Table 1 presents the specific incentivized quantity indicators and their corresponding fees. Services targeting child health include the number of fully vaccinated children under the age of 13 months, and growth monitoring and nutrition services for children under 5. Services targeting women include timeliness and quantity of ANC visits, postnatal care consultations, and contraceptive use. Hypertension services target increasing the number of diagnoses, and the number of hypertensive patients provided treatment. The selected incentives deliberately emphasize preventative rather than curative services to avoid excessive or unwarranted service provision. Following the midterm review of the project, the indicators and corresponding fees were revised starting in 2017.

The quality bonus is a share of the quantity payment calculated according to the quality score and the matrix presented in Table A1 in the appendix. Facilities that scored less than 55% on the quality checklist did not receive any quality bonus while facilities with high scores can double the quantity payments. The quality score is determined by a quality checklist containing both structural quality and clinical indicators measured through review of registries and medical records. The quality checklist for the RHCs contains 93 separate indicators in ten categories.[3] Health house checklists are a subset, containing 60 indicators in 8 categories. The quality checklist includes indicators directly related to the incentivized services and quality indicators related to non-incentivized services such as diarrhea management and curative care for acute respiratory infections.

3. Methods and Data

Difference-in-differences

We employ a difference-in-differences approach to identify the effects on the PBF program as the districts implementing the program were purposively selected. In March 2012, eight rural districts were chosen to be included in the World Bank supported HSIP project, four in each of the Sughd and Khatlon regions. The objective was to select districts that were near regional averages in terms

[3] The quality checklist categories are: administration, facility infrastructure, hygiene and sanitation, laboratory services, medication management, supplies for emergencies, health management information system, child health, maternal health and noncommunicable diseases.

of maternal and child health outcomes, the capacity of primary health care personnel, geographic location, health care budget, and physical infrastructure of facilities. In one district of Sughd region, a pre-pilot of the PBF program was implemented, and it was subsequently removed from the study sample.[4]

The research team selected control districts in the two regions. The selection was guided by two goals: 1) selection of districts that are similar in terms of the number of facilities and doctors per capita, and 2) achieving the same number of catchment areas in each region. Two control districts were chosen in the Sughd region and seven in the Khatlon region, as can be seen in the map presented in Figure 1.[5] Additional districts were added to the project after 2015, but the assignment of the control districts was maintained.[6]

The Tajikistan Demographic and Health Survey 2012 was used to test whether the parallel trends assumption can be rejected between 2008 to 2012. We created annual indicators for timely antenatal care, receiving at least four antenatal consultations and child vaccination coverage. We failed to reject the parallel trend hypothesis for all indicators. This analysis is presented in Appendix 2. In Appendix Table A2, we present baseline outcome by treatment group and tests for differences between the groups. The tests show that the PBF and control districts were overall similar before the launch of the PBF intervention.

Data

Large-scale household and facility-based surveys were conducted before the launch of the program and after about three years of implementation. Appendix 3 provides more detailed information on the sampling and survey design. The baseline survey was conducted from November 2014 to July 2015.[7] Facility-based surveys were conducted in 108 rural health centers in the seven PBF district

[4] Spitamen district was selected as a pre-pilot district. The remaining treated districts are Mastcho, J. Rasulov and Devashtich in Sughd region and Yavan, Farkhor, J. Balkhi and Kubodiyon in Khatlon region.
[5] The control districts in Sughd region are Asht and Konibodom. The selected districts in Khatlon region are Temurmalik, Vakhsh, Pyanj, Jilikul, A. Jomi, Khuroson and Kumsangir.
[6] We could not identify other interventions that were introduced in the control districts in the duration of the study through other sources.
[7] The facility-based survey was conducted in November to December 2014 prior to the launch of the PBF program. The household survey was conducted from March to July 2015 because of unforeseen delays related to procurement of tools for anthropometric measures and obtaining of administrative data needed for the sample framework. Respondents reported about care received in the preceding two years, so we expect the timing to have negligible

and 108 rural health centers in the nine control districts. The surveys included general facility assessments, health provider interviews, and direct clinical observation of consultations with children under-5 and adults aged 40 and above. When present, a single health house affiliated with each rural health center was randomly selected for assessment.

In the catchment area of each of the chosen health centers, two villages were randomly selected to be included as a sampling unit for the household survey.[8] Within these villages, 4,345 households with women who had live births in the 24 months before the survey were randomly selected for the household survey. In a third of the catchment areas in each district, households with adults aged 40 and above were also selected for a total of 1,668 such households.

A follow-up survey was conducted from March to July 2018 in the same health facilities and the same villages, using the same survey tools. The only change in survey methodology concerned the direct observations. In the baseline survey, less than 30% of the target number of consultations have been observed because few patients arrived in the facilities in the winter months, and care was mainly provided through home visits. Therefore, the follow-up facility-based survey was conducted in the spring and summer months, and the enumerators also observed consultations provided through home visits.

In six rural health centers included in the baseline sample, a follow-up survey was not conducted because they were under renovation, closed, or downgraded to health house status. Those catchment areas were removed from the analysis sample. Table A2 in the appendix presents sample characteristics and some key outcomes from the baseline survey of the analysis sample.

Empirical specification

Our primary empirical specification is the following difference-in-differences model used for estimating the impact of the PBF intervention on outcomes of interest:

$$y_{idt} = \beta_0 + \beta_1 * PBF_d + \beta_2 * Post_t + \beta_3 * PBF_d * Post_t + \gamma X_{idt} + \varepsilon_{idt}.$$

impact on results. If the program had immediate impacts, this would diminish the differences between baseline and follow-up surveys and our estimated impacts.

[8] In catchment areas with a health houses, one of the villages was selected from the list of villages directly served by the health house selected into the sample.

y_{idt} is the outcome for health facility/health provider/consultation/individual i in district d in period t. PBF_d takes value 1 if the observation belongs to a PBF district and zero otherwise. $Post_t$ takes value 1 if the observation is from the follow-up survey and 0 if it is from the baseline survey. X_{idt} is a vector of control variables, including catchment area fixed effects. The assignment into the PBF intervention was done at the district level. Because there are only 16 districts in the sample, clustering standard errors at that level is likely to lead to over-rejection. We, therefore, follow the methodology suggested by Cameron et al. (2008, 2011). We use a two-way standard error clustering by district and time, calculated using a wild-cluster bootstrap-t procedure.

For some outcomes of interest, we cannot employ the difference-in-difference approach. Some variables were not measured, or not measured correctly during the baseline survey. With respect to direct clinical observations, the baseline sample was too small, as described above. To assess the impact of the PBF program on the indicators for which we only have follow-up data, we employ a propensity score weighting approach (Hirano et al. 2003). In appendix 4, we provide more information on the procedure and show that the weighting achieved balance concerning baseline characteristics of the PBF and control groups.

4. Results

Utilization of health services

Panel A of Table 6 presents the results of vaccination and growth monitoring for children under-5. No significant impacts on vaccination rates are observed but coverage at baseline was already relatively high. There is also no measured impact on growth monitoring, despite low levels at baseline. In the follow-up survey, only about 30 percent of mothers reported that the growth of their child was measured in the past 6 months. This rate is surprisingly low given that in the direct observations of curative consultations in children, height and weight was measured in 85 percent of consultations. We cannot rule out that providers were more likely to conduct these measurements when they knew they were being watched. It could also be that the phrasing of the question was not clear to women.

In Panel B of Table 6, we show results on utilization of reproductive health services by recently pregnant women. In the baseline survey, 90 percent of women reported to have received at least

one antenatal consultation during their most recent pregnancy and the rate has increased to 96 percent in the follow-up survey. However, the rates of pregnant women who received the recommended number of antenatal consultations and timely initiation of care are substantially lower. In the follow up survey, 77 percent reported having received their first consultation during the first trimester and 70 reported having received at least four consultations. We do not find significant impacts of the PBF program on either of these outcomes.

We find a significant difference between the treatment arms in receiving postnatal care within the three days after being discharged after delivery. The baseline survey questionnaire did not distinguish between care received while women were still in the maternity ward and care received after they returned to their homes. Therefore, the results presented in the table come from an analysis using propensity score weighting approach. We find an effect of 18 percentage points that is statistically significant at the 99% level.

We do not find any impact of the PBF program on family planning. It is important to keep in mind that our sample is not representative of women of reproductive age. Households were selected if women were pregnant in the preceding two years. Family planning preferences and choices in this sample may be distinct from those of the general population of women of reproductive age. In the baseline survey, 60 percent of the women reported to use a family planning method or to be exclusively breastfeeding an infant. Twenty-seven percent of women reported using a modern method of contraception. Out of the women who wanted to stop or delay fertility, 67 percent were using any family planning method (including exclusive breastfeeding) and 34 percent used a modern method. As can be seen in Panel C of Table 6, the PBF intervention was not successful in reducing this unmet need for family planning. We tested the impact separately for all methods and for modern methods. We also conducted a separate test for the full sample and only those wishing to delay or stop fertility. None of the impact coefficients is statistically significant.

Lastly, we find a significant impact of the PBF intervention on the rate of adults aged 40 and above who have had their blood pressure measured in the preceding 12 months. The baseline rate in the PBF group was 50 percent and increased to 65 percent in the follow up survey. We estimate an impact of three percentage points, statistically significant at the 95% level.

Quality of care

The PBF program had statistically significant impacts on structural quality, facility infrastructure, and availability of equipment and other medical supplies. As can be seen in Table 2, all 18 indicators of structural quality in rural health centers had positive coefficients, with eight statistically significant at the 90% level. The PBF program substantially increased the likelihood that RHCs had designated reception areas and heating during the winter months. Multiple domains of infection prevention and control improved, including the likelihood that the facilities use improved source of water by 17 percentage points ($p = 0.076$), usage of proper biowaste disposal procedures by 54 percentage points ($p = 0.043$), and availability of containers for disposal of needles and sharps in patient consultation rooms by 11 percentage points ($p = 0.028$). The infrastructure of the health houses in the PBF districts also improved. Statistically significant coefficients were detected with respect to having a designated reception area, availability of heating, and availability of toilets. Unlike the RHCs, however, we do not find significant impacts on measures of infection prevention and control.

The PBF program also had statistically significant impacts on the availability of essential drugs and diagnostic test kits at the RHCs. For example, RHCs in the PBF district were more likely to have available units of amoxicillin by 52 percentage points, iron tablets by 21 percentage points, and paracetamol also by 21 percentage points (Table A3, $p<0.05$). The intervention did not significantly improve the availability of family planning products or vaccines. As can be seen in Table A3 in the appendix, there is low availability of these items in the facilities overall. Differences in procurement might explain these impact estimates; drugs and diagnostics kits are directly procured by health facilities whereas vaccines and family planning products may be distributed through donor-funded programs. We also find significant improvements in the presence of general equipment at the health house level but not in the RHCs. For both types of facilities, we find increases in the share of available protocols and medical guidelines which are statistically significant at the 99% level.

In Table 3, we present the PBF effects on facility administration. We find positive impacts on the number of external assessments of staff and facilities. These results are expected as increased external supervision is an activity embedded in the implementation of the PBF program. Within RHCs, we detect a larger impact on external evaluations of staff rather than on the overall facility,

while for the health houses we find a significant impact only on the evaluation of facilities. We do not find impacts on the internal administration of the facilities. The program did not affect the frequency of internal evaluations of staff, the number of staff meetings, or solicitation of patient opinions. We also do not find impacts on turnover as measured by whether staff left the facilities. Changes in staffing were measured over the 12 months preceding the follow-up survey. As the program had been ongoing for more than three years at the time of the follow-up survey, we cannot rule out impacts on staff retention earlier in the implementation period.

Next, we turn to examine the impacts of the PBF program on outcomes related to the health providers in the primary health facilities (Table 4). The analysis was conducted on a pooled sample of providers at the RHCs and HHs. The PBF program increased the monthly income of providers by 438 Tajik Somonis (about 48.6 USD at the time of the follow-up survey) ($p<0.001$). This difference corresponds to 62 percent of the income of the control group providers. We do not find an impact on self-reported absenteeism or number of hours worked. Providers in PBF facilities reported an average increase of 4.9 patients seen in the past day ($p = 0.065$).

Providers in the PBF districts reported significantly higher satisfaction. In the baseline survey, providers reported high levels of satisfaction with respect to the facility management, and relationship with the staff at the facility, district and ministry level. However, providers were mostly unsatisfied with respect to the physical condition of the facilities and the equipment and supplies available (Table A4 in the appendix). Of a list of 13 aspects related to their work and perception of the facility, providers in the PBF facilities were especially more likely to report satisfaction in the follow-up survey with respect to availability of medicine, equipment and other supplies in the facilities.

In Panel B of Table 4, we present impacts of the intervention on provider competence in case management of scenarios for children and adults, measured by clinical vignettes. For each of the 6 standardized scenarios, we coded whether the provider correctly diagnosed the case, the proportion of a recommended set of questions asked during clinical history taking, and the proportion of recommended examination procedures they would conduct. The definitions of these variables are presented in Appendix 5. We find an overall improvement in provider competence as a result of the project. Providers were more likely to correctly diagnose a case of high cardiovascular risk by 27 percentage points. For 4 of the 6 vignettes, the proportion of

recommended history taking items reported by providers was higher and statistically significant at least at the 90% level. With respect to recommended exam procedures, the rate reported by the PBF providers was higher in 5 of the 6 cases and statistically significant at least at the 90% level.

It is important to note that we cannot identify the channel through which provider competence improved. Because providers in the PBF districts were retrained in family medicine, these results may represent the impact of additional training rather than an effective of PBF incentives. Nevertheless, there are several mechanisms through which PBF might theoretically improve provider competence. As reported above, facilities were more likely to have medical protocols available and the facilities had more frequent supervision visits. Providers may have had an indirect incentive to improve their knowledge of the protocols because the medical records and registers are more likely to be reviewed, or a direct incentive to increase the quantity of services provided by improving service quality and attracting additional patients. Overall, we can only conclude that the combination of training and PBF was successful in improving provider competence. However, overall performance on the clinical vignettes was poor and there remains much room for improvement.

In Table 5, we present results on content of care measured through direct clinical observations of patient-provider interactions. These interactions were curative care consultations of children under-5 and adults aged 40 and above. The analysis reported in the table uses propensity score weighting analysis to account for low sample size in the baseline survey. Analysis of content of care is constrained due to variability in chief complaints and clinical presentation of patients. We therefore focus the analysis on the diagnosis process and a set of actions which should be performed in every consultation.

For the adult consultations, we do not find overall significant differences in content of care. Providers in the PBF facilities were statistically significantly more likely to inquire about patient history of consumption of alcohol and cigarettes, physical activity, diet and diabetes status (Appendix Table A5). However, once averaging over the 18 items which should be included in every patient history, the coefficient is not statistically significant ($p = 0.149$). The 10-year cardiovascular risk was calculated in 38 percent of consultations. In only 4 percent of consultations the cardiovascular risk was correctly calculated given the age, gender, diabetes, and smoking statuses of the patient. Measurement of blood pressure measurement is high overall, conducted in

90% of the observed consultations. Measurement is higher in the PBF group by 8 percentage points ($p = 0.107$).

For the child consultations, we find some significant differences in the conduct of physical examinations. Overall, weight and height of the children was measured in 85 percent of consultations. In the PBF facilities, these measurements were more likely to be conducted by 12 percentage points ($p = 0.037$). For the 13 core physical examination procedures, the share of completed procedures were higher by eleven percentage points in the PBF group. PBF providers were statistically significantly more likely to observe the children's ears, perform skin pinch tests for dehydration, check for lethargy, and count radial pulse. We do not find significant impacts on duration of these consultations, whether vaccination history was checked, and whether providers washed their hands prior to conducting the physical examinations.

Results from the analysis of the data collected at the health facilities show that the PBF project had a significant impact on the quality of care. As a result of the program, availability of equipment and medical supplies increased. We detect improvements in facility infrastructure and procedures related to infection prevention and control. We find positive impacts on provider competence and even some improvements in the content of curative care of children under-5. Although we find these positive impacts, substantive gaps in quality of care remain. For example, many providers could not correctly diagnose the clinical vignette cases and many providers did not properly wash their hands before conducting physical exams. There may be further improvements in quality with time, but additional intervention targeting quality of health may be needed. Next, we turn to analysis of the household data to evaluate whether the project was successful in increasing utilization of health services.

Health outcomes and health-related behaviors

In Table 7, we present results on the impact of the PBF intervention on health outcomes and health behaviors. As seen in Panel A, we do not find impacts on anthropometric measures of children. We also do not find a significant impact on the rate of adults over 40 with elevated blood pressure (Panel B). Adults in the PBF group were more likely to self-report being hypertensive prior to the measurement conducted by the survey teams ($p = 0.063$). This is likely to be a result of the increased measurement of blood pressure in the PBF group. Among adults over-40 with self-reported hypertension, we do not find any impact on rate of medication prescription or whether an

individual reported to have taken the medication in the preceding 24 hours, conditional on having such a prescription.

In the follow-up survey only, respondents were asked whether they use the services offered the RHC or HH serving the catchment area in which their village is located. Adults aged 40 and above in the PBF districts were significantly more likely to report using the service at the local primary level. With the propensity score weighting approach, we estimate an impact of 8 percentage points statistically significant at the 95% level. Without the weighting, the rate reporting using the local services is 85 percent versus 78 percent in the control districts. Among the sample of recently pregnant women, 92 of women in the PBF districts and 88 percent of women in control districts reported using the local primary services. With the propensity score weighting approach, the estimated impact is 6.4 percentage points ($p = 0.133$, presented in Panel C of Table 7).

Another indicator of interest is whether the PBF program affected bypassing of the local primary level facilities. In the baseline survey 28 percent of recently pregnant women reported receiving antenatal care at higher levels. The rate of women bypassing the local primary health facilities at the follow-up was 30 percent and as can be seen in Panel C of Table 7, the PBF intervention did not affect this bypassing behavior. It is important to keep in mind that while there were significant improvements in the quality of the RHCs, most of them still do not provide laboratory testing and even fewer have physicians or nurses with specialization in obstetrics.

Though the impact of the PBF program on utilization was moderate, we find that the program improved the perception of the population regarding the quality of care in the local RHCs. Perceptions regarding the quality of the health facilities was collected in the follow-up survey only and therefore the results in Table 8 are based on a propensity score weighting analysis. Overall, reported satisfaction by the population is very high and there is little variation. Over 90% of respondents indicated positive agreement for all statements about the local RHC, with the exception of whether the RHC had equipment needed to provide high quality services. Despite this almost universally reported satisfaction, we find some statistically significant impacts of the PBF project. Women with recent pregnancies in the PBF districts are more likely to agree that the staff is competent ($p = 0.070$), that the facility is in good physical state ($p = 0.042$), that the staff works closely with the community ($p= 0.040$) and that the staff listens to the opinions of the community ($p = 0.023$). When asked about changes over the three years of project implementation, women in

the PBF group were more likely to report improvements in attitude of health workers (p=0.033), quality of health services (p=0.030), and collaboration between the community and the health facility (p = 0.067). The adults 40 years and above in the PBF districts were significantly more likely to agree that the staff is competent (p = 0.002) and listens to the opinions of the community (p = 0.098). They were also more likely to report improvement in the attitude of health workers during the three preceding years (p = 0.015).

These results suggest that the population noticed the positive changes in the health facilities. Interestingly, both women with recent pregnancies and adults aged over 40 were more likely to agree that the staff listens to the community and that the attitude of health workers improved. In addition, women in the PBF group had better perceptions about the collaboration between the facility and the community. These findings might provide suggestive evidence that the health workers exerted more effort in their interaction with the population, even if these efforts resulted in limited changes in health seeking behaviors.

5. Conclusion

This study contributes to the global knowledge on PBF, which is mostly based on studies form Sub-Saharan Africa, by presenting the first evidence from a Central Asian country. Unlike most of the programs evaluated in Africa, the Tajikistan PBF pilot incentivizes performance solely at primary level facilities in rural areas, which do not offer inpatient services such as labor and delivery. It is also unique in incentivizing services related to noncommunicable diseases in addition to a package of reproductive, maternal and child health services. More generally, the political, economic and epidemiological contexts in Tajikistan are different. For example, the baseline health services coverage rates in Tajikistan were higher than those reported in the other studies.

Despite these differences in the contexts in which the programs were implemented, the results from the impact evaluation of PBF in Tajikistan are overall in line with those in previous studies. We find positive impacts on a range of measures of quality of care. Similar to the studies conducted in Rwanda, Afghanistan and Zambia, this study also shows that PBF can be effective not only in improving structural quality (e.g. infrastructure and equipment) but can also have positive impact on the content of care (Basinga et al. 2011; Engineer et al. 2016; Friedman et al. 2016). Consistent with most of the other studies, however, we only find impacts on utilization of few of the set of

incentivized services. The Tajikistan PBF pilot increased by 18 percentage points the rate of women who received timely postnatal care. In addition, the program improved by three percentage points the rate of adults aged 40 and above who had their blood pressure measured in the preceding year. Coverage rates of family planning, child vaccination and growth monitoring and the timing and number of antenatal consultations were not impacted.

Several mechanisms linked to the PBF reform can explain the positive impacts in quality of care. First, providers might have responded to the financial incentives by exerting more effort to improve quality and engage with the communities. Second, the program ensured resources reached the rural facilities, and these facilities decided how to invest the 30% of the PBF bonuses in their infrastructure. Third, providers may have been motivated by the increased payment and increased satisfaction with their working conditions. Fourth, the program introduced a multi-layered system of supervision and verification. Facilities were monitored more closely, and the PBF tools and action plans may have provided quality improvement guidance towards better performance. Lastly, many providers were retrained in the context of the PBF project. Some of the improvements, especially with respect to clinical knowledge and quality, might be directly linked to these trainings.

The more modest impacts on health care utilization might be explained by several factors. One explanation may be that behavioral change might take time to materialize. Although we did not find increased utilization of many of the targeted services, the population in the PBF districts reported higher satisfaction with the local primary care facilities and were more likely to report general use of their services. These outcomes might represent changes in the right direction and that with more time, utilization of the targeted services will increase. It is also important to note that for some of the indicators, the overall coverage was already high in baseline. Close to 90% of women received any ANC and the same proportion of children aged 12-23 months received all basic vaccinations. With respect to some indicators, we find overall positive trends even if we do not find impact of the PBF pilot when we compare with the control districts. Between the baseline and follow-up surveys, the rate of women in the control districts who received any antenatal consultation increased from 85% to 95%. The rate of women who initiated their ANC during the first trimester increased from 57% to 74%. The relatively high baseline coverage rates and the

overall positive trends might have limited the scope of the supply-side incentives to impact utilization.

Another explanation could be that the PBF pilot covered only primary health facilities. While the quality of these facilities improved, they still offer a narrower package of services relative to higher-level facilities. For example, only about a fifth of RHCs offer laboratory services. Pharmacies are often located near the higher-level facilities in areas with higher population density. Patients who are required to travel for prescription drugs may seek care at the nearby higher-level facilities and take advantage of the additional services. For example, about a third of women in our household sample used higher-level facilities for receiving ANC and the rate did not significantly change between baseline and follow-up. It could be that while the quality of HHs and RHCs improved, the choice of going to a higher-level facility is still more attractive for many. It could be that a more comprehensive intervention, that covers also higher levels of care, is needed to achieve stronger impacts on utilization.

To improve health outcomes, health system reforms should not only increase health service coverage but instead strive to improve *effective* coverage. There is a need to improve the rate of individuals receiving health services in a timely manner and at a level of quality necessary to obtain the desired potential health gains (Shengelia et al. 2005). Supply-side incentives in the form of PBF are effective in increasing the quality of care and therefore result in more effective coverage. However, the results in Tajikistan and elsewhere suggest that demand-side barriers might constrain the possible coverage gains through supply-side PBF incentives. Therefore, progress on effective coverage is likely to require that PBF programs are introduced together with interventions that reduce demand-side barriers. In Rwanda, demand-side in-kind incentives were effective in improving rates of timely antenatal and postnatal care although health facilities were already incentivized to improve these indicators through a PBF program (Shapira et al. 2018). There is a need to pilot more programs that combine supply-side and demand-side interventions in order to better understand the complementarities and synergies these interventions might have in achieving better population health outcomes.

References

Basinga, P., P. J. Gertler, A. Binagwaho, A. L. Soucat, J. Sturdy and C. M. Vermeersch (2011). "Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation." Lancet 377(9775): 1421-1428.

Bonfrer, I., R. Soeters, E. Van de Poel, O. Basenya, G. Longin, F. van de Looij and E. van Doorslaer (2014a). "Introduction of performance-based financing in burundi was associated with improvements in care and quality." Health Affairs 33(12): 2179-2187.

Bonfrer, I., E. Van de Poel and E. Van Doorslaer (2014b). "The effects of performance incentives on the utilization and quality of maternal and child care in Burundi." Social Science & Medicine. 123: 96-104.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*(3), 414-427.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, *29*(2), 238-249.

Celhay, P. A., P. J. Gertler, P. Giovagnoli and C. M. J. Vermeersch (2015). Long-run effects of temporary incentives on medical care productivity. The World Bank, Policy Research Working Paper Series, 7348.

Das, Jishnu, Jeffrey Hammer and Kenneth Leonard. 2008. "The Quality of Medical Advice in Low Income Countries". Journal of Economic Perspectives, 22(2): 93–114

Das, J., and J. Hammer. 2014. "Quality of Primary Care in Low-Income Countries: Facts and Economics." *Annual Review of Economics* 6 (1): 525–53.

de Walque, D., P. J. Gertler, S. Bautista-Arredondo, A. Kwan, C. Vermeersch, J. de Dieu Bizimana, A. Binagwaho and J. Condo (2015). "Using provider performance incentives to increase HIV testing and counseling services in Rwanda." Journal of Health Economics: 40: 1-9.

de Walque, Damien, Paul Jacob Robyn, Hamadou Saidou, Gaston Sorgho and Maria Steenland. 2017. Looking into the Performance-Based Financing Black Box. Evidence from an Impact Evaluation in the Health Sector in Cameroon. World Bank Policy Research Working Paper # 8162.

Engineer, C. Y., Dale, E., Agarwal, A., Agarwal, A., Alonge, O., Edward, A., ... & Peters, D. H. (2016). Effectiveness of a pay-for-performance intervention to improve maternal and child health services in Afghanistan: a cluster-randomized trial. *International journal of epidemiology*, *45*(2), 451-459.

Friedman, J. and R. M. Scheffler (2016). Pay for Performance in Health Systems: Theory, Evidence and Case Studies. World Scientific Handbook of Global Health Economics and Public Policy. R. M. Scheffler. Singapore, World Scientific Press

Friedman, Jed; Qamruddin, Jumana N.; Chansa, Collins; Das, Ashis Kumar. 2016. Impact evaluation of Zambia's health results-based financing pilot project. Washington, D.C. : World Bank Group.

Fritsche, G. B., Soeters, R., & Meessen, B. (2014). *Performance-based financing toolkit*. World Bank Publications.

Gertler, P., P. Giovagnoli and S. Martinez (2014). Rewarding provider performance to enable a healthy start to life: evidence from Argentina's Plan Nacer. The World Bank, Policy Research Working Paper Series, 6884.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161-1189.

Huillery, E., & Seban, J. (2019). Financial Incentives, Efforts, and Performances in the Health Sector: Experimental Evidence from the Democratic Republic of Congo.

Kandpal E. *Completed impact evaluations and emerging lessons from the health results innovation trust fund learning Portfolio*: World Bank, 2017. https://www. rbfhealth. org/ sites/ rbf/ files/ IE% 20and% 20emerging% 20lessons_ Eeshani% 20Kandpal. pdf \

Kruk, M. E., A. D. Gage, N. T. Joseph, G. Danaei, S. García-Saisó, and J. A. Salomon. 2018. "Mortality Due to Low-Quality Health Systems in the Universal Health Coverage Era: A Systematic Analysis of Amenable Deaths in 137 Countries." *The Lancet*.

Miller, G., & Babiarz, K. S. (2013). *Pay-for-performance incentives in low-and middle-income country health programs* (No. w18932). National Bureau of Economic Research.

Peabody, J. W., R. Shimkhada, S. Quimbo, O. Solon, X. Javier and C. McCulloch (2014). "The impact of performance incentives on child health outcomes: results from a cluster randomized controlled trial in the Philippines." Health Policy and Planning 29(5): 615-621.

Shapira, Gil, Ina Kalisa, Jeanine Condo, James Humuza, Cathy Mugeni, Denis Nkunda, Jeanette Walldorf. (2018). "Going beyond incentivizing formal health providers: Evidence from the Rwanda Community Performance-Based Financing program." *Health Economics* 1-20.

Shengelia, B., A. Tandon, O. B. Adams and C. J. L. Murray (2005). "Access, utilization, quality, and effective coverage: An integrated conceptual framework and measurement strategy." Social Science & Medicine 61(1): 97-109.

Sun, X., X. Liu, Q. Sun, W. Yip, A. Wagstaff and Q. Meng (2016). "The Impact of a Pay-for-Performance Scheme on Prescription Quality in Rural China." Health Economics: 25(6):706-22

Van de Poel, E., G. Flores, P. Ir and O. O'Donnell (2016). "Impact of Performance-Based Financing in a Low-Resource Setting: A Decade of Experience in Cambodia." Health Economics: 25(6): 688-705

Wagstaff, Adam, Caryn Bredenkamp, and Leander R. Buisman. 2014. Progress on Global Health Goals: Are the Poor Being Left Behind? *The World Bank Research Observer*; 29:137-162.

Witter, S., Fretheim, A., Kessy, F. L., & Lindahl, A. K. (2012). Paying for performance to improve the delivery of health interventions in low-and middle-income countries. *Cochrane Database Systematic Reviews*, *2*(2), CD007899.

Yip, W., T. Powell-Jackson, W. Chen, M. Hu, E. Fe, M. Hu, W. Jian, M. Lu, W. Han and W. C. Hsiao (2014). "Capitation combined with pay-for-performance improves antibiotic prescribing practices in rural China." Health Affairs 33(3): 502-510.

Tables and Figure

Table 1: PBF Quantity Indicators and corresponding financial reward

	January 2015 to December 2016		From January 2017	
	Indicator	**Fee [a]**	**Indicator**	**Fee [a]**
Child vaccination	Fully vaccinated children <13 months	41	Same	55
Nutrition	Detection of malnourished children <5	9	Growth monitoring for children < 2	2
	Treatment of malnourished children <5 years	27	Improvement in the status of malnourished children <5 years	31
Antenatal care	Initiation in first 12 weeks of pregnancy	18	Same	27
	At least 4 visits with the last one in the last 3 weeks of pregnancy	24	At least 4 visits with the last one within 2 weeks of expected delivery	27
Postnatal care	Consultation within 7 days after delivery	14	Postnatal home visit within 3 days after discharge from delivery facility	16
Family planning	Women 15-49 using modern FP methods	1	New users of modern FP methods	3
			Old users receiving additional pills or injection	3
Hypertension	Detection of hypertensive adults > 18	9	Same	10
	Treated hypertensive adults	3.5	Same	4

[a] Fee for service in Tajik Somoni. In January 2015, 1 USD equaled 5.1 Tajik Somoni. By the time of the indicators' change in January 2017, the conversion rate changed to 1 USD per 7.9 Tajik somoni.

Figure 1: Map of PBF and Control Districts

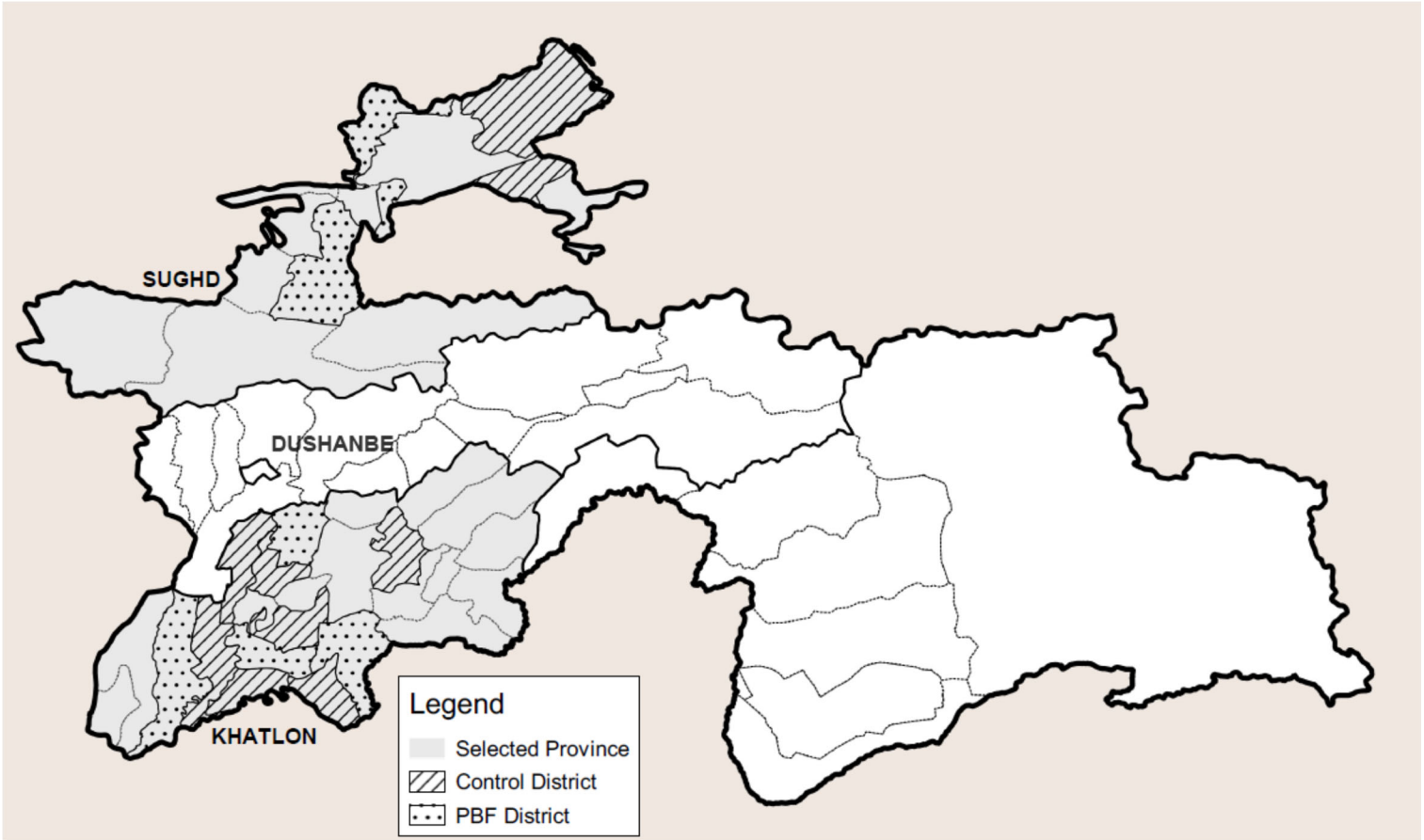

Table 2: Impact of PBF on Structural Quality Indicators

	Rural Health Centers					Health House				
Variable	**Baseline Mean in PBF districts**	**Trend**	**PBF impact**	**p-value**	**N**	**Baseline Mean in PBF districts**	**Trend**	**PBF impact**	**p-value**	**N**
Designated reception area	0.811	0.046	0.141***	0.002	420	0.714	0.134	0.143**	0.035	278
Heating	0.594	0.070	0.329**	0.041	420	0.486	0.275	0.314**	0.037	278
Toilets	0.802	0.021	0.075	0.874	420	0.414	0.129	0.240*	0.082	278
Separate gender toilets	0.358	-0.022	0.281	0.483	420	0.100	-0.012	0.113	0.229	278
Piped water into facility plot [a]	0.085	0.203	0.177	0.123	420	0.029	0.220	0.086	0.234	278
Improved source of water	0.717	0.075	0.172*	0.076	420	0.700	0.264	-0.226	0.776	278
Water in consultation rooms	0.420	0.153	0.179	0.523	282	0.455	0.084	0.135	0.422	130
Functional incinerator	0.858	-0.094	0.219	0.314	420	0.871	0.099	-0.042	0.506	278
Proper sterilization procedure [b]	0.189	0.172	0.122	0.114	420	0.086	0.089	0.023	0.314	278
Proper decontamination procedure [c]	0.189	0.033	0.169	0.275	420	0.229	0.186	-0.101	0.829	278
Proper biowaste disposal procedure [d]	0.585	-0.368	0.542**	0.043	420	0.571	-0.162	0.308	0.282	278
Containers for needles/sharps [e]	0.449	0.272	0.113**	0.028	282	0.182	0.287	0.166	0.746	130
Proportion of equipment items [f]	0.739	-0.010	0.166	0.157	420	0.584	-0.057	0.206***	0.001	278
Proportion of essential drugs [f]	0.257	0.388	0.279***	0.000	420					
Proportion of family planning products [f]	0.249	0.228	0.036	0.758	420					
Proportion of vaccines [f]	0.154	0.189	0.057	0.595	420					
Proportion of diagnostic tests [f]	0.109	0.005	0.220***	0.000	420					
Proportion of protocols and clinical guidelines [f]	0.565	0.086	0.254***	0.000	420	0.408	0.100	0.261***	0.001	278

Note: data from health facility assessments. All regressions are the difference-in-differences specification with facility fixed effects and controlling for exposure to the collaborative quality improvement and citizen scorecards interventions. Standard errors are clustered at the district and survey round level with wild bootstrapping. Availability of consumables was not measured at the health houses.

[a] Improved source of water includes piped water, public taps, tube wells, protected dug wells, protected springs, rainwater and bottled water.

[b] Proper sterilization procedure includes autoclaving, boiling, steam sterilization, chemical sterilization, and outsourcing

[c] Proper decontamination procedure is scrubbing, or cleaning followed by use of a disinfectant

[d] Proper biowaste disposal method includes burning or outsourcing

[e] Because of error in the questionnaire skip pattern, the variable is missing in the baseline survey data for many facilities.

[f] Availability was defined as existence of at least one unit, confirmed by direct observations of the survey enumerators. The items included in the calculation of availability of equipment, and consumables are listed in appendix Table A3.

Level of statistical significance: *p<0.10; **p<0.05; ***p<0.01

Table 3: Impact of PBF on Facility Administration

	Rural Health Centers					Health House				
Variable	**Baseline Mean in PBF districts**	**Trend**	**PBF impact**	**p-value**	**N**	**Baseline Mean in PBF districts**	**Trend**	**PBF impact**	**p-value**	**N**
Number of evaluations in past 12 months:										
External evaluations of facility	5.9	-1.8	0.1***	0.001	420	3.6	-0.5	0.6**	0.014	276
External evaluations of staff	3.6	-1.3	4.4***	0.000	420	3.8	0.5	0.8	0.156	274
Internal evaluations of staff	6.7	-4.0	8.5	0.265	418	5.9	3.6	1.3	0.350	262
Number of staff meetings in past 3 months	9.8	-1.6	4.2	0.518	420					
Solicitation of patient opinion	0.660	-0.078	0.246	0.425	420	0.371	-0.143	0.355	0.385	278
Staff leaving the facility in past 12 months:										
Number left	0.6	0.1	-0.3	0.782	420	0.4	0.2	-0.1	0.605	278
Any left	0.302	-0.091	0.008	0.939	420	0.129	0.138	0.091	0.695	278

Note: data from health facility assessments. All regressions are the difference-in-differences specification with facility fixed effects and controlling for exposure to the collaborative quality improvement and citizen scorecards interventions. Standard errors are clustered at the district and survey round level with wild bootstrapping. Level of statistical significance: *p<0.10; **p<0.05; ***p<0.01

Table 4: Impact of PBF on health provider outcomes

Variable	Baseline Mean in PBF districts	Trend	PBF impact	p-value	N
Panel A: General provider outcomes					
Monthly income (Salary + PBF bonus)	635.0	69.6	437.6***	0.000	2084
Number of days absent in past 30 days	0.3	0.0	0.3	0.361	2126
Number of hours worked in past week	22.4	-0.2	5.3	0.878	2126
Number of patients seen in past day	3.8	-2.2	4.9*	0.065	2125
Satisfaction score [a]	0.486	0.015	0.183**	0.019	2126
Panel B: Providers knowledge measured by clinical vignettes					
Severe dehydration vignette:					
Correct diagnosis	0.245	0.081	0.260	0.288	2095
Proportion of history taking	0.442	0.020	0.096*	0.093	2095
Proportion of exam procedures	0.516	-0.077	-0.005	0.932	2095
Pneumonia vignette:					
Correct diagnosis	0.355	0.030	0.043	0.672	2095
Proportion of history taking	0.421	0.021	0.088	0.107	2095
Proportion of exam procedures	0.345	-0.034	0.037**	0.040	2095
Acute respiratory infection vignette:					
Correct diagnosis	0.173	0.194	0.066	0.383	2094
Proportion of history taking	0.556	0.091	0.023**	0.020	2095
Proportion of exam procedures	0.377	-0.017	0.060*	0.056	2095
Malnutrition/Anemia vignette:					
Correct diagnosis	0.640	0.004	-0.003	0.608	2094
Proportion of history taking	0.470	0.024	0.087	0.119	2095
Proportion of exam procedures	0.316	-0.097	0.014**	0.034	2095
Moderate cardiovascular risk vignette:					
Correct diagnosis	0.315	0.017	0.135	0.395	2095
Proportion of history taking	0.336	-0.022	0.117*	0.066	2095
Proportion of exam procedures	0.373	-0.088	0.132**	0.040	2095
High cardiovascular risk vignette:					
Correct diagnosis	0.230	0.034	0.265**	0.021	2095
Proportion of history taking	0.385	-0.016	0.089*	0.079	2095
Proportion of exam procedures	0.371	-0.112	0.142**	0.011	2095

Note: data from interviews of health providers in both rural health centers and health houses. All regressions are the difference-in-differences specification with facility fixed effects. Standard errors are clustered at the district and survey round level with wild bootstrapping. All regressions include controls for facility type, exposure to the collaborative quality improvement and citizen scorecards interventions, age and gender of providers, position, a binary indicator for experience of over 20 years and an indicator for being born in the district. Standard errors are clustered at the district and survey round level with wild bootstrapping.

[a] Proportion of categories about which providers reported to be satisfied. The individual categories are listed in Appendix Table A4.

[b] The definition of all the indicators related to the clinical vignettes indicators are specified in Appendix 5.

Level of statistical significance: *p<0.10; **p<0.05; ***p<0.01

Table 5: Impact of PBF on content of care

Variable	Overall mean	PBF impact	p-value	N
Panel A: Adult consultations [a]				
Proportion of core clinical history items asked	0.397	0.102	0.149	2,614
CVD risk score calculated	0.384	-0.058	0.659	2,614
CVD risk score properly calculated	0.038	0.021	0.545	2,614
Clinician washed their hands before starting the exam	0.609	-0.148	0.396	2,614
Blood pressure measured	0.898	0.075	0.107	2,614
Average consultation time (minutes)	12.75	-2.275**	0.020	2,612
Panel B: Child consultations [b]				
Vaccination history checked	0.501	0.097	0.280	2,582
Clinician washed their hands before starting the exam	0.644	-0.091	0.551	2,582
Height and weight measured	0.847	0.120**	0.037	2,582
Proportion of core physical exam activities completed	0.341	0.107**	0.040	2,582
Average consultation time (minutes)	12.59	0.474	0.744	2,526

Note: data from direct clinical observations conducted at the rural health centers. The PBF impact is estimated using a propensity score weighting specification as only follow-up observations are being included in the analysis. Standard errors are clustered at the district level.

[a] Data from consultations of adults aged 40 and above. Regressions include controls for facility's exposure to the collaborative quality improvement and citizen scorecards interventions, Region, age and gender of patient, gender age and position of provider, and a binary indicator for experience of over 20 years.

[b] Data from consultations of children under-5. Regressions include controls for facility's exposure to the collaborative quality improvement and citizen scorecards interventions, Region, gender age and position of provider, and a binary indicator for experience of over 20 years.

Level of statistical significance: *p<0.10; **p<0.05; ***p<0.01

Table 6: Impact of PBF on coverage of incentivized health services

Variable	Baseline Mean in PBF districts	Trend	PBF impact	p-value	N
Panel A: Child health services [a]					
Under-5: Growth monitoring in 6 months	0.188	0.130	-0.042	0.424	13165
12-23 months: all basic vaccinations	0.853	-0.081	0.074	0.431	4277
12-23 months: any basic vaccination	0.894	-0.004	0.039	0.781	4277
24-35 months: All basic vaccination	0.803	-0.045	0.099	0.554	2550
24-35 months: Measles, Mumps and Rubella vaccine	0.805	-0.059	0.115	0.617	2403
24-35 months: any vaccination	0.892	0.013	0.062	0.546	2550
Panel B: Maternal health services [b]					
At least 4 antenatal consultations	0.526	0.058	0.007	0.648	6068
Timely initiation of antenatal care [e]	0.726	0.127	-0.098	0.940	5682
Postnatal consultation within 3 days after discharge from maternity [f]			0.18***	0.001	3565
Panel C: Family planning [c]					
All women: any method	0.633	-0.006	-0.046	0.930	9343
All women: Modern FP method [g]	0.290	0.025	-0.047	0.956	9343
Want to delay or stop: any method	0.698	0.006	-0.063	0.487	5958
Want to delay or stop: Modern method [g]	0.375	0.018	-0.058	0.862	5958
Panel D: Blood pressure measurement for adults aged 40 and above [d]					
Blood pressure measurement in past year	0.497	0.122	0.032**	0.034	9885

Note: data from household surveys. Unless otherwise noted, regressions are of the difference-in-differences specification with facility fixed effects. Standard errors are clustered at the district and survey round level with wild bootstrapping. All regressions include controls for catchment areas' exposure to the collaborative quality improvement and citizen scorecards interventions, region, household wealth quintile, and age.

[a] Sample of children under-5. Regressions include a control for the gender of the child.

[b] Sample of women with a live birth in the preceding two years. The indicators relate to care received during the most recent pregnancy. Regressions include controls for employment status, and education level.

[c] Sample of women age 15-49 who were pregnant in the preceding two years.

[d] Sample of adults 40 years and above. Regressions include controls for employment status, education level, gender, and marital status.

[e] First antenatal consultation within the first three months of the pregnancy.

[f] The PBF impact is estimated using a propensity score weighting specification as the baseline survey did not distinguish between postnatal care provided before and after women were discharged from the facilities where they gave birth. Standard errors are clustered at the district level.

[g] Modern methods include: sterilization, IUD, injectables, implants, pills, condoms, and diaphragms.

Level of statistical significance: *p<0.10; **p<0.05; ***p<0.01

Table 7: Impact of PBF on health outcomes and health behaviors

Variable	Baseline Mean in PBF districts	Trend	PBF impact	p-value	N
Panel A: Children under 5 [a]					
Weight-for-age					
Mean Z-score	-0.421	-0.396	0.028	0.990	12290
Percentage below -2 SD	0.181	-0.001	-0.007	0.901	12290
Percentage below -3 SD	0.092	-0.015	-0.017	0.598	12290
Weight-for-height					
Mean Z-score	0.183	-0.237	0.117	0.974	11236
Percentage below -2 SD	0.149	0.002	-0.033	0.652	11373
Percentage below -3 SD	0.082	-0.008	-0.018	0.659	11373
Height-for-age					
Mean Z-score	-1.106	-0.014	-0.159	0.891	12460
Percentage below -2 SD	0.283	0.029	0.016	0.989	12803
Percentage below -3 SD	0.137	0.028	-0.012	0.377	12803
Panel B: Adults aged 40 and above [b]					
Elevated blood pressure [c]	0.363	0.020	0.058	0.454	10535
Self-reported hypertension	0.264	0.028	0.063*	0.071	15450
Prescribed medication for hypertension	0.893	-0.019	0.044	0.128	3741
Took medication in past 24 hours	0.676	-0.034	0.039	0.292	3714
Use local RHC or HH [d]			0.076**	0.041	2002
Panel C: Recently pregnant women [c]					
Use local RHC of HH [d]			0.064	0.133	2879
ANC at the primary level [e]	0.769	0.029	-0.132	0.995	5,590

Note: data from household surveys. Unless otherwise noted, regressions are of the difference-in-differences specification with facility fixed effects. Standard errors are clustered at the district and survey round level with wild bootstrapping. All regressions include controls for catchment areas' exposure to the collaborative quality improvement and citizen scorecards interventions, region, household wealth quintile, and age.

[a] Sample of children under-5. Regressions include a control for the gender of the child.

[b] Sample of adults 40 years and above. Regressions include controls for employment status, education level, gender, and marital status.

[c] Sample of women with a live birth in the preceding two years. The indicators relate to care received during the most recent pregnancy. Regressions include controls for employment status, and education level.

[d] Respondents were asked whether they generally use the services offered by the staff of the local primary care facilities officially serving their catchment areas. The question was only included in the follow-up survey and the approach used for this outcome is the propensity score weighting specification. Standard errors are clustered at the district level.

[e] Women reported to receive antenatal care at a rural health center, health house or through home visits.

Level of statistical significance: *p<0.10; **p<0.05; ***p<0.01

Table 8: Impact of PBF on community satisfaction

Variable	Overall mean	PBF impact	p-value	N
Panel A: women with pregnancies in the preceding two years				
The staff is welcoming and respectful	0.988	0.005	0.592	3,033
The staff at the facility is competent	0.979	0.017*	0.070	2,955
The facility has the equipment needed to provide high quality health services	0.867	0.036	0.412	2,583
The facility is in good physical state to provide high quality health services	0.915	0.058**	0.042	2,669
The facility staff works closely with the community on health matters	0.970	0.040**	0.017	2,859
The staff listens to the opinions of the community	0.977	0.023**	0.049	2,842
Changes in the past three years: [c]				
Improved health facility infrastructure	0.949	0.027	0.325	2,168
Improved attitude of health workers	0.975	0.021**	0.033	2,442
Improved quality of health services	0.972	0.030*	0.069	2,433
Improved collaboration between community and health facility	0.978	0.018*	0.067	2,390
Panel B: Adults 40 years and above				
The staff is welcoming and respectful	0.979	0.008	0.221	2,024
The staff at the facility is competent	0.972	0.036***	0.002	1,956
The facility has the equipment needed to provide high quality health services	0.855	0.029	0.436	1,706
The facility is in good physical state to provide high quality health services	0.903	0.010	0.737	1,740
The facility staff works closely with the community on health matters	0.964	0.001	0.935	1,894
The staff listens to the opinions of the community	0.971	0.016*	0.098	1,889
Changes in the past three years: [c]				
Improved health facility infrastructure	0.939	0.007	0.640	1,496
Improved attitude of health workers	0.963	0.039**	0.015	1,656
Improved quality of health services	0.963	0.009	0.406	1,650
Improved collaboration between community and health facility	0.974	0.007	0.270	1,632

Note: data from household surveys. The PBF impact is estimated using a propensity score weighting specification as these data were only collected during the follow-up. Standard errors are clustered at the district level. All regressions include controls for age, gender, education level, employment status, household wealth quintile, province, number of health houses in the catchment area, an indicator for having any health houses in the catchment area, catchment population terciles, and an indicator for the catchment area's RHC being more than 10km away from the district hospital.
Level of statistical significance: *p<0.10; **p<0.05; ***p<0.01

Appendix 1: Appendix tables

Table A1: Quality bonus calculation matrix

Quality score	Quality bonus (% of quantity payment)	
	January 2015 to December 2016	**From January 2017**
90%+	150%	100%
85% - <90%	125%	90%
80% - <85%	100%	70%
75% - <80%	75%	50%
70% - <75%	50%	35%
65% - <70%	30%	20%
60% - <65%	20%	10%
55% - <60%	10%	5%
<55%	no bonus	no bonus

Note: the quality bonus is calculated as a percentage of the quantity bonus and is given in addition to the quantity bonus.

Table A2: Sample characteristics at baseline

	Variable	Means by treatment group		p-value for testing difference between PBF and control [a]	N
		PBF	Control		
Rural Health Centers	Khatlon region	0.70	0.68	0.95	210
	Any affiliated health houses	0.69	0.77	0.44	210
	Number of health houses	1.60	2.49	0.08*	210
	Catchment population	5030	5612	0.57	201
	Number of physicians	1.56	1.34	0.72	210
	Number of midwives	1.03	1.10	0.72	210
	Number of nurses	3.98	3.42	0.65	210
	Laboratory	0.17	0.24	0.54	210
	Improved source of water	0.72	0.73	0.89	210
	Share of general equipment	0.74	0.83	0.24	210
	Share of essential drugs	0.26	0.29	0.75	210
Health Workers	Male	0.39	0.38	0.89	1045
	Age	42.63	42.34	0.83	1045
	Physician	0.23	0.22	0.90	1045
	Midwife	0.11	0.11	0.89	1045
	Nurse	0.45	0.39	0.38	1045
	Years of experience	18.63	18.07	0.68	1045
	Monthly income (in TJS)	634	670	0.24	1037
	Days absent in past month	0.34	0.69	0.12	1045
	Hours worked in past week	22.37	26.53	0.37	1045
	Patients seen in past day	3.83	8.14	0.004***	1044
Recently pregnant women	age	26.49	26.58	0.73	2829
	Lowest wealth quintile	0.19	0.15	0.56	2829
	Low wealth quintile	0.21	0.16	0.30	2829
	Middle wealth quintile	0.23	0.18	0.05*	2829
	High wealth quintile	0.19	0.22	0.45	2829
	Highest wealth quintile	0.19	0.29	0.26	2829
	At least secondary education	0.60	0.60	0.99	2829
	married	0.98	0.97	0.30	2829
	Number of children	2.38	2.27	0.23	2733
	Received any ANC	0.93	0.86	0.25	2829
	Timely ANC	0.74	0.68	0.30	2509
	At least 4 ANC visits	0.52	0.62	0.49	2826
	Received PNC	0.57	0.67	0.14	2829
Adults aged 40 and above	Male	0.47	0.45	0.05*	5038
	Age	54.60	54.91	0.39	5038
	Lowest wealth quintile	0.22	0.14	0.28	5038
	Low wealth quintile	0.22	0.17	0.28	5038
	Middle wealth quintile	0.22	0.19	0.30	5038
	High wealth quintile	0.19	0.23	0.42	5038
	Highest wealth quintile	0.16	0.28	0.18	5038
	At least secondary education	0.65	0.62	0.71	5038
	Married	0.88	0.85	0.02**	5038
	Blood pressure measured in past year	0.50	0.45	0.50	5038

Data from Baseline survey.

[a] The tests are based on ordinary least square regressions with standard errors clustered at the district level.

Level of statistical significance: *p<0.10; **p<0.05; ***p<0.01

Table A3: Impact of PBF on availability of equipment and consumables

	Variable	Baseline Mean in PBF districts	Trend	PBF impact	p-value	N
Equipment	Clock	0.538	-0.096	0.290	0.501	394
	Child weight scale	0.858	0.051	0.091	0.194	410
	Height measurement tool	0.849	-0.010	0.142	0.251	418
	Tape measures	0.802	0.109	0.070	0.119	414
	Adult weight scale	0.774	0.059	0.158*	0.076	416
	Blood pressure cuff	0.774	0.038	0.188*	0.082	420
	Thermometer	0.811	0.048	0.141	0.190	420
	Stethoscope	0.755	0.097	0.139	0.222	418
	Fetoscope	0.802	-0.333	0.164	0.929	404
	Otoscope	0.387	-0.033	0.382	0.148	396
	Exam bed	0.783	-0.490	0.292	0.116	408
Essential drugs	Amoxicillin	0.283	0.115	0.526***	0.001	420
	Iron	0.066	0.644*	0.214**	0.018	420
	Oral rehydration solution (ORS)	0.377	0.452	0.171	0.184	420
	Paracetamol	0.302	0.490*	0.208**	0.010	420
Family planning products	Condoms	0.425	0.462*	-0.056	0.249	420
	Intrauterine device (IUD)	0.217	0.135	0.158	0.999	420
	Depot Medroxyprogesterone Acetate	0.179	0.202	0.053	0.875	420
	Implant	0.019	0.029	-0.010	0.813	420
	Oral contraceptive pills	0.406	0.462	0.038	0.864	420
Vaccines	Bacille Calmette-Guérin (BCG)	0.028	0.038	0.037	0.983	420
	Dyphteria Tetanus Pertussis (DTP)	0.170	0.317	0.192	0.242	420
	Hepatitis B Vaccine (HBV) Tetravalent	0.094	0.202	-0.060	0.876	420
	HiB vaccine	0.047	0.154	-0.088	0.440	420
	Measles vaccine	0.094	0.240	0.194	0.615	420
	Oral polio vaccine	0.330	0.269	-0.024	0.935	420
	Pentavalent vaccine	0.274	0.356	0.097	0.303	420
	Tetanus Toxoid (TT)	0.198	0.250	0.118	0.425	420
Diagnostic test kits	HIV test kit	0.113	-0.212	0.570***	0.000	420
	Pregnancy test kit	0.274	0.356	0.097	0.310	420
	Syphilis test kit	0.075	0.010	0.396**	0.030	420
	Urine protein & glucose testing kit	0.009	0.010	0.009	0.268	420

Note: data from rural health center assessments. All regressions are the difference-in-differences specification with facility fixed effects and controlling for exposure to the collaborative quality improvement and citizen scorecards interventions. Standard errors are clustered at the district and survey round level with wild bootstrapping.
Level of statistical significance: *p<0.10; **p<0.05; ***p<0.01

Table A4: Impact of PBF on Health Provider Satisfaction

Variable	Baseline Mean in PBF districts	Trend	PBF impact	p-value	N
Relationship with facility staff	0.944	0.016	0.019	0.973	2126
Relationship with district and ministry of health staff	0.905	-0.059	0.018	0.838	2126
Relationship with facility management	0.955	-0.023	-0.025	0.607	2126
Quality of facility management	0.862	-0.015	0.033	0.848	2126
Availability of medicine	0.226	0.052	0.408***	0.000	2126
Quality of medicine	0.416	0.033	0.334	0.216	2126
Availability of equipment	0.115	0.019	0.318**	0.039	2126
Quality of equipment	0.163	-0.020	0.330	0.155	2126
Availability of other supplies	0.064	0.048	0.320**	0.044	2126
Physical condition of health facility	0.265	-0.042	0.270	0.146	2126
Ability to provide high quality of care	0.681	-0.005	0.146	0.182	2126
Salary	0.068	0.081	0.124	0.443	2126
Overall satisfaction	0.650	0.115	0.088	0.265	2126

Note: data from interviews of health providers in both rural health centers and health houses. All regressions are the difference-in-differences specification with facility fixed effects. All regressions include controls for facility type, age and gender of providers, position, a binary indicator for experience of over 20 years and an indicator for being born in the district. Standard errors are clustered at the district and survey round level with wild bootstrapping.
For each category, health providers responded whether they are satisfied, unsatisfied or neither. The regressions are run on binary variables indicating the respondent reported satisfaction.
Level of statistical significance: *p<0.10; **p<0.05; ***p<0.01

Table A5: Impact of PBF on content of care

Variable	Overall mean	PBF impact	p-value	N
Panel A: Adult consultations history taking [a]				
Physical activity	0.36	0.22*	0.066	2,614
Alcohol	0.12	0.08*	0.085	2,614
Blood Pressure measurement	0.81	0.07	0.620	2,614
Chest pain	0.41	0.00	0.978	2,614
Diabetes status	0.44	0.18	0.120	2,614
Diet	0.60	0.19**	0.047	2,614
Family history of heart disease or stroke	0.31	0.21	0.124	2,614
Hypertension	0.55	0.09	0.558	2,614
Kidney disease history	0.26	0.11	0.209	2,614
Medication	0.31	0.04	0.624	2,614
Nausea	0.50	0.14	0.391	2,614
Oliguria	0.13	0.02	0.671	2,614
Smoking status	0.13	0.08**	0.048	2,614
Duration of symptoms	0.58	0.02	0.876	2,614
Symptoms	0.81	-0.02	0.767	2,614
Vomiting	0.35	0.10	0.450	2,614
Patient measures weight	0.33	0.16	0.213	2,614
Weight gain/loss	0.26	0.14	0.147	2,614
Panel A: child consultations: physical examinations [b]				
Assess ability to drink or breastfeed	0.32	0.10	0.265	2,582
Breathing	0.42	0.08	0.475	2,582
Examine ear	0.33	0.01	0.886	2,582
Check for ear infection	0.28	0.21***	0.010	2,582
Check for edema	0.10	0.06	0.241	2,582
Examine eye infection	0.21	0.04	0.474	2,582
Assess lethargy	0.13	0.12*	0.058	2,582
Check for mouth ulcers	0.27	-0.00	0.963	2,582
Measure radial pulse	0.39	0.31*	0.052	2,582
Pinch skin	0.63	0.17*	0.057	2,582
Stridor	0.25	0.13	0.146	2,582
Temperature	0.91	0.05	0.391	2,582
Check for visible wasting	0.19	0.11	0.131	2,582

Note: data from direct clinical observations conducted at the rural health centers. The PBF impact is estimated using a propensity score weighting specification as only follow-up observations are being included in the analysis. Standard errors are clustered at the district level.

[a] Data from consultations of adults aged 40 and above. Regressions include controls for facility's exposure to the collaborative quality improvement and citizen scorecards interventions, Region, age and gender of patient, gender age and position of provider, and a binary indicator for experience of over 20 years.

[b] Data from consultations of children under-5. Regressions include controls for facility's exposure to the collaborative quality improvement and citizen scorecards interventions, Region, gender age and position of provider, and a binary indicator for experience of over 20 years.

Level of statistical significance: *p<0.10; **p<0.05; ***p<0.01

Appendix 2: Parallel trends

The difference-in-difference approach relies on the assumption that in the absence of the PBF intervention, changes over time in the PBF and control districts would be similar. We used the Tajikistan Demographic and Health Survey (DHS) 2012 to test whether the parallel trends assumption could be rejected for the years prior to the launch of the PBF pilot. For the years 2008-2012, we constructed annual indicators for timely antenatal care, receiving at least four antenatal consultations and for vaccination coverage of children.

The results presented in the table below are from a difference-in-difference specification in which we define the period 2010-2012 to be a 'post intervention' period. As can be seen, for none of the indicators were we able to reject the parallel trends assumption. We clustered standard errors at the sampling cluster level, which provides a more conservative test in comparison to clustering at the district level as we do in our analysis. The results are robust to changing the pre- and post-intervention periods.

		Diff-in-diff coefficient	
Selected Indicators	N	β	SE
ANC Visits within 12 weeks	1163	-0.05	0.07
4 or more ANC visits	1463	-0.05	0.04
BCG vaccination date	2532	0.02	0.06
DPT 1 vaccination	2532	-0.01	0.06
Polio 1 vaccination	2532	0.01	0.08
DPT 2 vaccination	2532	-0.03	0.08
Polio 2 vaccination	2532	-0.02	0.06
DPT 3 vaccination	2532	-0.02	0.05
Polio 3 vaccination	2532	-0.01	0.05
Measles vaccination	2532	-0.03	0.06
Polio 0 vaccination	2532	-0.04	0.07

Several caveats, however, should be considered. First, we still rely on the assumption that the parallel trends would have persisted during the years of project implementation. Second, we can only perform these tests for a small set of outcomes of interest. Third, to preserve the anonymity of respondents, the location of DHS clusters is randomly displaced up to 2 km in urban areas and 5 km in rural areas, with 1% of clusters displaced up to 10km. Some clusters may be

misclassified between treatment and control if the displacement changes the district of the cluster, substantively altering the results of the analysis.

Appendix 3: Sampling and survey design

Catchment areas

The project implementation unit of the Health Services Improvement Project in the Tajikistan Ministry of Health and Social Protection provided the research team with the lists of rural health centers and the villages in their catchment areas. Few RHCs were randomly excluded from the sample when the total number of RHCs was not divisible by three. The study was set up to also evaluate the impact of two additional intervention, collaborative quality improvement and citizen scorecard, that were randomly introduced at the RHC level. Therefore, having the number of RHCs in each district divisible by three was desirable. The effects of these interventions are not discussed in this paper. The exposure to these additional interventions is balance across PBF and control districts and all regressions in our analysis include controls for exposure to these interventions.

In total, this selection process resulted in a sample of 216 RHCs/ catchment areas. There were 108 chosen catchment areas in the seven PBF districts and 108 areas in the nine control districts. In each treatment group, 33 areas were in the Sughd region and 75 in Khatlon.

Six catchment areas were removed from the follow-up survey because they were either closed for renovation, closed or downgraded to health house level. The catchment areas were removed from the analysis sample.

Health Facility Assessments

A detailed facility assessment was conducted in each selected rural health center. In catchment areas where there was at least one health house, a single health house was randomly selected for an abridged facility assessment. If the catchment area had more than one health house, all health houses had equal probability of being selected.

In each facility, interviews with health workers providing maternal and child health services or hypertension-related services were conducted. Four health workers were interviewed in each rural health center and 2 health workers in each health house.

Direct clinical observation of consultation of children under-5 and adults 40 and above were conducted in rural health centers. In the baseline survey, the target set for the survey teams was to observe 5 consultations of each type in each facility. The survey was conducted in November and

December 2014 and because of the winter there were very few consultations conducted at the facilities. The teams reached less than 30 percent of the target number of observations. Therefore, in the follow-up survey, the survey methodology was changed so that also home visits are observed, and the teams stayed for a longer duration in each health center. The target number of observations for each type of consultation was raised to 15.

Household Survey

Within each selected catchment area, two villages were randomly selected to serve as sampling units for the household survey. The survey teams conducted full listings of all households in the selected villages to identify households eligible for inclusion in the survey. In all catchment areas, households with a female member with a pregnancy in the preceding two years were selected. In a third of the catchment areas - those randomly selected to not implement neither the collaborative quality improvement nor the citizen scorecard intervention - households with a member 40 years old and above were also selected.

The same villages visited in the baseline survey were also selected for the follow-up survey, but a new set of households was identified by the listing exercise.

Survey tools and data

The baseline survey tools and data sets can be accessed with the following links:

Household survey: http://microdata.worldbank.org/index.php/catalog/2798

Facility-based survey: http://microdata.worldbank.org/index.php/catalog/2799

Appendix 4: Propensity score weighting

We use a logistic regression model to estimate the propensity that a catchment area falls in either the PBF or control group, given its baseline characteristics. The following baseline variables were used in the regression specification:

- Khatlon region
- Whether the catchment area has any health house, with region interaction
- Number of health houses in the catchment area, with region interaction
- Dummies for catchment population size terciles, with region interaction
- Availability of a laboratory, with region interaction
- Number of RHC providers in different positions: physicians, midwives, nurses
- Facility infrastructure indicators: dedicated reception area, separate toilet for women and men, waiting area, separate waiting area for women, heating, water piped into facility plot
- Infection prevention and control indicators: improved source of water, appropriate procedures for biowaste disposal, decontamination and sterilization
- More than 10 kilometers distance between RHC and the district hospital, with region interaction
- Share of recently pregnant women who received any antenatal care
- Share of recently pregnant women who received any postnatal care
- Share of the catchment area households in different wealth quintiles

Thirty-three out of the 210 catchment areas had propensity scores outside the common support and were dropped from the sample. The propensity scores are then used to reweight the observations. For catchment areas in the PBF districts, the weight is equal to the inverse of the predicted propensity while for catchment areas in the control districts, the weight equals the inverse of the predicted probability of *not* being in the treatment group.

The table below shows that the reweighting using the propensity score achieved balance in baseline characteristics between the PBF and control group. We ran OLS regression of the different baseline indicators on the PBF dummy using the propensity score weights and clustering standard errors at the district level.

Sample	Variable	PBF	Standard Error	N
Rural Health Centers	Khatlon region	0.04	(0.28)	177
	Health houses	-0.21	(0.53)	177
	Any health house	-0.01	(0.12)	177
	Catchment size medium	0.10	(0.13)	177
	Catchment size big	-0.13	(0.15)	177
	dist_hospital_10km	0.09	(0.15)	177
	Reception area	0.04	(0.09)	177
	Heating	0.09	(0.15)	177
	Water piped into plot	0.01	(0.09)	177
	Improved source of water	0.02	(0.10)	177
	Incinerator	-0.08	(0.05)	177
	Sterilization procedure	0.06	(0.07)	177
	decontamination procedure	0.02	(0.11)	177
	Biowaste disposal procedure	-0.02	(0.16)	177
	Container for sharps and needles	0.38**	(0.18)	122
	Share of general equipment	-0.00	(0.09)	177
	Share of essential drugs	0.09	(0.14)	177
	Share of family planning products	0.09	(0.12)	177
	Share of vaccines	0.06	(0.11)	177
	Share of diagnostic kits	0.03	(0.08)	177
	Share of protocols	0.02	(0.09)	177
	External evaluations of facility	1.17	(0.80)	177
	External evaluations of staff	-0.11	(0.91)	177
	Internal evaluations of staff	-2.71	(2.46)	176
	Staff meetings in past 3 months	-0.44	(1.35)	177
	Solicit patient opinion	-0.28*	(0.16)	177
	Staff left in past 12 months	0.25	(0.26)	177
	Any left in past 12 months	0.05	(0.14)	177
	Laboratory available	0.04	(0.10)	177
Health providers	age	0.62	(1.55)	646
	Male	-0.04	(0.07)	646
	Physician	0.02	(0.11)	646
	Midwife	0.01	(0.01)	646
	Nurse	0.04	(0.05)	646
	20 years of experience	-0.01	(0.06)	646
	Born in district	-0.06	(0.06)	646
	Monthly income	-24.62	(41.12)	642
	Days absent in past month	-0.04	(0.32)	646
	Hours worked past week	-7.72	(4.74)	646
	Patient seen in past day	-4.63***	(1.32)	645
	Satisfaction score	-0.01	(0.06)	646
	Age	0.15	(0.27)	2,466
	Primary education	-0.01	(0.04)	2,466
	Secondary education	-0.00	(0.03)	2,466
	2nd wealth quintile	0.02	(0.06)	2,466

Recently pregnant	3rd wealth quintile	0.01	(0.03)	2,466
	4th wealth quintile	-0.00	(0.05)	2,466
Recently pregnant women	5th wealth quintile	-0.01	(0.10)	2,466
	Received any ANC	0.06	(0.05)	2,466
	At least 4 ANC visits	-0.02	(0.13)	2,463
	Timely ANC	0.07	(0.05)	2,164
	ANC at primary level	0.11	(0.10)	2,179
	Received any PNC	-0.09	(0.08)	2,466
Children under-5	Age	0.01	(0.06)	6,935
	Male	0.01	(0.02)	6,935
	Growth monitoring in 6 months	0.10	(0.06)	5,706
	12-23 months: all basic vaccinations	-0.01	(0.03)	1,668
	12-23 months: any basic vaccination	-0.03	(0.03)	1,668
	24-35 months: All basic vaccination	-0.05	(0.06)	1,370
	24-35 months: MMR vaccine	-0.09*	(0.05)	1,290
	24-35 months: any vaccination	-0.04	(0.04)	1,370
	Weight for age z-score	-0.05	(0.12)	5,170
	Weight for height z-score	-0.18	(0.21)	4,544
	Height for age z-score	0.34*	(0.20)	5,419
Adults	Blood pressure measured in past year	0.07	(0.06)	13,494
	Elevated blood pressure	-0.05	(0.03)	9,822
	Self-reported hypertension	0.00	(0.01)	16,126
	Prescription for hypertension	-0.03	(0.03)	1,565
	Took medicine in past 24 hours	-0.12	(0.08)	1,531

Appendix 5: Variable definition

Clinical vignettes

Clinical vignettes were given to medical providers in both PBF and control arms to test knowledge on best clinical practices for a subset of conditions. Vignettes presented standardized patients, and each vignette was separated into three indicators; 1) a nominal variable indicating the correct diagnosis was reached, 2) the proportion of recommended items queried during the medical history, and 3) the proportion of recommended procedures mentioned to be conducted during the physical examination. For all three indicators, answers were solicited but no answer choices were presented. The set of recommended items for the clinical history and examination followed from a combination Integrated Management of Childhood Illness clinical guidelines, locally appropriate non-communicable disease guidelines, and basic patient engagement indicators (solicit patient name, ask chief complaint, etc.).

Vignette 1 tests provider knowledge on case management, diagnosis, and treatment of dehydration in a child of 2 years of age. The set of recommended items to be taken during the history are asking the mother's and child's name, asking the parent about the chief complaint, asking about breathing difficulty, convulsions, temperature, difficult feeding, diarrhea, and vomiting. Physical examination procedures included measurement of weight and height, temperature, breathing, assessment of mucous membranes, rash, lethargy, and assessment of soft spots.

Vignette 2 tests provider knowledge on case management of pneumonia in a 6-month old child. Clinical history items included asking the mother's and child's name, inquiring about the chief complaint, assessing measles infection in the past month, and assessing history of cough, vomiting, convulsions, and difficulty breathing. The physical examination criteria included measurement of weight and height, temperature, comparison of z-scores, lethargy, convulsions, breathing rate, stiff neck, runny nose, rash, red eyes, wheezing, and respiratory distress.

Vignette 3 presents a standardized case of severe illness. Primary information was given with the case, and history items were limited to patient engagement – asking the mother and child's name and inquiring about the chief complaint. Physical examination items included measurement of weight and height, temperature, breathing rate, respiratory distress, nasal flaring, soft spots, ear infection, umbilical infection, lethargy, and pustules.

Vignette 4 assesses knowledge on identification and treatment of malnutrition among children. The clinical history included asking the mother and child's name, the age of the child, chief complaint, assessing history of vomiting, convulsions, difficulty breathing, diarrhea, and inquiring about feeding frequency, food variety, vitamin A supplement, micronutrient supplementation, and deworming. The recommended set of physician examinations for the standardized case included measurement of weight, height and mid-upper arm circumference, comparison against growth charts, breastfeeding/drinking, temperature, lethargy, breathing rate, respiratory distress, wheezing, edema on feet, skin pallor, feeding, and running blood tests and parasite screens.

Vignettes 5 and 6 presented an adult standardized patient which tested provider ability to correctly assess high blood pressure and calculate cardiovascular risk. Clinical history included asking the patient's age, more detail on the chief complaint, symptom onset period, history of nausea, vomiting, oliguria, visual aura during headache, chest pain; inquire about the patient's smoking status, alcohol intake, family history of heart disease and stroke, diabetes status, lifestyle and physical activity, food intake, weight change, and whether the patient is currently taking antihypertensives or other medications. The physical examination items included the measurement of vitals (blood pressure, weight and height, pulse), and ordering of urine (specifically creatinine ratio and hematuria) and blood tests (specifically for plasma glucose, electrolytes, creatinine, filtration rate, cholesterol) and an electrocardiogram. The provider was then given the necessary information to calculate the cardiovascular risk. The items for Vignettes 5 and 6 were nearly identical; Vignette 6 did not include a clinical history item asking about weight gain.

Wealth Index

The wealth index estimates the living standard of the household or patients, relative to the living standard of the other households in the sample. In this study, the wealth index for households was calculated by an index using self-reported data including consumer item ownership, amount and value of owned land, the value of rentals, household infrastructure, the source of water, heating, and electricity, and number household members per room. The resulting index is calculated from the standardized first component of principal component analysis, apportioned into quintiles. This method is also used by the Demographic and Health Survey; however, results will differ as the population in this survey is entirely rural. The wealth index was used as a covariate in the household survey indicator regressions.

Anthropometrics

Standardized weight-for-age, weight-for-height, and height-for-age were calculated using the WHO child growth STATA package. The resulting z-scores compares each child to a standardized global population, accounting for gender and age. A z-score of 0 implies the child falls directly on the global mean of all children of the same gender and age or height, whereas a z-score of 1 or -1 implies the child is one standard deviation from the mean. Standard deviations are used as benchmarks; children below two standard deviations of weight-for-age are considered wasted, and below three standard deviations is severely wasted. Children below two standard deviations of height-for-age are considered stunted, whereas children below three standard deviations are considered severely stunted. Prevalence of wasting and stunting is compared against the expected prevalence for each of these categories based on a standard normal distribution.

Elevated blood pressure

Elevated blood pressure is defined as a systolic blood pressure greater than or equal to 140 mmHg, or diastolic blood pressure higher than 90 mmHg, based on the average of three readings.