WPS6100

Policy Research Working Paper 6100

Impact Evaluation Series No. 60

Using Performance Incentives to Improve Health Outcomes

Paul Gertler
Christel Vermeersch

The World Bank
Human Development Network
Health, Nutrition and Population Team
June 2012

Abstract

This study examines the effect of performance incentives for health care providers to provide more and higher quality care in Rwanda on child health outcomes. The authors find that the incentives had a large and significant effect on the weight-for-age of children 0–11 months and on the height-for-age of children 24–49 months. They attribute this improvement to increases in the use and quality of prenatal and postnatal care. Consistent with theory, They find larger effects of incentives on services where monetary rewards and the marginal return to effort are higher. The also find that incentives reduced the gap between provider knowledge and practice of appropriate clinical procedures by 20 percent, implying a large gain in efficiency. Finally, they find evidence of a strong complementarity between performance incentives and provider skill.

This paper is a product of the Health, Nutrition and Population Team, Human Development Network. 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://econ.worldbank.org. The author may be contacted at cvermeersch@worldbank.org.

Using Performance Incentives to Improve Health Outcomes

Paul Gertler and Christel Vermeersch

Acknowledgments: We are indebted to Claude Sekabaraga, Louis Rusa, Rigobert Mpendwanzi, Agnes Binagwaho, Louis Munyakazi, Kathy Kantengwa, Gyuri Fritsche, Willy Janssen, Werner Vandenbulcke, Bruno Meesen, Agnes Soucat, Jennifer Sturdy, Paulin Basinga, and Damien de Walque for useful comments. We also thank the Rwanda Ministry of Health, Rwanda National Institute of Statistics, Rwanda Ministry of Finance, Belgian Technical Cooperation, Cordaid, GTZ, Healthnet, USAID, and the World Bank and their staff for cooperating in the implementation of the Performance Based Financing rollout plan and supporting the evaluation. We also gratefully acknowledge Elena Samonte-Hickley, Sigrid Vivo Guzman, Rosangela Bando, Rafael Ramirez and Ricardo Cavazos for research assistance. We gratefully acknowledge funding from the World Bank's Bank-Netherlands Partnership Program, the British Economic and Social Research Council, the Government of Rwanda through a Japanese PHRD grant, and the World Bank's Spanish Impact Evaluation Fund. The findings and conclusions represent the opinion of the authors and do not necessarily represent the opinion of the World Bank, its executive directors or its member countries.

Keywords: performance incentives, results-based financing, pay-for-performance, child health, health services, provider payment, impact evaluation

JEL classification codes: D12, D22, D29, I12, I15, J33

Sector board: Health, Nutrition and Population (HNP)

Affiliations and contact information: Gertler – University of California Berkeley: Gertler@haas.berkeley.edu; Vermeersch – World Bank: cvermeersch@worldbank.org.

1. INTRODUCTION

Improving health outcomes is a high priority in the vast majority of low-income countries. However, despite dramatic increases in public spending on health in the last decade, most low-income countries have made little progress and will not reach the 2015 Millennium Development Goal health targets (United Nations 2010 and 2011).[1] One reason is the poor performance of medical care providers. Absenteeism among providers is widespread (Chaudhury et al. 2006; Reinikka and Svensson 2009). Provider knowledge of proper clinic procedures for basic care is low (Das and Hammer 2004). And providers give a standard of care that is well below the state their clinical knowledge (Das and Gertler 2007; Das, Hammer and Leonard 2008; Leonard and Masatu 2010a and 2010b). In this context, finding ways to increase health provider productivity in terms of supplying higher quality care to more patients is critical.

One promising intervention designed to improve health provider productivity is to pay them for performance. So-called Pay-for-Performance (P4P) schemes give bonuses to providers that give higher quality of care to more patients. P4P affects health care provision in two ways; first, through incentives for providers to expend more effort in specific activities; and second, through an increase the amount of financial resources. Proponents highlight P4P incentives and argue that they strengthen the link between productivity and rewards (Levine and Eichler 2009).

In this paper we provide evidence on the effect of P4P incentives on health outcomes and the associated pathways through which the incentives work. The evaluation uses data from Rwanda, one of the poorest countries in the world, where we nested a prospective evaluation into the national rollout of P4P. In order to identify the P4P incentive effect separately from the increase in resources, the program agreed to hold constant the level of resources constant across treatment and comparison facilities. It did so by increasing the traditional input-based budgets of the comparison group by the average amount of P4P payments to treatment facilities. As a result, while treatment and comparison facilities had the same average level of resources, a portion of the treatment facilities' resources was determined based on their performance whereas the comparison facilities' resources were not.

[1] Among the 67 countries with highest child mortality rates, only 10 are on track to reduce mortality by two thirds. And the rate of decline in maternal mortality for all is well short of the 5.5 % needed to achieve 2015 MDG target.

Our study is, to our knowledge, the first rigorous evaluation of P4P incentives to providers on health outcomes in a low-income setting, and it is the first to isolate the impact of P4P incentives from the associated increase in resources. This is important because if P4P achieves its results from increased financial resources rather than incentives, then the same results could be achieved from an increase in traditional input-based budgets. Then there would be no reason to incur the administrative costs associated with P4P.

We find that the performance incentives significantly increased the use and quality of prenatal and postnatal medical services, and that these effects translated into large and significant improvements in child health outcomes. Specifically, we find that P4P led to an increase of 0.53 standard- deviations in the weight-for-age of children 0-11 months and 0.25 standard deviations in the height-for-age of children 24-49 months. We also find, consistent with our conceptual framework, that there were larger effects of incentives on services where monetary rewards and the marginal return to effort are higher.

These results were associated with a substantial improvement in provider productivity. The incentives reduced the gap between provider knowledge and actual practice of the appropriate clinical procedures by approximately 20 percent. The difference between knowledge and practice is a measure of technical inefficiency. Hence, the incentive effects can be interpreted as generating substantial improvements in provider efficiency and hence productivity.

Finally, there are large complementarities between provider skill (knowledge of appropriate clinical procedures) and performance incentives in the production of quality. Specifically, we show that higher skilled providers increased quality more than lower skilled providers in response to the incentives. This means that P4P is more effective in a context where providers have higher skills. It also suggests that traditional training interventions may yield higher results with incentives than without.

Our work contributes to a small literature on the effects of paying medical care providers for performance in developing countries.[2,3] Most of the studies do not have control groups and

[2] There is, however, a growing literature on P4P for medical care in the U.S. and the U.K. with mixed results. Recent examples include Fleetcroft et al (2012), Jha et al (2012), Van Herck et al. (2010), Lindenauer (2007), Doran et al (2006), and Perterson et al (2006).

[3] There is a related literature on performance pay for teachers in low-income countries. For example see Muralidharan and Sundararaman (2011) for an application to India and Neal (2012) for a general review.

estimate the impact of P4P as jumps in time trends of the amount of services providers by treatment facilities.[4] There are, however, three well-identified evaluations. Hospital-based physicians in the Philippines who received extra bonus pay based in part on knowledge of clinical appropriate clinical procedures, reported increases in clinical knowledge (Peabody et al, 2011). In Indonesia, performance incentives to *villages* for improvements health and education outcomes led to an increase in labor supply from health providers (Olken et al. 2011). Finally, this paper builds on our previous work that demonstrated P4P in Rwandan increased prenatal care quality, institutional delivery, and use of preventive child medical care (Basinga et al 2011).

2. INSTITUTIONAL CONTEXT

Rwanda is one of the poorest countries in the world with a GDP of US$340 per capita in 2007 (World Bank, 2008). Since the end of the 1994 genocide, Rwanda has made remarkable progress in improving maternal and child health. Between 2000 and 2005, infant mortality fell from 107 to 86 deaths per 1,000 live births, while maternal mortality fell from 1,071 to 750 per 100,000 live births (Institut National de la Statistique du Rwanda and ORC Macro 2006). Despite this progress, the health system continues to grapple with serious shortages of qualified personnel and low levels of service delivery, especially in rural areas (World Bank 2010b).

a. P4P

In 2005, the Rwandan Ministry of Health (MoH) used the opportunity of an increase in the health sector budget to scale up nationally a P4P scheme for maternal and child health care services (Logie et al 2008; World Bank 2010a). The decision was based on positive reports of P4P schemes in two Provinces that had been developed by a number of NGO's (Kalk et al. 2005; Soeters et al. 2005 and 2006). The P4P scheme provides bonus payments to primary care facilities based on provision of various types of services and the quality of those services. P4P payments go directly to facilities and are used at each facility's discretion. The overall amount of P4P payments is large in comparison to facilities' budgets: a study of 68 facilities receiving P4P payments shows that P4P payments represent an increase in funding of 24.6% above the base budget. On average, 77 percent of P4P funds were used to compensate personnel resulting in an increase of 38 percent in staff compensation (Basinga et al 2011).

[4] See Witter et al., (2012) for a recent systematic review of health care P4P in low and middle income countries

b. The Payment Formula

The scheme pays for 14 maternal and child healthcare services conditioned on an overall facility quality assessment score. The formula used for payment to facility *i* in month *t* is:

$$Payment_{it} = \left(\sum_{j} P_j U_{jit} \right) \times Q_{it} \quad \text{with} \quad 0 \leq Q_{it} \leq 1,$$

where P_j is the payment per service unit *j* (e.g. institutional delivery or child preventive care visit), U_{ijt} is the number of patients using service j in facility *i* in period *t*, and Q_{it} is the overall quality index of facility *i* in period *t*.

The 14 service indicators (U_{jit}) and associated payment rates (P_j) are listed in Table 1. The first 7 indicators consist of the number of visits to the facility for various types of service such as prenatal care and institutional delivery, while the second set of 7 indicators refers to the content of care provided during those visits. They include the number of children who were fully vaccinated, the number of pregnant women who received tetanus vaccines and malaria prophylaxis during prenatal care, the number of at-risk pregnancies that were referred to hospitals for delivery during prenatal care, the number of severely malnourished children who were referred to treatment facilities, and the number of general emergencies that were referred to the appropriate place for care. The health literature considers these to be measures of aspects of the process or clinical quality of care (Donabedian, 1988). The Rwanda Ministry of Health (MoH) defined these indicators and payments based on national health priorities, available budget and previous NGO experience with P4P (Ministère de la Santé du Rwanda 2006 and 2008).

As for the payment rates (P_j), the largest rates are for deliveries at health facilities and emergency referrals of women to hospitals for obstetric services; both services amount to \$4.59 per case. The next largest rate, at \$1.83, is for new contraceptive user visits, referral of at-risk pregnancies to hospitals for delivery and referral of malnourished children to higher-level facilities for treatment. Facilities are paid about half as much, or \$0.93, for a child who is fully vaccinated on time. Facilities are paid about half as much again, or \$0.43, for pregnant women who receive the tetanus vaccine and the same for malaria prophylaxis. Curative and contraceptive re-supply visits are paid the small sum of \$0.18 per visit. Finally, prenatal care visits are paid only \$0.09 per visit with a bonus of \$0.39 for every woman who completes 4 visits.

The actual amount paid to the facility is adjusted based on the overall quality, Q_{it} ,of the facility. Specifically, the facility's overall quality enters the payment formula as a multiplicative factor that raises or lowers the payment for the 14 output indicators. The quality index is bounded between zero and one. If the facility meets all of the quality criteria, then the index equals one and the facility receives the full P4P payment. However, if the facility is deficient in some of the quality criteria, then all of the payments are discounted. For example, if the facility only scores 0.80 on the quality index, then it only receives 80 percent of the payment for the 14 output indicators. In this way, the P4P scheme pays for both facility output and facility quality.

The quality index Q_{it} is a function of structural and process measures of quality specified in the Rwandan preventive and clinical practice guidelines (Ministère de la Santé du Rwanda 1993, 1997, and 2003). Structural measures are the extent to which the facility has the equipment, drugs, medical supplies and personnel necessary to be able to deliver a specific medical service. Process measures are the clinical content of care provided for specific services. [5]

The formula for the quality index is:

$$Q_{it} = \sum_k \omega_k S_{kit} \quad \text{with} \quad \sum_k \omega_k = 1,$$

where S_{ikt} is the share of indicators for service k that are met by facility i in period t, and ω_k is the weight for service k. If a facility has perfect structural and process quality, then all the S_{ikt} take on value one and the overall quality index is equal to one; in this case, the facility is paid the maximum possible bonus for the services provided. By contrast, if the quality index is less than one, P4P payments are discounted for *all* services provided at the facility. Table 2 details the services that are included in the quality index, their weights and the relative importance of structural and process indicators in the computation of the score for each service k.

[5] Clinical practice guidelines are "systematically developed statements to assist practitioners and patient decisions about appropriate health care for specific circumstances." (Field and Lohr, 1990)

c. Administration of Payments

P4P payments are administered by district steering committees comprised of members of the District Government, Providers and civil society (Ministry of Health Republic of Rwanda, 2006; Fritsche et al 2010). Facilities submit monthly activity reports (U_{jit}) and quarterly requests for payment to the district steering committee, which is responsible for verifying the data and authorizing payment. For the referral indicators, the facility must also submit verification from the hospital that the referral was appropriate and the referred patient was treated. The committee verifies the reports by sending auditors to facilities on a quarterly basis on an unannounced randomly chosen day. The auditors review the utilization registry and facility records to verify the data reported is the same as the data recorded in facility records. During the 2006-2008 period, MoH conducted a survey of face-to-face interviews with approximately 1000 patients to verify the accuracy of the records. This survey found that false reporting on patients or services was less than 5 percent (HDP 2008).

Information used to compute each facility's overall quality score is collected under the existing national monitoring system that requires all district hospitals to monitor and supervise the quality of health centers in their districts. Every quarter, a district hospital team from different services (e.g. prenatal, curative care, preventive care) visits each facility on an unannounced randomly chosen day to assess the facility's quality through direct observation and review of patient records using a standardized tool. At the end of the visit, the team discusses their findings with the facility's personnel and provides recommendations to improve quality. In P4P districts, the data are used to construct the overall quality score for the facility each quarter.

3. CONCEPTUAL FRAMEWORK

a. Model

In this section we use a simple behavioral model to hypothesize how the introduction of P4P would likely affect provider behavior. We have in mind a rural clinic with no close substitutes locally. We assume for simplicity that a facility acts as one single decision-maker that we call the provider. Key to this discussion is the provider's objective function. We assume that medical care providers typically value their patients' health as well as the income they earn from the services they provide to treat patients. We take into account this ethical aspect of preferences by assuming that providers treat all patients who show up for care and provide them with at least a minimum level of care as defined by their ethical standards.

We begin by considering the case where the facility is paid a fixed amount for staff costs and has a fixed budget for non-personnel costs, and assume that the non-personnel budget cannot be reallocated for staff costs. In this case, seeing more patients and providing them with better care does not affect the provider's income. Hence, the provider treats all patients who show up and provides them with the minimum level of care.

The P4P scheme introduces a new dimension to the provider's optimization problem by linking part of the facility's income to the provision of certain services and to quality of care. For simplicity, we assume that the provider allocates effort to two types of patient services (e.g. prenatal care and delivery) and quality of care. Taking into account the basic structure of the P4P formula, we can write the new profit function as

$$V = I + \left[P_1 U_1 \left(\varepsilon_1 \right) + P_2 U_2 \left(\varepsilon_2 \right) \right] Q\left(\varepsilon_q \right) - C\left(\varepsilon \right) \tag{3}$$

where I is the fixed salary, P_i is the P4P payment for service i, U_i is the total quantity of service i provided to patients, Q is the overall quality of care, and $C(*)$ is the cost of effort. Recall the the U_i's are listed in Table 1 and Q is an index constructed based on the items in Table 2.

The provider chooses effort levels ε_1 and ε_2, to increase the quantity of services provided above the minimum levels necessary to treat patients who show up, as well as effort ε_q, to improve the quality of care above the minimum ethical standards.[6] The service production functions $U_i(.)$ and the quality production function $Q(.)$ are increasing in effort, but at a decreasing rate. Finally, $C(.)$ is a function of total effort (i.e., $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_q$) and is convex.

The provider then chooses effort levels to maximize income subject to effort levels being weakly positive. In the case of an interior solution, effort is allocated in such a way that marginal revenue of effort is equalized across the three types of effort and that it is equal to the marginal cost of effort:

$$P_1 U_1' \left(\varepsilon_1 \right) = P_2 U_2' \left(\varepsilon_2 \right) = \left[P_1 U_1 \left(\varepsilon_1 \right) + P_2 U_2 \left(\varepsilon_2 \right) \right] Q'\left(\varepsilon_q \right) = C'\left(\varepsilon \right) \tag{4}$$

[6] In this way, we effectively normalize the minimum effort levels to zero.

Note that the marginal return to effort supplied to each service depends not just on its own price but also on the price of the other service, as does the marginal return to effort supplied to quality depends on both prices. Hence, an increase in any of the two prices always raises the return to effort supplied to quality. Effort supplied to anything raises the marginal cost of effort because the cost of effort is a function of total effort.

The relative amount of effort allocated to the two types of services satisfies the following condition:

$$\frac{U_2'(\varepsilon_2)}{U_1'(\varepsilon_1)} = \frac{P_1}{P_2} \tag{5}$$

i.e. the ratio of the marginal returns to effort in delivering the services should equal the ratio of the payment rates for those services. Hence, more effort will be allocated to the service that has the highest price and the higher marginal productivity of effort.

b. Comparative Statics

We can discuss the likely effects of introducing P4P in terms of a comparative static of price increases, whereby the original level of P and ε are close to zero. Consider an increase in P_1, the payment for service 1. This will raise the marginal revenue from supplying effort to service 1 and to the provision of quality, and therefore is an incentive to supply more effort to that service and quality. Because the increased effort raises the marginal cost of total effort, the provider will reduce effort to service 2. As a result, the increase in effort for service 1 and for quality comes at the cost of both reduced effort for the other service and reduced leisure. Hence, while the total amount of effort increases, the relative allocation of effort increases to service 1 and quality and falls to service 2. If the price increase is large enough, the optimal effort allocated to service 2 will fall below the minimum ethical constraint and, as a result, the constraint will bind.

However, the comparative static analysis of a single price change is not exactly applicable to the introduction of a P4P scheme as the P4P scheme changes all prices simultaneously. Before the price increase, all effort levels are at the minimum ethical constraint. Increases in the prices of the services will increase the allocation of effort to quality because increases in any and all prices raise the marginal return to supplying effort to quality. The largest allocations of effort to a service will be to those services for which the relative price increases are the largest and the marginal productivity of effort is the highest. Analogously, the smallest allocations of effort will be to those services that get the smallest relative price increase and have the lowest marginal return to effort. In fact, if for a particular service the relative price increase is small enough and the marginal productivity of effort low enough, the provider will not supply any more effort to that service despite the absolute increase in price. In this case, the supply of effort will remain at the minimum ethical bound.

Hence, the effect of the introduction of the P4P payments depends not only on the relative payment rates, but also on how hard it is to increase the levels of services. In general, we argue that it takes more work to increase services that depend on patient choices than services that are completely in the provider's control. For example, it takes more work to convince a pregnant woman to come to the clinic for prenatal care than to give the woman a tetanus shot once she is there. Hence, even if payments were equal for an additional patient visit as for a tetanus shot, one would expect to see larger increases in the number of tetanus shots (which is under the control of the provider) than in the number of visits to the facility (which is largely under the control of the patients). Moreover, we argue that initiation of care takes more effort than its continuation. For example, it will take a provider substantial amounts of effort to go out to the community to find pregnant women, especially in the first trimester of pregnancy, and bring them in for prenatal care. By contrast, it is a relatively easier task to use an existing prenatal care visit to lobby women already in prenatal care to deliver in the facility.

c. Full Prices

The previous discussion assumes that the prices of the services enter in the profit function in a simple linear fashion as presented in equation 2. In reality, the payment scheme is more complicated and the services listed in Table 1 are made up of both primary reasons to visit a clinic as well as services provided conditional on such a visit. While they are all U_i's, the services provided during the visits also enter the quality index Q. Moreover, the payment P for seeing a patient depends on the services provided during that visit. Consider the payment for prenatal care. Providers receive \$0.18 for every pregnant women who starts prenatal care, an additional \$0.37 if the women completes at least 4 visits, an additional \$0.92 if they give the patient a tetanus shot and malaria prophylaxis during a prenatal care visit, and an additional

$1.83 if they assess the delivery to likely be risky and refer the mother to deliver at the district hospital. Hence, payments for prenatal care depend not only on the number of pregnant women coming for care and the number of times they visit, but also on the content of the care provided during those visits.

In fact, payment rates for visits are much higher if the provider supplies better content of care. As we discussed, a provider will receive $0.55 for four prenatal care visits of low quality versus $1.47 for providing high quality. If the provider detects a high-risk pregnancy and refers the woman to the hospital for delivery, payments for this high-quality care even increase to $3.30. In the case of growth monitoring, the payment to the provider is $0.18 per visit plus an additional $1.83 if the child is malnourished and she refers her to the hospital for treatment. Since 45 percent of Rwandan children under age five have moderate chronic malnutrition, and 19 percent have severe chronic malnutrition,[7] (Institut National de la Statistique du Rwanda and ORC Macro 2006), the expected payment for a high quality growth-monitoring visit is quite high. Overall, the incentive structure focuses not just on treating more patients, but on providing more patients with higher quality of care; this happens through both the multiplicative scaling factor Q and by direct payment for content of care services in the U_i's.

d. Empirical Implications

This discussion provides us with a number of empirical predictions. First, increases in payments will be more effective for services for which the relative price increase is highest and for those that have the highest relative marginal return to effort. Second, increases in payments will not necessary increase all services. There may be no effect on services for which payment rates and the marginal return to effort is low. Third, payment rate for a service depends not only on the number of patients treated, but also the content of care provided during a visit and it is this payment rate that matters for the allocation of effort. Finally, we expect the introduction of P4P to increase quality Q, the multiplicative factor in the payment formula.

[7] Moderate (severe) chronic malnutrition corresponds to height-for-age below -2 (resp. -3) standard deviations from the median of the reference population (Institut National de la Statistique du Rwanda and ORC Macro 2006).

4. IDENTIFICATION

a. Experimental Design

The evaluation design took advantage of the phased implementation of the program at the district level over a 23-month period. Rwanda manages its health care system at the district level and P4P is no exception. As a result the government mandated that all facilities in a district must be incorporated into the P4P scheme at the same time. Hence, the evaluation employed a stratified cluster randomized designed where districts were first grouped into blocks with common characteristics and then randomly assigned to treatment and comparison groups. Administrative districts with pre-existing P4P schemes were excluded from the experimental design. The remaining districts were grouped into blocks based on similar characteristics for relief, rainfall, and predominant livelihoods as per the 2002 Census.

However, just before implementation, administrative district boundaries were redrawn in the context of a government-wide decentralization effort (MINALOC 2004). As a result, some of the experimental areas were combined into new districts with areas that already had the pilot P4P schemes. Because P4P could not be "removed" from health facilities that were already implementing the schemes, and because P4P is managed at the district level, the MoH required that all facilities within those new districts be in the treatment group. This led the evaluation team to switch the assignment of treatment and comparison for eight districts. In the end, the study included 19 districts, of which 10 belonged to the treatment group and 9 to the comparison group.

The sample included 166 of Rwanda's 401 primary care facilities, 80 in treatment districts and 86 in comparison districts. The facilities in the treatment group started receiving P4P in 2006, while the facilities in the comparison group continued with traditional input-based financing for an additional 23 months. P4P was introduced in treatment districts over a 5-month period, yielding a minimum program exposure of at least 18 months.

b. The Incentive Effect

Since a primary objective of the evaluation was to isolate the impact of the P4P incentives separately from the effect of increased resources, it was necessary to hold the level of resources constant across treatment and comparison facilities. To accomplish this, comparison facilities' traditional input based budgets were increased by the average amount of P4P payments to treatment facilities on a quarterly basis during the entire 23-month treatment window. As a result, treatment and comparison facilities had the same levels of financial resources on average throughout the study. Therefore, the differences in outcomes between the two groups at follow-up must be attributed to the difference in incentive structures and not to a difference in available financial resources.

c. Estimation

Given the reassignment of districts between the treatment and comparison groups before the start of the study, and the limited number of districts that could be assigned to the treatment and comparison groups, we view our study as quasi-experimental. While the sample is balanced at baseline on outcomes and characteristics, it is possible that the reassignment of districts was correlated with something unobservable to us and related to health outcomes. However, redistricting took place within the context of a decentralization agenda that was led by the Ministry of Local Government, and we find no evidence that it was driven by or related to health outcomes (MINALOC 2004).[8] Therefore, we think it is likely that any relevant unobservable factors were likely to be invariant over the time period of the intervention.

Therefore, we will use the difference-in-differences methods that control for unobserved time invariant characteristics. This method compares the change in outcomes in the treatment group to the change in outcomes in the comparison group. By comparing changes, we control for observed and unobserved time-invariant characteristics as well as for time-varying factors that are common to the treatment and comparison groups. We show below that the final assignment to the treatment and comparison groups is orthogonal to pre-intervention observable variables, leading us to assert that there should be no correlation between this assignment and non-observables that drive program effects.

All of the individual outcomes relate to pregnancies and many women do not give birth in both waves of the survey. Hence, we treat the 2006 and 2008 household surveys, described below, as repeated cross-sections and estimate the following regression specification of the difference-in-difference model for individual outcomes:

[8] According to MINALOC 2004, the objective of the decentralization was to enhance institutional development and capacity building for responsive local governance., to develop an efficient transparent and accountable fiscal and financial management system at local government and grassroots level.

$$Y_{ijt} = \alpha_j + \gamma_{2008} + \beta \cdot P4P_j \cdot I_{2008} + \sum_k \lambda_k X_{kijt} + \varepsilon_{ijt} \quad (6)$$

where Y_{ijt} is the outcome of interest for individual i living in facility j's catchment area in year t; $P4P_j$ is a dummy variable that takes value 1 if facility j belongs to Phase I (i.e. started receiving P4P in 2006) and 0 otherwise; α_j is a facility fixed effect; γ_{2008} is a fixed effect for 2008; I_{2008} is a dummy variable that takes value 1 if the year of observation is 2008 and 0 otherwise; the X_{kijt} are individual characteristics; and ε_{ijt} is a zero mean error term. We compute robust standard errors clustered at the district level to correct for possible heteroskedasticity and correlation of the error terms within districts over time. We estimate each regression both with and without individual characteristics.

d. Statistical Inference

A major limitation of our design is the limited number of districts. Since the unit of assignment to treatment and comparison was the district and not the facility, there may be inter-cluster correlation in the error terms. The asymptotic justification for inference with cluster-robust standard errors assumes that the number of clusters goes to infinity. Yet in our application there are too few clusters for this assumption to hold. Therefore we base our statistical inference on randomization inference hypothesis tests that use WILD bootstrapping Monte Carlo methods as recommended in Cameron et al (2008). These tests return a p-value for the hypothesis rather than a standard error. Hence we report the estimated coefficient and the p-value for the one-sided test of significance.

5. DATA

a. Surveys

We surveyed all 166 facilities plus a random sample of households in the catchment area of each facility. The surveys were conducted at baseline in 2006 prior to the implementation of P4P in treatment facilities and again approximately 23 months later, before the comparison facilities were incorporated in the program. The surveys were conducted independently from the operation of the P4P program. Payment was based on administrative records and reports from the facilities and never on the evaluation surveys.

The facility survey collected information on staffing, expenditures, medical equipment, drug availability from the facility administrator, and provider knowledge about the appropriate clinical procedures for quality prenatal care. As part of survey, enumerators also conducted exit interviews with approximately 10 women who visited the facility to collect information on the actual clinical services (quality) provided during their prenatal care visit. The sample of facilities was well balanced at baseline. There were no differences in means for the 17 baseline characteristics presented in Table 3.

The household survey consists of a random sample of 13 households living in each facility's catchment area, for a total sample of 2,158 households. To build the sample, we first sampled 13 census zones (each containing approximately 15 to 20 households) from each facility's official list of zones in their catchment area. We then physically listed all households in the sampled zones and randomly selected one household with at least one child under 5 years old from each zone. The surveys were conducted by trained enumerators hired by external firms specialized in data collection. Interviewers did not inform respondents that the survey was associated with the P4P program, but rather that the survey was being conducted to assess health conditions.

Response rates were high as only 2 percent of sampled households refused to participate in the interview. In the follow-up survey, 88 percent of the baseline households were re-interviewed. The rate of attrition from the baseline sample was not statistically different between the treatment and comparison groups (12 percent each). Households that could not be found or interviewed were replaced with randomly selected households from the same zones. In addition, the household data were well balanced between treatment and comparison groups at baseline. Only 2 of the 30 characteristics reported in Table 4 were significantly different at the 5% level between treatment and control groups at baseline.

b. Provider Resources

Health clinics in rural Rwanda are staffed with about 6 nurses and 4 medical technicians or midwives (Table 3 Panel A). There are almost no physicians practicing in these clinics. The availability of equipment and drugs needed to provide quality care is reasonable for rural Africa (Table 3 Panel B). The structural quality indices are the share of drugs and equipment available at the facility among those that the Ministry of Health guidelines define as necessary in order to deliver each type of care (Ministère de la Santé du Rwanda 1993, 1997, 2003 and 2009). At baseline clinics had about 80 percent of the drugs and equipment necessary to provide curative care services, 78 percent of that needed for deliveries, 96 percent for prenatal care, and 94 percent for immunizations.

Finally, we find no significant differences between the treatment and comparison groups in terms of log total expenditures, allocation of the budget across medical personnel, medical supplies and non-medical purposes, the number of physicians, nurses and other types of personnel, in 2006 before the introduction of P4P in the treatment facilities (Table 3 Panel D). More importantly, there is no difference in mean log expenditures in 2008 between treatment and comparison groups, which supports our interpretation that any differences in outcomes are caused by the P4P performance incentives as opposed to resource differences.

c. Provider Skill (Knowledge) and Quality of Care

A key measure of medical care provider's skill and capability is their knowledge of appropriate clinic protocol. At baseline, medical providers knew about two-thirds of the prescribed clinical protocol for prenatal care (Table 3 Panel D). Our measure of provider knowledge is the share of the appropriate prenatal care clinical procedures specified in the official Rwandan Clinical Practice Guidelines (CPG) for prenatal care (Ministry of Health Contractual Approach Unit, Rwanda 2006, and Ministère de la Santé du Rwanda 2008).[9] The 24 specific clinical services collected are listed in Table 4. The items cover collecting information about previous pregnancies, medical history, current pregnancy status, physical exams, laboratory tests, and case-risk management.

We used a clinical "vignette" which is a standardized hypothetical patient case with a specific medical history. We presented the vignette to a randomly selected health worker who regularly provides prenatal care and asked the health worker to describe the clinical protocol that

[9] CPGs are a recommended set of clinical procedures conducted during the prenatal care visit that maximize the probability of good health outcomes based on the clinical literature and expert opinion. First developed in the United States (Field and Lohr 1990), the Rwandan guidelines are based on the US version adjusted to local resource constraints. The Rwandan CPG for prenatal care includes medical and pregnancy history questions, physical examinations, lab tests and follow-up procedures (Table 4).

she would apply. We then used the answers to compute the share of official CPG clinical content items that the provider mentioned without prompting from the interviewer.[10]

Providers, however, do not necessarily deliver clinical services up to their level of knowledge. We define quality of care as the actual clinical services delivered to patients. We collected information on actual clinical services provided during prenatal care from exit interviews of patients leaving facilities and from the household surveys. We measured the quality of prenatal care by computing the share of actual clinical content items delivered during a prenatal care consultation to the items that should compose a typical prenatal consultation as recommended in the Rwandan CPGs.[11] For the subsequent analysis we also break out tetanus vaccine as it is also one of the directly priced services (Table 1).

The quality index and the knowledge index cover the same items and therefore are comparable. While providers know two-thirds of the appropriate protocols, they only deliver about 45 percent of the protocols (Table 3 Panel E). This implies a gap about 18 percentage points between knowledge and practice, which translated into providers delivering about two-thirds of clinical services they know they should deliver. It implies that there is substantial distance from the provider's production possibility frontier in terms of technical efficiency.

d. Maternal and Child Utilization

Our utilization measures conform as closely as possible to the paid indicators in Table 1. At baseline over 95 percent of mothers had some prenatal care during the previous pregnancy, but most did not begin their care until after the first trimester and few completed 4 or more visits (Table 5 Panel B). In addition, only about one-third of women delivered in health care institutions and only about 1 in 10 were using modern contraception at the time of the survey.

We separate children's preventive care utilization into ages 0-23 months and 24-59 months as the younger group is expected to have higher preventive visit rates (Table 5 Panels E and F). Child preventive services cover immunization, vitamin A, distribution of mosquito nets and child growth monitoring with referral of malnourished children to higher levels of care for treatment. Younger children have about .2 visits in the month prior to the survey consistent with

---

[10] This measure of competency was used in Das and Hammer (2004), Kak et al (2001) and Peabody et al (2004).

[11] This measure enters the payment formula through the general facility quality index *Q* as specified in Table 2. It has also been used extensively in the literature to measure quality. (Barber 2006, Barber et al 2007, Peabody and Gertler 1997, Das and Hammer 2007).

a checkup every six months. Older children have about .1 visits. We examine curative care conditional on the child's experiencing a symptom, including fever, diarrhea, digestive track problems, respiratory infection, malaria, and skin rash. About one in four children who experienced illness symptoms sought treatment.

e. Child Health Outcomes

We measure child health outcomes for two age groups: children 0-11 months and 24-47 months. Children in the 0-11 months range in the treatment group in the 2008 survey would have been exposed to P4P during the full prenatal period and during the full time after birth. Children between 24 and 47 months in the treatment group in the 2008 survey would have been exposed to P4P for 18-23 months during the early stages of life, but not during the prenatal period.

We consider two measures of health: (i) height-for-age z-score and (ii) weight-for-age z-score. A child's height results from her genetic potential, adjusted for insufficient nutrient intake and inability to absorb nutrition because of illness. Therefore, height is a summary measure of health and nutrition since conception. By contrast, weight is an indicator of current nutrition and illness status and does not represent factors that accumulate over the lifetime. Better prenatal care, which includes nutritional advice to mothers and the diagnosis and treatment of maternal illness, could in principle increase both infant height and weight. Better child preventive care, which includes vaccination and growth monitoring, and child curative care, which limits the duration and severity of illness, also have the potential to affect height and weight.

We measured child height and weight using standard international procedures and portable scales and stadiometers, which were recalibrated on a twice-weekly basis in the field. As part of the quality control procedures, all children were measured twice during the visit. We first report height and weight in centimeters and kilograms respectively. Then, we standardize these measurements into height-for-age and weight-for-age z-scores in accordance with World Health Organization guidelines. The z-scores measure the number of standard deviations from age-sex standardized height of a healthy (U.S.) reference population. Finally, we classify children with a height-for-age or weight-for-age below -2 SD of the corresponding reference median (Z score < or = -2) are classified as stunted or underweight, respectively.

6. RESULTS

In this section, we begin by reporting the results on P4P on the ultimate outcome of interest, namely health. Once we establish a relationship between P4P and health, then we examine the pathways through which P4P works. Specifically, we consider the effect of P4P on utilization and on quality of care.

a. Child Health Outcomes

We consider two age groups: (1) children 0-11 months at endline, and (2) children 24-47 months at endline. In the treatment group, the younger group fully benefited from any improvement prenatal care and post-natal preventive services due to P4P as they were conceived during the experimental period. The older group only could have benefited from access to improved post-natal preventive care, as they were already born at baseline. These older children benefited from any improved post-natal services for their entire lifetimes. Children in-between benefited only partially from prenatal care and only partially from post-natal care.

We estimated versions of equation (6): one without controls and a second with controls. The controls include the child's age and sex, maternal height, mother's age, whether the mother has completed primary school, whether the father lives in the household, whether the family is a member of a Mutuelle, the total number of household members, the number of household members under the age of 6, whether the household owns land, and dummy variables for quartiles of the household asset value.[12] Age was entered as a series of dummy variables that represent one-month increments.

The estimated effects of P4P on child health outcomes are reported in Table 6. Among the 0-11 month old children who benefited from P4P since conception, we find large and significant positive effects on weight for age. Infants in the treatment areas gained 0.53 of a standard deviation in weight as a result of P4P. Among 24-47 months old, who benefited from the program for 23 months starting between age 1 and 24 months after birth, the program led to a gain of over 0.25 of a standard deviation in height for age. Note that the size of the effect on height for age of the younger group is 0.22 without controls and 0.16 with controls. However, the sample size for the younger group is less than half of the sample size for the older group.

[12] The household asset value was constructed on the basis of the value of household assets including owned houses, household durable goods, farm animals, farm equipment and micro-enterprise equipment.

While effects on health outcomes are large for both age groups, we see effects on weight for the younger group and on height for the older group. This difference can be explained by the role of breast-feeding on weight gain and height growth. Micronutrients that are not present in large amounts in breast milk are critical for improvements in height (Dewey and Adu-Afarwuah, 2008). Hence, during the period of exclusive breast-feeding children tend to gain weight rather than height. Reductions in illnesses that impede a child's absorption of the nutrition in breast milk tend to lead to gains in weight. Once supplemental foods that contain more micronutrients are introduced into the child's diet, reductions in illnesses that improve a child's ability to absorb nutrition are manifest in height. According to the 2005 Demographic and Health Survey, 88% of children less than 6 months old are exclusively breastfed. And 31 percent of children age 6 to 9 months did not receive supplementary foods (Institut National de la Statistique du Rwanda and ORC Macro 2006).

b. Pathways

We hypothesize that the observed impacts on child health are due to changes in their utilization and quality of prenatal and postnatal care induced by the introduction of P4P incentives. In this section, we examine the impact of P4P on utilization and quality of care and thereby document that the observed changes in health outcomes are plausibly caused by P4P.[13] In order to validate that the results on the effect of P4P on utilization and quality are plausible, we present and discuss the results in the context of our theoretical conceptual framework.

Our framework predicts that the introduction of P4P will lead to larger increases in services (the *U*'s) for which the relative price increase is highest and for those that have the highest relative marginal return to effort; conversely, there may be no effect on services for which payment rates and the marginal return to effort is low. Our theoretical discussion also predicted that the introduction of P4P would unambiguously increase quality *Q*, the multiplicative factor in the payment formula. Our results are very consistent with this framework.

First, we find significant effects on the quality of prenatal care. Specifically, we estimate that P4P is associated with a 0.16 standard deviation increase in prenatal care quality (Table 7).

[13] Many of the estimated impacts in Table 6 were previously reported Basinga et al. (2011). However, that paper used traditional asymptotic inference statistical tests based large samples. In this paper, we use randomization inference that is robust to small numbers of clusters and discuss the results in the context of the theory laid out in section 3.

Improving the quality of prenatal care has a low marginal cost and relatively high payoff for providers. To increase the quality of prenatal care, providers did not need to go into the community to motivate women to change their behavior; it was sufficient for providers to change their own practices, with the women that came in for prenatal care anyways. And the facility's quality index Q provides high-powered incentives, as the quality index is a scaling factor that is applied to all payments for all of the facility's services, and not only to payment of prenatal care services.

In addition, elements of prenatal care quality entered the payment formula both as U_{jit}, and within the quality index. For instance, P4P was associated with an increase in tetanus toxoid vaccines delivered during prenatal care visits. The tetanus vaccine is priced moderately at $0.46, but also contributes the multiplicative Q index and requires relatively low effort on the part of the provider as it is mostly in the control of the provider.

Second, as predicted, we find mixed effects on the services with some increasing and others unaffected. The services that increased were those with higher prices as well as more in the provider's control and less in the patient's control. While institutional delivery had of the highest payment rate of $4.59, institutional deliveries require mothers (families) to decide to plan to deliver in an institution and come in when it is time. Qualitative evidence from field visits suggests that in order to capture this rent, facilities paid community health workers to locate women in the late stages of pregnancy and then encourage and assist them to deliver at the facility. The health centers' strategy of teaming up with community health workers likely had higher chances of success because the community health workers have ongoing relationships with mothers.

Well-child care also responded strongly to the P4P scheme, even though its unit payment is relatively low at $0.18 per visit. However, the payment rate jumps substantially to $2.03 if the provider identifies a malnourished child and gets them into treatment. This becomes a very strong incentive since over half the children in Rwanda were stunted in 2005 (Institut National de la Statistique du Rwanda and ORC Macro 2006). Similar to institutional deliveries, facilities employ a strategy of working with community health workers to get infants and small children into care. In addition, since the P4P scheme led to higher rates of institutional delivery, treatment facilities would have more complete lists of infants and children in their books than comparison facilities making follow-up easier and cheaper. For the same reason, treatment facilities also had higher direct access to new mothers to try to convince them to come in for well-baby care.

On the other hand, P4P did not increase the initiation of prenatal care, the number of women who completed four visits. The reason that P4P did not increase prenatal care and contraceptive usage is that these services had a high marginal cost and low payoff for providers. The payment for prenatal care is $0.09 for women who initiate care; an additional $0.37 for woman completes 4 visits. At baseline 95 percent of women had at least one prenatal visit. It would be very costly and difficult for providers to go out into the community and find the remaining 5% of women and convince them to get care. Reaching out to women in early stages of pregnancy is a costly activity for providers, not only because of the travel and time that are involved, but also because the targeting of any information campaigns would necessarily be poor. In addition, getting women in for early prenatal care involves changing women's behavior, which is difficult in any case. It is unlikely that providers put much effort into reaching out to these women given the $0.09 return.

Similarly, getting women to complete 4 visits is also difficult. For women to complete four prenatal care visits, it would probably be necessary that they start prenatal care before the fourth month of pregnancy, a time at which pregnancy is usually hard to observe for an outsider. Culturally, Rwandan social norms dictate that women do not acknowledge pregnancies before they have to. Our local collaborators reported that women are concerned that the evil eye could be used to cause the women to spontaneously abort a fetus and this is the reason that women hid their pregnancies as long as possible. Hence, finding and getting women to initiate care before they are showing is almost impossible. Once they initiate care late, then it is hard to get them to return multiple times in a short period of time. From the providers perspective all this effort only yields them an additional $0.37. Moreover, providers can still get the payments for providing tetanus vaccine and malaria prophylaxis as long as women get any care.

Finally, we find no evidence that P4P led to increases in contraceptive prevalence rate, despite the relatively high payment rate for new acceptors. This may be due to the fact that we measured actual use at the time of the household survey, which is influenced more by resupply rates than by initial use rates, while P4P payments were made for initial acceptors of contraceptives only.

c. Knowledge

We have argued that P4P incentives induced providers to put more effort into delivering higher quality services for more patients. An alternative explanation might be that P4P affected provider knowledge and through this induced a change in provider practice. This could have happened in two different ways. First, facilities could have changed the personnel mix, substituting better trained personnel for lower trained ones. Second, there could have been an increase in knowledge of the existing personnel. As mentioned previously, during the P4P evaluation process the district hospital supervisors discuss the results of the quality score with health facility personnel, providing practical recommendations to improve the quality of care where needed. This by itself could have increased provider knowledge on the care protocol, by providing regular reminders of what the proper protocol is. However, these visits were applied to both treatment and comparison facilities, so the opportunities to learn provided by these visits were no different between treatment and comparison groups. But there may have been different amounts of learning from the feedback provided, because treatment facilities knew that the score would directly influence their payment rates; hence they may have been more attentive to advice from the district hospital teams, and they may have learnt about better practices as a result.

To ascertain this potential causal path, we first estimate the impact of the P4P program on provider knowledge of the prenatal care protocol. We find that health worker knowledge likely did improve as a result of P4P. We estimate a 0.40 standard deviation increase in knowledge, but not statistically significant (Table 8 Model 1). The lack of significance is likely to small sample sizes. We therefore re-run the regressions of prenatal care controlling for knowledge. However, we find that controlling for knowledge does not alter the estimated impact of P4P on quality (Table 8 Model 2). Therefore, knowledge is not likely to be the main path through which P4P improved practice; rather, P4P almost surely increased quality through incentives.

This analysis does allow us to examine the direct effect of knowledge on quality. In fact, one of the most common interventions to improve quality of care is by training medical care providers in proper clinical procedures. However, we find no effect of increased knowledge on quality of care (Table 8, Model 2).

d. Complementarities

While changes in knowledge do not seem to directly impact quality, we hypothesize that knowledge may be complementary to P4P in the sense that P4P may be more effective when providers are relatively high-skilled. More knowledgeable providers can more easily exploit the P4P incentives for gain than less knowledgeable providers. To test this hypothesis, we included an interaction between P4P and provider knowledge being in the top half of the knowledge distribution. In this model, we do find differences in the program's impact on prenatal care practice between facilities that had health workers with above-the-median levels of knowledge, as opposed to below the median (Table 8 model 3).

7. EFFICIENCY

Another interpretation of how P4P works is based on the idea that providers are not delivering services up their full ability (knowledge). There is indeed evidence of this efficiency argument as provider deliver of clinical services during prenatal care is substantially lower than their knowledge of appropriate clinical procedures. Recall that providers on average know 63 percent of appropriate procedures, but deliver only 45 percent. This leaves an 18 percentage point difference between knowledge and practice. If we consider a provider's knowledge as their production possibilities frontier, then one can interpret the gap between knowledge and practice as a measure of technical inefficiency. The P4P incentives are intended to reduce technical inefficiency.

We present the efficiency gap in figure where skill is represented on the horizontal access as the share of prenatal CPG recommended clinical services that the provider knows and the vertical access represents quality delivered as the share of prenatal CPG recommended clinical services actually provided. The 45^{o} line is the production possibility frontier (PPF) where providers deliver clinical quality care to the best of their knowledge. If providers deliver a quality of care below their level of knowledge, then they would be preforming inside the PPF. The vertical distance between the frontier and the performance point is a measure of technical inefficiency.

We also included in figure 1 the actual performance curves of the providers in our data set. The curves are bivariate nonparametric regressions of quality against knowledge separately for treatment and comparison groups at endline. Notice that both lines are well inside the PPF implying substantial levels of technical inefficiency at all skill levels. In addition, while the performance curves are upwards sloping, they are flatter than the PPF. This implies that while knowledge improves performance, the efficiency gap increases with knowledge. Finally, the performance curve for the treatment group is above and steeper sloped than the curve for the comparison group. This implies that P4P reduced the efficiency gap and reduces it more for more skilled providers.

We now estimate the order of magnitude of the impact of P4P on the efficiency gap. We measure the efficiency gap as the share of CPG clinical services the provider knows minus the share of CPG clinical services delivered. We find that P4P reduces the efficiency gap by 3.5 percentage points or about 20 percent of the gap on average (Table 9 Model 1). When we control for provider knowledge, the effect of P4P on efficiency increases slightly to 4 percentage points (Table 9 Model 2). In this model higher knowledge is actually associated with a larger efficiency gap. In other words, while increases in provider knowledge improve the quality of care, the improvement in quality is less than the improvement in knowledge. Finally, we estimate that P4P has a much larger effect on efficiency for more knowledgeable providers. We find no increase in efficiency for providers below the knowledge median, but we find a 6 percentage point improvement among providers above the knowledge median (Table 9 Model 3).

8. CONCLUSION

This study examines the effect of performance incentives to health clinics in rural Rwanda on child health outcomes, maternal and child health care utilization, and prenatal care quality. We show that the incentives improved access to higher quality care that resulted in substantial improvements in child health outcomes. Moreover, we find that provider incentives led to a 20 percent improvement in efficiency. These findings lend strong support for the use of provider performance incentives to improve health outcomes.

The results also have implications for the structure of the incentives. The incentives in Rwanda took the form of 14 different prices that are scaled by the overall quality of the facility. Consistent with economic theory, we found large effects of the incentives for services that had relatively higher prices and required lower effort on the part of the provider. Services that required lower provider effort such as quality were the ones that providers had direct control over, as opposed to those that were dependent on patient behavioral choices such as the decision to seek prenatal care.

In fact the program did not manage to get women into prenatal care earlier or more frequently, due to low unit payments and high costs to the providers. While it may be tempting to increase the unit payments for prenatal care, this would not resolve the issue of high provider costs. We think it is worth exploring alternative ways of getting women into the health center. Given the information asymmetries in knowledge about early pregnancy, it would be worth exploring the use of conditional cash transfers to poor women for attending prenatal care prior to the fifth month of pregnancy. In general, it is worth considering using provider incentives for services such as quality that are more in the provider's control and conditional cash transfers to patients for utilization decisions that are more in their control.

Finally, we find evidence of complementarity between the P4P incentive and the knowledge (skill) of health care providers. This suggests that effects of P4P incentives would be higher if completed with interventions that improve provider skill, such as training, and the training would have a great impact in settings with performance incentives.

REFERENCES

Barber, Sarah (2006), "Does the Quality of Prenatal Care Matter in Promoting Skilled Institutional Delivery? A Study in Rural Mexico," *Maternal and Child Health Journal*, Vol. 10, pp. 419–425.

Barber Sarah, Stefano Bertozzi and Paul Gertler (2007), "Variations in prenatal care quality for the rural poor in Mexico", *Health Affairs*, Vol. 26(3), pp.w310-23.

Basinga, Paulin, Paul Gertler, Agnes Binagwaho, Agnes Soucat, Jennifer Sturdy and Christel Vermeersch (2011), "Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation," *The Lancet*, Vol 377, pp. 1421-28.

Cameron, Colin, with Jonah Gelbach and Douglas Miller (2008)"Bootstrap-Based Improvements for Inference with Clustered Errors", *Review of Economics and Statistics*, Vol. 90, 414-427.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan and F. Halsey Rogers (2006), "Missing in Action: Teacher and Health Worker Absence in Developing Countries," *The Journal of Economic Perspectives*, Vol. 20, No. 1, pp. 91-116.

Das, Jishnu and Paul Gertler (2007) "Variations In Practice Quality In Five Low-Income Countries," *Health Affairs*, Vol. 26: pp.296-309

Das, Jishnu and Jeffrey Hammer (2004), "Which Doctor? Combining Vignettes and Item Response to Measure Clinical Competence," *Journal of Development Economics*, Vol. 78, pp. 348-383.

Das, Jishnu and Jeffrey Hammer (2007), "Money for Nothing: The Dire Straits of Medical Practice in Delhi, India," *Journal of Development Economics*, Vol. 83(1), pp. 1-36.

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.

Dewey, Kathryn and Seth Adu-Afarwuah (2008) "Systematic Review of the Efficacy and Effectiveness of Complementary Feeding Interventions in Developing Countries," *Maternal and Child Nutrition*, Vol. 4, pp.24-85.

Donabeidian, A. (1988), "The Quality of Care. How Can it be Assessed?" *The Journal of the American Medical Association*, Vol. 260(12), pp. 1743-1748.

Doran Tim, Catherine Fullwood, Hugh Gravelle, David Reeves, Evangelos Kontopantelis,Urara Hiroeh, and Martin Roland. (2006), "Pay-for-performance programs in family practices in the United Kingdom," *New England Journal of Medicine*, Vol. 355, pp. 375-384.

Field, Marilyn and Kathleen Lohr (1990) Clinical Practice Guidelines: Directions for a New Program, Institute of Medicine, National Academy Press, Washington D.C.

Fleetcroft R., N. Steel, R. Cookson, S. Walker and A. Howe (2012) "Incentive payments are not related to expected health gain in the pay for performance scheme for UK primary care: cross-sectional analysis.", *BMC Health Serv Res.*, Vol. 16, pp. 12-94.

Fritsche, György, Louis Rusa, Rigobert mpendwanzi, Agnes Soucat, Claude Sekabaranga, Bruno Meesen (2010), "The National Rollout of Performance-Based Financing for Health Services in Rwanda: How It Was Done," World Bank Working Paper.

Glickman, S., F. Ou, E. DeLong, M. Roe, B. Lytle, J. Mulgund, J. Rumsfeld, B. Gibler, M. Ohman, K. Schulman, E. Peterson (2007) "Pay for Performance, Quality of Care, and Outcomes in Acute Myocardial Infarction," *JAMA*, Vol. 297(21), pp. 2373-2380.

Health, Development and Performance (HDP) (2008), "Rapport d' Enquête de Contre-Vérification par la Communauté dans les Districts de Nyamasheke, Nyaruguru et Rulindo," Kigali, Rwanda, online: http://www.pbfrwanda.org.rw/index.php?option=com_docman&task=cat_view&gid=24&Itemid=29&limitstart=35.

Institut National de la Statistique du Rwanda (INSR) and ORC Macro (2006), *Rwanda Demographic and Health Survey 2005.* Calverton, Maryland, U.S.A.: INSR and ORC Macro.

Jha, A., K. Joynt, J. Orav and A. Epstien (2012) "The Long-Term Effect of Premier Pay for Performance on Patient Outcomes," *New England Journal of Medicine,* Vol. 366(17), pp. 1606-1615.

Kak, Neeraj, Bart Burkhalter and Merri-Ann Cooper (2001), "Measuring the Competence of Health Care Providers," *QAP Issues Paper*, Vol. 2(1), pp. 1-28.

Kalk, Andreas, Jean Kagubare Mayindo, Laurent Musango, Gerard Foulon. (2005), "Paying for Health in two Rwandan Provinces: Financial Flows and Flaws," *Tropical Medicine and International Health*, Vol. 10(9), pp. 872-878.

Leonard, K.L. and M.C. Masatu, "Using the Hawthorne Effect to examine the gap between a doctors best possible practice and actual practice," *Journal of Development Economics*. 93 (2): 226-243 (2010a)

Leonard, K.L. and M.C. Masatu, "Professionalism and the Know-Do Gap: Exploring Intrinsic Motivation among Health Workers in Tanzania," *Health Economics* 19 (12): 1461-1477 (2010b).

Levine, Ruth and Rena Eichler, Eds. (2009), *Performance Incentives for Global Health*, Brookings Institution Press, Washington, DC.

Lindenauer P., D. Remus, S. Roman et al. (2007), "Pay for performance in hospital quality improvement," *New England Journal of Medicine*, Vol. 356, pp. 486-496.

Logie, Dorothy, Michael Rowson and Felix Ndagije (2008), "Innovations in Rwanda's health system: looking to the future," *The Lancet*, Vol. 372, pp. 256-61.

Ministère de la Santé du Rwanda (1993), *Standards de Prestation des Services au Centre de Santé. Soins Préventifs en SMI/PF /Nutrition*. Volume 1, Première Edition.

Ministère de la Santé du Rwanda (1997), *Normes du District Sanitaire au Rwanda*, Kigali, Rwanda.

Ministère de la Santé du Rwanda (2003) *Normes du District de Santé au Rwanda*, Kigali, Rwanda.

Ministère de la Santé du Rwanda (2006), *Fiche Technique du Programme Elargi de Vaccination*, Kigali, Rwanda.

Ministère de la Santé du Rwanda (2008), *Module de Référence de Formation Continue en Planification familiale. A l'Usage des Formateurs, Superviseurs et Prestataires au Niveau des Formations Sanitaires*, Mars, Kigali, Rwanda.

Ministry of Health. Contractual Approach Unit, Republic of Rwanda. (2006) *Guide for Performance Based Financing. Training module for actors involved in the implementation of the PBF program*. Kigali: Rwanda.

Ministry of Local Government, Community Development and Social Affairs (MINALOC), Repuublic of Rwanda (2004), *Five Year Decentralization Implementation Programme* 2004-2008, available at http://www.minaloc.gov.rw.

Muralidharan, Karthik and Venkatesh Sundararaman (2011), "Teacher Performance Pay: Empirical Evidence from India," *Journal of Political Economy*, Vol. 119(1), pp. 39-77.

Olken, Benjamin, Junko Onishi, and Susan Wong (2011), "Should Aid Reward Performance? Evidence from a field experiment on health and education in Indonesia," Working Paper, MIT Department of Economics.

Peabody John and Paul Gertler (1997), "Are clinical criteria just proxies for socioeconomic status? A study of low birth weight in Jamaica," *J.Epidemiol.Community Health*, Vol. 51(1), pp. 90-95.

Peabody John, J. Luck, P. Glassman, S. Jain, J. Hansen, M. Spell, et al. (2004), "Measuring the quality of physician practice by using clinical vignettes: a prospective validation study," *Annuals Internal Medicine*, Vol. 141(10), pp. 771-780.

Peabody, John, Riti Shimkhada, Stella Quimbo, Jhiedon Florentino, Marife Bacate, Charles E. McCulloch, and Orville Solon (2011), "Financial Inventives and measurement Improved Physicians' Quality of Care In the Philippines," *Health Affairs*, Vol. 30(4), pp.773-781.

Petersen Laura A., LeChauncy D. Woodard, Tracy Urech, Christina Daw and Supicha Sookanan (2006), "Does pay-for-performance improve the quality of health care?" *Annals of Internal Medicine*, Vol. 145, pp. 265-272.

Reinikka, Ritva and Jakob Svensson (2010), "Working for God? Evidence From A Change in Financing of Nonprofit Health Care Providers in Uganda," *Journal of the European Economic Association*, Vol. 8(6), pp. 1159-78.

Soeters, Robert, Laurent Musango and Bruno Meessen (2005). *Comparison of two output based schemes in Butare and Cyangugu provinces with two control provinces in Rwanda*, Global Partnership on Output-Based Aid (GPOBA), The World Bank, Washington, DC.

Soeters, Robert, Christian Habineza and Peter Bob Peerenboom (2006) "Performance based financing and changing the district health system: experience from Rwanda," *Bulletin of the World Health Organization*, November, Vol. 8(11).

United Nations (2010), *The Millennium Development Goals Report*. New York: United Nations.

United Nations (2011), *The Millennium Development Goals Peport*. New York: United Nations.

Van Herck, Pieter, Delphine de Smedt, Lieven Annemans, Roy Remmen, Meredith B Rosenthal and Walter Serneus, "Systematic Review: Effects, Design Choices, and Context of Pay-for-Performance in Health Care," BMC Health Services Research, Vol. 10, pp. 247-260.

Witter Sophie, Atle Fretheim, Flora L Kessy, Karin Lindahl (2012), *Paying for performance to improve the delivery of health interventions in low- and middle-income countries*, *Cochrane Database of Systematic Reviews* 2012, Issue 2. Art. No.: CD007899. DOI: 10.1002/14651858.CD007899.pub2.

World Bank (2008), *World Development Indicators Database, September*, Washington D.C.

World Bank (2010a), "Budgeting for Effectiveness in Rwanda – From Reconstruction to Reform," World Bank Working Paper, No. 205, Washington, DC.

World Bank (2010b), *Rwanda A Country Status Report on Health and Poverty*, Washington, DC.

Table 1: Rwanda P4P Performance Indicators and Payment Rates

		Amount paid per unit (US$)
Visit and Outreach Indicators		
1	Curative care visits	0.18
2	First prenatal care visits	0.09
3	Women who completed 4 prenatal care visits	0.37
4	First time family planning visits (new contraceptive users)	1.83
5	One-month contraceptive resupply visits	0.18
6	Deliveries in the facility	4.59
7	Child (0-59 months) growth monitoring/preventive care visits	0.18
Content of Care Indicators		
8	Children who completed vaccinations on time	0.92
9	Appropriate tetanus vaccine during prenatal care[+]	0.46
10	2nd dose of malaria prophylaxis during prenatal care	0.46
11	Appropriate referral to hospital for delivery[++]	1.83
12	Appropriate Emergency transfers to hospital during delivery[++]	4.59
13	Malnourished child referrals to hospital during preventive care[++]	1.83
14	Other emergency referrals during curative treatment[++]	1.83

[+] Appropriate is defined to any woman who obtains her second, third, fourth or fifth tetanus shot.

[++] Referrals must be confirmed by hospital that patient was treated and referral was necessary.

Source: Ministry of Health. Republic of Rwanda. Guide for Performance Based Financing. Training module for actors involved in the implementation of the PBF program. 2006.

Table 2: Services and Weights Used to Construct the Quality Index for P4P Formula

	Service	Weight	Share of weight allocated to structural measures+	Share of weight allocated to process measures++	Means of assessment
1	Curative care	0.170	0.23	0.77	Medical record review
2	Delivery	0.130	0.40	0.60	Medical record review
3	Prenatal care	0.126	0.12	0.88	Direct observation
3	Family planning	0.114	0.22	0.78	Medical record review
4	Immunization	0.070	0.40	0.60	Direct observation
6	Growth monitoring	0.052	0.15	0.85	Direct observation
7	HIV services	0.090	1.00	0.00	Direct observation
8	Tuberculosis service	0.028	0.28	0.72	Direct observation
9	Laboratory	0.030	1.00	0.00	Direct observation
10	Facility cleanliness	0.028	1.00	0.00	Direct observation
11	Pharmacy management	0.060	1.00	0.00	Direct observation
12	General administration	0.052	1.00	0.00	Direct observation
13	Financial management	0.050	1.00	0.00	Direct observation
	Total	1.000			

+ Structural measures are the extent to which the facility has the equipment, drugs, medical supplies and personnel necessary to deliver the listed service.

++ Process measures capture the clinical content of care provided for the listed service.

Source: Ministry of Health, Republic of Rwanda (2006) "Guide for Performance Based Financing and Training Module for Implementation of the PBF Program." Kigali, Rwanda

Table 3: Facility Characteristics

	Treatment (N=80)		Comparison (N=86)		Difference in Means	P-value
	Mean	St.Dev.	Mean	St.Dev.		
A. Staffing (2006)						
Medical Doctors	0.05	*(0.23)*	0.05	*(0.27)*	0.00	0.94
Nurses	6.31	*(6.90)*	5.48	*(3.30)*	0.83	0.41
Other Clinical Staff	4.13	*(3.09)*	4.47	*(4.05)*	-0.34	0.55
Non-clinical Staff	5.25	*(3.56)*	5.33	*(5.09)*	-0.08	0.90
B. Structural Quality Indices (2006)						
Curative Care	0.80	*(0.07)*	0.81	*(0.07)*	-0.01	0.58
Delivery	0.78	*(0.11)*	0.79	*(0.10)*	0.00	0.84
Prenatal Care	0.96	*(0.15)*	0.97	*(0.11)*	-0.01	0.29
Immunization	0.94	*(0.17)*	0.94	*(0.15)*	0.00	0.90
Laboratory	0.49	*(0.32)*	0.43	*(0.32)*	0.06	0.40
C. Expenditures						
Log Total Expenditures (2006)	15.81	*(1.04)*	15.61	*(1.01)*	0.20	0.42
Log Total Expenditures (2008)	16.91	*(0.71)*	16.99	*(1.08)*	-0.08	0.57
Personnel Budget Share (2006)	0.46	*(0.23)*	0.49	*(0.26)*	-0.03	0.56
Medical Supplies Budget Share (2006)	0.22	*(0.19)*	0.20	*(0.19)*	0.01	0.71
Non-medical Budget Share (2006)	0.32	*(0.25)*	0.30	*(0.22)*	0.02	0.72
D. Prenatal Care Clinical Knowledge (2006)						
Share prenatal care protocol known	0.63	*(0.10)*	0.65	*(0.09)*	-0.02	0.33
Knowledge of protocol (z-score)	0.02	*(0.81)*	0.15	*(0.77)*	-0.13	0.33
E. Prenatal Quality of Care (2006)						
Tetanus vaccine (=1)	0.71	----	0.67	----	0.04	0.33
Share of protocol provided	0.45	*(0.39)*	0.46	*(0.43)*	-0.01	0.66
Clinical protocol provided z-score	-0.13	*(1.49)*	-0.10	*(1.63)*	-0.02	0.76

Notes: P-values are for two-sided tests of the null hypothesis that the difference in means is zero and were calculated using WILD bootstrap with 999 draws. Except for prenatal care quality, the unit of observation is the facility. For prenatal care quality, the unit of observation is the patient visit and the results are based in 1584 observations.

Table 4: Rwandan CPG Prenatal Care Items Collected

1.	**MEDICAL HISTORY**
	• High blood pressure • Sexually transmitted infections including HIV • Tetanus immunizations • Pap smear test • Tobacco use • Alcohol use
2.	**PRIOR PREGNANCIES INFORMATION**
	• Number of previous pregnancies • Number of previous miscarriages and stillbirths
3.	**CURRENT PREGNANCY STATUS**
	• Last menstrual date • Health problems or concerns during pregnancy • Bleeding • Weight loss, nausea, or vomiting • Medications
4.	**PHYSICAL EXAMINATION**
	• Height • Weight • Blood pressure • Examine abdomen • Checked for swelling or water retention
5.	**LABORATORY EXAMINATIONS**
	• Take a blood sample for anemia test • Take a urine sample for gestational diabetes test • Test for current STI including HIV
6.	**PREVENTION AND CASE MANAGEMENT**
	• Tetanus toxoid injection • Iron/vitamin pills • Plan delivery

Table 5: Household and Individual Baseline (2006) Characteristics

	Treatment		Comparison		Diff	P-value+
	Mean	St.Dev.	Mean	St.Dev.		
A. Household characteristics						
Health insurance (=1)	0.54	---	0.51	---	0.04	0.58
Number of household members	4.92	*(4.40)*	5.00	*(5.16)*	-0.07	0.73
Household-Facility distance (in Km)	3.31	*(6.89)*	3.32	*(8.20)*	-0.02	0.97
Ownership of land (=1)	0.91	---	0.87	---	0.04	0.23
Value of assets	11.20	*(26.74)*	12.59	*(29.44)*	-1.39	0.35
B. Maternal Characteristics and Utilization						
Age < 20 years (=1)	0.03	---	0.02	---	0.01	0.32
Age > 35 years (=1)	0.29	---	0.31	---	-0.02	0.57
Primary education or more (=1)	0.10	---	0.11	---	-0.02	0.47
Living with Partner (=1)	0.94	---	0.91	---	0.04	0.21
Number of pregnancies (Parity)	4.32	*(4.58)*	4.33	*(5.24)*	-0.01	0.97
Any prenatal care (=1)	0.95	---	0.96	---	-0.01	0.77
Made 4 or more prenatal care visits (=1)	0.18	---	0.11	---	0.07	0.03
Number of prenatal care visits made	2.76	*(1.58)*	2.62	*(1.80)*	0.14	0.18
1st Prenatal care visit in 1st trimester (=1)	0.11	---	0.09	---	0.02	0.55
Institutional Delivery (=1)	0.35	---	0.36	---	-0.01	0.80
Use modern contraceptive (=1)	0.09	---	0.13	---	-0.03	0.16
D. Children's Demographic Characteristics						
Age (months)	25.86	*(11.01)*	26.64	*(8.38)*	-0.78	0.06
Female (=1)	0.50	---	0.50	---	0.00	0.87
Maternal height (cms)	157.84	*(14.36)*	158.15	*(16.45)*	-0.31	0.63
Mother's age (years)	31.07	*(12.26)*	31.28	*(13.48)*	-0.22	0.68
Mother completed primary school (=1)	0.09	---	0.12	---	-0.03	0.15
Father present (=1)	0.91	---	0.88	---	0.03	0.33
E. Medical Care Utilization in Last 4 Weeks by Children Age 0-23 Months						
Preventive visit (=1)	0.213	---	0.238	---	-0.025	0.556
Curative visit conditional on illness (=1)	0.305	---	0.266	---	0.039	0.530
F. Medical Care Utilization in Last 4 Weeks by Children Age 24-47 Months						
Preventive visit (=1)	0.084	---	0.140	---	-0.056	0.116
Curative visit conditional on illness (=1)	0.201	---	0.283	---	-0.082	0.124
G. Health outcomes of Children Age 0-11 Months						
Standardized height for age (z-score)	-0.03	*(2.50)*	-0.07	*(2.46)*	0.04	0.87
Standardized weight for age (z-score)	-0.31	*(1.91)*	-0.12	*(2.00)*	-0.19	0.32
H. Health Outcomes of Children Age 24-47 Months						
Standardized height for age (z-score)	-1.95	*(1.79)*	-1.80	*(1.94)*	-0.15	0.22
Standardized weight for age (z-score)	-0.75	*(1.32)*	-0.71	*(1.44)*	-0.04	0.66

Notes: P-values are for two-sided tests of the null hypothesis that the difference in means is zero and were calculated using WILD bootstrap with 999 draws.

Table 6: Impact of P4P on Child Health Outcomes

	Control Mean (2008)	No Controls		With Controls		N
		β	P-Value	β	P-Value	
Children Age 0-11 months						
Height for age Z-Score	-0.20	0.22	0.29	0.16	0.38	800
Weight for age Z-Score	-0.18	0.54	0.02	0.53	0.03	800
Children Age 24-47 Months						
Height for age Z-Score	-1.80	0.23	0.03	0.25	0.00	1957
Weight for age Z-Score	0.69	0.01	0.39	0.03	0.34	1957

Notes: P-Values are for one-sided tests of the null hypothesis that $\beta = 0$ and are calculated based on a WILD bootstrap with 999 draws. Controls include the child's age and sex, maternal height, mother's age, whether the mother has completed primary school, whether the father lives in the household, whether the family is a member of a Mutuelle, the total number of household members, the number of household members under the age of 6, whether the household owns land, and dummy variables for quartiles of the household asset value. Age was entered as a series of dummy variables that represent one-month increments.

Table 7: Impact of P4P on Health Care Utilization and Quality of Care

	Control Mean (2008)	No Controls		With Controls		N
		β	P-Value	β	P-Value	
Maternal Health Care Utilization						
Any Prenatal Care (=1)	0.98	0.00	0.46	0.00	0.45	2309
4 or more Prenatal Care Visits (=1)	0.25	0.01	0.44	0.01	0.44	2223
Institutional Delivery (=1)	0.50	0.07	0.06	0.08	0.04	2108
Use Modern Contraception (=1)	0.35	0.02	0.32	0.02	0.27	3154
Utilization by Children Age 0-23 Months						
Preventive Care Visit (=1)	0.48	0.13	0.02	0.12	0.03	1971
Curative Care Visit (=1)	0.34	-0.00	0.50	-0.02	0.30	986
Utilization by Children Age 24-47 Months						
Preventive Care Visit (=1)	.24	0.11	0.00	0.11	0.00	2902
Curative Care Visit (=1)	0.42	0.10	0.13	0.08	0.18	1229
Quality of Prenatal Care						
Tetanus Vaccine (=1)	0.66	0.05	0.06	0.05	0.07	2856
Clinical Protocol Z-Score	0.00	0.16	0.06	0.16	0.04	3826

Notes: P-Values are for one-sided tests of the null hypothesis that $\beta = 0$ and are calculated based on a WILD bootstrap with 999 draws. Controls for child utilization include the child's age and sex, maternal height, mother's age, whether the mother has completed primary school, whether the father lives in the household, whether the family has health insurance, the total number of household members, the number of household members under the age of 6, whether the household owns land, and dummy variables for quartiles of the household asset value. Age was entered as a series of dummy variables that represent one-month increments. Controls for the maternal utilization and quality regressions include whether the woman is younger than 20 years, is older than 35, has at least primary education, is currently married, in union or living with the partner. Controls also include the number of household members, the number of prior pregnancies, and the distance from the household to the health facility whether the household had health insurance, whether the household owns any land, and quartiles for the household asset value.

Table 8: Impact of P4P on Knowledge and Quality of Care

	Knowledge Z-Score (Standardized)		Quality of Care Z-Score (Standardized)			
	β	P-Value	β	P-Value	β	P-Value
P4P	0.40	0.12	0.13	0.01	0.07	0.14
Knowledge Z-Score			0.03	0.21	0.02	0.36
P4P * Knowledge in Top 50%					0.142	0.06
N Observations	294		3,709		3,709	

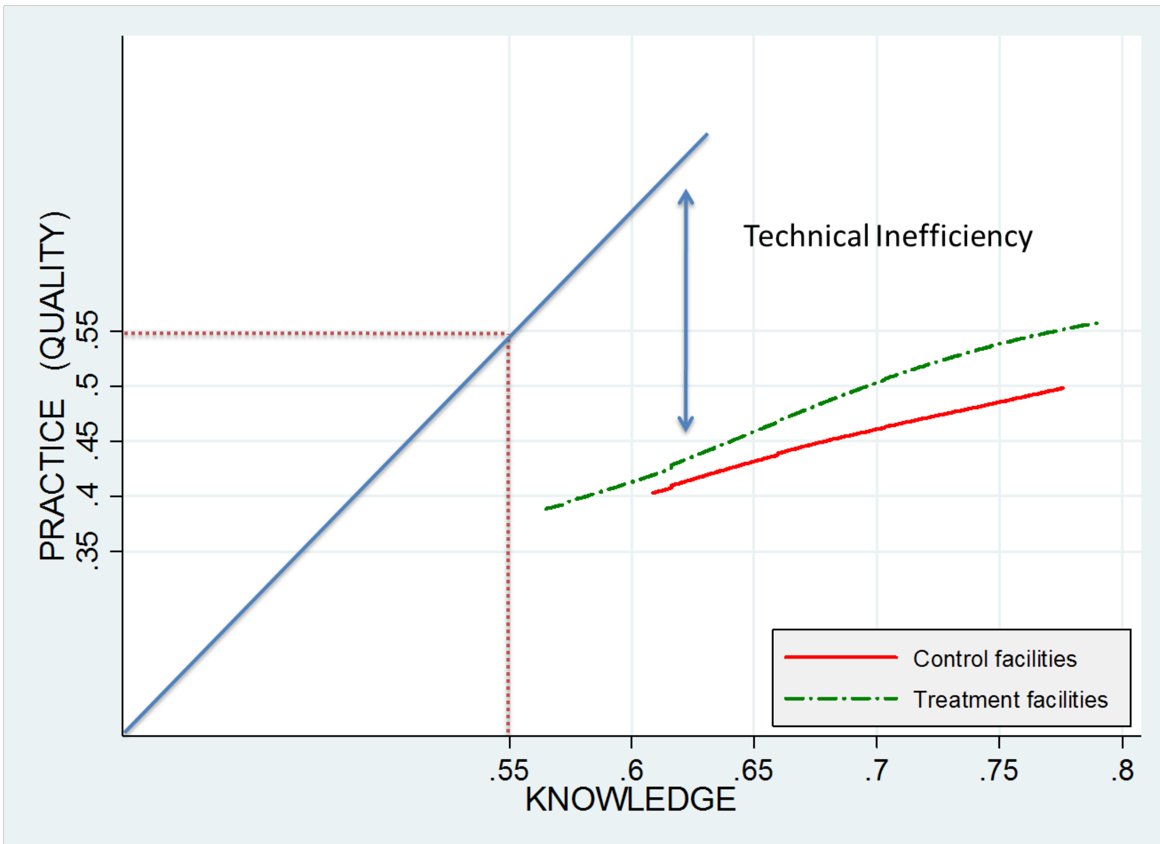
Notes: P-Values are for one-sided tests of the null hypothesis that β = 0 and are calculated based on a WILD bootstrap with 999 draws. Knowledge share is a continuous variable of the share of the CPG items that the provider knew and is bounded between zero and 1.

Table 9: Impact of P4P on Efficiency Gap (Knowledge Share – Quality Share)

	β	P-Value	β	P-Value	β	P-Value
P4P (=1)	-0.035	0.00	-0.04	0.03	-0.02	0.24
Knowledge Share			0.16	0.00	0.21	0.00
P4P * Knowledge in Top 50%					-0.06	0.01
N Observations	3709		3709		3709	

Notes: P-Values are for one-sided tests of the null hypothesis that β = 0 and are calculated based on a WILD bootstrap with 999 draws. Knowledge Share is the share of CPG protocol items correctly identified by the provider during the administration of the vignette. Quality Share is the percentage of CPG protocol items that were delivered during prenatal care, as reported in patient exit interviews and in household surveys.

Figure 1: The Knowledge-Quality Efficiency Gap for Prenatal Care (2008)



Notes: The horizontal axis is Knowledge expressed as the percentage of CPG protocol items correctly identified by the provider during the administration of the vignette. The horizontal axis is the percentage of CPG protocol items that were delivered during prenatal care, as reported in patient exit interviews and in household surveys.