Health Policy and Planning Advance Access published July 19, 2016

Health Policy and Planning, 2016, 1–10
doi: 10.1093/heapol/czw091
Original Manuscript

Health system changes under pay-for-performance: the effects of Rwanda's national programme on facility inputs

Diana KL Ngo,[1,*] **Tisamarie B Sherry**[2] **and Sebastian Bauhoff**[3]

[1]Department of Economics Occidental College, Fowler 223, 1600 Campus Rd, Los Angeles, CA 90041, USA, [2]Department of Medicine Brigham and Women's Hospital, 75 Francis St, Boston, MA 02115, USA and [3]Center for Global Development, 2055 L Street NW, Fifth Floor, Washington, DC 20036, USA

*Corresponding author. Department of Economics Occidental College, Fowler 223, 1600 Campus Rd, Los Angeles, CA 90041, USA. E-mail: dngo@oxy.edu

Accepted on 16 June 2016

Abstract

Pay-for-performance (P4P) programmes have been introduced in numerous developing countries with the goal of increasing the provision and quality of health services through financial incentives. Despite the popularity of P4P, there is limited evidence on how providers achieve performance gains and how P4P affects health system quality by changing structural inputs. We explore these two questions in the context of Rwanda's 2006 national P4P programme by examining the programme's impact on structural quality measures drawn from international and national guidelines. Given the programme's previously documented success at increasing institutional delivery rates, we focus on a set of delivery-specific and more general structural inputs. Using the programme's quasi-randomized roll-out, we apply multivariate regression analysis to short-run facility data from the 2007 Service Provision Assessment. We find positive programme effects on the presence of maternity-related staff, the presence of covered waiting areas and a management indicator and a negative programme effect on delivery statistics monitoring. We find no effects on a set of other delivery-specific physical resources, delivery-specific human resources, delivery-specific operations, general physical resources and general human resources. Using mediation analysis, we find that the positive input differences explain a small and insignificant fraction of P4P's impact on institutional delivery rates. The results suggest that P4P increases provider availability and facility operations but is only weakly linked with short-run structural health system improvements overall.

Key words: Health care financing, health care systems, incentive-based financing, pay-for-performance, Rwanda

Key Messages

- Provider presence and management improved in response to P4P but only partially account for the observed effects of P4P. Other unobserved inputs, such as provider effort, may be important and would be helpful additions to facility monitoring instruments.
- P4P may not be as successful in improving structural aspects of the health system, such as the availability of basic supplies and equipment.
- P4P may shift resources away from certain inputs that may be important for quality improvement, as shown by a negative effect on delivery statistics monitoring.

Introduction

Improving the quality of health care services at sustainable cost remains an important priority in middle- and low-income countries. One approach that has been gaining popularity in recent years is paying-for-performance (P4P) (Honda 2013; Miller and Babiarz 2013). Under P4P, health care providers receive financial rewards for delivering contracted services or achieving targeted health outcomes. The specific approaches to achieving these targets are left to providers' discretion (Eichler 2006; Miller and Babiarz 2013). Under a Donabedian (1988) framework for quality improvement, P4P could foster changes in both the structural (i.e. setting characteristics, such as equipment and management) and process (i.e. patient and provider interactions, such as giving a diagnosis or recommending treatment) dimensions of health care that contribute to health outcomes.

To date, the evidence on P4P is limited but growing and has focused primarily on the impacts on targeted measures (Honda 2013). Recently, there has been research on programme effects in resource-limited settings that shows positive effects on many, but not all, incentivized measures, e.g. in Indonesia (Olken 2012), the Philippines (Peabody *et al.* 2011, 2014), the Democratic Republic of the Congo (DRC) (Huillery and Seban 2013), Cambodia (Van de Poel *et al.* 2015), Burundi (Bonfrer *et al.* 2014), Tanzania (Borghi *et al.* 2015) and Rwanda (Basinga *et al.* 2011; De Walque *et al.* 2015; Sherry *et al.* 2015). P4P programmes are likely to have impacts that extend beyond targeted services and outcomes, yet these impacts on untargeted dimensions of the health system remain relatively unexplored. In theory, by rewarding a subset of services, P4P programmes may divert inputs from unrewarded services, a phenomenon known as 'multitasking' (Holmstrom and Milgrom, 1991; Sherry, 2015). Conversely, P4P can potentially stimulate changes in health care inputs that have positive spillovers across rewarded and unrewarded services, thereby strengthening health systems more broadly (Mullen *et al.* 2010). Within this health systems framework, experts have called for more rigorous evidence on (1) how providers modify health care inputs to achieve performance gains (Eldridge and Palmer 2009; Ireland *et al.* 2011) and (2) how P4P affects health system quality, including untargeted structural dimensions of care (Lagarde *et al.* 2010; Witter *et al.* 2012).

Empirical evidence on factors associated with P4P's success and on P4P's impacts on broader organizational change is limited. Existing studies focus on a small subset of structural and process inputs. Olken *et al.* (2012) find that quality improvements in Indonesia's P4P programme were mediated in part by more efficient spending and increased labour supply of providers. Huillery and Seban (2013) find that in the DRC's P4P programme, quality improvements were mediated by increased provider effort (e.g. preventive health sessions and community outreach) and decreased absenteeism. In Cambodia, improvements in targeted maternal and child health measures were associated with improvements in health centre management, notably increased availability of 24-h medical care, increased supplies and equipment, decreased provider absenteeism and increased supervisory visits (Bloom 2006). In Rwanda, Gertler and Vermeersch (2012) shows that P4P improved adherence to clinical guidelines, while Kalk *et al.* (2010) find both positive and negative reports on infrastructure and management quality from semi-structured interviews with key informants. Finally, under both Rwanda and Nicaragua's P4P schemes, providers have used community outreach campaigns to recruit patients and increase service provision (Regalia and Castro 2009). Evidence to date therefore suggests that P4P may prompt modifications to both structural attributes of care settings and health care processes, but the evidence remains limited, particularly along the structural dimensions of care.

In this article, we examine P4P's impacts on a large set of structural health system inputs drawn from international and national guidelines, using Rwanda's national programme as a case study. In doing so, we address two key questions raised in the literature. First, how do providers modify health inputs to achieve performance gains? Second, what are P4P's effects on broader health systems, including both targeted and untargeted dimensions of care? We extend the previous literature on P4P's impacts on health services and outcomes (Basinga *et al.* 2011; Gertler and Vermeersch 2012; De Walque *et al.* 2015; Okeke and Chari, 2015; Sherry *et al.* 2015;), by focusing on impacts on both targeted and untargeted health system inputs. The analysis informs policy by identifying potential mediators of P4P's impacts and examining P4P's potential to enhance or detract from overall health system quality, information which can in turn be used to strengthen the design of P4P programmes and their interaction with other elements of a complex health system.

Background: Rwanda's P4P programme

As one of the earliest large-scale P4P programmes in a developing country, Rwanda's 2006 national programme offers a valuable case study for examining provider responses to P4P. Rwanda's programme was designed as one of three primary 'quality strategies' pursued by the Rwandan Ministry of Health (MOH); the other strategies included a quality assurance programme and a community-based health insurance programme, implemented in 1998 and 2000, respectively (National Institute of Statistics 2008; Saksena 2010).

Rwanda's P4P programme has two components, as described in Basinga *et al.* (2011). First, the programme rewards facilities with varying unit payments for a set of 24 health service indicators in the domains of maternal health, child health, family planning and HIV/AIDS. Second, bonus payments are adjusted according to overall facility quality using a quality multiplier, which is constructed by measuring facility performance across several domains, including general administration, financial management and hygiene and sanitation. Within each of these domains, performance assessments take into account the availability of key inputs (i.e. structural attributes) and adherence to clinical protocols (i.e. process measures). In this way, Rwanda's P4P programme incentivizes overall facility quality (Sherry *et al.* 2015). Details on the P4P incentive payments and quality multiplier components are presented in the Supplementary Appendix. Payments are disbursed at the facility level and are used at the discretion of each facility. The magnitude of disbursements in the initial phases was large, with payments resulting in an average increase in expenditures of 22% above 2006 levels (Basinga *et al.* 2011).

The P4P programme was instituted by the Rwandan MOH through a quasi-randomized, phased roll-out. In 2006, districts without existing P4P pilot programmes were assigned to either treatment or control status by block-randomization. A government redistricting process occurred shortly before the roll-out of P4P, resulting in the merging of several control and P4P pilot districts. Consequently, these districts were reassigned by the MOH to treatment status. The final quasi-randomization resulted in 12 treatment and 7 control districts. The P4P programme was launched in treatment districts in 2006 and expanded to control districts in 2008 but from 2006 to 2008 control facilities received budget increases equal to the average payments paid out to treatment facilities. This allowed policymakers to distinguish the effect of incentives (payments conditional on performance) from unconditional increases in facility budgets by a similar amount.

Previous studies on the impact of Rwanda's P4P programme have found mixed impacts on health service provision and outcomes. The programme increased more generously rewarded services such as institutional deliveries, contraceptive supply and HIV testing but had no impact on less generously rewarded services (Basinga *et al.* 2011; De Walque *et al.* 2015; Sherry *et al.* 2015). Several unrewarded prenatal care services also increased, and there were no negative spillovers to other unrewarded services (Sherry *et al.* 2015). Among health outcomes, Gertler and Vermeersch (2012) find improvements in infant weight-for-age and young children's height-for-age, whereas Okeke and Chari (2015) are unable to reject a null effect of the programme on infant mortality.

Together, the programme's mixed effects raise the question of how Rwandan providers and the health system as a whole responded to P4P incentives. The successes suggest that providers effectively modified certain dimensions of care, while the null effects suggest limited effects on other potentially important inputs.

Methods

To analyse how providers achieved performance gains and how P4P affects broader health systems, we select a set of structural inputs and identify how these inputs responded to P4P. We use a simple differences linear regression analysis, regressing the inputs of interest on P4P treatment, controlling for facility characteristics that are unlikely to change in the short-run. Under randomization, the differences between treatment and control facilities in the post-period should represent causal effects of the P4P programme. Specifically, each input is the dependent variable in a separate regression, where the independent variable of interest is an indicator for being in a P4P treatment district and the controls are the log catchment population and indicators for managing authority, province, health post, clinic type and funding type. We perform the analysis using Stata. Files for replicating the analysis are available at https://dataverse.harvard.edu/dataverse/cgdev.

We select the inputs based on availability in the data using guidance from the components of the Rwandan P4P quality multiplier and the World Health Organization (WHO) Service Availability and Readiness Assessment (Ministry of Health Contractual Approach Unit 2008; World Health Organization 2012). We also combine all inputs in substantively similar categories into indices to elicit broader investment patterns. Analytically, combining individual inputs into indices reduces the number of tests we perform and is a strategy used by others to address multiple inference concerns (Kling *et al.* 2007). Conceptually, the indices capture provider responses that are common within categories even though responses may be heterogeneous for specific inputs. To construct the indices, we equally weight the normalized individual inputs following the method used by Kling *et al.* (2007).

Delivery-specific inputs

Among the rewarded services, institutional delivery rates showed the largest positive response to P4P (Basinga *et al.* 2011; Sherry *et al.* 2015), providing a natural case study for examining the first question of how providers achieve performance gains. Specifically, we analyse structural inputs directly associated with deliveries, using these inputs as the dependent variables in the regressions specified earlier.

Within delivery-specific inputs, we group variables into physical resources, human resources and operations. Within delivery-specific physical resources, we analyse the number of maternity beds, the availability of emergency transport and an indicator for delivery equipment. This delivery equipment indicator is constructed based on the availability of 17 items: exam table, exam light, infant scale, sterilized instruments, neonatal aspirator, obstetrical stethoscope, suture thread, ophthalmic ointment, local anaesthesia, sterile gloves, umbilical cord clamp, skin disinfectant, injectable diazepam, intravenous solution with infusion set, injectable antibiotic, injectable magnesium sulphate and injectable uterotonic. Within delivery-specific human resources, we analyse variables for maternity staff availability, maternity community health workers, number of midwives and availability of doctors and nurses. Within delivery-specific operations, we analyse delivery room privacy, the availability of antenatal services and an indicator for delivery statistics monitoring. The delivery statistics monitoring indicator combines evidence of delivery statistics monitoring, meetings to discuss delivery statistics and meetings to discuss adverse deliveries. For each input constructed from multiple items, we use principal components analysis (PCA) to extract the most variation from the data (Dunteman 1989). PCA is commonly used to combine related variables into summary measures (Filmer and Pritchett 2001).

From the delivery-specific inputs, we generate three indices for delivery-specific physical resources, human resources and operations. For each index, we combine all inputs within the relevant category using the equal weighting method described earlier. For example, we combine maternity beds, emergency transport and delivery equipment/medication into an index for delivery-specific physical resources and repeat the procedure for delivery-specific human resources and operations.

General structural inputs

To address the second question concerning the effects of Rwanda's P4P programme on broader health system performance, we examine impacts on a set of more general health care inputs, again grouping variables into physical resources, human resources and operations. Within general physical resources, we analyse an indicator for sanitation supplies, constructed from soap and disinfectant availability and an indicator for basic equipment, constructed based on the availability of six items: electricity, water dispensers, a clean water source, a functioning incinerator, functioning sterilizing equipment and beds per capita. Within general human resources, we analyse the total number of staff, staff availability and hours worked. Within general operations, we analyse clinic cleanliness, the presence of covered waiting areas and overall management. Clinic cleanliness is constructed from the availability of trash bins, the availability of sharps containers and cleanliness of surfaces. The management indicator is constructed from full-time service provision, minutes for monthly management meetings, medical systems reports, quality assurance reports, minutes for monthly community meetings and routine equipment maintenance.

We also combine the general inputs into broader indices for the categories of physical resources, human resources and operations. In total, we test 18 inputs and 6 associated indices for delivery-specific physical resources, delivery-specific human resources, delivery-specific operations, general physical resources, general human resources and general operations.

Robustness tests

We employ additional tests to determine the sensitivity of our results to our variable definitions and model specifications. First, we compute *P* values that account for the overall number of inputs we examine to address concerns about statistical inference with multiple comparisons. We follow the method proposed by Benjamini *et al.* (2006) to control the false discovery rate (FDR). Details of the rationale and algorithm for the FDR-control correction procedure and the corresponding adjusted *P* values for each input are shown in the Supplementary Appendix.

Second, we test a variation of the six composite indices, using PCA instead of equal weighting to combine the individual inputs. We also run additional regression specifications with different subsets of controls.

Third, we address the concern that we are unable to control for each facilities' baseline values in a difference-in-difference framework. We use Demographic Health Survey (DHS) household data to assess whether our simple differencing approach leads to similar findings as double-differencing. Specifically, we compare treatment effects for institutional deliveries using a difference-in-differences regression to the effect identified by a post-treatment simple-difference regression.

Fourth, we address the concern that the short-run dataset may limit our ability to identify programme impacts if facilities responded slowly. The facility data were collected a year after programme exposure, in contrast to the 18- to 24-month-exposure associated with the household data used in other studies (Basinga *et al.* 2011; Sherry *et al.* 2015). As a robustness check, we identify the effect of P4P on institutional delivery rates using the facility data and compare it with previous studies. If provider responses occurred within the first year and remained stable thereafter, the increase in deliveries identified in the facility data should correspond closely to the impact identified using later follow-up data.

Mediation analysis

After identifying which inputs change in response to P4P incentives, we perform a mediation analysis to quantify how much the input differences contribute to differences in institutional delivery rates. We use the method developed by Imai *et al.* (2011) to calculate the average causal mediation effects (ACMEs) of intermediate mechanisms on final outcomes. The mediation analysis is done by fitting two models, a model of the mediator as a function of treatment and covariates and a model of the outcome as a function of the mediator, treatment and covariates. The mediation effects are computed as the difference between the outcome predicted under treatment when the mediator is also predicted under treatment and the outcome predicted under treatment when the mediator is predicted using control conditions. Average mediation effects are computed by repeating the prediction step under different values of the model parameters drawn from the parameter distribution estimated in the initial step. We present the basic equations in the Supplementary Appendix. The method assumes that there is no reverse causality between the mediator and outcome and that there are no unmeasured factors that affect both the mediator and the outcome. Compared with the traditional linear structural equation modelling framework popularized by Baron and Kenny (1986), Imai *et al.* (2011) provides an estimation strategy with fewer parametric assumptions and additional methods for sensitivity analyses.

To date, this method has been used to examine mediators underlying voter behaviour (Karpowitz *et al.* 2012; Nyhan *et al.* 2012), mental health outcomes (Varese *et al.* 2012; Walters 2011; Sagatun *et al.* 2014;) and health disparities (Litzelman *et al.* 2014; Andersen *et al.* 2015; Garawi *et al.* 2015). Linden and Karlson (2013) advocate for the increased use of mediation analysis within health systems research, and in a comparison of alternative mediation analysis methods, they find that the Imai *et al.* method is among the best performing.

We also apply the sensitivity analysis provided by Imai *et al.* (2011) to measure how the ACME is likely to change in the presence of a confounding omitted variable. For example, if an unobserved factor like intrinsic motivation is positively related to both a mediator (general operations) and the outcome (institutional delivery rates), the estimated ACME will be biased. The associated sensitivity parameter captures how much an omitted variable must alter the R-squared statistic of the two models to result in a statistically insignificant ACME. Specifically, the parameter equals the product of the proportions of total variance in the mediator and outcome models explained by the hypothesized unobserved confounder that would result in a true ACME of zero. An ACME that is no longer statistically significant under small changes in the R-squared statistics is less robust.

Data

We use data from the 2007 Rwanda Service Provision Assessment (SPA) Survey. The SPA instruments are designed to monitor health care systems in developing countries by assessing service availability, facility readiness, adherence to clinical protocols and client satisfaction. Key topics include infrastructure, management systems and provision of services related to maternal and child health, family planning, sexually transmitted infections and communicable and non-communicable diseases. The 2007 SPA represents a census of all public health facilities, a census of all private facilities with five or more staff and a sample of private facilities with three or more staff (National Institute of Statistics 2008). Data collection occurred from June through October 2007, roughly a year after P4P was introduced in treatment areas but before it was introduced in control districts. This allows us to identify short-run differences in health care inputs between treatment and control areas.

The SPA public-release files do not contain geographic identifiers for facilities, but the Measure DHS team provided us with each facility's treatment category based on its geographic location (i.e. in treatment vs control districts), specifically for this analysis. Treatment assignment occurs within provinces, and we include province identifiers in our analyses. We exclude hospitals, which were subject to a different incentive scheme, and facilities from the P4P pilot districts. Our final sample consists of 201 treatment facilities and 101 control facilities. Because the SPA includes all public health facilities, our sample is larger than that previously analysed by Basinga *et al.* (2011) and Kalk *et al.* (2010), who surveyed a subsample of facilities among these.

Table 1 provides summary statistics for the analysed sample. Sixty-five percent of the sampled facilities are publicly managed, while the remaining facilities are classified as private, non-governmental or community-run. The majority of facilities is polyclinics, which are expected to provide a full range of basic services. The remaining facilities are health posts, situated in more remote areas and offering fewer services (National Institute of Statistics 2008). The vast majority of both polyclinics and health posts have general outpatient care clinics, while only 5.1% of the facilities have inpatient medical clinics. There is large variation in catchment populations and per capita spending within each facility type. Polyclinics, health posts and public and private facilities were subject to the same P4P incentives.

Balance tests

To address district-level differences and potential bias introduced by redistricting, we want to ensure balance between treatment and control facilities on pre-treatment characteristics. Since this data are not available to us, we perform balance tests on post-treatment characteristics that are unlikely to have changed in the first year of the programme. Table 2 presents the *t*-statistics for each characteristic as well as Hotelling's *t*-squared and associated *P* value for multivariate tests on groups of related variables. The results indicate that long-run facility characteristics are comparable across treatment status for ownership, clinic availability, funding sources, catchment population and 2006 spending. Of the 21 characteristics we test, we find differences in two: the fraction of health posts and facilities in Northern Province; we account for these differences by including all fixed facility characteristics as regression controls.

Table 1. Sample summary statistics

	Polyclinic	Health post	Total
Number of facilities			
Management type			
Public	185	12	197
Private/non-governmental organization/community	71	34	105
Province			
Northern	60	9	69
Southern	50	3	53
Eastern	78	12	90
Western	67	22	89
Kigali City	1	0	1
Clinic availability			
Has general outpatient clinic	249	46	295
Has inpatient medical clinic	16	0	16
Group means and standard deviations			
Catchment population	21 088	16 856	20 738
	[11 042]	[31 674]	[13 887]
Spending/catchment population	2696	10 210	3327
	[17 235]	[29 308]	[18 583]

Standard deviations in square brackets.

Results

We present the programme impacts on facility inputs in Table 3. Among delivery-specific inputs, P4P had no significant impact on indices capturing physical resources, human resources and operations. Similarly, the treatment effects on the delivery-specific index components are statistically insignificant with two exceptions. P4P increased maternity-related staff presence by 29% (increase of 0.28 relative to control mean of 0.97; $P = 0.02$) and decreased delivery statistics monitoring by 11% ($P = 0.02$) relative to the control mean, with FDR-adjusted *P* values of $P = 0.10$ for both inputs.

Among general structural inputs, P4P increased the general operations index by 9% over the control mean ($P < 0.01$) but had no significant effect on indices capturing general physical resources and general human resources. The impact on general operations is driven by a 7% increase in the presence of covered waiting areas ($P = 0.01$) and an 11% increase of the general management indicator ($P = 0.01$) relative to the control mean. The FDR-adjusted *P* values for covered waiting areas and general management are $P = 0.10$. The treatment effects for the remaining general inputs are statistically insignificant and of mixed signs.

Supplementary analyses on the impacts of P4P on facility staff show that P4P decreased the number of non-medical, non-managerial support staff by 47% relative to the control mean ($P < 0.01$). This group accounted for approximately one-fifth of the total staff in control areas and was ~50% lower in treated facilities.

The robustness analyses using indices generated with alternative weighting methods and different subsets of controls suggest that the results are not sensitive to the regression specifications or the

Table 2. Facility characteristics: balance tests

	Control mean	Difference	*t*-stat	Hotelling's *t*-squared
Adjacent to main facility	0.07	0.06	1.47	
Catchment population	20 492.43	376.46	0.22	
Health post	0.07	0.12	2.88	
Per capita spending, 2006	1557.57	2681.32	1.08	
Private/non-governmental organization/community	0.33	0.03	0.54	
Clinic availability				5.06 ($P = 0.42$)
Antenatal care	0.18	−0.01	−0.31	
General outpatient	0.96	0.02	1.34	
Inpatient medical	0.05	0.01	0.19	
Inpatient/outpatient TB	0.11	−0.05	−1.71	
voluntary counseling and testing/HIV/special diagnoses	0.21	−0.03	−0.60	
Funding source				10.13 ($P = 0.13$)
Employer	0.06	0.05	1.41	
Equity fund for poor	0.15	0.09	1.73	
Government risk pool	0.70	−0.09	−1.48	
Insurance	0.56	−0.09	−1.42	
Other	0.06	−0.01	−0.55	
User fees only	0.12	0.07	1.45	
Province				11.76 ($P = 0.02$)
Eastern	0.32	−0.03	−0.51	
Kigali city	0.01	−0.01	−1.41	
Northern	0.13	0.15	2.93	
Southern	0.23	−0.08	−1.69	
Western	0.32	−0.03	−0.60	

Table 3. Treatment effects on structural inputs

	Control mean	β	se	P-val	N
Indices (components later, equally weighted)					
Delivery physical resources	0.31	0.01	0.02	0.73	178
Delivery human resources	0.25	0.01	0.01	0.55	238
Delivery operations	0.62	−0.02	0.02	0.35	233
General physical resources	0.63	−0.02	0.03	0.53	213
General human resources	0.23	−0.00	0.01	0.83	273
General operations	0.69	0.06	0.02	0.00	250
Delivery-specific					
Physical Resources					
Maternity beds/1000 pregnant women	10.96	1.57	1.08	0.15	239
Transport for obstetric emergencies	0.13	−0.05	0.03	0.18	255
Delivery equipment and medication indicator[a]	0.73	0.03	0.03	0.27	178
Human resources					
No. maternity-related staff present today/10 000 people	0.97	0.28	0.12	0.02	275
Has community health worker, delivery	0.78	0.05	0.05	0.32	275
No. midwives/10 000 people	0.03	0.00	0.02	0.94	275
Doctor/A1 nurse present for deliveries	0.05	−0.02	0.02	0.46	238
Operations: management and responsiveness					
Private delivery room (auditory and visual)	0.82	0.02	0.06	0.66	236
Antenatal care services: days per month provided	7.37	0.15	0.52	0.78	266
Delivery stats monitoring indicator[a]	0.81	−0.09	0.04	0.02	237
General					
Physical resources					
Sanitation supplies indicator[a]	0.64	0.01	0.03	0.81	275
Basic equipment indicator[a]	0.61	−0.04	0.03	0.28	213
Human resources					
No. staff/10 000 people	9.95	−0.64	0.83	0.44	275
No. staff present today/10 000 people	5.74	0.57	0.56	0.31	275
Average hours/week worked in facility[b]	48.12	0.82	0.84	0.33	896
Operations: management and responsiveness					
Cleanliness of clinics[a]	0.58	0.03	0.02	0.21	275
Covered waiting areas	0.87	0.06	0.02	0.01	275
Management indicator[a]	0.61	0.07	0.03	0.01	250

Regressions controls include: log catchment population and indicators for managing authority, province, health post, clinic type and funding type.

[a] Components described in Supplementary Appendix and combined using PCA.

[b] Hours worked from health worker interviews; regression for hours worked includes individual covariates and clustering by facility.

Table 4. Robustness tests: institutionalized deliveries across datasets

	Mean (control)	β	se	P-val	N	Data
DHS, simple-difference (post-period only)	0.42	0.11	0.03		1183	Individual
DHS, difference-in-differences (Sherry *et al.*)	0.30	0.10	0.04		5657	Individual
Basinga *et al.*	0.36	0.07		0.03	2108	Individual
SPA facility data	0.40	0.10	0.03	0.002	237	Facility

Number of pregnancies calculated for SPA data from catchment population using WHO service availability indicator guidelines. Mean control reported for pretreatment year in Sherry *et al.* 2015 and Basinga *et al.* 2011 and post-treatment year in the SPA data and the DHS simple-difference model.

variable definitions. The supplementary analyses and results of these robustness checks are shown in the Supplementary Appendix.

Table 4 presents the results from the robustness checks analysing the effects of P4P on institutional delivery rates. The simple-difference DHS household results are statistically indistinguishable from the difference-in-differences analysis. Similarly, the simple-difference SPA result corresponds closely with the independent estimates from household surveys (Basinga *et al.* 2011; Sherry *et al.* 2015), showing that P4P increased institutional delivery rates by 10.4 % points or 26% ($P < 0.01$).

Table 5 shows the ACMEs for the three inputs associated with positive and significant treatment effects: daily presence of maternity-related staff, covered waiting areas and the management indicator. The results are small and statistically insignificant for all three inputs. The ACME for the general operations index is marginally significant at the 10% level and indicates that adjustments in general operations account for 14% of the increase in institutional delivery rates. However, the sensitivity analysis suggests that the result is very sensitive to the presence of unobserved confounding variables. Specifically, the sensitivity parameter implies that an unobserved confounder that explains a small proportion of the variance in the general operations index and a small proportion of the variance in facility delivery rates would result in a true ACME of zero. The sensitivity parameter is the product of these two

Table 5. Mediating increases in institutional deliveries

Mediating factor	Total effect (95% CI)	ACME (95% CI)	Frac. of total effect mediated (95% CI)	Sensitivity parameter[a]
Staff presence[b]	0.110 (0.045, 0.171)	−0.003 (−0.014, 0.007)	Not applicable[c]	0.001
Covered waiting areas	0.110 (0.045, 0.171)	0.000 (−0.010, 0.009)	Not applicable[c]	0.000
Management indicator	0.104 (0.035, 0.169)	0.002 (−0.012, 0.016)	0.016 (0.010, 0.048)	0.000
General operations index	0.104 (0.035, 0.172)	0.015 (0.00004, 0.034)	0.143 (0.086, 0.423)	0.012

[a] The sensitivity parameter captures the robustness of the ACME to an unobserved confounding variable. The parameter equals the product of the proportions of total variance in the mediator and outcome models explained by the hypothesized unobserved confounder that would result in a true ACME of zero; smaller parameters indicate higher sensitivity.

[b] Number of staff providing maternity-related services present today/10 000 people.

[c] The mediators are negatively associated with institutional deliveries and do not explain positive fractions of the total effect.

proportions; the computed value of 0.012 implies that a confounder explaining 10–12% of the variance in both models would result in an ACME of zero. This number is small relative to the examples presented in Imai *et al.* (2011), suggesting that the ACME is not robust to unobserved confounders.

Discussion

This study examines the impact of Rwanda's P4P programme on a range of structural inputs, providing a more complete view of P4P's mechanisms and effects on the broader health system. We find positive and significant effects of P4P on a general operations index but no effects on indices capturing delivery-specific inputs, general human resources or general physical resources. P4P had positive impacts on three underlying inputs: the daily presence of maternity-related staff, the presence of covered waiting areas and facility management. The mediation analysis shows that the differences in each of these three inputs explain a small and statistically insignificant fraction of the P4P impact on institutional delivery rates. The corresponding mediation analysis for the general operations index indicates that general operations may account for a moderate fraction of the increase in delivery rates but this estimate appears sensitive to the presence of relatively weak confounding factors. Finally, P4P had a negative impact on the monitoring of delivery statistics.

Our results strengthen and extend the existing evidence on P4P's effects on mediating factors and untargeted inputs. First, the findings suggests that P4P improves provider availability and facility management, supporting conclusions from earlier studies that P4P improved other dimensions of provider effort (Gertler and Vermeersch 2012) and may play a role in addressing the problems of low provider effort and high absenteeism in developing country health systems (Chaudhury *et al.* 2006; Meessen *et al.* 2007; Olken, 2012; Huillery and Seban, 2013). The results indicate that indirect targeting can be successful, as aspects of management were included in the quality multiplier but also implies that untargeted inputs like provider availability will respond if deemed important by providers. However, the mediation analysis indicates that changes in provider presence and management only partially account for the observed effects of P4P. Thus, while these may be potential mediators, factors other than those observed in the SPA may be important for programme success.

Second, this study is among the first to quantify P4P's effects on a broader range of untargeted inputs. The null effects suggests that P4P may have had a limited role in promoting structural quality improvements beyond targeted services, an issue raised by Honda (2013) and Lagarde *et al.* (2010). In contrast to the Salud Mesoamerica Initiative that included direct incentives for specific inputs such as equipment availability (Mokdad *et al.* 2015) and achieved large increases in these inputs, Rwanda's programme incentivized structural improvements more indirectly through the quality multiplier, leading to fewer structural responses. Even with institutional deliveries, we find that delivery-related input levels remain low overall in Rwandan facilities. Only 54% of facilities meet the WHO target of 10 maternity beds per thousand pregnant women, and on average, facilities have only 36% of suggested delivery instruments and medications.

The analysis on untargeted inputs also suggests that P4P might have had detrimental effects on some aspects of health care quality. P4P's negative effect on the monitoring of delivery statistics suggests that the programme might have shifted resources away from monitoring and evaluation of delivery outcomes in the interests of recruiting more clients and performing more deliveries. Similarly, the programme's negative effect on non-medical and non-managerial support staff supports the notion that resources were shifted away from less incentivized areas. These negative effects support qualitative findings reported by Kalk *et al.* (2010) but contrast with the majority of P4P studies that report positive impacts on a small subset of untargeted measures (Bloom 2006; Gertler and Vermeersch 2012; Olken 2012; Huillery and Seban 2013; Regalia and Castro 2009).

Our study has several limitations. First, treatment reassignments forced by redistricting are a potential source of bias. However, in the robustness analysis, the correspondence between the simple-difference and difference-in-differences models using the DHS household data supports the assumption of pre-treatment balance underlying our simple-difference approach. Moreover, other studies of Rwanda's P4P programme have shown pre-treatment balance in population and facility characteristics. Using population data, Basinga *et al.* (2011) find no differences in maternal and household demographics, quality of prenatal care and maternal care utilization, with the exception of an indicator for four or more prenatal visits. Using the DHS population data, Sherry *et al.* (2015) also find pre-treatment balance for a large set of rewarded services, unrewarded services and health outcomes. At the facility level, Gertler and Vermeersch (2012) find balance in pre-treatment staffing levels, budget shares towards personnel and supplies, structural quality indicators for various services and total expenditure levels. Together, the balance tests from the SPA and multiple independent datasets provide consistent evidence that the treatment and control areas were balanced on pre-treatment measures, despite the redistricting.

Second, the short study time span may limit our ability to identify impacts if programme effects are gradual. However, in the robustness analysis, the impact on institutional delivery rates using the SPA data is similar to that seen using other datasets with longer programme exposure. Moreover, other studies find short-run responses to P4P after 18 months, identifying improvements in health processes in Indonesia's programme (Olken, 2012) and gains in structural attributes in the Mesoamerican programme (Mokdad 2015). In Mexico, these short-run input responses were large; the fraction of health facilities with inputs and equipment for pre- and postnatal care (e.g. gynecological exam tables and obstetric tape) increased from 3.6 to 45.8%. Similar changes were observed in Nicaragua, Belize, El Salvador, Panama and Honduras for many of the same inputs included in this study (Mokdad 2015), demonstrating that some structural inputs changes can be achieved quickly. Together, these considerations suggest that the programme exposure period is sufficient for identifying some short-run effects.

Third, the sample sizes may limit the statistical power of our analysis. However, one-half of the coefficients on the indices are negative and the remaining positive coefficients are small in magnitude. Specifically, the upper bounds of the 95% confidence intervals (CIs) reject treatment effects larger than 13, 15, 4, 6, 4 and 4% above the control mean for the indices capturing delivery-specific physical resources, delivery-specific human resources, delivery-specific operations, general physical resources and general human resources, respectively. This suggests that the effects are small, at best, and that increased precision would unlikely change the general conclusions of the analysis. Related, we cannot cluster standard errors within districts due to the lack of district identifiers in the data. Clustering would likely increase the standard errors on the treatment coefficients since management is similar within districts. This suggests that our results are less conservative estimates of programme impacts. However, since the majority of our results is insignificant, these would likely remain insignificant with clustering.

Fourth, our analysis is limited to identifying systematic behaviour changes across providers. Fundamentally, P4P allows providers to choose their own approach, so it is possible that providers had idiosyncratic responses to the incentives. On the other hand, many facilities face similar constraints, such as shortages of trained health care providers (National Institute of Statistics 2008), and may respond similarly to common challenges. Furthermore, our indices capture broad categories to account for heterogeneity in provider adjustments of specific inputs.

Finally, the assumptions underlying the mediation analysis cannot be tested and limit our ability to interpret the mediation effects as causal. Since we use cross-sectional data, reverse causality is possible, as higher institutional delivery rates may feed back into system changes. However, the short time span of the data decrease the likelihood that we observe the downstream effects of increased institutional deliveries. It is also possible that unobserved factors such as intrinsic motivation or provider knowledge affect both institutional delivery rates and the identified mediators. Although we employ the sensitivity analysis to address these concerns, these potential violations limit our ability to interpret the mediation effects as causal. However, we apply the analysis to broadly quantify the relationships between mechanisms and outcomes. This is important given the dearth of studies exploring mechanisms underlying P4P and can encourage additional mediation studies when more detailed, longitudinal data become available.

Together, the results have several implications for policymakers and researchers interested in implementing P4P programmes in low-resource settings. Our results suggest that management trainings and mechanisms for decreasing provider absenteeism can potentially have synergistic effects when combined with P4P programmes, and align with the suggestion by Oxman and Fretheim 2009 that such programmes incorporate appropriate capacity support. This has been built into a recent programme in the DRC, which provides health facility managers with strategic support from consultants (Soeters *et al.* 2011). At the same time, programme design could also promote monitoring and evaluation of health care quality, as suggested by Eldridge and Palmer (2009), to protect against detrimental effects on untargeted aspects of quality. Alternatively, P4P could incentivize health outcomes rather than service provision, though more recent studies have identified important limitations to this outcome-based approach as well (Mohanan *et al.* 2015). Furthermore, incentives for broader system improvements, complementary investments or other policy measures may be needed to address the persistent challenge of low health system readiness.

For researchers, the limited structural input responses suggest that additional research is necessary to systematically assess P4P's impact on other aspects of the health system. Although the SPA includes comprehensive information on structural attributes, it collects limited information on processes like provider effort and patient–provider interactions. As noted earlier, the hypothesis that P4P alters health care processes is consistent with previous research that finds positive P4P impacts on processes such as community outreach effort and adherence to clinical protocols (Regalia and Castro, 2009; Gertler and Vermeersch 2012; Huillery and Seban 2013). In Rwanda, specifically, Basinga *et al.* (2011) report that a large fraction of the bonus payments went to increased personnel compensation, again suggesting that P4P may work by shifting provider behaviour. Future research is therefore necessary to systematically assess P4P's effects on care processes. For this, health system monitoring instruments like the SPA surveys could be modified to include detailed information on process inputs.

Conclusion

Our analysis of facility input responses provides evidence on how P4P affects facility resources and investment patterns, showing that P4P can improve management and provider presence but may do little to influence many other structural inputs, at least in the short-run. Although much remains to be learned about the health care production function, the results suggest that P4P programmes can be helpful strategies to increase the delivery of specific health care services but are unlikely to strengthen all aspects of the health system. To address this broader challenge, other health policy strategies will be needed to supplement P4P.

Funding

Funding for this work was provided by the RAND Corporation Bing Center for Health Economics.

Ethics approval

This work was done using publicly available data collected by the Rwandan Ministry of Health and Macro DHS. Geographical identifiers were removed from the data prior to analysis due to confidentiality concerns.

Supplementary Data

Supplementary data are available at *HEAPOL* online.

Conflict of interest statement. None declared.

References

Andersen JP, Zou C, Blosnich J. 2015. Multiple early victimization experiences as a pathway to explain physical health disparities among sexual minority and heterosexual individuals. *Social Science & Medicine* **133**: 111–9.

Baron RM, Kenny DA. 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51**: 1173–82.

Basinga P, Gertler PJ, Binagwaho A *et al*. 2011. Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. *The Lancet* **377**: 1421–8.

Benjamini Y, Krieger AM, Yekutieli D. 2006. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**: 491–507.

Bloom E., Bushan I, Clingingsmith D *et al*. 2006. Contracting for health: evidence from Cambodia. Brookings Institution, Washington, DC.

Bonfrer I, Soeters R, Van de Poel E *et al*. 2014. Introduction of performance-based financing in Burundi was associated with improvements in care and quality. *Health Affairs* **33**: 2179–87.

Borghi J, Little R, Binyaruka P, Patouillard E, Kuwawenaruwa A. 2015. In Tanzania, the many costs of pay-for-performance leave open to debate whether the strategy is cost-effective. *Health Affairs* **34**: 406–14.

Chaudhury N, Hammer J, Kremer M, Muralidharan K, Rogers FH. 2006. Missing in action: teacher and health worker absence in developing countries. *The Journal of Economic Perspectives* **20**: 91–116.

De Walque D, Gertler PJ, Bautista-Arredondo S *et al*. 2015. Using provider performance incentives to increase HIV testing and counseling services in Rwanda. *Journal of Health Economics* **40**: 1–9.

Donabedian A. 1988. The quality of care: how can it be assessed?. *JAMA* **260**: 1743–8.

Dunteman GH. 1989.*Principal Components Analysis*, Vol. 69. Newbury Park, CA: Sage.

Eichler R. 2006. Can pay for performance increase utilization by the poor and improve the quality of health services. In: *Background papers for the Working Group on Performance Based Incentives* Center for Global Development, Washington, DC.

Eldridge C, Palmer N. 2009. Performance-based payment: some reflections on the discourse, evidence and unanswered questions. *Health Policy and Planning* **24**: 160–6.

Filmer D, Pritchett LH. 2001. Estimating wealth effects without expenditure data or tears: an application to educational enrollments in states of India*. *Demography* **38**: 115–32.

Garawi F, Ploubidis GB, Devries K, Al-Hamdan N, Uauy R. 2015. Do routinely measured risk factors for obesity explain the sex gap in its prevalence? Observations from Saudi Arabia. *BMC Public Health* **15**: 254.

Gertler P, Vermeersch C. 2012. Using Performance Incentives To Improve Health Outcomes. In: *World Bank Policy Research Working Paper, Impact Evaluation Series*, 6100.60. The World Bank, Washington, DC.

Holmstrom B, Milgrom P 1991. Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization* **7**: 24–52.

Honda A. 2013. 10 best resources on pay for performance in low-and middle-income countries. *Health Policy and Planning* **28**: 454–7.

Huillery E, Seban J. 2013. Performance-Based Financing for Health: Experimental Evidence from the Democratic Republic of Congo. In: *Working paper*. ENS-France Working Paper.

Imai K, Keele L, Tingley D, Yamamoto T. 2011. Unpacking the black box of causality: learning about causal mechanisms from experimental and observational studies. *American Political Science Review* **105**: 765–89.

Ireland M, Paul E, Dujardin B. 2011. Can performance-based financing be used to reform health systems in developing countries?. *Bulletin of the World Health Organization* **89**: 695–8.

Kalk A, Paul FA, Grabosch E. 2010. Paying for performance in Rwanda: does it pay off?. *Tropical Medicine & International Health* **15**: 182–90.

Karpowitz CF, Mendelberg T, Shaker L. 2012. Gender inequality in deliberative participation. *American Political Science Review* **106**: 533–47.

Kling JR, Liebman JB, Katz LF. 2007. Experimental analysis of neighborhood effects. *Econometrica* **75**: 83–119.

Lagarde M, Powell-Jackson T, Blaauw D. 2010. Managing incentives for health providers and patients in the move towards universal coverage. Background paper for the Global Symposium on Health Systems Research., 16–19 November 2010. Montreux, Switzerland.

Linden A, Karlson KB. 2013. Using mediation analysis to identify causal mechanisms in disease management interventions. *Health Services and Outcomes Research Methodology* **13**: 86–108.

Litzelman K, Skinner HG, Gangnon RE *et al*. 2014. Role of global stress in the health-related quality of life of caregivers: evidence from the Survey of the Health of Wisconsin. *Quality of Life Research* **23**: 1569–78.

Meessen B, Kashala J-PI, Musango L. 2007. Output-based payment to boost staff productivity in public health centres: contracting in Kabutare district, Rwanda. *Bulletin of the World Health Organization* **85**: 108–15.

Miller G, Babiarz KS.2013. Pay-for-performance incentives in low-and middle-income country health programs. *Tech. rep*. National Bureau of Economic Research.

Ministry of Health Contractual Approach Unit. 2008. Quarterly quality assessment grid for health centers. *Tech. rep*. Rwanda Ministry of Health.

Mohanan M, Miller G, Donato K, Truskinovsky Y *et al*. 2015. Input-based versus output-based incentive contracts in health care: experimental evidence from in India. Conference presentation. iHEA: International Health Economics Association, Milan.

Mokdad AH, Colson KE, Zúñga-Brenes P *et al*. 2015. Salud Mesoamérica 2015 initiative: design, implementation, and baseline findings. *Population Health Metrics* **13**: 3.

Mokdad AH. 2015. Salud Mesoamérica 2015 Initiative: Data for Better Health. http://www.healthdata.org/presentation/salud-mesoam%C3%A9rica-2015-initiative-data-better-health, accessed 1 November 2016.

Mullen KJ, Frank RG, Rosenthal MB. 2010. Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. *The Rand Journal of Economics* **41**: 64–91.

National Institute of Statistics (NIC) [Rwanda], Ministry of Health (MOH) [Rwanda], Macro International Inc. 2008. Rwanda Service Provision Assessment Survey 2007. *Tech. rep*. NIS, MOH, Macro International Inc.

Nyhan B, McGhee E, Sides J, Masket S, Greene S. 2012. One vote out of step? The effects of salient roll call votes in the 2010 election. *American Politics Research* **40**: 844–79.

Okeke EN, Chari AV. 2015. Can Institutional Deliveries Reduce Newborn Mortality? RAND Corporation Working Paper

Olken BA, Onishi J, Wong S. 2012. Should aid reward performance? Evidence from a field experiment on health and education in Indonesia. *Tech. rep*. National Bureau of Economic Research.

Oxman AD, Fretheim A. 2009. Can paying for results help to achieve the millennium development goals? Overview of the effectiveness of results-based financing. *Journal of Evidence-Based Medicine* **2**: 70–83.

Peabody J, Shimkhada R, Quimbo S *et al*. 2011. Financial incentives and measurement improved physicians quality of care in the Philippines. *Health Affairs* **30**: 773–81.

Peabody JW, Shimkhada R, Quimbo S. 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**: 615–21.

Regalia F, Castro L. 2009. Nicaragua: combining demand-and supply-side incentives. In: The Performance-Based Incentives Working Group, Eichler R, Levine R (eds). Performance Incentives for Global Health: Potential and Pitfalls. Washington, DC: Center for Global Development, 215–35

Sagatun Å, Heyerdahl S, Wentzel-Larsen T, Lien L. 2014. Mental health problems in the 10th grade and non-completion of upper secondary school: the mediating role of grades in a population-based longitudinal study. *BMC Public Health* **14**: 16.

Saksena P, Antunes AF, Xu K, Musango L, Carrin G. 2011. Impact of mutual health insurance on access to health care and financial risk protection in Rwanda.

Sherry TB, Bauhoff S, Mohanan M. 2016. Multitasking and Heterogeneous Treatment Effects in Pay for Performance in Health Care: Evidence from Rwanda. Forthcoming in American Journal of Health Economics.

Sherry TB. 2016. A note on the comparative statics of pay-for-performance in health care. *Health Economics* **25**(5): 637–644.

Soeters R, Peerenboom PB, Mushagalusa P, Kimanuka C. 2011. Performance-based financing experiment improved health care in the Democratic Republic of Congo. *Health Affairs* **30**: 1518–27.

Van de Poel E, Flores G, Ir P, O'Donnell O. 2015. Impact of performance-based financing in a low-resource setting: a decade of experience in Cambodia. *Health Economics* **25**: 688–705. ISSN: 1099-1050. DOI: 10.1002/hec.3219.

Varese F, Barkus E, Bentall RP. 2012. Dissociation mediates the relationship between childhood trauma and hallucination-proneness. *Psychological Medicine* **42**: 1025.

Walters GD. 2011. Criminal thinking as a mediator of the mental illness–prison violence relationship: a path analytic study and causal mediation analysis. *Psychological Services* **8**: 189.

Witter S, Fretheim A, Kessy FL, Lindahl AK. 2012. Paying for performance to improve the delivery of health interventions in low-and middle-income countries. *Cochrane Database of Systematic Reviews* **2**: CD007899.

World Health Organization. 2012. Measuring service availability and readiness: a health facility assessment methodology for monitoring health system strengthening. *Tech*. *rep*. World Health Organization.