Payment-by-Results for health interventions in low- and middle-income countries: A critical review

Maren Duvendack

Structured Abstract

Motivation: Payment-by-results (PbR) is now an important form of conditionality, whereby donors disburse development aid on achieving a pre-agreed measure.

Purpose: The article presents a critical review of PbR for health interventions, aiming to draw out lessons about the implementation and impact of PbR in international development programmes.

Approach and Methods: An extensive search identified 81 studies that varied widely in terms of health sub-themes, geographical regions and methodological quality.

Findings: Employing the Measure-Agent-Principal (MAP) framework we find that governments are the dominant principal most of these studies, and health facilities and individuals the dominant agent; 60% of the evidence reports a wide range of heterogeneous output-level measures.

Policy implications: We assess PbR in the context of health to better understand whether it is an effective aid-delivery mechanism more broadly, and find that positive and significant effects dominate. We include evidence maps to highlight research gaps in the PbR literature.

Key words: aid conditionality; aid delivery; critical review; health; payment-by-results

1 INTRODUCTION

Aid-delivery mechanisms have been debated for many years. Conditionality was one of the most characteristic features of aid policy in the 1980s and 1990s (Riddell, 1995). Many studies have examined the usefulness of conditionality, arguing that it has not improved aid effectiveness and failed to support government reform processes in low- and middle-income countries (LMICs) (Collier, 1997; Killick, 1997; 1998; Mosley, 1991). The main reasons for the so-called failure of aid conditionality lie with donors' unwillingness to link incentives, either punitive or rewards-based, to aid-effectiveness and programme outcomes. Hence, there have been arguments for better selectivity, such as giving aid to countries with an established record of accomplishing reforms (Collier, 1997). Given that is not entirely possible to avoid misappropriation, discussions about different means of aid delivery, including better monitoring, have not abated. This is the context for the rise of Performance-based Financing schemes (PbF), which are based on a concept that builds on aid conditionality and selectivity by providing supply-side financial incentives in the form of payments that depend on delivering a defined quantity of services (Soucat et al., 2017). Most PbF schemes focus on health interventions in

LMICs. More recently, donor agencies have promoted the use of Payment-by-Results (PbR), whereby aid is disbursed upon achieving a pre-agreed measure (Clist, 2016).

A recent example of a PbR project is the Girls' Education Challenge (GEC) funded by the former UK Department for International Development (DFID)[1]. GEC is a large and complex programme with 37 non-government organization (NGO)-led projects in 18 countries with a total budget of £300 million. For 15 of the 37 GEC PbR projects, 10% to 20% of their total programme cost is 'at risk', because it is conditional upon achievement of pre-agreed outcomes. These affect mostly girls' learning and attendance outcomes. Projects would lose the entire amount 'at risk' if they failed to achieve their target outcomes, while a bonus would be paid to reward over-achievement (Holden & Patch, 2017; see Clist (2019) for a review of the DFID experience with PbR schemes).

PbR has rapidly gained in popularity among donors funding work in LMICs, the World Bank and DFID being prominent proponents (World Bank, 2011). While there have been more general reviews on PbF (e.g. Bertone et al., 2018; Paul et al., 2018; Renmans et al., 2016; Turcotte-Tremblay et al., 2016), many of these have been surprisingly thin in discussing how PbF can or has been adapted to different contexts and how these may also influence PbF schemes. In other words, much remains unknown about the exact mechanisms that underpin PbF (Renmans et al., 2016). There have been even fewer reviews of PbR, and those that exist acknowledge the slim evidence base (e.g. Oxman & Fretheim, 2009; Perrin, 2013). The current evidence base for PbR appears to be primarily theoretical rather than empirical (e.g. Clist, 2016), as its history is rather short. Furthermore, existing reviews are often unsystematic and broad, especially in terms of their geographical scope. A review by the National Audit Office (Mason et al., 2015) finds that health-specific studies dominate the PbR evidence base.

This article aims to synthesize more recent PbR evidence with a focus on aid-funded health interventions in LMICs, drawing on a systematic review methodology. There has been no systematic investigation of a similar nature to date. The purpose of this review is to understand the most recent and most innovative forms of micro-level PbR initiatives that align with the definition of PbR outlined above. This review will allow us to generate lessons about what works for the implementation and impact of PbR mechanisms in LMIC government- and NGO-led health programmes. Some of these insights may also prove useful for the non-health sector. We relate this to the experience of PbF schemes, which have not been uncontested (e.g. Soucat et al., 2017). With this in mind, we also reflect on the future of PbR and whether it will experience a renewed interest in the context of the COVID-

[1] DFID has been merged with the Foreign Office as of September 2020 creating the Foreign, Commonwealth and Development Office (FCDO).

19 pandemic, with tighter aid budgets and more selective use of funds. Finally, we explore whether PbR is an effective aid-delivery mechanism by assessing its success and failure and by understanding the circumstances in which what type of PbR instruments work best.

Our method as to adapt the systematic review process and adjust it in terms of breadth and depth of search and screening processes, quality assessment and synthesis. Our approach was inspired by the Rapid Evidence Assessment (REA) toolkit developed by the UK's Government Social Research Service (GSR) and the EPPI Centre.[2] To complement our critical review we include maps of evidence gaps to highlight some of the research gaps in the PbR health literature (Gough et al., 2013).

In the next section we first outline the search methodology; then the screening, search results and data extraction. The next step involves an assessment of validity and quality of the evidence before presenting the evidence gap maps. Finally, an in-depth synthesis, which is guided by the Measure-Agent-Principal (MAP) framework developed by Clist (2016), is used to assess success and failure of PbR measures, using a 'vote counting' procedure where studies are categorized according to their findings, i.e. positive or negative, statistically significant or not (discussed in more depth below).

Search methodology

The search process followed Cochrane Collaboration guidelines (Higgins & Green, 2011, chapter 6) and began with a snowballing approach that focused on five existing reviews of PbR (Eijkenaar, 2012; Mason et al., 2015; Oxman & Fretheim, 2009; Perrin, 2013; Webster, 2016). The results of the snowballing exercise were complemented by extensive electronic searches of academic (e.g. Web of Science and Science Direct) and institutional repositories (e.g. DFID, World Bank Open Knowledge Repository, Research4Development, OECD iLibrary, websites of the African, Asian and Inter-American Development Banks) as well as Google Scholar, to create an initial database of relevant key evidence, which was then reviewed to derive a list of key terms[3] that best describe PbR. The list of key terms informed the search strings used in the electronic searches of the academic and institutional repositories (see Appendices 1–3).

[2] http://webarchive.nationalarchives.gov.uk/20140402163101/http://www.civilservice.gov.uk/networks/gsr/resources-and-guidance/rapid-evidence-assessment/how-to-do-a-rea

[3] Examples of some of the key terms used: big results now, cash on delivery aid, output based aid, outcome-based contracting, outcome-based payment, payment for performance, pay for performance, payment by outcome, payment by results, paying for results, performance-based contracting, performance-based aid, results-based financing etc.

Screening and data extraction

The documents in the database were initially screened by research assistants with independent cross-checking by the principal researcher and co-researcher. The inclusion criteria guiding the screening process were: at least one lower/middle income country;[4] quantitative empirical data, e.g. primary and/or secondary survey data drawing on a range of research designs such as randomized controlled trials (RCTs), quasi-experiments, basic surveys, etc.; and payment according to a pre-agreed measure for aid-funded projects. We excluded documents that examined sector-wide approaches (SWAps) and/or other forms of aid disbursement such as grants or loans that may have used a pre-agreed measure as this was beyond the scope of this review and would not have aligned with our definition of PbR.

We used a PRISMA diagram (Figure 1) to present the different stages of the search and screening process. The initial snowballing approach, complemented by searching academic and institutional repositories, led us to identify 5,458 records. These records were screened by title and abstract using the inclusion criteria listed above. We excluded 4,944 records (including 23 duplicates) as they did not meet at least one of our criteria. Of these 5,458 records, 491 did not allow us to make an assessment based on screening title and abstract and hence we had to read the full text to make a judgment, of which 393 did not meet the criteria.[5] This left 98 records, plus 23 that were identified in a search of one of DFID's internal databases.[6] In total, 121 studies met the inclusion criteria. However, given the nature of our key search terms, some of the evidence we identified related to non-health sectors, notably education. To be able to focus on aid-funded health interventions, we discarded the non-health evidence and were left with 81 studies published between 2000 and 2020.

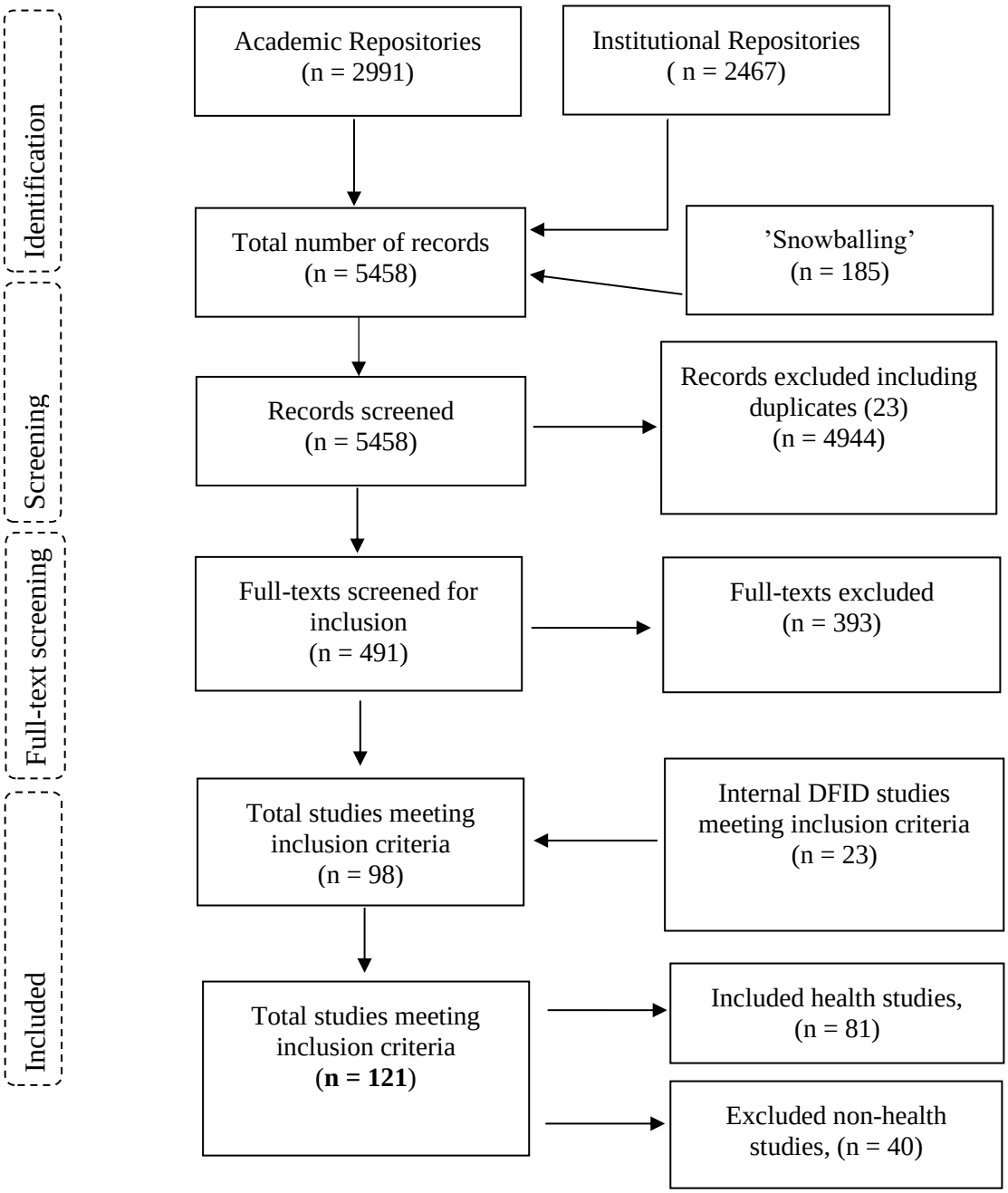
[4]The World Bank definitions of lower/middle income countries were used: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

[5] This is not an unusual additional step in the screening process as many studies in the social science and international development context are often not written in uniform ways and this creates challenges from a systematic review perspective (see Mallet et al, 2012 for a discussion).

[6] Given DFID is a key adopter of PbR mechanisms as well as the funder of this study, we approached them to enquire about any unpublished studies held internally and we obtained access to search one of their internal databases for relevant DFID funded PbR studies – this database is not accessible to the public.

Figure 1

PRISMA Diagram



Exclusion reasons:

1. Not lower/middle income country: 2,579 studies
2. No quantitative empirical data (primary and/or secondary data): 2,276 studies
3. No payment according to a pre-agreed measure for aid-funded projects: 482 studies
4. Duplicates: 23 studies
5. Non-health study: 40 studies
6. Total studies excluded: 5,400

These 81 studies were then read in detail to extract information for quality assessment as well as synthesis. We extracted descriptive information from each of the studies using the adapted data-extraction form designed by Boaz et al. (2002) (Table 1).

Table 1

Basic Data Extraction Tool

Details of publication
Author
Title
Source (journal, conference etc.)
Year
Context and population
Country
Sector
Donor
Who gets paid?
What is the measure?
How is payment calculated?
How much was disbursed?
Any treatment effect sizes reported?
Study design
Research design category
Type of secondary data used
Data source
Analysis
Statistical techniques used

Source: Adapted from Boaz et al. (2002).

These data were used in the validity and quality appraisal outlined in the next section.

Validity and quality appraisal

Systematic reviews commonly adopt criteria for assessing each piece of evidence according to whether it is reliable within its own methodological paradigm and aims (Waddington et al., 2012). The quality-appraisal process can and should be used to validate what constitutes evidence in relation to the specific question(s) to which a particular study seeks to find answers.

We adopt the scoring scheme developed by Duvendack et al. (2011; 2012) to assess the quality of the 81 studies we included in this review. This scheme categorizes each study by its reported research design and analytical method, assigning scores which are then combined into an index. A cut-off point of 2 is applied, e.g. a study with a score of 2 and above is considered to have high threats to internal validity—this would correspond to a study based on a basic survey reporting descriptive statistics. Studies with scores of less than 2 have lower threats to internal validity, e.g. an RCT using

econometric techniques would have a score of less than 2 and would be considered to be reasonably reliable (see Table 2).

Table 2

Potential Risk of Bias in PbR Studies

	Statistical methods of analysis		
Research design	Difference-in-difference, Propensity score matching, Instrumental variables, Regression discontinuity	Multivariate (or bivariate with covariate means tests)	Tabulation
Randomised Controlled Trial	8	16	3
Panel	6	0	0
Cross section: before/after or with/without	19	8	9
Basic survey	1	4	7

Low score	30
Medium score	23
High Score	28

Source: Adapted from Duvendack et al. (2012).

Table 2 indicates that in our sample of 81 studies, 30 have a low score, indicating a low risk of bias,[7] 23 have a medium score indicting a medium risk of bias[8] and 28 have a high score meaning they have a high risk of bias.[9]

High risk of bias does not mean that studies do not contribute in significant ways either substantively or methodologically, only that they have shortcomings in how they deal with threats to internal and external validity and thus we should treat their findings with caution. In the systematic review literature there is a debate on whether to include studies with low methodological quality, i.e. high risk of bias, in the synthesis. We feel that lessons can be learnt from studies of high, medium as

[7] Low risk of bias studies in our sample include RCTs or panel data designs using advanced econometric techniques for analysis.
[8] Medium risk of bias studies draw on cross-sectional research designs using a mixture of advanced econometric multivariate techniques for analysis.
[9] High risk of bias studies in our sample adopt cross-section designs or basic surveys in combination with less rigorous analytical techniques such as multivariate statistics or basic descriptive statistics.

well as low risk of bias and so include all studies irrespective of their methodological quality in the maps of evidence gaps as well as in the in-depth synthesis of the health evidence.

Evidence gap maps

Evidence gap maps are a thematic collection of the evidence base summarizing what we do or do not know about a particular topic. They are presented in the form of a matrix outlining the types of evidence available, and thus identifying gaps, given pre-defined parameters that reflect types of evidence or certain characteristics of the evidence base (Sniltsveit et al., 2017). The parameters of the map can be adjusted according to the users' priorities. The users are often policymakers, researchers and/or donors. As mentioned above, our evidence gap maps and in-depth synthesis are guided by the MAP framework (Clist, 2016). This framework builds on the classic Principal-Agent model (e.g. Akerlof, 1982), which is often used in contract theory, asserting that the payoff to a principal depends on the actions taken by an agent. In the context of this review, the principal refers to the donor or funding body which commits to pay for aid on a pre-agreed measure, the agent is the party being paid by the principal to produce results and the measure is the pre-agreed measure on which PbR contracts are based.

Principal

As principals, aid donors, we empirically identified the following categories from the 81 studies we included:

- World Bank
- NGOs
- Governments (this includes DFID and similar government departments based in high-income countries)
- Development financing organizations (e.g. IADB)
- Public–private partnerships (PPPs) (e.g. GAVI).

Of the 81 health-related studies, 68% list governments as the main principal, followed by NGOs (9%) and the World Bank (6%); 14% of studies provided no information on the principal—this lack of reporting of key information is a recognized challenge in systematic reviews, hence the development of the CONSORT statement[10] to improve the quality of reporting of studies (with a focus on RCTs).

[10] http://www.consort-statement.org/

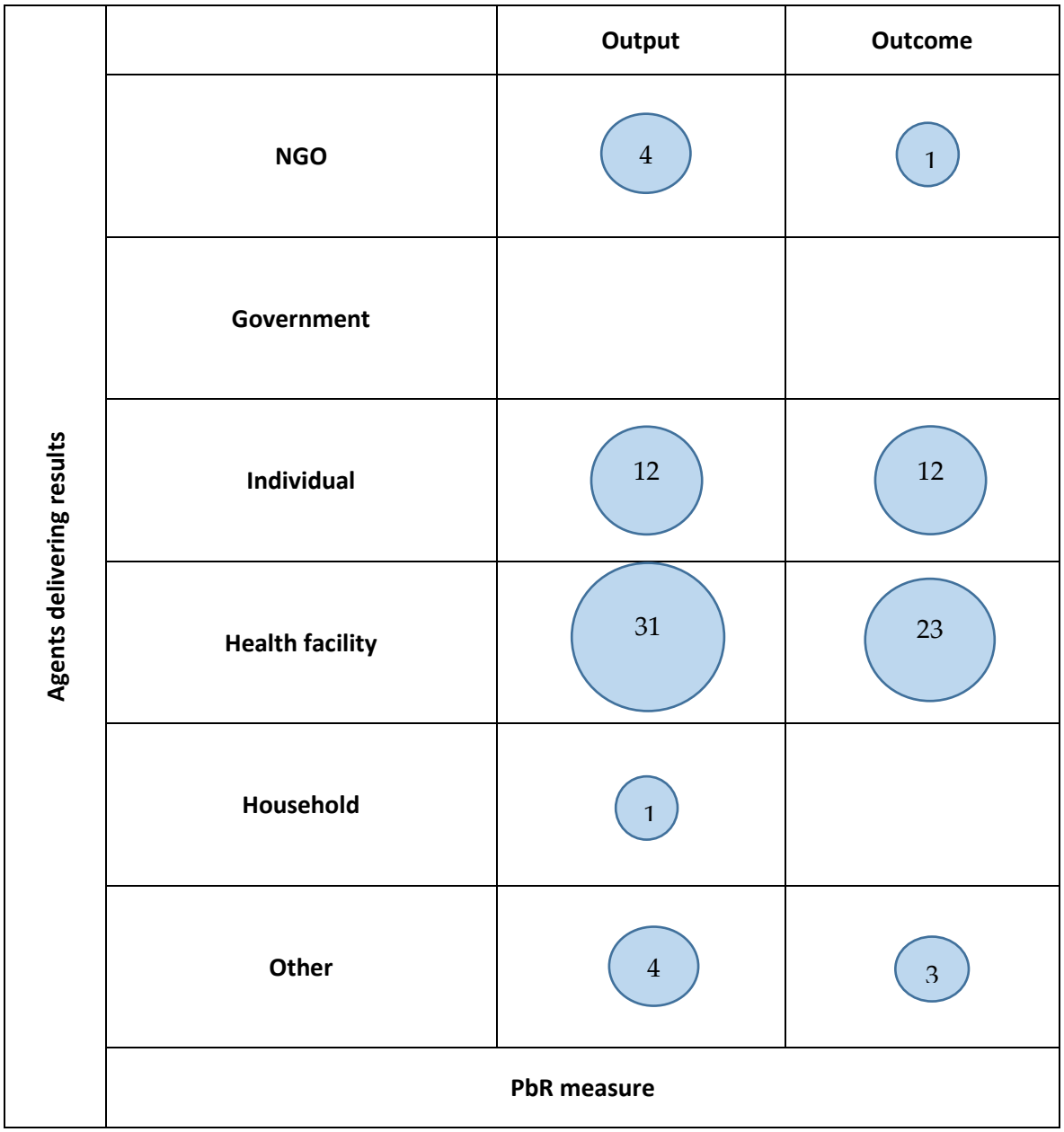
Agent

The picture is more mixed in relation to the domain of the agent, i.e. 'Who gets paid?'. The health evidence is dominated by two agents: Health facilities (47% of studies) and individuals (26% of studies). The remaining 27% of studies present a diverse set of agents such as households (3%), international NGOs (4%), local/national NGOs (4%), schools (3%) and others ranging from provincial governments to villages and central medical stores. As a result of these findings, our synthesis focuses largely on the studies reporting health facilities and individuals as agents, which captures most of the of the health evidence (73%).

Measure

The most important element in the MAP framework is the measure, which forms the basis for the PbR contract, Map 1 describes the health evidence by PbR measure (categorized as output and outcome measures) and by agent. We define outputs as tangible goods and services that are delivered by the project, e.g. how many children are vaccinated, number of condoms sold—the implementing agency has direct control over these outputs. Outcomes build on outputs, they are realized once beneficiaries have used the project outputs, they relate to the project's intended medium-term goals. Measuring outcomes is important in trying to answer cause-and-effect questions allowing an assessment of the overall difference a project made, while outputs will only be able to answer normative questions to assess project activities and whether they achieved short-term targets. Outputs are delivered while outcomes come from expected behaviour changes among project beneficiaries (Gertler, 2011).

Map 1 *Health Studies by PbR Measure and by Agent, all Risk of Bias Levels*



Notes: 79 rather than 81 health studies are captured because nine provide no information on "Who gets paid?"—the agent—while seven report two different agents (these appear twice in the map), hence 79 studies. Fifteen studies report no information on the pre-agreed measure but of the 66 studies reporting pre-agreed measures 42 report two or even three measures. The section 'Other' includes very study specific agents, e.g. villages, co-operatives, central medical stores, schools or school principals, local civic organizations.

For the purpose of Map 1, we grouped the 26 different PbR measures that we found in our studies into output and outcome measures[11] to grapple with the high level of heterogeneity we

[11] Applying the definition of outputs and outcomes outlined above and considering the broad results chain of each of the health sub-categories we identified, we categorise the PbR measures as follows: **Outputs** are number of vaccine given, condoms sold, number of people attending health services, number and type of health services

identified. It is clear from the map that most PbR measures are output-related (60% of all health studies across all risk of bias levels), especially when the agent is an individual or a health facility, as indicated by the size of the bubbles in Map 1. We know relatively little about the type of PbR measures used when the agent is a government agency or a household (blank spaces or small bubbles indicate few or no studies engaging with this issue). Furthermore, studies examining PbR programmes with individuals and health facilities as agents appear to lack methodological rigour as indicated by Map 2, which suggests that most suffer from either high or medium risk of bias (the size of the bubble with the actual number of studies given within the bubble is the key indicator here).

Map 2 *Health Studies by Risk of Bias levels and by agent*

Agents delivering results		High	Medium	Low
	NGO	3	1	2
	Government		1	
	Individual	9	7	9
	Health facility	12	13	14
	Household		3	
	Other	1	1	3
	Risk of bias level			

delivered, number of new patients, percentage of discharges, number of screening and referring/caring malnutrition cases, length of hospital stay, number and type of STI treatment, number of women receiving PNC/ANC, number of referral of pregnant women to health facility, number of prescriptions given, number of TB treatment/referral provided. **Outcomes** are negative test for disease, giving birth at a facility, birth with skilled attendant, health insurance coverage, contraceptive prevalence rate, cases of malnutrition, investing in health and education expenditure, clinical performance vignettes, provision of family planning services.

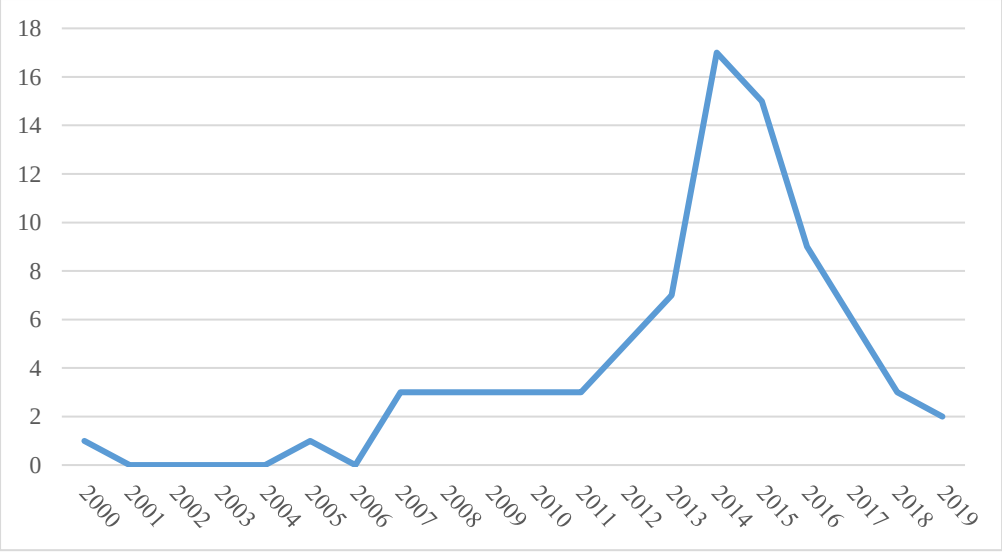
Notes: Nine studies provide no information on "Who gets paid?"—the agent; five report two different agents (these appear twice in the map), hence 67 studies. "Other" includes very study-specific agents, e.g. villages, co-operatives, central medical stores, schools or school principals, local civic organizations.

2 SYNTHESIS—EXPLORING TRENDS AND HETEROGENEITY

Of the 81 health studies included, 24% have already been captured in other systematic and unsystematic reviews that touch on PbR themes (e.g. Blacklock et al., 2016; Eichler et al., 2013; Eldridge & Palmer, 2009; Glassmann et al, 2007; Lindsay et al., 2011; Oxman & Fretheim, 2009; Witter et al., 2012); hence, our synthesis focuses on exploring trends and demonstrating heterogeneity. Given the high levels of heterogeneity in our study sample, we are unable to conduct meta-analysis, which requires studies to be comparable on a conceptual level with similar constructs and relationships and similar study designs, as well as statistical approaches to avoid the "apples and oranges" problem (Lipsey and Wilson, 2001). As this is not the case here, we have chosen a narrative synthesis approach that involves textual as well as descriptive analyses of the quantitative evidence.

As discussed above, PbR is a relatively recent phenomenon, reflecting the changing nature of aid-delivery modalities; this is illustrated in Figure 2, which suggests that the publication of health-related PbR studies has steadily risen since 2011, implying an increased use of PbR. It appears, however, that a peak was reached in 2014–2015 with a clear downwards trend from 2016 onwards. This raises the question of whether PbR has run its course or whether a revival is imminent, especially in the context of COVID-19 when budgets are getting tighter and considerations of value for money and results agendas dominating more than ever.

Figure 2 *Number of PbR Studies Published by Year*



Our sample of health studies is relatively small and highly heterogeneous in terms of health sub-themes, geographical regions and methodological quality. For examplemost of the health evidence is dominated by maternal and child health (MCH) interventions (47.4%), this is followed by interventions on HIV and Sexually Transmitted Infections (STIs) (13.6%) and nutrition (5.1%). A third of the health evidence (33.9%) cannot be categorized quite so easily as a wide range of health sub-themes is covered. covered. InIn terms of geographical regions, close to 52% of the studies focus on sub-Saharan Africa, followed by Latin America (17.3%), South Asia (13.6%) and East Asia (8.6%). Roughly 6% of studies focus on Southeast Asia and almost 2.5% of studies on Middle Eastern countries. Overall, 30 countries are captured in our sample of 81 studies. Given PbR mechanisms are highly context-specific, it is challenging to generate meaningful lessons when faced with such a highly heterogeneous evidence base. To summarize, we identified 81 health studies covering 30 countries, employing 26 different PbR measures, nine types of agents across many health sub-themes and displaying a considerable variation in methodological quality (as indicated in Table 2).

3 DISCUSSION—UNDERSTANDING SUCCESS AND FAILURE OF PBR MEASURES

We see from Map 1 that the majority (60%) of the PbR measures identified in the health literature is at output levels. Following Eldridge and Palmer (2009), we now explore what these output—and outcome—measures allow us to conclude with regard to the success and failure of PbR contracts. Depending on the main mechanism of each programme, success should be defined at the level (i.e. input, output, outcome or impact) appropriate to the particular theory of change (ToC) embodied in the PbR programme, but often a ToC is not reported, making it difficult to unpack the underpinning mechanisms. A similar observation can be made regarding the broader literature on PbF (e.g. Renmans et al. (2016) argue that the "black box" allowing us to understand PbF mechanisms has yet to be fully opened). Bertone et al. (2018) acknowledge that much of the PbF literature emphasizes the importance of understanding context, but there has been little scrutiny on understanding how context shapes the implementation of PbF schemes, possibly because each scheme needs to be understood within its unique contextual framing, the nature of the programme being implemented and so on. This makes it particularly difficult to unpack the underlying mechanisms of PbR in health interventions given the highly heterogenous nature of the evidence base we identified, which does not allow us to derive generalizable findings about clear-cut mechanisms that may apply to PbR schemes more broadly. Therefore, we aim to better understand the impact of PbR by defining its success and failure

in admittedly naïve terms (as discussed above) by counting[12] the signs and levels of significance for both output- and outcome-level measures, where positive and significant measures would suggest success (in relation to doing nothing—the 81 studies we reviewed suggest that the counterfactual to PbR is no PbR rather than the next best alternative, such as alternative health-financing schemes). In some cases, however, a negative measure implies success, e.g. a nutrition programme may measure success in terms of reducing cases of malnutrition. This would be expressed in the form of a negative effect size, which we would code as positive because a reduction in this case suggests success. Similarly, where positive measures imply failure, e.g. an increase in STIs, we would code them as negative. To help with our explorations, we revisit all 81 health studies irrespective of their risk of bias level to extract statistical information on each of the pre-agreed PbR measures reported in each . We extract data on whether the results reported for the pre-agreed PbR measure were positive or negative and whether they were statistically significant. Table 3 presents the results for all 26 PbR measures across output and outcome levels.

Table 3 *Success and Failure of all PbR Measures (Output and Outcome), across all Risk of Bias Levels*

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	53	9	62
Not significant	41	10	51
Total	94	19	113

Notes: Health studies only, 57 studies rather than 81 as not all studies reported information on success or failure, but some reported multiple effects owing to using multiple measures, all risk of bias levels.

The majority of PbR measures report positive results that are statistically significant, which would suggest success—although it pays to be cautious about reaching such hasty conclusions. If we disaggregate the information presented in Table 3 by output and outcome-level measures, we find that output measures are slightly more likely to find positive and significant effects (41% of the studies) than outcome measures (31% of the studies). We caution against drawing firm conclusions on success or failure of PbR measures based on these findings without considering methodological quality or issues related to publication bias.

As discussed above, we initially hypothesized that high risk of bias means that studies may be more likely to report positive and significant effects but, interestingly, we find that this is not the case here. However, many studies with a high risk of bias do not report levels of significance as they present

[12] We adopt a 'vote counting' procedure, which was first discussed by Light and Smith (1971), where studies are sorted into categories: studies yielding significant or not significant, positive or negative results.

only basic descriptive statistics or graphics to demonstrate PbR effects. Furthermore, we explore whether publication bias may play a role in explaining the dominance of positive and significant effects presented in Table 3. Publication bias is a well-recognized and serious issue in systematic reviews. It is argued that studies reporting statistically significant findings are more likely to be published in peer-reviewed journals than those reporting statistically non-significant findings (Borenstein et al., 2009). In our sample, 59% are published in peer-reviewed journals, of which 92% present positive estimates with 66% reporting positive as well as significant findings. Similarly, among the 41% of unpublished studies (in the systematic review literature, unpublished studies refer to the grey literature, i.e. studies that are identified through institutional repositories or websites), we find that 97% report positive estimates with 64% of those being statistically significant. This suggests that publication bias may not play a major role since reporting rates of positive and significant effects are similar across both published and unpublished studies, especially in terms of statistical significance.[13]

Understanding failures

To better understand success or failure of PbR measures, we examine some of the successes and failures in more depth. We start with some of the failures listed in Table 3, admittedly few, negative PbR estimates across output and outcome-level measures. For example, we find that three of the studies reporting negative output-level PbR effects are suffering from high risk of bias, which suggests low methodological quality. Across these three studies, a range of PbR measures are used, such as type of health services provided, number of patients referred, and number of anaemic students treated, to be able to measure success. The principal was the government in most of these studies and the agents largely individuals or health facilities.

The picture is similar for the outcome-level measures reporting negative effects. Though there is a notable difference in terms of PbR measures used, e.g. negative tests for disease and giving birth at health facility; and the methodological quality which was either low or medium risk of bias. To explain the negative effects, we examine the geographical locations of each of the studies to identify any potential patterns. We find that half of the negative outcomes are reported in fragile states such as Afghanistan and the Democratic Republic of the Congo (DRC). This may suggest issues related to data quality and/or measurement errors, or to issues related to governance, security, quality of institutions among other things, but we cannot be certain given the limitations of the available evidence base.

[13] Publication bias is widespread in published academic research, in particular in the social and medical sciences. Rosenthal (1979) first used the term 'file drawer problem' to draw attention to biases in published research.

Understanding successes

For 58% of the studies reporting positive and statistically significant *output-level* effects, the principal is the government. The agent is mainly a health facility (64% of studies) followed by individual, e.g. such as a health worker and/or a patient (18%)—these findings hold when taking out the high risk of bias studies. Eleven output measures are reported by 34 studies—such as number of vaccinations administered, condoms sold, individuals attending health services, new patients, percentage of patient discharges, type of health service provided, number of patients for STI treatment, number of women receiving ante- and post-natal care. The number of vaccinations is the most frequent output (reported in 14 studies) followed by type of health service (reported in nine studies). One could argue that it may not be surprising that these measures dominate the positive and statistically significant effects as they are easily measurable and changes can be observed within short time periods. While it is important to have information on the success of short-term output-level health measures, this does not say much about quality, appropriateness or long-term effects of such health measures. In fact, the recipients of the measures discussed here, e.g. health facilities and individuals, may not be encouraged to embark on activities that affect or change long-term health behaviour. In terms of measuring quality, it is worth noting that PbR has contributed to bringing quality concerns to the fore by linking payments to data and thus creating more demand for investments in health information systems that would lead to improvements in data quality and therefore enable more timely and accurate payments (Soucat et al., 2017). An important issue in this context is the process of verification, i.e. how trustworthy are the data derived from the health information system used to make decisions on payments. In the absence of solid internal auditing mechanisms, there is a high risk of overreporting on a pre-agreed measure (e.g. Jerven, 2013, and Soucet et al., 2017 arguing that PbR creates moral hazards) and needs to be counteracted by establishing administrative systems that allow checking for data validity (Soucat et al., 2017). However, these systems can be expensive and time-consuming to implement, thus leading to delayed payments (Dale, 2014). Antony (2017) cites an example of Benin where the costs of verification activities were equivalent to almost 40% of the bonus payments made to health facilities. Hence, the key to successful verification measures is to ensure that they pose a credible threat against false reporting while at the same time being cost-effective (Soucat et al., 2017). Furthermore, to better understand the success of PbR, it may also be worth paying closer attention to the qualitative evidence and encourage the pursuit of more holistic evaluations that encompass a range of methodological and analytical approaches that facilitate unpacking the drivers involved in PbR mechanisms. We observe no variations across geographical regions, but 69% of the studies reporting on vaccinations administered and type of health services provided suffer from high risk of bias.

As for the positive and statistically significant *outcome-level* measures, we find that in 64% of all cases the government is listed as the principal, followed by the World Bank (16%). The dominant agents are again health facility (in 76% of all cases) and individuals (20%)—as above, these findings hold when taking out the studies of high risk of bias. Seven different PbR outcome measures are reported across 25 studies, e.g. giving birth at health facility, provision of contraceptive services, health insurance coverage, contraceptive prevalence rate, and cases of malnutrition. The frequent measures are giving birth at health facility (in 84% of studies), contraception (20%), and birth with skilled attendant (16%)—some studies report multiple measures hence the total percentages do not add up to 100. No geographical trends can be detected across the positive and significant outcome measures. An interesting finding here is the type of data used to verify the three dominant outcomes, in 20% of all cases Demographic and Health Survey (DHS) data are used to verify the birth and contraception-type outcomes. DHS data are of good quality in most cases but Schoumaker (2014) raises quality concerns for some of the historical birth and fertility data reported in some DHS datasets.[14] Furthermore, DHS data, like so many surveys, are subject to common biases such as respondent and authority biases, and often no ex-post enumeration verification or quality assessment is conducted in the context of DHS. Clist (2016) raises the issue of being able to draw on unbiased and unincentivized data sources to accurately assess the value of PbR mechanisms—one way to achieve this could be to increasingly involve independent evaluators in the data-collection process, allowing them to pay attention to data quality as well as to appropriate methods and types of analysis. Interestingly, in the studies identified from 2017 onwards, we observe a trend towards more primary data collection, i.e. adoption of RCTs and collection of panel data employing more sophisticated econometric techniques.

What can we learn from the PbR evidence in the field of health in terms of advancing PbR mechanisms? First, from Map 1, it is clear that 60% of PbR measures reported in the health evidence are output-related (across all risk of bias levels). This is unsurprising as outputs are easier to measure than outcomes and successes can be achieved within shorter time horizons. This observation is linked to the short-term nature of PbR contracts favouring measures that can be verified relatively quickly and more easily at low cost (Clist, 2016). However, outputs may not always be a reliable measure of success or failure; thus, programme commissioners should seek to think more carefully about the PbR measures they adopt and how they relate to the mechanisms and goals set out in their particular

[14] Schoumaker (2014) raises concerns on some of the birth and fertility metrics found in DHS data, further discussed below.

ToC—the above discussion indicated that the PbF and PbR literature on unpacking the underlying mechanisms of such schemes is underdeveloped, which suggests scope for further research.

Second, to better understand some of the successes and failures of PbR initiatives, it may be worth examining the role of the agent in more depth. From our evidence base, it is clear that individuals and health facilities are the dominant agents, but which agent characteristics are the most important to ensure success and avoid failure of PbR mechanisms? An interesting example in this context is health workers' motivation, but given the heterogeneous nature of the evidence base there is no clear pattern. We identified two studies dedicated to exploring health workers' motivation: Huillery and Sebanz (2014) and Robyn et al. (2014). The former conduct an experiment in DRC, finding that financial incentives led to better efforts from health workers. They also find a shift from intrinsic to extrinsic motivation but caution that the capacity of the health provider needs to be taken into account if PbR measures are to succeed. Robyn et al. (2014) investigate the case of a health insurance scheme in Burkina Faso where the quality of health care declined due to health workers' dissatisfaction with the payment method, e.g. capitation payments were insufficient, payments were infrequent and there were no mechanisms to reimburse service fees. These two studies indicate that motivational and behavioural aspects can affect the quality of the health care provided within a PbR framework, and this might be worth exploring further.

Third, to understand the failures of some of the PbR mechanisms, we examined the geographical locations for each of the studies and find that half of the failures in outcome measures were reported in fragile states. This may suggest issues related to data quality and/or measurement errors, lack of local research capacity, or concerns about governance and security casting doubts on the reliability of the PbR measures used. Output-level measures may be easier to measure and verify in fragile states, but this may not always be useful from a PbR perspective since the ultimate goal of sustainable development programmes should be to create long-lasting outcome-level effects brought about by behavioural changes but which may also require structural and institutional changes. However, we have very few data points to reach any firm conclusions on success and/or failure of PbR measures in fragile states. Furthermore, the evidence provides no details on the costs of measuring outputs versus outcomes and we do not yet fully understand what sort of PbR measures work best in fragile states.

Limitations

This study is not a full systematic review as the standard review process has been adapted to address resource constraints. The search process was more limited in terms of depth and breadth especially in relation to searching grey literature sources. Qualitative studies have not been considered, although these may be particularly useful to understand the underlying causal mechanisms that underpin PbR schemes. In addition, the review does not reflect on the experiences of PbR initiatives in high-income countries and focuses only on the health sector.

4 CONCLUSION

We report the findings of a critical review to synthesize the PbR evidence on aid-funded health interventions in LMICs in order to gain a better understanding of whether PbR schemes are an effective aid-delivery tool and under what circumstances what type of PbR measures may work best. We hope these insights may also prove useful beyond the health sector. We identified 81 PbR studies that focus on health interventions, with high levels of heterogeneity in terms of health sub-themes, number and type of agents and PbR measures (26 different measures with a focus on output-level measures), geographical scope (30 countries), and conclude that a third of the studies suffer from high risk of bias, i.e. their methodological quality is poor.

We assess success and failure of PbR measures using a vote-counting procedure and observe that while positive and significant effects dominate, these successes can be misleading. We also know relatively little about the type of PbR measures used when the agent is a government or a household. It is not possible to tell from the evidence reviewed whether PbR measures do or do not work and whether PbR schemes are a more effective aid-delivery mechanism than other health-financing schemes. We reviewed the PbF literature in the hope of obtaining greater clarity on the workings of PbR but we found that this literature is equally beset with challenges, i.e. PbF mechanisms are still not fully understood and the focus on measures may obscure the bigger picture, such as reforming the wider health system, and examine how these systems interact with the health measures of interest (Soucet et al., 2017). Nor do we know how sustainable PbR measures are as the evidence points towards a preference for short-term measures that focus on the initial stages of the results chain (input to output levels). This could amount to fool's gold (Clist, 2016); outcome-level measures capturing longer time horizons may be a better choice—assuming the ultimate goal is to seek behaviour changes, which could lead to more sustainable effects of development programmes. One could of course argue that measuring outcomes is not realistic, especially in fragile and conflicted-affected areas. The evidence base we reviewed indicates that output measures were preferred in fragile environments; or where outcome measures were used within fragile and conflict-affected areas, admittedly very few, these were negative in half of the cases. As argued above, this could be

due to measurement errors, low data quality, a lack of sufficient local research capacity and/or poor governance, security concerns etc. Given the many unknowns surrounding PbR schemes, the question remains whether there may be renewed interest in PbR in the context of the COVID-19 pandemic. It is too early to say, but future research could attempt to pinpoint the linkages between tighter aid budgets and a further rise in popularity of PbR schemes among donors.

Furthermore, to better understand PbR mechanisms, it may be worth examining the existing qualitative evidence in more depth and to conduct more individual country-level studies (such as the one conducted by de Walque et al., 2017) that would pay meticulous attention to the unique context of a given country, with a focus on isolating each and every component of a PbR scheme. This would potentially allow us to unpack the "black box" that conceals PbR mechanisms. Encouraging more RCTs on this topic may not be useful unless they are embedded in a theory-based evaluation approach and replicated across countries. We should also be looking at sectors other than health and beyond LMICs to be able to identify good PbR measures. Furthermore, in order to conduct better reviews, we need more detailed reporting and investigation on the type of PbR measures that were adopted in different contexts, including fragile and conflict-affected areas. Mallet et al. (2012) argue that many empirical studies in the social sciences are not written in a uniform fashion or from the perspective of aiming to be included in a systematic review, hence studies may report only limited detail on method, impact, data or value for money—all areas of importance from the perspectives of a systematic review and PbR lessons. Finally, we need to be able to draw on unbiased data sources collected by independent evaluators who pay attention to selecting appropriate and rigorous types of methods and analyses depending on the particular context.

First submitted June 2020

Final draft accepted November 2020

Corresponding author: Maren Duvendack University of East Anglia, School of International Development, UK. ORCID: 0000-0002-8125-9115. **Email: m.duvendack@uea.ac.uk**

Acknowledgements: The author would like to thank Sonja Marzi and Asta Hansen for excellent research assistance and Paul Clist for comments on this review. This work was funded by the UK's Department for International Development (DFID).

REFERENCES

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *The Quarterly Journal of Economics*, *97*(4), 543–569. https://doi.org/10.2307/1885099

Antony, M., Bertone, M. P., & Barthes, O. (2017). Exploring implementation practices in results-based financing: The case of the verification in Benin. *BMC health services research*, *17*(1), 204. https://doi.org/10.1186/s12913-017-2148-9

Bertone, M. P., Jacobs, E., Toonen, J., Akwataghibe, N., & Witter, S. (2018). Performance-based financing in three humanitarian settings: Principles and pragmatism. *Conflict and Health*, *12*(1), 28. https://doi.org/10.1186/s13031-018-0166-9

Blacklock, C., MacPepple, E., Kunutsor, S., & Witter, S. (2016). Paying for performance to improve the delivery and uptake of family planning in low and middle income countries: A systematic review. *Studies in Family Planning*, *47*(4), 309–324. https://doi.org/10.1111/sifp.12001

Boaz, A., Ashby, D., & Young, K. (2002). *Systematic reviews: What have they got to offer evidence based policy and practice?* (ESRC UK Centre for Evidence Based Policy and Practice: Working Paper 2). https://www.kcl.ac.uk/sspp/departments/politicaleconomy/research/cep/pubs/papers/assets/wp2.pdf

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Clist, P. (2019). Payment by results in international development: Evidence from the first decade. *Development Policy Review*, *37*(6), 719–734. https://doi.org/10.1111/dpr.12405

Clist, P. (2016). Payments by results in development aid: All that glitters is not gold. *The World Bank Research Observer*, 1–24. http://hdl.handle.net/10986/29310

Clist, P., & Verschoor, A. (2014). The conceptual basis of payments by results. https://assets.publishing.service.gov.uk/media/57a089bb40f0b64974000230/61214-The_Conceptual_Basis_of_Payment_by_Results_FinalReport_P1.pdf

Collier, P. (1997). The failure of conditionality. In C. Gwin & J. Nelson (Eds.), *Perspectives on aid and development*. Overseas Development Council.

Dale, E. (2014). Performance-based payments, provider motivation and quality of care in Afghanistan. Doctoral dissertation, Johns Hopkins University. https://jscholarship.library.jhu.edu/bitstream/handle/1774.2/37010/dale-dissertation-2014.pdf

Duvendack, M., Palmer-Jones, R., Copestake, J.G., Hooper, L., Loke, Y., & Rao, N. (2011). *What is the evidence of the impact of microfinance on the well-being of poor people?* EPPICentre Social Science Research Unit Institute of Education, University of London.

Duvendack, M., García Hombrados, J., Palmer-Jones, R., & Waddington, H. (2012). Assessing 'what works' in international development: Meta-analysis for sophisticated dummies. *Journal of Development Effectiveness*, *4*(3), 456–471. https://doi.org/10.1080/19439342.2012.710642

Eichler, R., Agarwal, K., Askew, I., Iriarte, E., Morgan, L., & Watson, J. (2013). Performance-based incentives to improve health status of mothers and newborns: What does the evidence show? *Journal of Health Popul Nutrition*, *31*(4 Suppl 2), S36–S47.

Eijkenaar, F. (2012). Pay for performance in health care: An international overview of initiatives. *Medical Care Research and Review*, *69*(3), 251–276. https://doi.org/10.1177%2F1077558711432891

Eldridge, C., & Palmer, N. (2009). Performance-based payment: Some reflections on the discourse, evidence and unanswered questions. *Health Policy Plan*, *24*(3), 160–166. https://doi.org/10.1093/heapol/czp002

Gertler, P. J., Martínez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2011). *Impact evaluation in practice.* World Bank.

Glassman, A., Todd, J., & Gaarder, M. (2007). Performance-based incentives for health: Conditional cash transfer programs in Latin America and the Caribbean. CGD Working Paper No 120. https://www.cgdev.org/publication/performance-based-incentives-health-conditional-cash-transfer-programs-latin-america-and

Gough, D., Oliver, S., & Thomas, J. (2013). Learning from research: Systematic review for informing policy decisions. a quick guide. A Paper for the Alliance for Useful Evidence. Nesta.

Higgins, J.P.T., & Green S. (2011). Cochrane handbook for systematic reviews of interventions. Version 5.1.0 [updated March 2011]. www.cochrane.handbook.org

Holden, J., & Patch, J. (2017). Does skin in the game improve the level of play? The experience of payment by results (PbR) on the Girls' Education Challenge (GEC) programme. https://www.pwc.com/gx/en/government-public-sector-research/assets/skin-in-the-game-pbr-on-the-gec-final.pdf

Jerven, M. (2013). *Poor numbers: How we are misled by African development statistics and what to do about it*. Cornell University Press.

Killick, T. (1997). Principal agents and the failings of conditionality. *Journal of International Development*, *9*(4), 483–495. https://doi.org/10.1002/(SICI)1099-1328(199706)9:4%3C483::AID-JID458%3E3.0.CO;2-S

Killick, T. (1998). *Aid and the political economy of policy change*. Routledge.

Light, R.J. & Smith, P.V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, *41*, 429–471. https://doi.org/10.17763/haer.41.4.437714870334w144

Lindsay, M., Beith, A., & Eichler, R. (2011). *Performance-based incentives for maternal health: Taking stock of current programs and future potentials*. USAID Health Systems 20/20 Report. http://abtassociates.com/AbtAssociates/files/96/96b67395-90b9-4bfe-bbcf-b808fdf0a2f3.pdf

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. *Applied social research methods*. Sage Publications.

Mallet, R., Hagen-Zanker, J., Slater, R., & Duvendack, M. (2012). The benefits and challenges of using systematic reviews in international development research. *Journal of Development Effectiveness*, *4*(3), 445–455. https://doi.org/10.1080/19439342.2012.711342

Mason, P., Fullwood, Y., Singh, K., & Battye, F. (2015). Payment by results: Learning from the literature. ICF International. https://www.nao.org.uk/wp-content/uploads/2015/06/Payment-by-Results-Learning-from-the-Literature.pdf

Mosley, P. (1991). Kenya. In J. Harrigan, P. Mosley, & J. Toye (Eds.), *Aid and power: The World Bank and policy-based lending, Volume 2.* Routledge.

Oxman, A. D., & 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*(2), 70–83. https://doi.org/10.1111/j.1756-5391.2009.01020.x

Paul, E., Albert, L., Bisala, B. N. S., Bodson, O., Bonnet, E., Bossyns, P., Colombo, S., De Brouwere, V., Dumont, A., Eclou, D. S., & Gyselinck, K. (2018). Performance-based financing in low-income and middle-income countries: Isn't it time for a rethink?. *BMJ Global Health*, *3*(1), e000664.

Perrin, B. (2013). *Evaluation of payment by results (PBR): Current approaches, future needs* (DFID Working Paper No 39). http://www.dev-practitioners.eu/fileadmin/Redaktion/Documents/TG_RBA/payment-results-current-approaches-future-needs.pdf.pdf

Renmans, D., Holvoet, N., Orach, C. G., & Criel, B. (2016). Opening the 'black box' of performance-based financing in low-and lower middle-income countries: A review of the literature. *Health Policy and Planning*, *31*(9), 1297–1309. https://doi.org/10.1093/heapol/czw045

Riddell, R. (1995). *A new framework for international development cooperation: Towards a research agenda.* Overseas Development Institute.

Rosenthal, R. (1979). File drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. https://psycnet.apa.org/doi/10.1037/0033-2909.86.3.638

Schoumaker, B. (2014). *Quality and consistency of DHS fertility estimates, 1990 to 2012* (DHS methodological reports No. 12). ICF International.

Snilstveit, B., Bhatia, R., Rankin, K., & Leach, B. (2017). *3ie evidence gap maps: A starting point for strategic evidence production and use* (3ie Working Paper 28). International Initiative for Impact Evaluation (3ie).

Soucat, A., Dale, E., Mathauer, I., & Kutzin, J. (2017). Pay-for-performance debate: Not seeing the forest for the trees. *Health Systems & Reform*, *3*(2), 74–79. https://doi.org/10.1080/23288604.2017.1302902

Turcotte-Tremblay, A. M., Spagnolo, J., De Allegri, M., & Ridde, V. (2016). Does performance-based financing increase value for money in low-and middle-income countries? A systematic review. *Health Economics Review*, *6*(1),30.

Waddington, H., White, H., Snilstveit, B., Hombrados, J. G., Vojtkova et al. (2012). How to do a good systematic review of effects in international development: A tool kit. *Journal of Development Effectiveness*, *4*(3), 359–387. https://doi.org/10.1080/19439342.2012.711765

Webster, R. (2016). *Payment by results: Lessons from the literature*. Russell Webster. http://russellwebster.com/Lessons%20from%20the%20Payment%20by%20Results%20literature%20Russell%20Webster%202016.pdf

Witter, S., Fretheim, A., Kessy, F., & Lindahl, A. (2012). Paying for performance to improve the delivery of health interventions in low- and middle-income countries. Cochrane Database of Systematic Reviews, 2. http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD007899.pub2/full

World Bank (2011). *A new instrument to advance development effectiveness: Program-for-results financing.* World Bank.

List of included studies

1. Anselmi, L., Binyaruka, P., & Borghi, J. (2017). Understanding causal pathways within health systems policy evaluation through mediation analysis: an application to payment for performance (P4P) in Tanzania. *Implementation Science, 12*(1), 1–18.

2. Alonge, O., Gupta, S., Engineer, C., Salehi, A. & Peters, D. (2015). Assessing the pro-poor effect of different contracting schemes for health services on health facilities in rural Afghanistan. *Health Policy and Planning*, *30*(10), 1229–1242.

3. Ashraf, N., Bandiera, O., & Jack, K. (2013). No margin, no mission? Evaluating the role of incentives in the distribution of public goods in Zambia, 3ie Impact Evaluation Report 9. http://www.3ieimpact.org/media/filer_public/2014/02/20/ie_9-ashraf-no_margin_no_mission_web.pdf

4. Barham, T. (2011). A healthier start: The effect of conditional cash transfers on neonatal and infant mortality in rural Mexico. *Journal of Development Economics*, 94(1), 74–85.

5. Basinga, P., Gertler, P. J., Binagwaho, A., Soucat, A. L. B., Sturdy, J., & Vermeersch, C. M. J. (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*(9775), 1421–1428.

6. Binagwaho, A., Condo, J., Wagner, C., Ngabo, F., Karema, C., Kanters, S., Forrest, J. I., & Bizimana, J. d. D. (2014). Impact of implementing performance-based financing on childhood malnutrition in Rwanda. *BMC Public Health*, *14*(1132).

7. Bonfrer, I., Soeters, R., Van de Poel, E., Basenya, O., Longin, G., van de Looij, F., & van Doorslaer, E. (2014a). Introduction Of Performance Based Financing In Burundi Was Associated With Improvements In Care And Quality. *Health Affairs*, *33*(12), 2179–2187.

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

9. Bossuroy, T., Delavallade, C., & Pons, V. (2016). Fighting tuberculosis through community based counsellors: a randomized evaluation of performance based incentives in India. 3ie Grantee Final Report. http://www.3ieimpact.org/media/filer_public/2016/07/28/gfr-ow31218-tb-health-worker.pdf

10. Bowser, D. M., Figueroab, R., Natiqc, L., & Okunogbec, A. (2013). A preliminary assessment of financial stability, efficiency, health systems and health outcomes using performance-based contracts in Belize. *Global Public Health*, *8*(9),1063–1074.

11. Celhay, P., Gertler, P., Giovagnoli, P. & Vermeersch, C. (2015). Long-Run Effects of Temporary Incentives on Medical Care Productivity, World Bank Policy Research Working Paper 7348. https://openknowledge.worldbank.org/bitstream/handle/10986/22228/Long0run0effec0al0care0productivity.pdf?sequence=1&isAllowed=y

12. Chansa, C., Das, A., Qamruddin, J., Friedman, J., Mkandawire, A., & Vledder, M. (2015). Linking Results to Performance : Evidence from a Results Based Financing Pre-Pilot Project in Katete District, Zambia, World Bank Health, Nutrition, and Population (HNP) discussion paper.

https://openknowledge.worldbank.org/bitstream/handle/10986/22390/Linking0result0ete0District00Zambia.pdf?sequence=1&isAllowed=y

13. Cornejo-Ovalle, M., Brignardello-Petersen, R., & Pérez, G. (2015). Pay-for-performance and efficiency in primary oral health care practices in Chile. *Revista Clínica de Periodoncia, Implantología y Rehabilitación Oral*, *8*(1**),** 60–66.

14. de Walque, D., Dow, W., & Nathan, R. (2014). Rewarding safer sex: Condlitional cash transfers for HIV/STI prevention, World Bank Policy Research Working Paper 7099. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2522733

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

16. de Walque, D., Robyn, P.J., Saidou, H., Sorgho, G., & Steenland, M. (2017). Looking into the performance-based financing black box: evidence from an impact evaluation in the health sector in Cameroon. The World Bank. http://documents1.worldbank.org/curated/ru/834601502391015068/pdf/WPS8162.pdf

17. Duchoslav, J., & Cecchi, F. (2019). Do incentives matter when working for god? The impact of performance-based financing on faith-based healthcare in Uganda. *World Development*, *113*, 309–319.

18. Eichler, R., Auxila, P., Antoine, U., & Desmangles, B. (2007). Performance-based incentives for health: Six years of results from supply-side programs in Haiti, Centre for Global Development Working Paper 121. http://www.cgdev.org/sites/default/files/13543_file_Haiti_Incentives.pdf

19. Engineer, C., Dale, E., Agarwal, A., Agarwal, A., Alonge, O., Edward, A., Gupta, S., Schuh, H., Burnham, G., & Peters, D. (2016). Effectiveness of a pay-for-performance intervention to improve maternal and child health services in Afghanistan: A cluster-randomized trial. *International Journal of Epidemiology*, *45*(2), 451–459.

20. Falisse, J.-B., Ndayishimiye, J., Kamenyero, V., & Bossuyt, M. (2015). Performance-based financing in the context of selective free health-care: an evaluation of its effects on the use of primary health-care services in Burundi using routine data. *Health Policy and Planning*, *30*(10), 1251–1260.

21. Fox, S., Witter, S., Wylde, E., Mafuta, E., & Lievens, T. (2014). Paying health workers for performance in a fragmented, fragile state: Reflections from Katanga Province, Democratic Republic of Congo. *Health Policy and Planning*, *29*(1), 96–105.

22. Garcia Prado, A., & Lao Peña, C. (2010). Contracting and Providing Basic Health Care Services in Honduras : A Comparison of Traditional and Alternative Service Delivery Models, World Bank Health, Nutrition and Population (HNP) discussion paper. https://openknowledge.worldbank.org/bitstream/handle/10986/13609/560080WP0Box341ContractingProviding.pdf?sequence=1&isAllowed=*y*

23. Gertler, P. (2000). The impact of PROGRESA on health, International Food Policy Institute Final Report. http://www.ifpri.org/publication/impact-progresa-health

24. Gertler, P., Giovagnoli, P., & Martinez, S. (2014). Rewarding provider performance to enable a healthy start to life: Evidence from Argentina's Plan Nacer, World Bank Policy Research Working Paper 6884. https://elibrary.worldbank.org/doi/pdf/10.1596/1813-9450-6884

25. Gertler, P., & Vermeersch, C. (2012). Using performance incentives to improve health outcomes, World Bank Policy Research Working Paper 6100. https://openknowledge.worldbank.org/handle/10986/9316

26. Gopalan, S. S., & Varatharajan, D. (2012). Addressing maternal healthcare through demand side financial incentives: Experience of Janani Suraksha Yojana program in India. *BMC Health Services Research*, *12*(319).

27. Grover, D., Bauhoff, S., & Friedman, J. (2018). Using supervised learning to select audit targets in performance-based financing in health: An example from Zambia. https://olc.worldbank.org/system/files/RBF%20Zambia.pdf

28. Huillery, E., & Sebanz, J. (2014). Performance based financing, motivation and final output in the health sector: Experimental evidence from the Democratic Republic of Congo. http://spire.sciencespo.fr/hdl:/2441/4pmvo3bm7m9claao2gl0337ip4/resources/2014-12.pdf

29. Ir, P., Korachais, C., Chheng, K., Horemans, D., Van Damme, W., & Meessen, B. (2015). Boosting facility deliveries with results-based financing: A mixed-methods evaluation of the government midwifery incentive scheme in Cambodia. *BMC Pregnancy and Childbirth*, *15*(170).

30. Janisch, C., Albrecht, M., Wolfschuetz, A., Kundu, F., & Klein, S. (2010). Vouchers for health: A demand side output-based aid approach to reproductive health services in Kenya. *Global Public Health*, *5*(6), 578–594.

31. Janssen, W., Ngirabega, J., Matungwa, M., & Van Bastelaere, S. (2015). Improving quality through performance-based financing in district hospitals in Rwanda between 2006 and 2010: A 5-year experience. *Tropical Doctor*, *45*(1), 27–35.

32. Kohler, H.-P., & Thornton, R. L. (2012). Conditional Cash Transfers and HIV/AIDS Prevention : Unconditionally Promising? *The World Bank Economic Review*, *26*(2), 165–190.

33. Kumar, M., Lehmann, J., Rucogoza, A., Kayobotsi, C., Das, A., & Schneidman, M. (2016). East Africa public health laboratory networking project : Evaluation of performance-based financing for public health laboratories in Rwanda, World Bank Health, Nutrition and Population (HNP) discussion paper. https://openknowledge.worldbank.org/bitstream/handle/10986/24400/East0Africa0000boratories0in0Rwanda.pdf?sequence=1&isAllowed=y

34. Kuule, Y., Dobson, A.E., Woldeyohannes, D., Zolfo, M., Najjemba, R., Edwin, B.M.R., Haven, N., Verdonck, K., Owiti, P., & Wilkinson, E., (2017). Community health volunteers in primary healthcare in rural Uganda: Factors influencing performance. *Frontiers in Public Health*, *5*, 62.

35. Lannes, L. (2015). Improving health worker performance: The patient-perspective from a PBF program in Rwanda. *Social Science & Medicine*, 138, 1–11.

36. Lannes, L., Meessen, B., Soucat, A., & Basinga, P. (2015). Can performance-based financing help reaching the poor with maternal and child health services? The experience of rural Rwanda. *The International Journal of Health Planning and Management*, *31*, 309–348.

37. Leroy, J. L., Garcia-Guerra, A., Garcia, R., Dominguez, C., Rivera, J., & Neufeld, L. M. (2008). The Oportunidades program increases the linear growth of children enrolled at young ages in urban Mexico. *Journal of Nutrition*, *138*(4), 793–798.

38. Lim, S. S., Dandona, L., Hoisington, J.A., James, S.L., Hogan, M.C., & Gakidou, E. (2010). India's Janani Suraksha Yojana, a conditional cash transfer programme to increase births in health facilities: an impact evaluation. *The Lancet*, *375* (9730), 2009–2023.

39. Liu, X., & Mills, A. (2005). The effect of performance-related pay of hospital doctors on hospital behaviour: A case study from Shandong, China. *Human Resources for Health*, *3*(11).

40. Mac Arthur, I., Nelson, J., & Woodye, M. (2014). Quality improvement of health care in Belize: focusing on results, Inter-American Development Bank Technical Note 661. https://publications.iadb.org/bitstream/handle/11319/6468/Quality%20Improvement%20of%20Health%20Care%20in%20Belize%20_%20Focusing%20on%20Results.pdf?sequence=1

41. Matsuoka, S., Obara, H., Nagai, M., Murakami, H., & Lon, R. (2014). Performance-based financing with GAVI health system strengthening funding in rural Cambodia: A brief assessment of the impact. *Health Policy and Planning*, *29*(4), 456–465.

42. Moyo, I., Gandidzanwa, C., Tsikira, T., Mabhena, T., Dieleman, M., & Kane, S. (2015). Process monitoring and evaluation II of Zimbabwe's results-based financing project: The case of Mutoko, Chiredzi, Nkayi and Kariba Districts, DFID.

43. Mohanan, M., Miller, G., Donato, K., Truskinovsky, Y., & Vera-Hernández, M. (2017). Different strokes for different folks: Experimental evidence on the effectiveness of input and output incentive contracts for health care providers with different levels of skills. https://kingcenter.stanford.edu/sites/default/files/publications/WP1025.pdf

44. Mussah, V.G., Mapleh, L., Ade, S., Harries, A.D., Bhat, P., Kateh, F. & Dahn, B. (2017). Performance-based financing contributes to the resilience of health services affected by the Liberian Ebola outbreak. *Public Health Action*, *7*(1), S100–S105.

45. World Bank, (2016). Rewarding provider performance to improve quality and coverage of maternal and child health outcomes—Evidence to inform policy and management decisions, DFID Report. *Available.*

46. na, na, Rwanda Community Performance-Based Financing Impact Evaluation, DFID

47. Nahimana, E., Iyer, H., Manzi, A., Uwingabiye, A., Gupta, N., Uwilingiyemungu, N., Drobac, P., & Hirschhorn, L. (2015). The race to the top initiative: towards excellence in health-care service delivery. *Global Health Action*, 9.

48. Ngo, D., & Bauhoff, S. (2018). The medium-run and scale-up effects of performance-based financing: An extension of Rwanda's 2006 trial using secondary data, Center for Global Development Working Paper 497. https://www.cgdev.org/publication/medium-run-and-scale-effects-performance-based-financing-extension-rwandas-2006-trial

49. Obare, F., Okwero, P., Villegas, L., Mills, S., & Bellows, B. (2016). Increased coverage of maternal health services among the poor in Western Uganda in an output-based aid voucher scheme, World Bank Policy Research Working Paper 7709. https://openknowledge.worldbank.org/bitstream/handle/10986/24626/Increased0cove0d0aid0voucher0scheme.pdf?sequence=1&isAllowed=y

50. Olken, B. A., Onishi, J., & Wong, S. (2014). Should aid reward performance? Evidence from a field experiment on health and education in Indonesia. *American Economic Journal: Applied Economics*, *6*(4),1–34.

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

52. Powell-Jackson, T., & Hanson, K. (2012). Financial incentives for maternal health: Impact of a national programme in Nepal. *Journal of Health Economics*, *31*(1), 271–284.

53. Powell-Jackson, T., Sumit Mazumdar, & Mills, A. (2015). Financial incentives in health: New evidence from India's Janani Suraksha Yojana. *Journal of Health Economics*, 43154-169.

54. Regalía, F. & Castro, L. (2007). Performance-based incentives for health: demand- and supply-side incentives in the Nicaraguan Red de Protección Social, Center for Global Development Working Paper 119. http://www.cgdev.org/sites/default/files/13541_file_Nicaragua_final.pdf

55. Renaud, A., & Semasaka, J.-P. (2014). Verification of performance in results-based financing: The case of community and demand-side RBF in Rwanda, World Bank Health, Nutrition and Population (HNP) Discussion paper. https://openknowledge.worldbank.org/bitstream/handle/10986/20791/917720WP0Verif00Box385343B00PUBLIC0.pdf?sequence=1&isAllowed=y

56. Robyn, P. J., Bärnighausen, T., Souares, A., Traoré, A., Bicaba, B., Sié, A., & Sauerborn, R. (2014). Provider payment methods and health worker motivation in community-based health insurance: A mixed-methods study. *Social Science & Medicine*, *108*, 223–236.

57. Rusa, L., Ngirabega, J., Janssen, W., Van Bastelaere, S., Porignon, D., & Vandenbulcke, W. (2009a). Performance-based financing for better quality of services in Rwandan health centres: 3-year experience. *Tropical Medicine & International Health*, *14*(7), 830–837.

58. Rusa, L., Schneidman, M., Fritsche, G., & Musango, L. (2009b). Rwanda: Performance-based financing in the public sector, Center for Global Development Case Study. http://www.nvag.nl/afbeeldingen/Bibliotheek/Rwanda/Rwanda%20PBF.pdf

59. Sherry, T. B., Sebastian Bauhoff & Mohanan, M. 2017. Multitasking and heterogeneous treatment effects in pay-for-performance in health care: Evidence from Rwanda. *American Journal of Health Economics*, *3*(2), 196–226.

60. Skiles, M. P., Curtis, S. L., Basinga, P., & Angeles, G. (2012). An equity analysis of performance-based financing in Rwanda: Are services reaching the poorest women? *Health Policy and Planning*, 28(8), 825–837.

61. Skiles, M. P., S.L. Curtis, P. Basinga, G. Angeles, & Thirumurthy, H. (2015). The effect of performance-based financing on illness, care-seeking and treatment among children: An impact evaluation in Rwanda. *BMC Health Services Research*, *15*(375).

62. Soares, F. V., Ribas, R. P., & Hirata, G. I. (2008). Achievements and shortfalls of conditional cash transfers: Impact evaluation of Paraguay's Tekopora Programme, UNDP—International Poverty Centre Evaluation Note Number 3. http://www.ipc-undp.org/pub/IPCEvaluationNote3.pdf

63. Soeters, R., Habineza, C., Peerenboom, P., & Rietsema, A. (2007). Performance-based financing and changing the district health system: Experience from Rwanda. *Tropical Medicine & International Health*, 12.

64. Soeters, R., Peerenboom, P., Mushagalusa, P. & Kimanuka, C. 2011. Performance-based financing experiment improved health care in the Democratic Republic of Congo. *Health Affairs*, *30*(8), 1518–1527.

65. Sood, N., Bendavid, E., Mukherji, A., Wagner, Z., Nagpal, S., & Mullen, P. (2014). Government health insurance for people below poverty line in India: Quasi-experimental evaluation of insurance and health outcomes. *BMJ*, *349*(g5114), 1–13.

66. Spisak, C., Morgan, L., Eichler, R., Rosen, J., Serumaga, B., & Wang, A. (2016). Results-based financing in Mozambique's central medical store: A review after 1 year. *Global Health Science and Practice*, *4*(1), 165–177.

67. Sun, X., Xiaoyun Liu, Qiang Sun, Winnie Yip, Adam Wagstaff, & Meng, Q. (2016). The impact of a pay-for-performance scheme on prescription quality in rural China. *Health Economics*, *25*(6), 706–722.

68. Sylvia, S., Luo, R., Zhang, L., Shi, Y., Medina, A., & Rozelle, S. (2013). Do you get what you pay for with school-based health programs? Evidence from a child nutrition experiment in rural China. *Economics of Education Review*, *37*, 1–12.

69. Tawfiq, E., Desai, J., & Hyslop, D. (2019). Effects of results-based financing of maternal and child health services on patient satisfaction in Afghanistan. *Journal of health services research & policy*, *24*(1), 4–10.

70. Urquieta, J., Angeles, G., Mroz, T., Lamadrid-Figueroa, H., & Hernández, B. (2009). Impact of Oportunidades on skilled attendance at delivery in rural areas. *Economic Development and Cultural Change*, *57*(3), 539–558.

71. Valadez, J., Jeffery, C., Brant, T., Vargas, W., & Pagano, M. (2015). Final impact assessment of the results-based financing programme for Northern Uganda, DFID Final Report.

72. Van de Poel, E., Flores, G., Ir, P., & O'Donnell, O. (2016). Impact of performance-based financing in a low-resource setting: A decade of experience in Cambodia. *Health Economics*, *25*(6), 688–705.

73. Wei, X., Zou, G., Yin, J., Walley, J., Yang, H., Kliner, M., & Mei, J. (2012). Providing financial incentives to rural-to-urban tuberculosis migrants in Shanghai: An intervention study. *Infectious Diseases of Poverty*, *1*(9).

74. Witter, S., Zaman, R., Scott, M., & Misty, R., 2016. Delivering Reproductive Health Results Through Non-State Providers in Pakistan, DFID Impact Evaluation Report.

75. World Bank. (2013). Turkey—performance based contracting scheme in family medicine : Design and achievements, World Bank Report. https://openknowledge.worldbank.org/bitstream/handle/10986/16532/770290Revised0box377292B00PUBLIC00.pdf?sequence=1&isAllowed=y

76. Yao, H., Wei, X., Liu, J., Zhao, J., Hu, D., & Walley, J. (2008). Evaluating the effects of providing financial incentives to tuberculosis patients and health providers in China. *The International Journal of Tuberculosis and Lung Disease*, *12*(10), 1166–1172.

77. Yip, W., Powell-Jackson, T., Wen Chen, Min Hu, Eduardo Fe, Mu Hu, Weiyan Jian, Ming Lu, Wei Han, & Hsiao, W. C. (2014). Capitation combined with pay-for-performance improves antibiotic prescribing practices in rural China. *Health Affairs*, *33*(3), 502–510.

78. Zeng, W., Cros, M., Wright, K. & Shepard, D. 2013. Impact of performance-based financing on primary health care services in Haiti. *Health Policy and Planning*, *28*(6), 596–605.

79. Zeng, W., Rwiyereka, A., Amico, P., Avila-Figueroa, C., & Shepard, D. (2014). Efficiency of HIV/AIDS health centers and effect of community-based health insurance and performance-based financing on HIV/AIDS service delivery in Rwanda. *American Journal of Tropical Medicine and Hygiene*, *90*(4), 740–746.

80. Zeng, W., Shepard, D.S., Nguyen, H., Chansa, C., Das, A.K., Qamruddin, J., & Friedman, J. (2018). Cost–effectiveness of results-based financing, Zambia: a cluster randomized trial. *Bulletin of the World Health Organization*, *96*(11), 760.

81. Zhang, L., Rozelle, S. ,& Shi, Y. (2013). Paying for performance in China's battle against anaemia, 3ie Impact Evaluation Report 8. http://www.3ieimpact.org/media/filer_public/2014/02/20/ie_8-zhang-china_anaemia_final.pdf